Frontiers in Artificial Intelligence and Applications

ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

Proceedings of the 25th International Conference of the Catalan Association for Artificial Intelligence

Edited by
Ismael Sanz
Raquel Ros
Jordi Nin

IOS Press

ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

Artificial intelligence is no longer solely the preserve of computer scientists and researchers; it is now a part of all our lives, and hardly a day goes by without discussion and debate about the implications of its many applications in the mainstream media.

This book presents the proceedings of CCIA 2023, the 25th International Conference of the Catalan Association for Artificial Intelligence, held from 25 - 27 October 2023 in Barcelona, Spain. CCIA serves as an annual forum welcoming participants from around the globe. The theme of the 2023 conference was Supportive AI, the main goals of which are to strengthen collaboration between research and industry by sharing the latest advances in artificial intelligence, and opening discussion about how AI can better support the current needs of industry. A total of 54 submissions were received for the conference, of which the 26 full papers, 18 short papers and 6 abstracts included here were selected after peer review. The papers cover a wide range of topics in Artificial Intelligence, including machine learning, deep learning, social media evaluation, consensus-building, data science, recommender systems, and decision support systems, together with crucial applications of AI in fields such as health, education, disaster response, and the ethical impact of AI on society. The book also includes abstracts of the keynotes delivered by Professor Aida Kamišalić and Dr. Lluis Formiga.

Providing a useful overview of some of the latest developments in artificial intelligence, the book will be of interest to all those working in the field.

ISBN 978-1-64368-448-2 (print)
ISBN 978-1-64368-449-9 (online)
ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, selected doctoral dissertations, handbooks and proceedings volumes. The FAIA series contains several sub-series, including 'Information Modelling and Knowledge Bases' and 'Knowledge-Based Intelligent Engineering Systems'. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. The series has become a highly visible platform for the publication and dissemination of original research in this field. Volumes are selected for inclusion by an international editorial board of well-known scholars in the field of AI. All contributions to the volumes in the series have been peer reviewed.
The FAIA series is indexed in ACM Digital Library; DBLP; EI Compendex; Google Scholar; Scopus; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Zentralblatt MATH.

Series Editors:
Nicola Guarino, Pascal Hitzler, Joost N. Kok, Jiming Liu, Ramon López de Mántaras,
Riichiro Mizoguchi, Mark Musen, Sankar K. Pal, Ning Zhong

Volume 375

Recently published in this series

Vol. 374. J.-L. Kim (Ed.), Machine Learning and Artificial Intelligence – Proceedings of MLIS 2023
Vol. 373. G. Grigoras and P. Lorenz (Eds.), Advances in Artificial Intelligence, Big Data and Algorithms – Proceedings of the 3rd International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA 2023)
Vol. 372. K. Gal, A. Nowé, G.J. Nalepa, R. Fairstein and R. Rădulescu (Eds.), ECAI 2023 – 26th European Conference on Artificial Intelligence – September 30–October 4, 2023, Kraków, Poland
Vol. 371. H. Fujita and G. Guizzi (Eds.), New Trends in Intelligent Software Methodologies, Tools and Techniques – Proceedings of the 22nd International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_23)
Vol. 370. A.J. Tallón-Ballesteros (Ed.), Modern Management Based on Big Data IV – Proceedings of MMBD 2023
Vol. 369. P. Hitzler, M.K. Sarker and A. Eberhart (Eds.), Compendium of Neurosymbolic Artificial Intelligence
Vol. 368. P. Lukowicz, S. Mayer, J. Koch, J. Shawe-Taylor and I. Tiddi (Eds.), HHAI 2023: Augmenting Human Intellect – Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

Artificial Intelligence Research and Development

Proceedings of the 25th International Conference of the Catalan Association for Artificial Intelligence

Edited by

Ismael Sanz

Universitat Jaume I, Spain

Raquel Ros

PAL Robotics, Spain

and

Jordi Nin

Universitat Ramon Llull, ESADE, Spain

IOS Press

Amsterdam • Berlin • Washington, DC

© 2023 The Authors.

This book is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

ISBN 978-1-64368-448-2 (print)
ISBN 978-1-64368-449-9 (online)
Library of Congress Control Number: 2023948020
doi: 10.3233/FAIA375

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

LEGAL NOTICE
The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

Preface

The International Conference of the Catalan Association for Artificial Intelligence (CCIA) serves as a unique forum, bringing together not only researchers in Artificial Intelligence within the Catalan-speaking territories (spanning southern France, Catalonia, Valencia, the Balearic Islands, and Alghero in Italy) but also drawing participants from around the globe.

This volume represents the proceedings of the 25th edition of CCIA, which took place in Sant Fruitós de Bages in October 2023. Past editions of CCIA have been hosted in various captivating locations, including Tarragona (1998), Girona (1999), Vilanova i la Geltrú (2000), Barcelona (2001, 2004, 2014, 2016), Castelló de la Plana (2002), Mallorca (2003), Alghero (Sardinia) (2005), Perpignan (France) (2006), Andorra (2007), Sant Martí d'Empúries (2008), Cardona (2009), L'Espluga de Francolí (2010), Lleida (2011), Alacant (2012), Vic (2013), València (2015), Deltebre (2017), Roses (2018), Colònia de Sant Jordi (2019), Lleida (2021), and Sitges (2022). Notably, CCIA was regrettably canceled in 2020 due to the COVID-19 restrictions.

This compilation includes 26 long papers, 18 short papers and 6 long abstracts meticulously selected from a pool of 54 submissions, a process made possible through the dedicated efforts of a program committee consisting of 104 artificial intelligence experts. We extend our heartfelt gratitude to them for their diligent reviews and also wish to acknowledge the hard work of the authors of the 54 submissions.

The accepted papers cover a diverse range of artificial intelligence domains, including machine learning, deep learning, social media evaluation, consensus-building, data science, recommender systems, decision support systems or computer vision, its crucial applications of AI in fields such as health, education, disaster response and serious games, and its ethical impact on society. Furthermore, this proceedings includes the abstracts of the keynote addresses delivered by Professor Aida Kamišalić and Dr. Lluis Formiga.

We extend our sincere appreciation to the Catalan Association for Artificial Intelligence (ACIA), the Universitat Jaume I, PAL Robotics, and Universitat Ramon Llull ESADE for their invaluable support.

Sant Fruitós de Bages, Catalonia, October 2023

Ismail Sanz, Universitat Jaume I
Raquel Ros, PAL Robotics
Jordi Nin, Universitat Ramon Llull, ESADE

This page intentionally left blank

About the Conference

The CCIA 2023 conference was organized by the Associació Catalana d'Intel·ligència Artificial (ACIA), Universitat Jaume I, Pal Robotics and the Universitat Ramon Llull, ESADE.

General Chair

Ismael Sanz, Universitat Jaume I

Scientific Chair

Raquel Ros, PAL Robotics

Local Chair

Jordi Nin, Universitat Ramon Llull, ESADE

Local Organizing Committee

Núria Agell, Universitat Ramon Llull, ESADE
Bernat Salbanya, Universitat Ramon Llull, ESADE
Irene Unceta,Universitat Ramon Llull, ESADE

Programme Committee Members

In alphabetic order:

Maider Abad Vázquez (UOC)
Nuria Agell (ESADE)
Mohammed Ali (URV)
Rosa Ma Alsina-Pagès (La Salle-URL)
Cecilio Angulo (UPC)
Javier Antich Tobaruela (UIB)
Josep Argelich (UdL)
Eva Armengol (IIIA-CSIC)
Cristian Barrué (UPC)
Robert Benavente (CVC)
Ismael Benito-Altamirano (IIA-CSIC)
Ester Bernado (Tecnocampus)
Christian Blum (IIIA-CSIC)
Francisco Bonnín (UIB - SRV)
Vicent Botti (UPV)
Dídac Busquets (Immense Simulations)
Ramòn Béjar (Universitat de Lleida)
Isaac Cano Franco (IDIBAPS)
Vicente Casales-Garcia (Universidad Sevilla)
Jordi Casas-Roma (UOC)
Gustavo Casañ (UJI)
Jesus Cerquides (IIIA-CSIC)

Ana Cocho Bermejo (ARU)
Jordi Coll (Universitat de Girona)
Ulises Cortés (UPC)
Atia Cortés Martínez (BSC)
Vicent Costa (IIIA-CSIC)
Jordi Delgado (UPC)
Pilar Dellunde (UAB)
Sergio Escalera (UB)
Joan Espasa Arxer (University of St Andrews)
Vlad Estivill-Castro (UPF)
Zoe Falomir (UJI)
Francesc J. Ferri (UV)
Tommaso Flaminio (IIIA-CSIC)
Sergi Foix (IRI-CSIC)
Thiago Freitas Dos Santos (IIIA-CSIC)
Emilio Garcia-Fidalgo (UIB)
Dario Garcia-Gasulla (BSC)
Ricard Gavalda (UPC)
Victor Gimenez (UAB)
Jesús Giráldez-Cru (Universidad de Granada)
Lluis Godo (IIIA-CSIC)
Jordi Gonzàlez (UAB)
Manuel González-Hidalgo (UIB)
Francisco Grimaldo (UV)
Oriol Gómez (UB)
Mohamed Hassanien (ASWU)
Jerónimo Hernández-González (UB)
Gonzalo E. Imaz (IIIA-CSIC)
Jose M. Iñesta (UA)
Vicente Julian (URV)
Ashutosh Karna (UPC)
Tetiana Klymchuk (UIC)
Javier Larrosa (UPC)
Jordi Levy (IIIA-CSIC)
Josep Llados (UAB)
Maite Lopez-Sanchez (UB)
Beatriz López (UDG)
Santiago Macho (IIIA-CSIC)
Felip Manyà (IIIA-CSIC)
Mario Martin (UPC)
Eloy Martinez-de-Las-Heras (R2A)
Carles Mateu (UDL)
Joaquim Melendez (UDG)
Sergi Mercadé Laborda (I2CAT)
Antonio Moreno (URV)
Lledó Museros (UJI)
David Nettleton (UPF)
Jennifer Nguyen (ESADE)
Jordi Nin (ESDE)
Pablo Noriega (IIIA-CSIC)
Jonathan Ayebakuro Orama (URV)
Pere Pardo (UNILU)
Jordi Pascual Fontanilles (URV)
Ferran Prados (UOC)
Eloi Puertas (UB)
Oriol Pujol (UB)
Juan Vicente Riera (UIB)
Juan Antonio Rodriguez Aguilar (IIIA-CSIC)
Emma Rollon (UPC)
Eric Roselló (UB)
Albert Sabater Coll (UdG)
Jordi Sabater Mir (IIIA-CSIC)
Bernat Salbanya (ESADE)
Ismael Sanz (UJI)
Marco Schorlemmer (IIIA-CSIC)
Marc Serramia (King's College London)
Josefina Sierra (UPC)
Agusti Solanas (URV)
Joan Ramon Soler (IIIA-CSIC)
Laia Subirats (Eurecat)
Josep Suy (UdG)
Miquel Sànchez-Marrè (UPC)
Mónica Sánchez (UPC)
Aida Valls (URV)
Maria Vanrell (UAB)
Javier Varona (UIB)
Alfredo Vellido (UPC)
Carles Ventura (UPC)
Xavier Vilasís-Cardona (La Salle-URV)
Mateu Villaret (UdG)
Jordi Vitria (UB)
Franz Wotawa (TU Graz)

Organizing Institutions

This page intentionally left blank

Contents

Preface v
Ismail Sanz, Raquel Ros and Jordi Nin

About the Conference vii

Part 1. Invited Talks

Unlocking the Power of LLMs: Navigating Challenges for Industry in the AI Renaissance 3
Lluís Formiga

Formalization of Drug Dose Titration Procedures 4
Aida Kamišalić

Part 2. Machine Learning: Theory and Applications

Inter-Satellite Link Prediction with Supervised Learning Based on Kepler and SGP4 Orbits 7
Estel Ferrer, Joan A. Ruiz-De-Azua, Francesc Betorz and Josep Escrig Escrig

Hydranet: A Neural Network for the Estimation of Multi-Valued Treatment Effects 16
Borja Velasco-Regulez and Jesus Cerquides

Applying Generative Models and Transfer Learning to Physiological Data Classification 28
José Fernando Núñez, Jamie Arjona, Adrián Tormos, Darío García and Javier Béjar

Stacking Up for Success: A Cascade Network Model for Efficient Road Crack Segmentation 38
Ammar M. Okran, Adel Saleh, Domenec Puig and Hatem A. Rashwan

Internal Representation and Shepard's Law in an Artificial Haptic System 48
Claudia Serrano and David Miralles

Expert Finding for Citizen Science 59
Bjoern Komander, Jesus Cerquides, Jaume Piera, Jeffrey Chan and Azadeh Alavi

Clustering of Qualitative Time Series for 3D Printing Quality Management 70
Ashutosh Karna and Karina Gibert

Cognitive Similarity Through Bibliometric Analysis 80
Lucas Marco-Tordera, Daniel García-Costa and Francisco Grimaldo

Full-Range Yaw Prediction: A Multi-View Approach for 3D Head Model Pose Estimation Using Convolutional Neural Networks 90
Álvaro Heredia-Lidón, Neus Martínez-Abadías and Xavier Sevillano

A Deep Learning-Based Pipeline for Computing Brain Biomarkers from MRI 94
Jordi Malé, Yann Heuzé, Juan Fortea, Neus Martínez-Abadías and Xavier Sevillano

Identifying Patterns Between Acoustic Environment and Visual Landscape Through Semantic Segmentation Based on Deep Learning 98
Riccardo Giacalone, Carlos Guerrero-Mosquera and Xavier Sevillano

OPNet: A One-Shot Image Similarity Algorithm for Production Systems 102
Karl Fabian Svensson and Carlos Guerrero-Mosquera

Analyzing Car Configurator Impact Through Genetic Algorithm from a Regional Perspective 106
Juan Manuel García-Sánchez, Xavier Vilasís-Cardona, Álvaro García-Piquer and Alexandre Lerma-Martín

Navigating Black Swan Events in Algorithmic Trading: A Reinforcement Learning Perspective 110
Fernando Villamarín Díaz and Carlos Guerrero-Mosquera

Machine Learning in Particle Physics Experiments: The Case of LHCb 115
Sergi Bernet Andrés, Míriam Calvo Gómez, Àlvaro García Piquer, Elisabet Golobardes Ribé, Núria Valls Canudas and Xavier Vilasís-Cardona

Petri-Nets-Based Controllers Generation Using Genetic Programming 120
Carlos A. Garcia-Sanchez, Manel Velasco, Cecilio Angulo, Pau Martí and Antonio Camacho

Addressing Scale and Density Challenges in LiDAR Point Classification 124
Mariona Carós, Ariadna Just, Santi Seguí and Jordi Vitrià

A Bayesian Network Framework to Study Class Noise: Exploring the Filtering of Completely Random Noise 128
David Martínez-Galicia, Alejandro Guerra-Hernández, Xavier Limón, Nicandro Cruz-Ramírez and Francisco Grimaldo

EduSign: Real-Time Application for Spanish Sign Language Recognition 132
Víctor Vallés and Ester Vidaña-Vila

Automated Cleanliness Scoring and Digestive Content Segmentation for Capsule Endoscopy 134
Pere Gilabert, Carolina Malagelada, Hagen Wenzek, Jordi Vitrià and Santi Seguí

Development of an Automated CBCT Nerve and Teeth Segmentation Tool Based on Deep Learning 136
Vicent Caselles, Wenceslao Piedra, Jordi Casas, Xavier Paolo Burgos and Ferran Prados

Zero-Shot Prediction for Emotion Recognition Using Deep Spectrum Features 138
Elena Ortega Beltrán, Ismael Benito-Altamirano and Carles Ventura

Part 3. AI for Health: Theory and Applications

Bayesian Optimization with Additive Kernels for the Calibration of Simulation Models to Perform Cost-Effectiveness Analysis 143
David Gómez-Guillén, Mireia Díaz, Josep Lluís Arcos and Jesus Cerquides

Exploring the Role of Explainability in AI-Assisted Embryo Selection 153
Lucia Urcelay, Daniel Hinjos, Pablo Martin-Torres, Marta González, Marta Méndez, Salva Cívico, Sergio Álvarez-Napagao and Dario Garcia-Gasulla

Longitudinal Segmentation of Multiple Sclerosis Lesions Using nnU-Net Architecture 163
Eloy Martínez-Heras, Adrian Vicente-Gomez, Francesc Vivó, Marcos Diaz-Hurtado, Baris Kanber, Jordi Casas-Roma, Sara Llufriu and Ferran Prados

Resource Allocation in Home Care Services Using Reinforcement Learning 173
Bernat Salbanya, Irene Unceta and Jordi Nin

Assessing VTE Risk in Cancer Patients Using Deep Learning Synthetic Data Generation and Domain Adaptation Techniques 183
Sergi Bech, Bárbara Lobato, Oriol Pujol, José Manuel Soria and Andrés Muñoz

Hipertension Demand Forecasting Using Cross-Correlation and Lagged Multiple Linear Regression Models for Anticipating Health Resources Needs 193
Guillem Hernández Guillamet, Beatriz López, Oriol Estrada and Francesc López Seguí

Challenges in the Exploitation of Historical Clinical Data for the Classification of Diabetic Retinopathy Patients 204
Jordi Pascual-Fontanilles, Aida Valls, Antonio Moreno and Pedro Romero-Aroca

Part 4. AI for Good, Ethics, Society, Serious Games

A General Model for the Metainformation of Complex Questionnaires for Automatic Preprocessing and Reporting Under INSESS Methodology 211
Xavier Angerri and Karina Gibert

Respect for Autonomy in the Machine Learning Pipeline 221
Paula Subías-Beltrán, Oriol Pujol, Itziar De Lecuona and Vicent Ribas Ripoll

Sequence Pattern Mining for Citizens Behaviour Learning in Fair-Purpose Social Games. Some Preliminary Results 231
Beatriz López, Roger Campdepadros and Albert Sabater

Reversal Error in Mathematics: Training Against It Using $\overset{\curvearrowleft}{\varepsilon}$ *coach* 241
Vicent Santamarta, Pablo Garcia-Segarra, Noelia Ventura-Campos, Aida Moreno-Rus and Zoe Falomir

Recommenders for Improved Lesson Planning in Formal Education 250
Pere Garcia-Calvés, Jordi Sabater-Mir, Pablo Aramendia, Jordi Corominas and Bryan Miguel Perez Ruchat

Capturing Young People's Interest in Food Waste Reduction: An Approach Using Qualitative Reasoning 256
Konstantina Zacharaki, Jennifer Nguyen, Queralt Prat-I-Pubill and Núria Agell

Explainable Machine Learning Models for Predicting COVID-19 Cases in Catalonia Based on Wastewater Monitoring Data 262
Marc Ribalta, Josep Pueyo-Ros, Lluis Corominas, Ramón Béjar, Carles Mateu and Edgar Rubión

Exploring Learning Techniques for Developing Socially-Aware Service Robots: Best Practices for Social Comfort 267
Silvia Guillén Ruiz, Juan P. Bandera Rubio, Alejandro Hidalgo-Paniagua and Antonio Bandera

Adding Edge Local Differential Privacy to the Dynamic Stochastic Block Model 273
Sudipta Paul, Julián Salas and Vicenç Torra

Reconstruction of the LHCb Calorimeter Using Machine Learning: Lessons Learned 277
Núria Valls Canudas, Xavier Vilasís-Cardona, Míriam Calvo Gómez and Elisabet Golobardes Ribé

A Real-World Dataset for Benchmarking False Alarm Rate in Keyword Spotting 283
Sergi Sánchez Deutsch, Ivan Huerta Casado and Josep Escrig Escrig

Predicting the Perceptual Rating of a Soundscape Using Artificial Intelligence 287
Daniel Bonet-Solà, Ester Vidaña-Vila and Rosa Ma. Alsina-Pagès

Part 5. Consensus, Social Media Evaluation

An Application to Measure Consensus on Customers' Ratings Using Hesitant Fuzzy Linguistic Term Sets 291
Walaa Abuasaker, Jennifer Nguyen, Núria Agell, Francisco J. Ruiz and Mónica Sánchez

Mathematical and Computational Models for Crowdsourced Geolocation 301
Rocco Ballester, Yanis Labeyrie, Mehmet Oguz Mulayim, Jose Luis Fernandez-Marquez and Jesus Cerquides

Sentiment Analysis of Gastronomic Posts from Colour Palettes and Narrative Content 311
V. Casales-Garcia, I. Sanz, L. Museros, Z. Falomir and L. Gonzalez-Abril

Improving Disaster Response by Combining Automated Text Information Extraction from Images and Text on Social Media 320
Hafiz Budi Firmansyah, Carlo Alberto Bono, Valerio Lorini, Jesus Cerquides and Jose Luis Fernandez-Marquez

Text Characterisation for the Fight Against Political Misinformation 330
Elena Álvarez-García, Daniel García-Costa, Joan Vila-Francés and Francisco Grimaldo

Part 6. Boolean Satisfiability Problems

SAT-IT: The Interactive SAT Tracer 337
Marc Cané, Jordi Coll, Marc Rojo and Mateu Villaret

General Boolean Formula Minimization with QBF Solvers 347
Eduardo Calò and Jordi Levy

A Complete Tableau Calculus for the Regular MaxSAT Problem 359
Jordi Coll, Chu-Min Li, Felip Manyà and Elifnaz Yangin

On the Density of States of Boolean Formulas 369
Carlos Ansótegui, María Luisa Bonet and Jordi Levy

Subject Index 383

Author Index 387

This page intentionally left blank

Part 1

Invited Talks

This page intentionally left blank

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230651

Unlocking the Power of LLMs: Navigating Challenges for Industry in the AI Renaissance

Lluís FORMIGA [1]
Verbio Technologies, S.L., Barcelona, Spain

Keywords. MML, speech technology, industry, AI impact

During the last year, AI has undergone its most disruptive transformation in half a century, largely due to OpenAI's innovations and cloud democratization. ChatGPT, stable diffusion, and Dall-E have reshaped our AI understanding and applications across society. Despite ethical concerns, it's vital to examine this shift's impact on resource-constrained entities like small and medium-sized enterprises and research groups.

At its core, this transformation relies on Large Language Models (LLMs) built upon Transformer and GPT architectures. Yet, it's essential to recognize that training these models demands significant hardware budgets and access to curated datasets. Equally crucial is the labor-intensive data curation, supervision, and refinement often overlooked.

Beyond training challenges, we must address inference cost, latency, concurrency, and scalability—affecting real-world viability and the quality decline in ChatGPT due to high request loads and its business model while maintaining profitability.

In this keynote, we introduce LLM fundamentals and explore challenges. We delve into how this AI revolution impacts medium-sized enterprises, providing speech technology as a use case, focusing on technical and business aspects. We'll emphasize customer-centric customization, domain specialization, and value addition in process automation. We will try to understand the profound implications of this AI renaissance for enterprises of all sizes.

[1]Corresponding Author: Lluís Formiga, lformiga@verbio.com

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230652

Formalization of Drug Dose Titration Procedures

Aida KAMIŠALIĆ [1]
University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia

Keywords. Drug dose titration, Drug prescription, Medication errors, Knowledge representation, Evidence-based medicine, Decision support

Drug prescription is a complex knowledge-based health-care professional task that may have several affecting factors such as the target disease, the patient's demographic group and comorbidity, the drug, dose, route and frequency of administration, accompanying drugs, etc. Multiple studies and publications reported the risks of wrong drug prescriptions, representing a major public health burden in the European Union.
Drug dose titration is the process of progressively adjusting the dose of a medication for the maximum benefit without adverse effects. It consists of prescribing an initial dose that is increased until the target medical issue is controlled (steady dose) or the maximum dose is reached. Dose titration can follow different procedures, depending on the disease, drug and patient condition. Once modelled, these procedures can serve for clinical homogenization, standardization, decision support and retrospective analysis.

This talk will provide an overview of the formal modelling of one- and two-drug dose titration procedures of chronic and acute cases as well as provide insight into alternative and supplementary drug dose titration procedures.

[1]Corresponding Author: Aida Kamišalić, aida.kamisalic@um.si

Part 2

Machine Learning: Theory and Applications

This page intentionally left blank

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230654

Inter-Satellite Link Prediction with Supervised Learning Based on Kepler and SGP4 Orbits

Estel FERRER [a,1], Joan A. RUIZ-DE-AZUA [a], Francesc BETORZ [a] and Josep ESCRIG ESCRIG [a]

[a] *i2CAT Foundation*

ORCiD ID: Estel Ferrer https://orcid.org/0000-0001-9999-2844, Joan A. Ruiz-de-Azua https://orcid.org/0000-0001-5892-3404, Francesc Betorz https://orcid.org/0009-0004-4022-6626, Josep Escrig Escrig https://orcid.org/0000-0002-0918-8148

Abstract.
Distributed Space Systems (DSS) are gaining prominence in the space industry due to their ability to increase mission performance by allowing cooperation and resource sharing between multiple satellites. In DSS where communication between heterogeneous satellites is necessary, achieving autonomous cooperation while minimizing energy consumption is a critical requirement, particularly in sparse constellations with nano-satellites. In order to minimize the functioning time and energy consumed by the Inter-Satellite Links (ISL) established for satellite-to-satellite communication, their temporal encounters must be anticipated. This work proposes an autonomous solution based on Supervised Learning that allows heterogeneous satellites in circular polar Low-Earth Orbits to predict their encounters, given the Orbital Elements (OE) and assuming isotropic antenna patterns. The model performance is evaluated and compared in two different scenarios: 1) a simplified scenario assuming that satellites follow Kepler orbits and 2) a realistic scenario assuming that satellites follow Simplified General Perturbations 4 (SGP4) orbits. This work could be considered the first stage of a promising and alternative approach in the field of DSS.

Keywords. Distributed Space Systems, Non-Terrestrial Networks, Inter-Satellite Links, Low-Earth Orbits, Kepler, Simplified General Perturbations, SGP4, Simplified Perturbations Models, 6G, Supervised Learning

1. Introduction

In recent years, the space industry has undergone a significant shift toward the implementation of Distributed Space Systems (DSS) [1]. As opposed to traditional monolithic systems, DSS consists of multiple satellites that communicate and coordinate their activities to increase the overall mission performance and provide global and continuous

[1] Corresponding Author: Estel Ferrer, estel.ferrer@i2cat.net

coverage for applications such as telecommunications, navigation, and remote sensing.

DSS requiring satellite-to-satellite communications gain importance with the definition of Non-Terrestrial Networks (NTN) and sixth generation (6G). Due to the satellites' dynamics, the connections established for satellite-to-satellite communications, known as Inter-Satellite Links (ISL), are characterized by a limited lifetime. In order to minimize the functioning time and energy consumed by the ISL, these temporal encounters must be anticipated, especially in sparse constellations where the connections can be sporadic.

The traditional solution for satellite encounter prediction is based on deterministic and centralized on-ground orbit propagation [2], [3]. This approach utilizes an orbital model to propagate the satellite's initial state, which is encoded in the Two-line Elements (TLE). Given the satellite trajectories and assuming isotropic antennas, the encounters can be scheduled defining a maximum threshold distance for communication. While on-ground propagation solutions can generate precise contact plans, the centralized nature of these solutions makes them unsuitable for DSS with heterogeneous satellites since it would require substantial cooperation efforts between different stakeholders. To overcome these limitations, new investigations are moving towards deterministic and decentralized solutions, providing autonomy to the satellites [4]. However, these techniques are still based on orbital propagation, which can result in a high processing cost for satellites with limited resources, such as nano-satellites. To avoid orbital propagation while keeping a decentralized and autonomous solution, authors in [5] have designed predictive algorithms to self-learn and self-construct the contact plan. However, in order to mathematically formulate the solution, they apply some linearizations only valid for circular and same-high satellites.

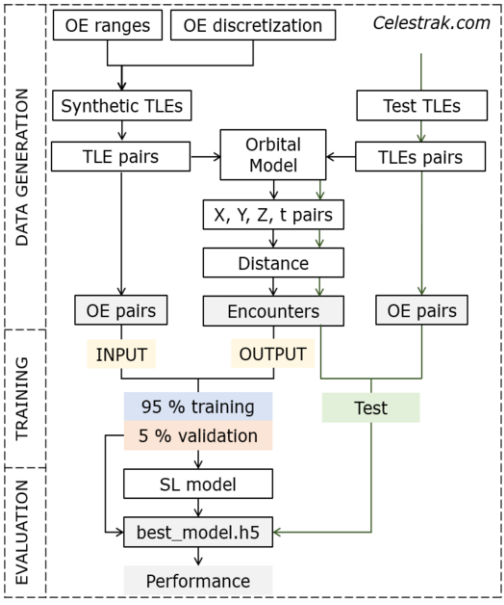
Recent studies have investigated the use of Graph Neural Networks (GNNs) together with Recurrent Neural Networks (RNNs) to model the temporal evolution of dynamic systems and predict future behavior based on past data [6–12]. The feasibility of treating a constellation as a time-evolving graph has been studied in [13] but discarded due to its complexity, poor performance, and strong dependency on past ground-truth data. However, new under-revision studies [14] are further investigating in this promising direction.

Building on a previous publication [15], the current work presents some advances in the proof of concept of using supervised learning to predict the contact plans between heterogeneous satellites. Concretely, this work presents and compares the results of predicting the encounters between Low Earth Orbit (LEO) satellites with isotropic antenna radiation patterns in two different scenarios: 1) ideal scenario assuming that satellites follow Kepler orbits, and 2) realistic scenario assuming that satellites follow Simplified General Perturbations 4 (SGP4) model, which accounts for perturbations caused by Earth's shape and atmospheric drag. In both scenarios, the same model architecture and input data are used. Even though the computation cost of the approach is a crucial point, this work focuses on obtaining good prediction performances. The analysis of the inference time and resources is left for future work.

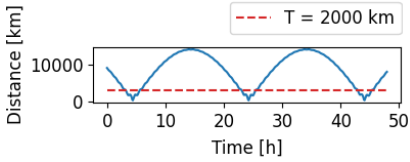
The remainder of the article is structured as follows. Section 2 defines the problem statement, introducing the datasets used for training, validation, and testing. Section 3 details the SL model architecture used to fit the input and output data. Section 4 presents the model performance in the validation and test set and compares the results obtained with the two different orbital models. Finally, Section 5 concludes the paper with a summary of keyfindings and suggestions for future research in this area.

2. Problem statement

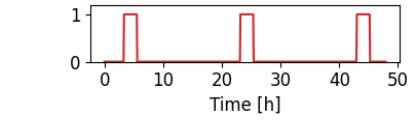
The encounter prediction problem consists in predicting the communication opportunities between any pair of satellites at any time. This problem is directly related to the communication features of the satellites and their trajectories. In this case, an isotropic antenna pattern is assumed for all the satellites. Therefore, having an encounter means having a satellite-to-satellite distance lower than a given threshold. Concerning satellite trajectories, both Kepler and SGP4 models are used and compared to propagate the initial state of the satellites.



(a) Block diagram showing the most important blocks used for data generation, model training and model evaluation



(b) Satellite-to-satellite distance and threshold distance.



(c) Output binary sequence of encounters with 576 samples corresponding to 48 hours with 5 minutes discretization.

Figure 1. Encounter prediction diagram, together with an example of the satellie-to-satellie distance and the derived binary sequence of encounters

The solution proposed in this work consists in training a SL model with a large set of input-output pairs. Figure 1a illustrates a simplified diagram of the three main steps that define the solution: data generation, model training, and model evaluation. The

role of the data generation part is to create the input and output data used for training, validation, and testing. On the one hand, the input data contains the Orbital Elements (OE) for different satellite pairs, a set of six parameters that locate a satellite by uniquely identifying its orbit shape, size, and orientation. Both training and validation input data are synthetically generated from 1000 different OE random values within the range for circular, polar, and low orbits defined in Table 1. Notice that, since all orbits are circular, both the eccentricity (e) and the argument of periapsis (ω) are always zero and can be excluded from the input set. As an example, Table 2 presents the first six rows of the input training data, featuring the satellite pairs 0-1, 0-2, 1-0, 1-2, 2-0, and 2-1. On the other hand, the output ground-truth data is the binary time sequence that determines when the satellite-to-satellite distance is below a given upper bound threshold set to 2000 kilometers (see Figs. 1b and 1c). To determine the satellite-to-satellite distance, it is assumed that the OE values are defined at the exact same moment in time, which coincides with the start of the encounter signal. Each encounter signal is characterized by a set of 576 binary output features, representing a total duration of 48 hours discretized into 5-minute intervals. It is important to note that these output data points are directly linked to both the input data and the orbital model used to propagate the satellites' initial positions.

Table 1. OE ranges for circular, polar, and low orbits. e: eccentricity, M: mean anomaly, i: inclination, ω: argument of periapsis, Ω: right ascending node, a: semi-major axis.

OE	e [-]	M [°]	i [°]	ω [°]	Ω [°]	a [km]
From	0	0	90	0	0	6800
To	0	360	100	0	360	7300

Table 2. First six samples of the input training set, containing the orbital necessary information for satellites 0, 1 and 2

Pair ID	$M1$ [rad]	$i1$ [°]	$\Omega1$ [rad]	$a1$ [km]	$M2$ [rad]	$i2$ [°]	$\Omega2$ [rad]	$a2$ [km]
0-1	4.40	95.2	6.05	7,335	3.04	97.7	2.88	6,906
0-2	4.40	95.2	6.05	7,335	3.11	96.0	4.50	6,746
1-0	3.04	97.7	2.88	6,906	4.40	95.2	6.05	7,335
1-2	3.04	97.7	2.88	6,906	3.11	96.0	4.50	6,746
2-0	3.11	96.0	4.50	6,746	4.40	95.2	6.05	7,335
2-1	3.11	96.0	4.50	6,746	3.04	97.7	2.88	6,906

In the second part of the diagram in Fig. 1a, the SL model is fitted with the training input-output sequences: 95 % of synthetic samples are used for training, while the remaining 5 % is used for validation to avoid overfitting. Details about the model architecture are presented in section 3. Finally, the model parameters learned during training are saved and used to predict both validation and test sets. The test set is a collection of 48 realistic satellites obtained from Celestrak [16] by gathering all active satellite TLE data and filtering those in circular, polar, and low orbits. The model performance, presented in section 4, is obtained by comparing the predicted encounters with the ground truth.

Notice that, to anticipate the encounter, this approach assumes that the satellite not only needs the trained model and its own OE values but also the OE values of the other satellite. This information is assumed to be known with enough anticipation.

3. Supervised Learning Model

Supervised Learning (SL) is a well-known machine-learning method used to solve problems where a large set of input-output pairs can be obtained yet the mathematical relationship between them is highly complex, costly, or unknown. While the training stage is usually costly, the simplicity of the operations in the inference part renders it a cost-effective approach, which could be suitable for resource-constrained systems like nanosatellites. However, it is important to clarify that this work focuses on finding the SL model architecture that best fits the input data. Future research will address tasks concerning resource optimization and trade-off analysis between performance and computational costs.

Figure 2 illustrates the Fully-Connected Neural Network (FCNN) architecture proposed to map the input OE pairs to the output sequence of encounters. The optimal number of layers and neurons per layer depends on the number of input and output features. As stated in section 2, the input set consists of 8 values, while the output describes the next 48 hours with 576 binary features. Since the number of input features is much lower than the number of output features, increasing the number of neurons while progressing deeper into the model is a good approach. Based on this reasoning, multiple architectures with varying numbers of hidden layers and neurons were evaluated. Among all the evaluated models, the one depicted in Fig. 2, with 8 layers and 270,048 trainable parameters, presents one of the best training curves with high generalization.

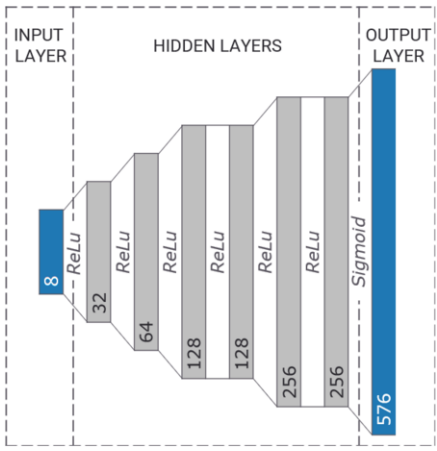


Figure 2. SL model architecture consisting of 8 layers. The values within each layer represent the number of neurons while the word in-between each layer is the applied activation function

Rectified Linear Unit is used as the activation function for the first seven layers. Contrarily, Sigmoid activation function is used at the output layer to ensure each neuron value ranges from 0 to 1. Notice that, since the ground truth is a binary signal, a decision threshold should be applied at the end to decide if the predicted values should be considered as no encounter (values close to 0) or encounter (values close to 1).

Other training settings include 500 epochs of training, a batch size of 3000 samples, the use of a binary cross-entropy loss function, the Adam optimizer with a learning rate of 0.01, and early stopping based on a patience of 200 epochs.

4. Results

This section presents the performance of the SL model presented in section 3. Two different scenarios are evaluated and compared: the ideal scenario and the realistic scenario. In the ideal scenarios, the output sequence of encounters used for training, validating, and testing are computed using Keplerian orbits, an ideal orbital model where external perturbations are not considered. Contrarily, in the realistic scenario, SGP4 orbital model, which includes Earth's shape and atmospheric drag, is assumed. Poorer results are expected with SGP4, since output-dependent information, concerning epoch time, satellite mass, and drag coefficient, is missing in the input data.

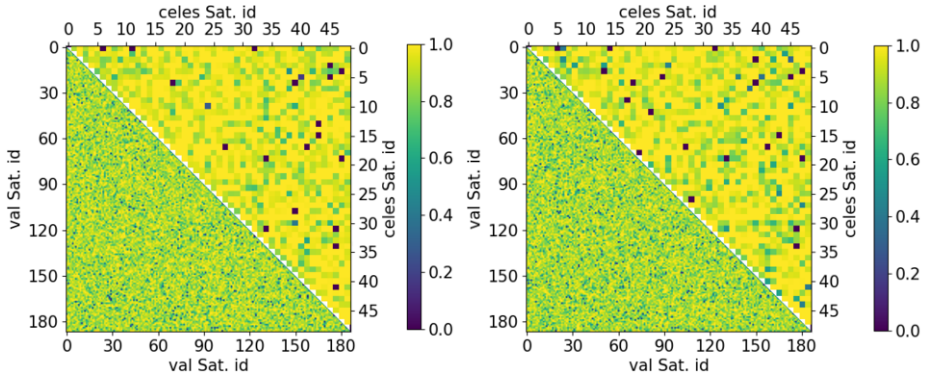
In both scenarios, the $F1$ score is used to quantify the model's performance for all satellite pairs. Equation (1) defines $F1$ as the harmonic mean of Precision (P) and Recall (R). In our scenario, P is the probability of estimating a contact inside the real timeslot (viewed as a measure of wasted energy in the prediction), while R assesses the proportion of successful encounters relative to the total available communication time and serves as an indicator of missed communication opportunities.

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

where TP, FP, and FN stand for True Positive, False Positive, and False Negative, respectively. Note that, to compute TP, FP, and FN, a decision threshold needs to be applied at the output of the model. For the following results, a threshold of 0.5 is used, meaning that all model-output values higher or equal to 0.5 are set to one, while values below this threshold are set to zero.

Figures 3a and 3b show the $F1$ score obtained for both validation satellite pairs (triangle below diagonal) and test satellite pairs (triangle above diagonal) assuming Kepler and SGP4 orbital models, respectively. Notice that some concrete pairs present zero $F1$. These pairs correspond to those with a limited number of encounters, in which having zero TP is almost unavoidable, and a single FP or FN leads to an overall $F1$ of zero. Given these observations, it is crucial to also consider other metrics, such as Balance Accuracy (BA) to gain a more comprehensive understanding of the model's performance. As observed in equation 2, BA can also be used for unbalanced datasets, and unlike $F1$, it is especially useful for critical situations where $TP = 0$.

$$BA = \frac{TPR + TNR}{2} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{2}$$



(a) Ideal scenario assuming Kepler orbits to compute the encounters

(b) Realistic scenario assuming SGP4 orbits to compute the encounters

Figure 3. F1-score for all validation and test satellite pairs. The triangle below the diagonal corresponds to the validation results while the triangle above the diagonal contains the test results

where TPR and TNR stand for True Positive Rate (also known as Recall or Sensitivity) and True Negative Rate (also known as Specificity), respectively.

Table 3 shows the mean BA, as well as the mean $F1$, mean P, and mean R for both scenarios. Unexpectedly, despite output-dependent information is missing in the realistic case with SGP4, the results for both scenarios are very similar, with a mean BA above 90 % and a mean $F1$ above 80 %. Based on this similarity, it can be inferred that, despite the stark difference between Keplerian and SGP4 trajectories, the satellite-to-satellite distance, and therefore, the output sequence of encounters. Consequently, only knowing the semi-major axis a, the mean anomaly M, and the right-ascending node Ω for two satellites in circular, polar, and low orbits, the future encounters under realistic SGP4 trajectories can be predicted with a BA of around 95 %. However, this is only true for small propagation times of less than a week: as time progresses, the influence of missing initial state information related to Earth's shape and drag (epoch, satellite mass, and drag coefficient) gains exponential importance.

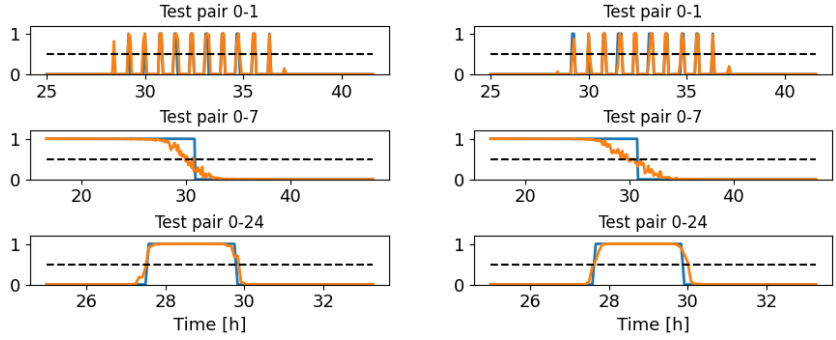
Although the results presented in Table 3 are notably good, they also indicate a decrease in the validation performance compared to the test performance. This behavior can be attributed to the fact that the validation dataset has an inclination range between 90 ° and 100 °, whereas, in a typical satellite constellation, such as the test Celestrak dataset, the polar satellites are confined to a much smaller inclination range between 96 ° and 100 °, corresponding to the Sun-synchronous orbits.

Finally, Figure 4 presents a comparison between the encounter ground truth and the model prediction for three different satellite pairs in the Celestrak test set during a specific time range of interest. The results show that the model is able to correctly predict the encounters for unseen and realistic input data, even in complex scenarios where contacts are really short in time.

Notice that the decision threshold plays an important role in the performance of the model and should be used depending on the user interests: if the maximum number of

Table 3. Mean model performance obtained for both validation and test sets and for both Kepler and SGP4 scenarios

	Kepler		SGP4	
	Val.	Test	Val.	Test
BA [%]	92.34	96.54	91.91	95.69
*F*1 [%]	85.62	92.76	84.98	91.76
P [%]	89.26	94.58	89.33	94.96
R [%]	85.06	93.47	84.19	91.63



(a) Ideal scenario assuming Kepler orbits to compute the encounters

(b) Realistic scenario assuming SGP4 orbits to compute the encounters

Figure 4. Comparison of the true encounters (in blue) with the model prediction (in orange). A decision threshold of 0.5 (dashed back line) is used to post-process the results and obtain the predicted encounters

encounters want to be detected, the threshold should be reduced; otherwise if the false encounters want to be minimized, the threshold should be increased. In future works, the impact of different decision thresholds will be studied.

5. Conclusions

This work presents a supervised machine-learning approach for forecasting encounters between satellites in low, circular, and polar orbits. The model uses 8-dimensional synthetic input training data obtained by combining various pairs of orbital elements within the region of interest for low, circular, and polar orbits. The output data is a set of 48-hour binary time series that describe the encounters of different satellite pairs assuming isotropic antenna patterns with a maximum threshold distance of 2000 km. Two different orbital models are used to compute the output data used for training, validating, and testing: ideal Kepler model and realistic SGP4, which accounts for the most important perturbations in LEO orbits. A Fully-Connected Neural Network architecture is used for training and predicting the output given the input information. The model performance is analyzed for the two scenarios.

Results show that both scenarios present high and similar performance, with $F1$ and *BA* of around 80 % and 90 %, respectively. On the one hand, these results demonstrate that a supervised learning model can be used to predict the encounters in polar, circular,

low, and realistic SGP4 orbits only knowing the relative inclination, the semi-major axis a, the mean anomaly M, and the right-ascending node Ω. On the other hand, it can be inferred that, despite the highly notable variation between Keplerian and SGP4 trajectories, the satellite-to-satellite distance, and therefore, the output sequence of encounters, remains almost invariant for a propagation time of only two days.

Future work must encompass more challenging frameworks, moving from low, circular, and polar orbits to low orbits with different eccentricities and a broader range of inclinations. We expect that larger training sets and more complex Machine Learning models will be required, providing novel insights for a future onboard AI-assisted device. In this sense, the computational cost of the model inference should be analyzed and optimized in the near future to be used in a real application.

References

[1] C. A. L opez. In pursuit of autonomous distributed satellite systems. Ph.D. dissertation, Polytechnic University of Catalonia, 2019

[2] D. Fischer, D. Basin, K. Eckstein, and T. Engel. Predictable mobile routing for spacecraft networks. IEEE Transactions on Mobile Computing, vol. 12, no. 6, pp. 1174–1187, 2012.

[3] J. A. Fraire and J. M. Finochietto. Design challenges in contact plans for disruption-tolerant satellite networks. IEEE Communications Magazine, vol. 53, no. 5, pp. 163–169, 2015

[4] O. C. FSu, J. B. Malaer and K. Suh. Using mobility prediction to enhance network routing in leo crosslink network. Proc. Int. Astron. Congr. (IAC), p. 1, 2019.

[5] J. A. Ruiz-De-Azua, V. Ramırez, H. Park, A. Calveras, and A. Camps. Assessment of satellite contacts using predictive algorithms for autonomous satellite networks. IEEE access, vol. 8, pp. 100 732–100 748, 2020.

[6] F. Manessi, A. Rozza, and M. Manzo. Dynamic graph convolutional networks. Pattern Recognition, vol. 97, p. 107000, 2020.

[7] J. Chen, X. Wang, and X. Xu. Gc-lstm: Graph convolution embedded lstm for dynamic link prediction. arXiv preprint arXiv:1812.04206, 2018

[8] J. Chen, J. Zhang, X. Xu, C. Fu, D. Zhang, Q. Zhang, and Q. Xuan. E-lstm-d: A deep learning framework for dynamic network link prediction. IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 6, pp. 3699–3712, 2019.

[9] A. Taheri, K. Gimpel, and T. Berger-Wolf. Learning to represent the evolution of dynamic graphs with recurrent models. Companion proceedings of the 2019 world wide web conference, 2019, pp. 301–307.

[10] L. Qu, H. Zhu, Q. Duan, and Y. Shi. Continuous-time link prediction via temporal dependent graph neural network. Proceedings of The Web Conference 2020, 2020, pp. 3026–3032.

[11] P. Goyal, S. R. Chhetri, and A. Canedo. dyngraph2vec: Capturing network dynamics using dynamic graph representation learning. Knowledge-Based Systems, vol. 187, p. 104816, 2020.

[12] K. Lei, M. Qin, B. Bai, G. Zhang, and M. Yang, "Gcn-gan: A non-linear temporal link prediction model for weighted dynamic networks. IEEE INFOCOM 2019-IEEE Conference on Computer Communications. IEEE, 2019, pp. 388–396

[13] E. Ferrer. Graph neural networks for optimization in telecommunications satellite constellations. Master's thesis, Polytechnic University of Catalonia, 2022.

[14] G. Casadesus and E. Alarcon. Toward autonomous cooperation in heterogeneous nanosatellite constellations using dynamic graph neural networks. International Astronautical Congress, 2023.

[15] E. Ferrer, J. Escrig and J. A. Ruiz-De-Azua. Inter-Satellite Link Prediction for Non-Terrestrial Networks Using Supervised Learning. European Conference on Networks and Communications (EuCNC) Congress, 2023.

[16] D. T. Kelso. (1985) Celestrak. [Online]. Available: https://celestrak.org/

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230655

Hydranet: A Neural Network for the Estimation of Multi-Valued Treatment Effects

Borja VELASCO-REGULEZ [a,1], Jesus CERQUIDES [a]
[a] *Artificial Intelligence Research Institute (IIIA-CSIC)*
ORCiD ID: Borja Velasco-Regulez https://orcid.org/0000-0003-4718-3388, Jesus Cerquides https://orcid.org/0000-0002-3752-644X

Abstract.
Neural network-based treatment effect estimation algorithms are well-known in the causal inference community. Many works propose new designs and architectures and report performance metrics over benchmarking data sets, in a *Machine Learning manner*. Nevertheless, most authors focus solely on binary treatment scenarios. This is a limitation, as many real-world scenarios have a multivalued treatment. In this work, we present a novel approach where we generalize a top-performing, neural network-based algorithm for binary treatment effect estimation to a multi-valued treatment setting. Our approach yields an estimator with desirable asymptotic properties, that delivers very good results in a wide range of experiments. To the best of our knowledge, this work is opening ground for the benchmarking of neural network-based algorithms for multi-valued treatment effect estimation.

Keywords. Causal Inference, Multi-valued Treatment Effect Estimation, Neural Networks

1. Introduction

Machine learning and neural networks are becoming a common choice for performing causal analysis tasks (causal inference, causal discovery) due to their power and flexibility for modelling complex functions, especially when dimensionality of the data is high[1]. Several authors have investigated specific network architectures, loss functions, regularization methods, etc. to tackle the task of inferring causal quantities using neural networks[2][3][4]. The performance of those algorithms is being benchmarked in the scientific literature, by using specific data sets and common metrics to achieve comparable results [5][6]. These advancements are happening almost exclusively in binary treatment scenarios. Nevertheless, often real-life applications have multiple-valued treatments (for instance, multi-armed clinical trials) or continuous treatments that can be discretized to multiple values [7][8]. This highlights the need to explore neural network-based causal inference methods for multi-valued treatments, both at the theoretical and empirical levels. In the present work, we select a top performance, neural network-based, binary average

[1]Corresponding Author: Borja Velasco-Regulez, bvelasco@iiia.csic.es

treatment effect estimation algorithm named Dragonnet[9] and test its generalizability to n-valued treatment settings. To the best of our knowledge, this[10] is the first attempt to establish a benchmark for this type of algorithm in the aforementioned setting. We present the theoretical and mathematical formulation, we develop a framework for experiments, and we provide the results obtained in different scenarios.

2. Problem Statement

Let the treatment of interest be a discrete random variable $T \in [0..k]$ that can take $k+1$ different values. Let the outcome be a continuous random variable $Y \in \mathbb{R}$, and let the covariates (i.e. the variables affecting both the treatment and the outcome) be a random vector $X \in \mathbb{R}^j$. Thus, our set of data points is $(Y_i, T_i, X_i),\ i \in [1..N]$, generated independently and identically. This set of data points constitutes our body of observational data. We define the causal effect of the treatment t over the outcome Y as, $\mu_t = \mathbb{E}[Y|do(T=t)]$, using Pearl's *do-calculus* notation [11], which denotes intervention. It can be shown that, if our data meets certain conditions, we can estimate causal (interventional) quantities based on observational data. Those conditions are known as the identifiability conditions: positivity, consistency and "no hidden confounder" conditions. For a more detailed explanation, see [11]. Under such conditions, $\mu_t = \mathbb{E}[Y|X=x, T=t]$, which is a quantity that is inferrable from our body of observational data. Along the rest of the section we assume that the identifiability conditions are fulfilled.

We define the conditional outcome as the expectation of the outcome given the treatment and the covariates, $Q(t,x) = \mathbb{E}[Y|t,x]$. Based on Q, we can construct a simple estimator $\hat{\mu}_t$ of μ_t as $\hat{\mu}_t = \frac{1}{N}\sum_i Q(t,x_i)$. In the following, we will be interested in approximating Q. Let $\hat{Q}$ be an approximation of Q. We define $\mu_t^{\hat{Q}} = \frac{1}{N}\sum_i \hat{Q}(t,x_i)$ as the estimator of μ_t obtained replacing Q by its estimation $\hat{Q}$. Furthermore, we define the Generalized Propensity Score (GPS[12]), expressed as $\mathbf{G}(x) = [g_0(x), g_1(x), \ldots, g_k(x)] \in \mathbb{R}^{k+1}$, with $g_t(x) = P(T=t|x)$.

In a binary treatment setting, under the identifiability conditions, the Average Treatment Effect (ATE) is one of the most common causal quantities of interest, and it is defined as $\psi = \mu_1 - \mu_0$. Given an approximation $\hat{Q}$ of Q, we could easily estimate ψ as $\psi^{\hat{Q}} = \mu_1^{\hat{Q}} - \mu_0^{\hat{Q}}$. In a multi-valued treatment setting, a wider class of causal quantities of interest can be defined, and all the conditional outcomes must be computed together in order to obtain valid estimates of those quantities[12]. In this work, we define such quantities of interest as the pair-wise average differences between the several treatments and a treatment considered the control (note that, in practice, the control treatment does not necessarily mean absence of treatment). Thus, we define a vector of ATEs $\psi \in \mathbb{R}^k$, $\psi = [\psi_1, \psi_2, \ldots, \psi_k]$, with $\psi_t = \mu_t - \mu_0$. We can approximate these quantities in a similar fashion as shown before, the t-th element of the vector being $\psi_t^{\hat{Q}} = \mu_t^{\hat{Q}} - \mu_0^{\hat{Q}}$. Note that if the causal quantity of interest was $\psi_{i,j} = \mu_i - \mu_j$, we could easily compute it based on the previous definition, as $\psi_{i,j} = \psi_i - \psi_j$, due to the linearity of the expectation operator.

The subject of interest in this the paper is the estimation of the vector of ATEs ψ. In the next section we generalize the estimation method provided in [9], which has the objective of estimating the ATE in the binary case, to the estimation of ψ in the multivalued treatment case presented above.

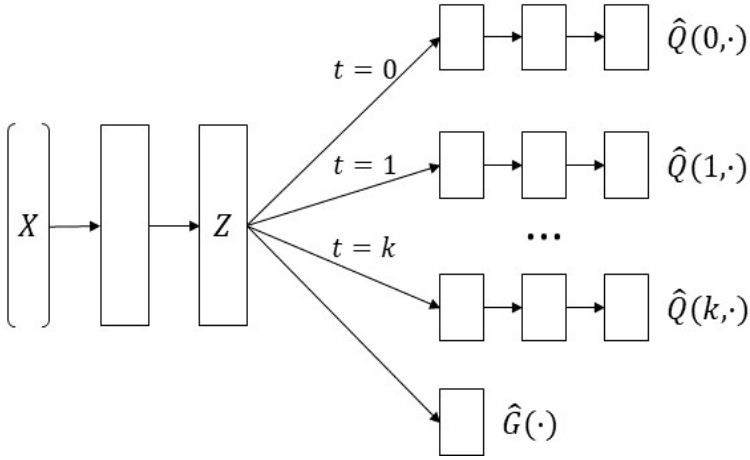


Figure 1. Hydranet architecture, where Z is the representation layer, and the $k+2$ heads correspond to the $k+1$ potential outcomes, $\hat{Q}(k,\cdot)$, and the Generalized Propensity Score, $\hat{G}(\cdot)$.

3. From Dragonnet to Hydranet

Dragonnet is a high-capacity, end-to-end neural network architecture for estimating binary treatment effects[9]. We present here the variation of the architecture, mathematical formulations and proofs for adapting Dragonnet to a multivalued treatment setting. We call this adaptation Hydranet.

3.1. Architecture

The architecture of Hydranet can be seen in Figure 1. It consists of two parts: the representation part, formed by the input layer and two hidden layers, and the heads, formed by $k+2$ ends. Out of those, $k+1$ correspond to the conditional outcomes, and are formed by two more hidden layers plus the output layer. The remaining head corresponds to the GPS, $\mathbf{G}(x) = [g_0(x), g_1(x), \ldots, g_k(x)] \in \mathbb{R}^{k+1}$, with $g_t(x) = P(T = t|x)$, consisting on just the output layer. All layers are fully connected. Recall that we approximate the t-th element of the vector of ATEs as $\psi_t^{\hat{Q}} = \frac{1}{N}\sum_i \hat{Q}(t,x_i) - \hat{Q}(0,x_i)$.

The baseline objective function has the shape

$$\hat{R}(\theta) = \frac{1}{N}\sum_i [(Q^{nn}(t_i,x_i;\theta) - y_i)^2 + \alpha \text{CrossEntropy}(g_t^{nn}(x_i;\theta),t_i)] \tag{1}$$

where the quadratic term relates to the errors of the potential outcomes' heads and the cross entropy term relates to the errors of the propensity score's head. The model parameters are

$$\hat{\theta} = \arg\min_{\theta}[\hat{R}(\theta)] \tag{2}$$

3.2. Targeted Regularization

Now, following the reasoning in [9], we present targeted regularization. Targeted regularization is a modification of the objective function that introduces an extra parameter, epsilon. In our setting, ε is a vector in $\mathbb{R}^k, \varepsilon = (\varepsilon_1, \varepsilon_2, ...\varepsilon_k)$, and the new objective function is

$$\bar{F}(\theta, \varepsilon) = \hat{R}(\theta) + \beta \frac{1}{N} \sum_i \gamma_i(y_i, t_i, x_i; \theta, \varepsilon), \text{ where} \tag{3}$$

$$\gamma_i(y_i, t_i, x_i; \theta, \varepsilon) = (y_i - \bar{Q}_i(\theta, \varepsilon))^2, \text{ and} \tag{4}$$

$$\bar{Q}_i(\theta, \varepsilon) = Q^{nn}(t_i, x_i) + \varepsilon_1 \left(\frac{\mathbf{I}(T=1)}{g_1^{nn}(x_i)} - \frac{\mathbf{I}(T=0)}{g_0^{nn}(x_i)} \right) + \ldots + \varepsilon_k \left(\frac{\mathbf{I}(T=k)}{g_k^{nn}(x_i)} - \frac{\mathbf{I}(T=0)}{g_0^{nn}(x_i)} \right) \tag{5}$$

with $\mathbf{I}(T=t)$ the indicator function, and thus the sought model parameters are defined by

$$\hat{\theta}, \hat{\varepsilon} = \arg\min_{\theta, \varepsilon} [\hat{R}(\theta) + \beta \frac{1}{N} \sum_i \gamma_i(y_i, t_i, x_i; \theta, \varepsilon)] \tag{6}$$

But why this modification in the first place? The answer lies in semiparametric estimation theory (SET) and targeted maximum likelihood estimation (TMLE). Very generally, SET provides us with conditions that ensure desirable properties of our estimator ψ when they are fulfilled, and TMLE is an efficient method to achieve the fulfillment of those conditions. The conditions are the set of non-parametric estimating equations, defined as

$$\mathbf{0} = \left[\frac{1}{N} \sum_i \varphi_{i,1}, \frac{1}{N} \sum_i \varphi_{i,2}, ... \frac{1}{N} \sum_i \varphi_{i,k} \right] \tag{7}$$

and they employ the elements of the vector of efficient influence curves, defined as $\varphi \in \mathbb{R}^k, \varphi = [\varphi_1, \varphi_2, ...\varphi_k]$, with

$$\varphi_{i,t} = Q^{nn}(t, x_i) - Q^{nn}(0, x_i) + \left(\frac{\mathbf{I}(T=t)}{g_t^{nn}(x_i)} - \frac{\mathbf{I}(T=0)}{g_0^{nn}(x_i)} \right) (y_i - Q^{nn}(t, x_i)) - \psi_t \tag{8}$$

Finally, recall that what we want is that the minimization of the modified objective function ensures the fulfillment of the non-parametric estimation equations. This can be expressed mathematically as

$$\mathbf{0} = \nabla \bar{F}|_{\hat{\varepsilon}} = \left[\frac{\partial \bar{F}}{\partial \varepsilon_1}, \frac{\partial \bar{F}}{\partial \varepsilon_2}, \ldots \frac{\partial \bar{F}}{\partial \varepsilon_k}\right]\Bigg|_{\hat{\varepsilon}} = \left[\frac{\beta}{N}\sum_i \varphi_{i,1}, \frac{\beta}{N}\sum_i \varphi_{i,2}, \ldots \frac{\beta}{N}\sum_i \varphi_{i,k}\right] \tag{9}$$

and the proof can be found in the supplementary material. This warrants the aforementioned desirable properties of the estimator ψ, i.e. double robustness, fast convergence, and efficiency.

4. The Data and the Metrics

We have tested Hydranet in two datasets, a fully synthetic one and a semi-synthetic one. In the remainder of the text we will refer to them as the synthetic dataset (or SynD for short) and the IHDP dataset, respectively. In order to generate these datasets we have designed and implemented algorithms mimicking different Data Generating Processes (DGP). For the synthetic dataset, the covariates, treatments and outcomes have been synthetically generated, and we have taken inspiration from [13]. For the IHDP dataset, the covariates are taken from a study with real participants, while the treatments and outcomes are synthetically generated. Those real covariates were collected for a Randomized Controlled Trial (RCT) carried out in 1985 [14], and are routinely used for benchmarking causal inference algorithms, usually following the configuration in[15]. We have followed a similar strategy but adapting the DGP to our needs (a multi-valued treatment scenario). With both datasets the number of treatments has been set to 5. In the remainder of this section we provide a more detailed explanation of the DGP of each dataset and its output.

4.1. Synthetic Dataset DGP

For generating fully synthetic data, we have designed DGPs with tunable parameters of dataset size D, bias size B and number of confounders NC. The number of treatments has been set to 5. The potential covariates are constituted by vectors $\mathbf{x} \in \mathbb{R}^{30}$ with each value sampled from a uniform distribution $\mathscr{U}(-1,1)$. The number of such vectors is equal to the data size parameter D, forming a matrix $\mathbf{X} \in \mathbb{R}^{Dx30}$. The actual confounders, i.e., the variables that participate in the determination of both the treatment and the outcome, are the first NC (number of confounders) elements of each covariate vector, thus forming a matrix $\mathbf{C} \in \mathbb{R}^{DxNC}$. The treatment for each datapoint has been obtained in two steps. First, we square the confounder vector element-wise and sum the elements, apply a $min-max$ scaler to the range $[0,4]$ (for 5 treatments), and round to the closest integer. Then, in order to fulfill the positivity condition, we draw the final treatment value from a categorical distribution such that

$$p(t|\mathbf{c}) = \begin{cases} 0.8, & \text{if } t = m(\mathbf{c}) \\ \frac{0.2}{k-1}, & \text{otherwise} \end{cases}$$

with $m(\cdot)$ the operation defined in the first step. Finally, for computing the potential outcomes, we have defined three outcome functions $(l_a(t,\mathbf{x}), l_b(t,\mathbf{x}), l_c(t,\mathbf{x}))$ that map a combination of the covariates and the treatment to the output space, for each datapoint. The outcome functions have the shape

$$l_a(t,\mathbf{x}) = 30\mathbf{v}_0^T\mathbf{x} + 10\,t^2\,\mathbf{v}_t^T\mathbf{x} + \varepsilon$$
$$l_b(t,\mathbf{x}) = 20\mathbf{v}_0^T\mathbf{x} + 5\,B\,t\,\mathbf{v}_t^T\mathbf{x} + \varepsilon$$
$$l_c(t,\mathbf{x}) = 10\mathbf{v}_0^T\mathbf{x} + 5log(|B\,t\,\mathbf{v}_t^T\mathbf{x}|) + \varepsilon$$

with B the bias parameter, $\mathbf{v}_0$ the baseline effect parameter, defined as $\mathbf{u}_0/||\mathbf{u}_0||$ with $||\cdot||$ the euclidean norm and $\mathbf{u}_0 \sim \mathscr{U}(0,1)$ a randomly sampled vector ($\mathbf{u}_0 \in \mathbb{R}^{30}$), and $\varepsilon \sim \mathscr{N}(0,1)$. Recall that a potential outcome, denoted $\mathbf{y}^t$, is the outcome that a datapoint would have had, had it been treated with a particular treatment t. The matrix of potential outcomes $\mathbf{Y} \in \mathbb{R}^{Dx5}$ is defined as $\mathbf{Y} = [\mathbf{Y}^0, \mathbf{Y}^1, \mathbf{Y}^2, \mathbf{Y}^3, \mathbf{Y}^4] = [l_a(\mathbf{0},\mathbf{X})^T,\ l_b(\mathbf{1},\mathbf{X})^T,\ l_c(\mathbf{2},\mathbf{X})^T,\ l_b(\mathbf{3},\mathbf{X})^T,\ l_a(\mathbf{4},\mathbf{X})^T]$, with $\mathbf{0} = (0,0,...0) \in \mathbb{R}^D$, $\mathbf{1} = (1,1,...1) \in \mathbb{R}^D$, etc.

We have generated datasets under varying values of the three parameters of interest, bias size $B = [2,5,10,30]$, dataset size $D = [1000,2000,5000,10000]$ and number of confounders $NC = [2,5,10,18]$, varying one parameter at a time. When kept fixed, the values have been set to $B = 20$, $D = 2000$ and $NC = 2$.

4.2. IHDP Dataset DGP

For generating the IHDP dataset, we have followed a similar strategy, but fixing $NC = 2$ and $B = 10$, and $D = 985$ being the size of the original IHDP covariate set. The treatment assignment function is based in two variables present in the set, mom ethnicity and weeks preterm. We assign treatment 0 to individuals with mom ethnicity equalling "black", treatment 1 to individuals with mom ethnicity equalling "white", treatment 2 to individuals with mom ethnicity equalling "hispanic", treatment 3 to individuals with mom ethnicity equalling "hispanic" and weeks preterm being bigger than 6, and treatment 4 to individuals with mom ethnicity equalling "black" and weeks preterm smaller than 6. Note that this setting is completely made up and has no connection with any real-life situation. Then, the final treatment is sampled from a probability distribution as explained in the previous section. The outcome functions are defined as

$$l_1(t,\mathbf{x}) = \exp(\mathbf{x}\beta) + B*MB + t^2 + \varepsilon$$
$$l_2(t,\mathbf{x}) = log(|\mathbf{x}\beta|) + B*MW*t + \varepsilon$$
$$l_3(t,\mathbf{x}) = \mathbf{x}\beta + B*MH + t^2 + \varepsilon$$
$$l_4(t,\mathbf{x}) = \exp(\mathbf{x}\beta) + B*WP + t + \varepsilon$$
$$l_5(t,\mathbf{x}) = log(|\mathbf{x}\beta|) + B*WP*t + \varepsilon$$

where β is a vector of parameters, B is the bias parameter, MB, MW and MH are the components of the one-hot encoding of mom ethnicity, and WP is weeks preterm.

4.3. Metrics

For performance benchmarking purposes, we have employed the sum of errors of the vector of ATEs. This is computed as the sum of the absolute values of the differences of

all estimated ATE components with respect to their true values, $E = \sum_{t=1}^{k} |\psi_t - \hat{\psi}_t|$. This choice allows to have a single real number as a final result, making comparisons simpler. All values have been computed as averages across 20 dataset realizations to increase the robustness of the results, and 95% confidence intervals have been computed with Bootstrapping.

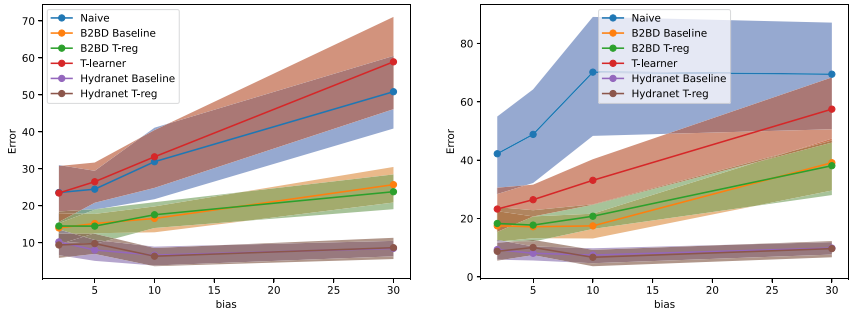
5. Experiments and Results

In the case of binary treatment settings, there are *de facto* benchmarking datasets and metrics, i.e., datasets and metrics that are widely used in the literature and thus serve for algorithmic performance comparison purposes. The IHDP dataset and the metrics presented in[15] are an example of this. This is not the case in multi-valued treatment settings, where comparators are scarce. Nevertheless, we have developed and implemented algorithms that can be considered comparable to Hydranet, to benchmark its performance. Thus, in every experiment, we present the results of the following algorithms: 1) **Naive**, a naive estimator of the treatment effect that employs only the observable data, without controlling, and serves to visualize the impact of confounding 2) **B2BD**, back to back Dragonnets, a strategy that uses 4 Dragonnets (with the same setup as in [9]), each one estimating one element of the vector of ATEs ψ, 3) **T-learner**, a T-learner [16] estimator that employs a gradient boosting machine (GBM) model[2], and finally 4) **Hydranet**, both in its baseline form and with targeted regularization. Hydranet performs well in all the tested scenarios and outperforms the comparators, both with in-sample (train set) data and with out-sample (test set) data, reaching low or very low error values for different dataset sizes, bias sizes and number of confounders. The employed training scheme consists of a first stage with the ADAM optimizer and a second stage with the Stochastic Gradient Descent (SGD) optimizer, with hyperparameters similar to [9].

5.1. Synthetic Data Experiments

Figure 2 and Table 1 show the error of the different algorithms for varying values of bias. As it can be seen, Hydranet performs better than the comparators. Hydranet is only mildly affected by the increasing size of the bias, while the comparators show bigger error increases. Similarly, Figure 3 and Table 2 show the performance of the algorithms for varying dataset sizes. As expected, all algorithms reduce their error with bigger data set sizes, but Hydranet with targeted regularization outperforms the rest, and shows a smaller error even for small dataset sizes, proving its (data) efficiency. Note that in this experiment, the output of Baseline Hydranet has been plugged into an Augmented Inverse Probability Weighting (A-IPTW) estimator, instead of the simple estimator presented in Section 1. The purpose of this change was to test the performance of Baseline Hydranet for delivering a plug-in estimator that is doubly robust. As it can be seen, this estimator fails for smaller dataset sizes, due to a well-reported phenomenon of finite-sample instability. This fact motivates the development and implementation of Targeted Regularization-equipped Hydranet. Finally, experiments with varying number of confounders show similar results, with Hydranet outperforming the alternatives.

[2] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html



(a) In sample (b) Out sample

Figure 2. Errors w.r.t. bias size

Table 1. Errors of the different algorithms w.r.t. bias size

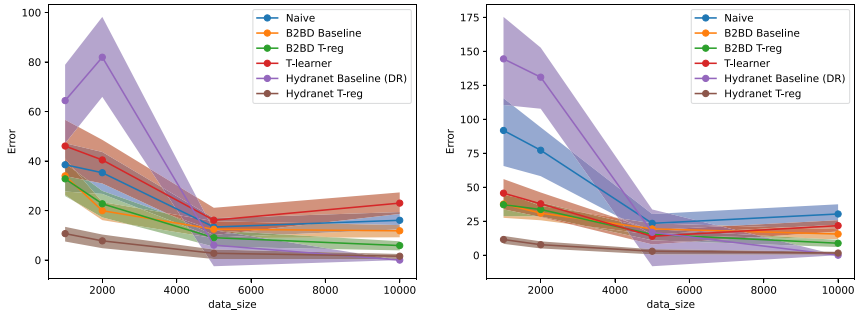
Bias	5		10		30	
	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
Naive	48.83 ± 16.04	24.47 ± 5.27	70.16 ± 20.39	31.91 ± 9.67	69.46 ± 18.28	50.84 ± 9.82
B2BD base.	17.18 ± 3.74	15.19 ± 2.61	17.47 ± 4.25	16.55 ± 3.5	39.12 ± 8.91	25.7 ± 4.81
B2BD t-reg.	14.48 ± 5.38	14.48 ± 4.7	17.56 ± 4.22	17.56 ± 3.49	23.78 ± 9.35	23.78 ± 4.7
T-learn	26.44 ± 5.45	26.49 ± 5.44	33.09 ± 7.8	33.23 ± 7.82	57.48 ± 11.17	58.93 ± 12.48
Hydranet base.	8.08 ± 2.41	8.01 ± 2.83	7.48 ± 2.56	6.64 ± 2.55	9.88 ± 2.13	8.46 ± 2.09
Hydranet t-reg.	9.75 ± 2.6	9.75 ± 2.74	6.34 ± 2.81	6.34 ± 2.46	8.64 ± 2.77	8.64 ± 2.87

Table 2. Errors of the different algorithms w.r.t. dataset size

Data Size	2000		5000		10000	
	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
Naive	77.23 ± 17.96	35.29 ± 8.46	23.45 ± 7.5	13.35 ± 3.93	30.34 ± 7.51	16.07 ± 3.5
B2BD base.	30.97 ± 5.3	19.95 ± 3.65	19.39 ± 4.2	12.46 ± 3.43	15.59 ± 3.73	11.83 ± 2.46
B2BD t-reg.	22.74 ± 5.11	22.74 ± 5.35	9.04 ± 4.15	9.04 ± 3.28	5.87 ± 2.63	5.87 ± 1.84
T-learn	37.85 ± 9.01	40.44 ± 8.81	13.95 ± 5.34	16.11 ± 5.52	21.76 ± 4.15	23.01 ± 4.34
Hydranet base.	130.98 ± 22.48	81.93 ± 16.15	16.83 ± 20.88	6.04 ± 7.29	2.43 ± 1.19	2.94 ± 1.10
Hydranet t-reg.	7.76 ± 2.61	7.76 ± 2.76	2.74 ± 1.9	2.74 ± 1.91	1.57 ± 1.07	1.57 ± 1.04

5.2. IHDP Data Experiments

Table 3 shows the error of the different algorithms with the IHDP dataset. Similarly as with synthetic data, Hydranet (both baseline and targeted regularization) outperform the comparators. The targeted regularization algorithm has a slightly bigger error than the baseline algorithm, but the difference is considered negligible. These results prove the efficacy of Hydranet with semi-synthetic data, showing its suitability for real-world scenarios.



(a) In sample (b) Out sample

Figure 3. Errors w.r.t. dataset size

Table 3. Performance with IHDP dataset

	Out-Sample	In-Sample
Naive	14.81 ± 0.97	17.51 ± 2.07
B2BD base.	22.75 ± 1.92	22.74 ± 2.38
B2BD t-reg.	22.22 ± 1.9	22.81 ± 2.68
T-learn	13.53 ± 1.22	13.7 ± 1.2
Hydranet base.	3.09 ± 0.56	3.03 ± 0.69
Hydranet t-reg.	3.82 ± 0.88	3.8 ± 0.87

6. Discussion

In this work, we have generalized a top-performing, neural network-based algorithm for ATE estimation from a binary treatment setting to a n-valued treatment setting. We have developed and implemented synthetic and semi-synthetic DGPs for algorithmic benchmarking purposes, and we have developed comparator algorithms for evaluating the performance of Hydranet. We show that Hydranet (both baseline and targeted regularization) performs well under different bias sizes, dataset sizes, and number of confounders, and we provide both theoretical and empirical evidence of the motivation for developing targeted regularization-equipped Hydranet. In addition, we show the good performance of the algorithm with semi-synthetic data. The direct generalizability of neural network-based algorithms for ATE estimation from binary settings to n-valued treatment settings is a common claim in the literature, but we show that it has its own challenges and that the behavior of the algorithms in each particular scenario requires its own interpretation. As far as we know, this paper is opening ground on the proposal of benchmarking results for neural network-based ATE estimation in multivalued treatment scenarios.

The main limitations of this work are twofold: on one hand, we have been forced to construct the competitor algorithms of Hydranet ourselves, due to the scarcity of

benchmarking data in the literature. On the other hand, we have only tested this algorithm for a 5-valued treatment scenario. It is a line of future work to adapt the algorithm and perform experiments for k-valued scenarios. Finally, we have run into some instabilities during neural network training in some of the experiments. It is also a line of future work to correct them and ensure the stability of the training process.

Acknowledgments

This work was supported by Doctorat Industrial funded by Generalitat de Catalunya [DI-2020-18] and by project CI-SUSTAIN funded by the Spanish Ministry of Science and Innovation [PID2019-104156GB-I00]. Borja Velasco-Regúlez is a PhD Student of the doctoral program in Computer Science at the Universitat Autonoma de Barcelona.

References

[1] Hernán MA, Robins JM. Causal Inference: What If;.
[2] Yoon J, Jordon J, Van Der Schaar M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In: International Conference on Learning Representations; 2018. .
[3] Johansson FD, Shalit U, Sontag D. Learning Representations for Counterfactual Inference. arXiv:160503661 [cs, stat]. 2018 Jun. ArXiv: 1605.03661. Available from: http://arxiv.org/abs/1605.03661.
[4] Nair N, Gurumoorthy KS, Mandalapu D. Individual Treatment Effect Estimation Through Controlled Neural Network Training in Two Stages. arXiv; 2022. Number: arXiv:2201.08559 arXiv:2201.08559 [cs]. Available from: http://arxiv.org/abs/2201.08559.
[5] Lin A, Merchant A, Sarkar SK, D'Amour A. Universal Causal Evaluation Engine: An API for empirically evaluating causal inference models:9.
[6] Shimoni Y, Yanover C, Karavani E, Goldschmnidt Y. Benchmarking Framework for Performance-Evaluation of Causal Inference Analysis. arXiv:180205046 [cs, stat]. 2018 Mar. ArXiv: 1802.05046. Available from: http://arxiv.org/abs/1802.05046.
[7] Cattaneo MD, Drukker DM, Holland AD. Estimation of Multivalued Treatment Effects under Conditional Independence. The Stata Journal: Promoting communications on statistics and Stata. 2013 Sep;13(3):407-50. Available from: http://journals.sagepub.com/doi/10.1177/1536867X1301300301.
[8] Li F, Li F. Propensity score weighting for causal inference with multiple treatments. The Annals of Applied Statistics. 2019 Dec;13(4):2389-415. Publisher: Institute of Mathematical Statistics. Available from: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-13/issue-4/Propensity-score-weighting-for-causal-inference-with-multiple-treatments/10.1214/19-AOAS1282.full.
[9] Shi C, Blei DM, Veitch V. Adapting Neural Networks for the Estimation of Treatment Effects. arXiv:190602120 [cs, stat]. 2019 Oct. ArXiv: 1906.02120. Available from: http://arxiv.org/abs/1906.02120.
[10] Velasco B, Cerquides J, Arcos JL. Hydranet: A Neural Network for the estimation of Multi-valued Treatment Effects. In: NeurIPS 2022 Workshop on Causality for Real-world Impact; 2022. Available from: https://openreview.net/forum?id=sJChORLuPHK.
[11] Pearl J, Madelyn Glymour,, Nicholas P Jewell. Causal Inference in Statistics. A Primer. John Wiley and Sons Ltd, United States; 2016.
[12] Cattaneo MD. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. Journal of Econometrics. 2010 Apr;155(2):138-54. Available from: https://linkinghub.elsevier.com/retrieve/pii/S030440760900236X.
[13] Kaddour J, Zhu Y, Liu Q, Kusner MJ, Silva R. Causal Effect Inference for Structured Treatments.
[14] Enhancing the Outcomes of Low-Birth-Weight, Premature Infants: A Multisite, Randomized Trial. JAMA. 1990 Jun;263(22):3035-42. _eprint:

https://jamanetwork.com/journals/jama/articlepdf/382131/jama_263_22_030.pdf. Available from: https://doi.org/10.1001/jama.1990.03440220059030.

[15] Dorie V, Hill J, Shalit U, Scott M, Cervone D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. arXiv:170702641 [stat]. 2018 Jul. ArXiv: 1707.02641. Available from: http://arxiv.org/abs/1707.02641.

[16] Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences. 2019;116(10):4156-65. _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1804597116. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1804597116.

A. Supplementary material

A.1. Hydranet

We want to prove that

$$\left.\frac{\partial \bar{F}}{\partial \varepsilon_t}\right|_{\hat{\varepsilon}_t} = \frac{1}{N}\sum_i \varphi_{i,t}, \qquad \forall\, t \text{ in}[0,k] \tag{10}$$

Proof. On one hand, using equations (3), (4) and (5) we get

$$\begin{aligned}
\left.\frac{\partial \bar{F}}{\partial \varepsilon_t}\right|_{\hat{\theta},\hat{\varepsilon}_t} &= \left.\frac{\partial}{\partial \varepsilon_t}\left(\hat{R}(\theta) + \beta \frac{1}{N}\sum_i \gamma_i(y_i, t_i, x_i; \theta, \varepsilon)\right)\right|_{\hat{\theta},\hat{\varepsilon}_t} \\
&= \left.\frac{\beta}{N}\sum_i \frac{\partial}{\partial \varepsilon_t}\gamma_i(\theta,\varepsilon)\right|_{\hat{\theta},\hat{\varepsilon}_t} \\
&= \left.\frac{2\beta}{N}\sum_i (y_i - \bar{Q}_i(\theta,\varepsilon))\frac{\partial \bar{Q}_i(\theta,\varepsilon)}{\partial \varepsilon_t}\right|_{\hat{\theta},\hat{\varepsilon}_t} \\
&= \left.\frac{2\beta}{N}\sum_i \left[(y_i - \bar{Q}_i(\theta,\varepsilon))\left(\frac{\mathbf{I}(T=t)}{g_t^{nn}(\theta)} - \frac{\mathbf{I}(T=0)}{g_0^{nn}(\theta)}\right)\right]\right|_{\hat{\theta},\hat{\varepsilon}_t} \\
&= \frac{2\beta}{N}\sum_i \left[(y_i - \hat{Q}(t,x_i)\left(\frac{\mathbf{I}(T=t)}{\hat{g}_t} - \frac{\mathbf{I}(T=0)}{\hat{g}_0}\right)\right] \text{(evaluate at } \hat{\theta},\hat{\varepsilon}) \\
&= \frac{2\beta}{N}\sum_i \left(\hat{Q}(t,x_i) - \hat{Q}(0,x_i)\right) - \frac{\beta}{N}\sum_i \left(\hat{Q}(t,x_i) - \hat{Q}(0,x_i)\right) + \\
&\frac{\beta}{N}\sum_i \left[(y_i - \hat{Q}(t,x_i))\left(\frac{\mathbf{I}(T=t)}{\hat{g}_t} - \frac{\mathbf{I}(T=0)}{\hat{g}_0}\right)\right] \text{(add and subtract term)} \\
&= \frac{2\beta}{N}\sum_i \left[\hat{Q}(t,x_i) - \hat{Q}(0,x_i) + (y_i - \hat{Q}(t,x_i))\left(\frac{\mathbf{I}(T=t)}{\hat{g}_t} - \frac{\mathbf{I}(T=0)}{\hat{g}_0}\right) - \hat{\psi}_t\right] \text{(group sums)}
\end{aligned}$$

On the other hand, by substituting the definition of the efficient influence curves (8) in the set of non-parametric estimation equations (9), multiplying by β and particularizing at $\hat{Q}, \hat{g}, \hat{\psi}$ (the functions modelled by the neural network at the optimal point of the parameter space), we obtain an expression equal to the one in the last line of the proof. Thus, the non-parametric estimation equations (9) are satisfied, and the proof is complete. □ □

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230656

Applying Generative Models and Transfer Learning to Physiological Data Classification

José Fernando Núñez [a,1] Jamie Arjona [a] Adrián Tormos [b] Darío García [b] Javier Béjar [a]
[a] *Computer Science Department, Universitat Politècnica de Catalunya (Barcelona Tech), Barcelona, Spain*
[b] *Barcelona Supercomputing Center (BSC), Spain*

Abstract. The scarcity and imbalance of datasets for training deep learning models in a specific task is a common problem. This is especially true in the physiological domain where many applications use complex data collection processes and protocols, and it is difficult to gather a significant number of subjects.

In this paper, we evaluate generative deep learning algorithms by training them to create data based on open physiological datasets and conduct a study on their potential for transfer learning. We measure the performance change of classifiers when the training data is augmented with the synthetic samples and also perform experiments in which we fine-tune classification models trained with the generated data adding increasing amounts of the real data to investigate the transfer learning capabilities of synthetic datasets.

Finally, we advise and provide the best option for researchers interested in augmenting ECG datasets using these algorithms and the best fine-tuning strategies that would generalize correctly when tested on new data from the same domain but for a different classification task.

Keywords. Synthetic Data, Transfer Learning, Time Series, Physiological Signals, ECG

1. Motivation

Given the limited availability of open physiological datasets due to regulations and data privacy, researchers face the problem of finding enough data to train data-hungry deep learning models. In addition, in many applications, there are severe class imbalances due to the scarcity of examples which are very often the class of interest for the diseases studied. One way to amend these problems is the use of synthetic data as a data augmentation technique, as opposed to the more *classical* techniques.

Many of these techniques for the time series [1] domain are inspired by augmentations used in image processing (scaling, transposition, cut-outs, etc.) or are domain-specific transformations (noise, jittering, warping, etc.), but it is difficult for this augmented data to both capture the attributes of the problem and to be more than simply

[1]Corresponding Author: José Fernando Núñez, Universitat Politècnica de Catalunya, E-mail: jose.fernando.nunez@upc.edu.

small alterations of the original data. Also, in the time series domain, applying these techniques results in data that does not capture the underlying dataset distribution.

Generative models provide us with a more interesting way to augment all types of data. From text-to-image generation to NLP tasks, generative data has been a hot topic for quite some time. However, training these models on time series proves harder than other types of data like images and text since there is a very limited literature on generative models in the domain of physiological time series (EEG, ECG, EMG, etc.).

The aim of this paper is to present a short study in this domain, focusing on the selection of a generative model and how to evaluate the performance of the synthetic data, both graphically and when used to augment existing physiological datasets by pretraining a classification model and thus perform transfer learning.

This research is being developed as part of the Artificial Intelligence for Healthy Aging (AI4HA) project. The goal of the project is to develop diagnostic models for aging-associated diseases, including Parkinson's disease, cognitive decline, heart failure, sarcopenia, hearing loss, etc. Most of the data used in the project come from a wide range of physiological time series such as electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), and inertial sensors (IMU).

2. Modeling data for synthetic generation

Modeling data in machine learning can be approached from two perspectives. The discriminative approach assumes that it is sufficient to train a model to associate the input data to a class label or a continuous value without capturing the inner process behind the data. The generative approach assumes that modeling a problem also involves capturing the process that defines the data, usually represented explicitly or implicitly as a probability distribution. This is a harder problem, but allows us to do sampling on the trained model to obtain new data samples.

Currently, there are many deep learning approaches for modeling data generatively as probability distributions. Some models define proper probability distributions directly from the attributes of the data (autoregressive models [2], normalizing flows [2]) or by learning latent variables models (VAEs [2], Denoising Diffusion [3]); others capture implicitly the data distribution with models that can be sampled from (Generative Adversarial Networks [4]). These models can be used to generate data and have been successfully applied in other domains, especially in natural language processing and for many computer vision tasks.

One advantage of these models is that they can generate data without conditioning, thus capturing a joint probability distribution, or conditionally given additional information that allows partitioning the data distribution accordingly. This allows, for instance, the usage of class labels, but the conditioning factor can be any supplementary information available for the data. In this work, we condition the generative models on the class labels, as we would like to be able to generate data resembling specific classes from the original datasets.

Time series appear in many applications and are a challenge for generative models, since many techniques applied in other areas are difficult to adapt or do not make sense. Literature on applying these models for time series is limited compared to other domains and is usually focused on specific applications and datasets. As a contribution to this

topic, in the next sections we will experiment with different generative models trained with physiological time series, propose methods for assessing the quality of the synthetic data using dimensionality reduction methods, use the trained models to generate data for solving classification tasks, and use the generated data to pretrain models to demonstrate the potential of transfer learning.

3. Models

For data generation, we will select models from the families of Generative Adversarial Networks (GANs) [4] and Denoising Diffusion [3]. The former models the problem of data generation as a game between a network that can obtain samples from a latent Gaussian vector (generator) and a network whose task is to distinguish these generated samples from the real ones (discriminator). The latter model assumes that a network can be trained to progressively add noise to the data samples, and the model then learns how to reverse this process so it can start from Gaussian noise and iteratively remove the noise until a sample with maximium likelihood is obtained.

We train these models conditionally based on the classification task that will be solved. In other words, we take into account the class labels of the dataset when fitting the generative models. For the GANs, we mix the class label with the latent noise that the model learns to convert to data, and we also add classification as an additional output for the discriminator to determine the class of the real and generated samples. This procedure biases the model so that the generator is able to transform the Gaussian latent on samples of the selected class. For the diffusion model, the conditioning appears as an embedded component added to the original noise prediction network. A cross-attention layer is used to bias the model so the path followed by the denoising process ends on sample of the desired class.

We chose to train two GANs, one based on CNNs with 1D convolutions [5] whose discriminator output is an array of numbers corresponding to different parts of the sequence instead of a single number measuring the realness of the data [6], and the other based on transformers [7]. We also chose a diffusion model based on DiffWave [8], a popular model used in audio synthesis. The selected models have been adapted to work with physiological time series and class conditioning, by modifying the architecture to work with single-channel time series and adding a conditional embedding representing the labels attributed to the corresponding datasets.

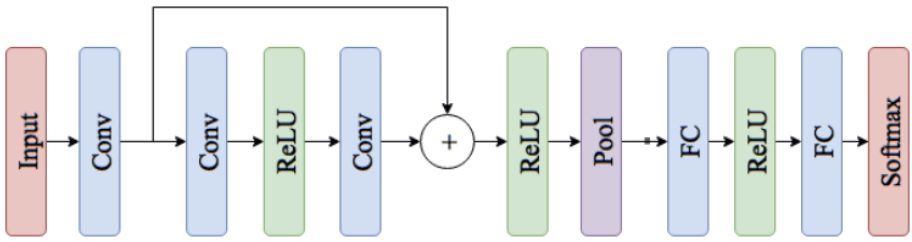


Figure 1. Deep residual convolutional classifier architecture.

For the classification tasks models, to study the capabilities of synthetic data, we have considered training a deep residual convolutional neural network model (Figure 1) under different settings.

4. Datasets

For creating the synthetic data, we used open source datasets from PhysioNet. These consist of ECG signals from various patients with different anomalies. Specifically, the MIT-BIH dataset [9] offers data related to cardiac arrhythmias (five classes based on the categories proposed by the AAMI EC57 standard, subdivided into healthy with 90,589 observations, atrial premature with 2,779 observations, ventricular malfunction with 7,236 observations, fusion of ventricular and normal with 803 observations and a last class of 8,039 unclassified observations), while the PTB dataset [10,11] contains ECG signals corresponding to patients diagnosed with myocardial infarction and healthy patients used as control data (two classes, the abnormal one with 10,505 observations, and the normal or healthy one with 4,045). For both datasets, we used the same preprocessing technique as presented in the work [12], where the ECG signals are divided into QRS segments of length 187.

5. Methodology and Evaluation

When it comes to methods for evaluating the generated data, there is no agreement on the correct way. The most common ones in the literature are the ones used for synthetic images, namely the Fréchet Inception Distance [4] and other similar metrics. In the case of images, this is measured based on the features obtained from a pre-trained network on a massive dataset of natural images. There is, however, no such concept of *natural* time series, nor pretrained networks that can be used for feature extraction, so this is not a viable approach. The evaluation of conditional generation is also an open problem given the difficulty in assessing how well the distribution of the generated data has captured the underlying information present in the real dataset.

In this work, we chose to perform two simple types of evaluation. The first is a 2D plot of the transformation of the data sets using the dimensionality reduction algorithms t-Stochastic Neighbor Embeding (t-SNE, [13]) and Locally Linear Embeding (LLE, [14]), which give visual insight into the distribution of the synthetic data in spaces of reduced dimensionality. Properly generated data would have an *overlapped* distribution with the original dataset, when fitting t-SNE and LLE to both the original and generated data.

The second type of evaluation is the change in the classification scores (*accuracy*, *precision*, and *recall*) when using the aforementioned classifier. We perform several kinds of evaluations on the synthetic quality of generated data: training on synthetic data and testing on real data (TSTR), training on real data and testing on synthetic data (TRTS), training and testing on real data (TRTR), training and testing on synthetic data (TSTS), training on a mix of real and synthetic data and testing on real data (TRSTR), and training on synthetic and adding increasing proportions of real data. These tests were carried out to see how different training affects the classification task, with the aim of improving the performance of the classifiers using only real data.

6. Generative and Classification models experiments

By training all the generative models with the original datasets, we have found the TTS-CGAN and Diffwave models to be the best choices, as they were both successful in different ways. The Diffwave model seems to be better at capturing the underlying probabilistic distribution and also generating an evenly distributed array of synthetic data, while the TTS-CGAN based model generates the best pretraining data for the MIT-BIH dataset. The GAN with 1D convolutions, arguably the simplest model, does not do a good job in capturing the probabilistic distribution of the real dataset with the generated samples, and seems to cluster the data in dense patches corresponding to the class labels.

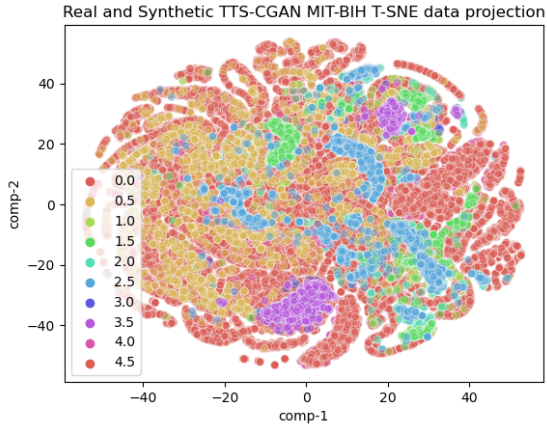


Figure 2. t-SNE fitted with real and synthetic data generated with TTS-CGAN model trained on the MIT-BIH dataset. Real classes are decimals ending in .0 and synthetic ones in .5 (e.g. healthy data: real=0.0, generated=0.5)

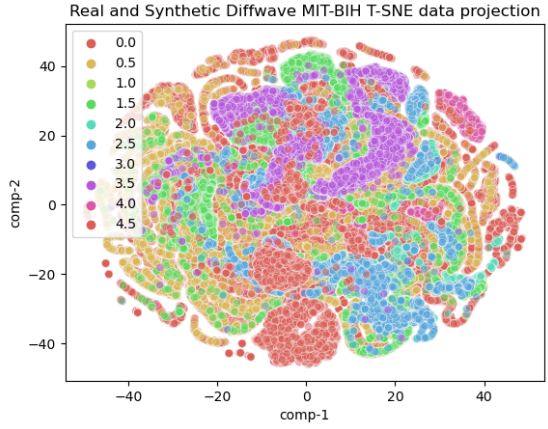


Figure 3. t-SNE fitted with real and synthetic data generated with Diffwave model trained on the MIT-BIH dataset. Real classes are decimals 0 and synthetic ones in .5 (e.g. healthy data: real=0.0, generated=0.5)

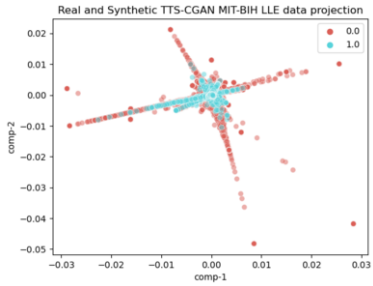


Figure 4. LLE fitted with the MIT-BIH dataset and TTS-CGAN generated data

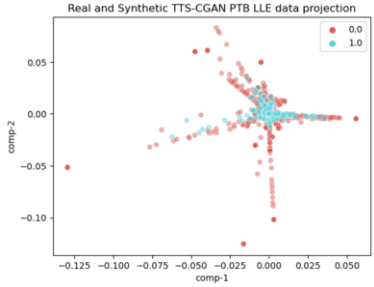


Figure 5. LLE fitted with the PTB dataset and TTS-CGAN generated data

As we can observe in figure 3, the data distribution generated by the Diffwave model trained on the MIT-BIH dataset almost completely overlaps the original data's distribution, informing us that the model learned to capture the underlying attributes of the data and is thus considered of good probabilistic quality. Furthermore, we can appreciate that the TTS-CGAN model struggled to do the same, with many gaps in the synthetic data's overlapped regions. However, we will show how this model could be of interest for augmenting datasets and pretraining classifiers. Furthermore, the LLE proved to be better at showing outliers, as the data was segmented into narrow shapes resembling lines in the reduced space (see figs 4, 5) and most of the data is found near the *center* of the shapes.

For a more practical evaluation, we have trained a deep residual convolutional neural network similar to the one presented in [9] but with only one residual block. The model has been trained and tested with the MIT-BIH and PTB datasets and has been evaluated through its performance on different classification scores, as presented in Table 1.

Next, with the goal of assessing how synthetic data can help in the training a Deep Learning model when data are scarce or highly unbalanced, we focus on the MIT-BIH dataset's minority classes (related to arrythmias), as there is an overwhelming prevalence of the healthy class compared to the others, making the dataset highly unbalanced. Once we have a trained model using a synthetic dataset consisting of 550k samples (110k from each class), we fine-tune it using different sizes of real data, starting with 20% of the real data until we use the whole dataset, with increments of 20%. Table 2 shows the evolution of the scores as more real data is used.

Another experiment performed to evaluate the model trained on synthetic data has been to fine-tune the model with the real data from the PTB dataset changing the task to myocardial infarction classification (binary classification). Similarly as before, we start with a pre-trained model trained with synthetic PTB data (112k observations for each class) then fine-tune the model adding increments of 20% of the real data in order to study the evolution of the classification scores. Results are shown in Table 3.

7. Results

On the basis of the previous results, we can conclude that the data generated by the TTS-CGAN model proved to be more useful and that the t-SNE plots were the most helpful measure of synthetic quality. For the experiments, we first evaluate the synthetic data and compare them with the real data, the results are shown in Table 1. These results

Table 1. Results from testing the models under different train and test combinations.

Dataset	Setting	Accuracy	Precision	Recall
MIT	**TSTR**	0.551	0.430	0.753
	TSTS	0.999	0.999	0.999
	TRTR	0.980	0.961	0.840
	TRSTR	0.971	0.933	0.783
PTB	**TSTR**	0.660	0.660	0.699
	TSTS	1.000	1.000	1.000
	TRTR	0.914	0.908	0.874

Table 2. Comparison table between trained model on synthetic data and fine-tuned on real data vs model trained with real data using different percentages of real data. The data used corresponds to the MIT dataset.

Percentage	Accuracy		Precision		Recall		Execution	
	Synthetic	*Real*	*Synthetic*	*Real*	*Synthetic*	*Real*	*Synthetic*	*Real*
0.2	0.9760	0.9664	0.9098	0.9153	0.8651	0.7700	33.5108	32.1254
0.4	0.9803	0.9741	0.9329	0.9476	0.8703	0.7854	50.4500	51.5347
0.6	0.9814	0.9763	0.9551	0.9586	0.8502	0.8074	65.3325	65.7416
0.8	0.9804	0.9783	0.9591	0.9599	0.8311	0.8162	81.9592	82.6904
1.0	0.9799	0.9751	0.9633	0.9612	0.8168	0.7920	100.0152	110.3127

Table 3. Comparison table between trained model on synthetic data and fine-tuned on real data vs model trained with real data using different percentages of real data. The data used corresponds to the PTB dataset.

Percentage	Accuracy		Precision		Recall		Execution	
	Synthetic	*Real*	*Synthetic*	*Real*	*Synthetic*	*Real*	*Synthetic*	*Real*
0.2	0.6627	0.7785	0.6810	0.7901	0.6684	0.7700	1.7065	7.3009
0.4	0.7888	0.8092	0.8166	0.8196	0.8052	0.7989	12.8148	16.2860
0.6	0.7961	0.8669	0.8131	0.8641	0.8234	0.8562	20.4001	26.8113
0.8	0.8248	0.8621	0.8187	0.8564	0.8504	0.8394	25.0002	34.6997
1.0	0.8221	0.8902	0.7885	0.8660	0.8352	0.8647	27.1476	42.3554

highlight that the synthetic data can be more easily classified (observations of each class are very similar, but very different from observations of other classes). Indeed, a classifier trained and tested with synthetic data has perfect performance, while models trained on synthetic data and tested on real data have very low performance. We added scores from training and testing on real data to give insight into the difficulty of the tasks. An insight to be drawn from this is that the generative models studied produce data maximizing the likelihood that it belongs to a specific class, as evidenced in figure 2 where the mass of the generated data sits well within the boundaries of the real data distribution. This could mean that the generated data is in a sense a *simplification* of the real data, but more work is to be done to find the actual explanation of this behavior.

For the second experiment, we assess the effect of the synthetic data when training a Deep Learning classification model when data are scarce or highly unbalanced. The results (see table 2) show that starting with a model trained with synthetic data provides better classification scores. This is of interest in particular for the case of *precision*

Table 4. Comparison table between trained model on synthetic MIT data and fine-tuned on real PTB data vs model trained with real PTB data using different percentages of real data.

Percentage	Accuracy		Precision		Recall		Execution	
	Synthetic	*Real*	*Synthetic*	*Real*	*Synthetic*	*Real*	*Synthetic*	*Real*
0.2	0.8362	0.7785	0.8551	0.7901	0.8441	0.7700	11.6971	7.3009
0.4	0.8926	0.8092	0.8999	0.8196	0.9036	0.7989	19.8847	16.2860
0.6	0.9452	0.8669	0.9396	0.8641	0.9507	0.8562	33.1194	26.8113
0.8	0.9466	0.8621	0.9359	0.8564	0.9542	0.8394	39.0200	34.6997
1.0	0.9641	0.8902	0.9490	0.8660	0.9655	0.8647	46.4486	42.3554

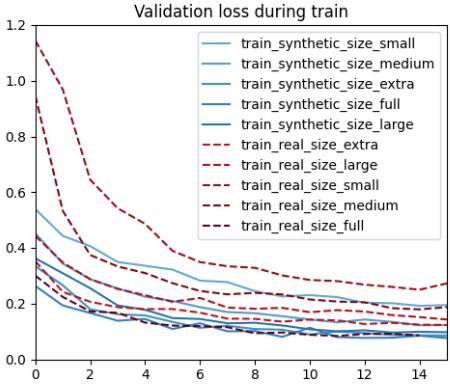


Figure 6. Loss curves on validation set of the fine tuned synthetic model (blue solid lines) and the real data model (red dashed lines) trained under different sizes of the MIT-BIH dataset. Lines become darker as more data has been used

and *recall*, which provide information about the biases in the classifier resulting from class imbalance. While *precision* is nearly the same for the fine-tuned model and the one trained with real data, it is clear that *precision* is better in the fine-tuned model, which improves the final accuracy. Furthermore, although the loss curves (Figure 6) show that models trained with synthetic data start with less loss, we can observe in the second-to-last column (time needed to train the model in seconds) that it is unclear whether the training time is influenced by fine-tuning the model. It is clear that as more data becomes available, less time is needed to train a fine-tuned model. Finally, observe that the scores in both models become more similar as the percentage of real data used increases, probably as an effect of the model becoming accustomed to the real data and *forgetting* patterns from the synthetic data.

The same experiment has been performed with the PTB dataset, where we pre-train the same classification model on PTB synthetic data and then fine-tune it with different sizes of real data from the same dataset. The results (see table 3), show that the synthetic data does not reflect the patterns of the real data with either model. This could be caused by the small amount of data available in the PTB dataset (roughly 1/10th the size of the MIT-BIH dataset), and the difficulty to train a generative model on such a scarce dataset.

Moreover, the third and last experiments consisted in using the classification model pre-trained with synthetic data for the arrythmia detection task (MIT-BIH dataset) and then fine-tuning it in order to solve the myocardial infarction detection task (PTB dataset). Results are shown in table 4, and it can be seen that the fine-tuned model has better overall scores with a considerable margin compared to the classifier trained only on real data. This indicates two things: arrhythmia detection and the myocardial infarction task have similar patterns. Additionally, the synthetic data generated using the MIT-BIH dataset are good enough to capture the patterns that are useful for both tasks. Therefore, in this case of data scarcity, having more data through synthetic generation has significantly improved the baseline.

8. Conclusions and future work

Our results show that synthetic data generation can be useful when classifying a scarce dataset or one with an overwhelming minority class, especially when the task highly depends on the data quantity and quality. This work also shows the need for a better synthetic time series evaluation method, as the t-SNE and LLE plots do not give all the probabilistic insight needed and the classification experiment does not tell us if we are truly capturing the patterns in the distribution of physiological time series.

As future work, it would be of interest to perform new experiments on the recently released PTB-XL data set [15], one of the largest clinical data sets focused on ECG signals, as well as to work with other physiological time series such as EEGs and EMGs. Also of great interest to us is to compare the presented generative models with newer ones coming from the diffusion family or those which incorporate attributes from different families of models. Lastly, we want to find other methods to evaluate the generative quality of these models (e.g. UMAP looks promising) [16].

Finally, beyond physiological time-series data, new generative models for images based on diffusion techniques can also be used to enhance classifiers related to aging-related illnesses, for instance tasks like colon polyp detection on colonoscopies, detection of prostate cancer using MRI scans, histopathology analysis, etc. which are tasks related to our current project as well.

9. Acknowledgments

This research has been financed by the Artificial Intelligence for Healthy Aging (AI4HA, MIA.2021.M02.0007) project from the Programa Misiones de I+D en Inteligencia Artificial 2021.

References

[1] Talavera E, Iglesias G, González-Prieto Á, Mozo A, Gómez-Canaval S. Data Augmentation techniques in time series domain: A survey and taxonomy. arXiv preprint arXiv:220613508. 2022.

[2] Bond-Taylor S, Leach A, Long Y, Willcocks CG. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. IEEE transactions on pattern analysis and machine intelligence. 2021.

[3] Yang L, Zhang Z, Hong S. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv preprint arXiv:220900796. 2022.
[4] Gui J, Sun Z, Wen Y, Tao D, Ye J. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. IEEE Transactions on Knowledge and Data Engineering. 2023;35(4):3313-32.
[5] Li Z, Ma C, Shi X, Zhang D, Li W, Wu L. TSA-GAN: A Robust Generative Adversarial Networks for Time Series Augmentation. In: IJCNN. IEEE; 2021. p. 1-8.
[6] Isola P, Zhu J, Zhou T, Efros AA. Image-to-Image Translation with Conditional Adversarial Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society; 2017. p. 5967-76. Available from: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.632.
[7] Li X, Ngu AHH, Metsis V. TTS-CGAN: A Transformer Time-Series Conditional GAN for Biosignal Data Augmentation. arXiv preprint arXiv:220613676. 2022.
[8] Kong Z, Ping W, Huang J, Zhao K, Catanzaro B. Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:200909761. 2020.
[9] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine. 2001;20(3):45-50.
[10] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik/Biomedical Engineering. 1995.
[11] Oeff M, Koch H, Bousseljot R, Kreiseler D. The PTB diagnostic ECG database. National Metrology Institute of Germany. 2012.
[12] Kachuee M, Fazeli S, Sarrafzadeh M. ECG Heartbeat Classification: A Deep Transferable Representation. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI); 2018. p. 443-4.
[13] van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9:2579-605. Available from: http://www.jmlr.org/papers/v9/vandermaaten08a.html.
[14] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. science. 2000;290(5500):2323-6.
[15] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data. 2020 May;7(1):154. Available from: https://doi.org/10.1038/s41597-020-0495-6.
[16] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. The open journal. 2018. Cite arxiv:1802.03426Comment: Reference implementation available at http://github.com/lmcinnes/umap. Available from: http://arxiv.org/abs/1802.03426.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230657

Stacking Up for Success: A Cascade Network Model for Efficient Road Crack Segmentation

Ammar M. OKRAN [a,1], Adel SALEH [b], Domenec PUIG [a] and Hatem A. RASHWAN [a]
[a] *DEIM, Rovira i Virgili University, 43007 Tarragona, Spain*
[b] *Gaist Solutions Ltd, U.K*
ORCiD ID: Ammar M. Okran https://orcid.org/0000-0001-9264-0301, Adel Saleh https://orcid.org/0000-0001-5502-100X, Domenec Puig https://orcid.org/0000-0002-0562-4205, Hatem A. Rashwan https://orcid.org/0000-0001-5421-1637

Abstract. This paper proposes an integrated framework for automatically segmenting road surface cracks that utilize a Multi-Attention-Network and a modified U-Net, combined through neural network stacking, to segment the crack regions accurately. To evaluate the effectiveness of the proposed framework, we introduce a road crack dataset containing complex environmental noise. We explore several stacking scenarios and perform thorough evaluations to assess the performance of the proposed model. Our results show that the proposed method improves the IOU score of 1.5% compared to the original network, indicating its effectiveness in segmenting road cracks. The proposed framework can be a valuable tool for road maintenance and inspection, enabling timely detection and repair of cracks and improving road safety and longevity. Our findings demonstrate the importance of exploring various stacking scenarios and performing comprehensive evaluations to establish the efficacy of the proposed framework.

Keywords. Road Cracks, Deep learning, Semantic segmentation, Cascade network

1. Introduction

Road networks are vital for society, but maintaining them is challenging. Cracks on road surfaces are a common problem, leading to accidents and economic losses. Spain had over 1,000 fatal accidents in 2022, many caused by road cracks. Factors like weather and traffic contribute to crack formation. Prompt repair using methods like asphalt or concrete filling is crucial for road safety[2].

These accidents caused human losses and inflicted considerable material damages and direct property destruction. Consequently, the Spanish government has allocated a

[1]Corresponding Author: Ammar M. Okran, ammar.okran@urv.cat.
[2]Directorate-General for Traffic (DGT): https://www.dgt.es/comunicacion/notas-de-prensa/1.145-personas-fallecieron-en-siniestros-de-trafico-durante-2022

significant budget of 2.75 billion euros for 2023 to address road construction, maintenance, and repairs, as reported by the Spanish government [3].

Accurate and efficient identification of road cracks is crucial for effective maintenance. Traditional visual inspections are subjective and time-consuming. Deep learning (DL) technologies, like CNNs, have emerged as essential tools for crack detection, such as Yolov7 [1] and semantic segmentation, UNet [2]. They can accurately identify and segment cracks by extracting high-level features from images. DL-based algorithms have improved crack detection performance, with solutions presented in the CRDDC'2022 [3] using one-stage detectors and ensembling approaches [4,5]. Ensembles of one-stage and two-stage detectors [6], have also been proposed. These advancements provide more efficient and accurate solutions for automating crack detection and characterization.

However, road crack detection may not be the most efficient method when it comes to cost calculation. Accurate cost estimation for road maintenance and repair requires precise segmentation of road cracks to determine their size and area, which provides information on the extent of damage, needed repairs, and associated material and labor costs [7,8]. Consequently, several studies have proposed various methods for crack segmentation. For instance, DeepCrack [9] employs an auto-encoder comprising encoder and decoder networks similar to SegNet [10] that the encoder and decoder networks generate convolutional features at the same scale, which are pairwise fused to obtain the final feature representations. In turn, the authors of [11] proposed using gated skip connections that enable the decoder layers to selectively incorporate crack-aware feature representations from the encoder layers. The gating mechanism assigns higher weights to crack-relevant features from the encoder layers and lower weights to irrelevant features.

The existing methods for crack segmentation can encounter challenges due to factors like illumination changes, texture issues, noise, occlusions, and complex object structures. Cascading and stacking DL networks offer a suitable solution by extracting hierarchical features [12], integrating contextual information [13] and improving robustness [14]. This approach ensures accurate and reliable segmentation results [15,16]. In the cascaded or stacked networks architecture, lower-level features capture fine-grained crack details and local information, while higher-level features capture abstract and global contextual information [17]. This hierarchical feature extraction aids in precise object boundary delineation and semantic understanding while handling noise, occlusions, and complex structures. Cascading and stacking DL networks enhance the robustness of crack segmentation systems [17].

The performance of a cascaded DL model heavily relies on the design choices, such as the number of stages, the architecture of each step, and the information flow between stages [18]. Finding an optimal design configuration can be challenging and require extensive experimentation, especially with road crack segmentation and its challenges. Thus, careful experimentation and validation involve finding the right balance regarding model capacity, receptive field size, and complexity for each stage [18]. Iterative refinement and tuning of the cascade architecture are often required to achieve the best trade-off between localization accuracy and segmentation refinement [18]. Consequently, this paper proposes a road crack segmentation model with two main contributions.

[3] Spanish Government: https://www.sepg.pap.hacienda.gob.es/sitios/sepg/es-ES/Presupuestos/PGE/ProyectoPGE2023/Documents/LIBROAMARILLO2023.pdf

- Introducing a new road crack segmentation dataset derived from the RDD2022 dataset, widely used for road damage detection assessment. This dataset provides a comprehensive and diverse set of crack images for training and evaluating crack segmentation models. The new dataset will provide a more diverse and challenging set of crack images for training and evaluating crack segmentation models.
- Proposing a novel cascade/stack network architecture is proposed for accurate road crack segmentation. It utilizes skip and dense connections to fuse features across layers, improving efficiency. Ensemble learning combines different models for enhanced performance and robustness in crack detection.

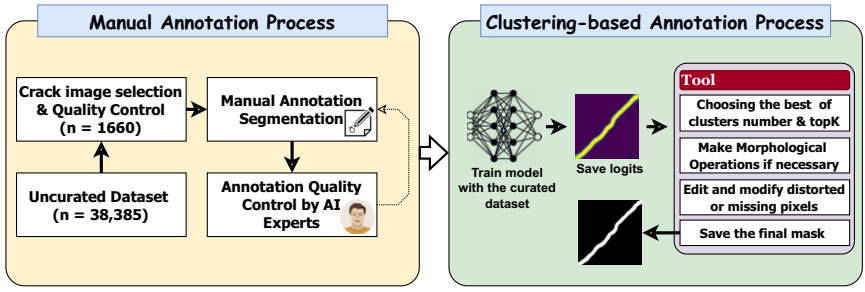


Figure 1. The two stages of the dataset construction process.

The structure of this work is as follows: Section 2 provides a detailed explanation of the dataset construction and the proposed road crack segmentation system. Section 3 presents the experimental results obtained from our proposed method. Finally, Section 4 concludes the paper by summarizing the findings and suggesting potential future research directions.

2. Methodology

The study methodology includes two subsections: dataset construction and road crack segmentation method. The dataset creation process and labeling procedure are detailed in the first subsection. The second subsection presents the innovative DL-based road crack segmentation method using a cascade network architecture. It covers components like the loss function and training procedure. The aim is to clearly understand the methodology, dataset, and proposed approach for road crack segmentation.

2.1. Road Crack Segmentation Dataset

The available road crack datasets often rely on manual annotations, which can introduce annotation variations and subjectivity. Different annotators may interpret crack boundaries differently, resulting in inconsistencies in the ground truth annotations. This variability can affect the training and evaluation of models and impact their performance. Thus, in this work, we implement standardized annotation protocols merging two stages of manual and automatic annotations to reduce annotation variations and improve the consistency of ground truth labels. Regular quality control checks and inter-annotator

agreement assessments can ensure higher annotation accuracy and reliability. The process of constructing the road crack segmentation dataset consists of two stages, an extension of our previous work [19], illustrated in Figure 1.

In the first stage, 1,660 sub-images were selected from the RDD2022 dataset [20], which includes 38,385 training images from diverse acquisition systems in six countries, i.e., Japan, the Czech Republic, India, the United States of America, Norway, and China. The selection process involved careful review to ensure acceptable image quality. Rejected images, such as those that were blurry or dark, were excluded. An experienced engineer provided manual annotations for each crack in the dataset to ensure accuracy and consistency. AI experts conducted quality control to identify errors or inconsistencies in the annotations. The engineer re-evaluated and manually corrected any rejected annotations to ensure a high-quality final dataset.

We trained a DL model in the second stage using the UNet [2] and DeepLabv3+ [14] networks. The best checkpoint was selected and saved for further use. Subsequently, the model was used to generate logit matrices for each new image after applying the sigmoid function, which produced values ranging from $[0, 1]$. The critical stage of the process involved filtering the results and refining them to create masks that closely resembled those generated manually. Thus, we cluster the segmented regions related to the crack for each image based on the K-means algorithm. We then select the optimal number of clusters by choosing the top-K clusters produced masks that closely resembled the manual annotations. As a postprocessing stage, we applied morphological operations as necessary and manually edited and refined the masks to ensure the highest level of accuracy. Finally, the masks were saved, completing the dataset creation process.

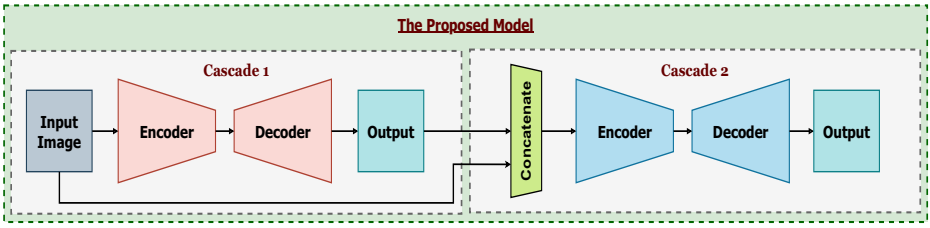


Figure 2. The proposed cascaded model.

In total, we obtained $6,246$ images. The training set comprised $5,041$ images taken from the training set of the RDD2022 dataset, while the testing set contained $1,205$ images taken from the test set of RDD2022. Each image is resized to 128×128.

2.2. *Road Crack Segmentation Method Using a Cascade Network Architecture*

This subsection outlines the methodology employed in our proposed model, which consists of two cascaded networks. The overall model architecture is illustrated in Figure 2. The first cascade network utilizes a Multi-Scale Attention Network (MA-Net) [21], a modular architecture designed to leverage the attention mechanism for capturing rich contextual dependencies. This network is instrumental in capturing comprehensive spatial information and contextual relationships. In turn, the second cascade network in our model adopts a Deeper-UNet architecture. It is fed by concatenating the output from the previous cascade network and the original image as input. The Deeper-UNet architecture is well-suited for capturing fine-grained details and refining the segmentation results.

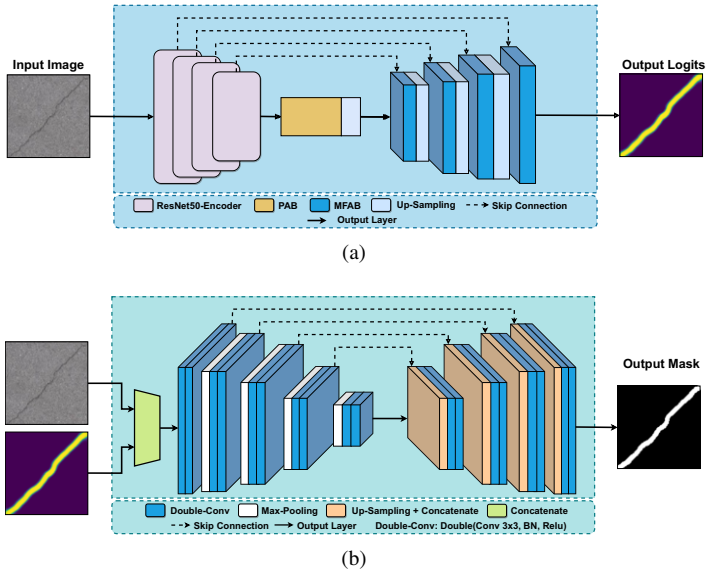


Figure 3. The architecture of the first (a) and the second (b) stages of the cascaded model.

2.2.1. ***Cascade 1:*** *The Multi-Scale Attention Network (MA-Net)*

We tested different state-of-the-art segmentation models to select the best model for Cascade 1. Among them, the MANet yields the best results for road crack segmentation. The MANet (cascade 1) [21], incorporates a ResNet-50 [22] encoder as its backbone, as illustrated in Fig 3-(a). The choice of ResNet-50 as the backbone is motivated by its reputation as a state-of-the-art architecture for different image segmentation tasks. The ResNet-50 network is particularly advantageous due to its implementation of residual connections, which address the issue of gradient vanishing. These connections allow the output from a preceding layer to be directly added to a subsequent layer, facilitating the flow of gradients during training. By mitigating the gradient vanishing problem, ResNet-50 enables the efficient propagation of information through the network, extracting meaningful features from the road crack images. The ResNet-50 in our model can encode the input road crack image and extract significant and abstract features at various scales or levels of detail. Cascade 1 also depends on two attention networks for accurate road crack segmentation.

Position-Wise Attention Block (PAB) To capture spatial dependencies between any two positions of feature maps, we employ a PAB [23] at the neck of our network. This approach enhances the model's capability to incorporate broader spatial contextual information across local feature maps. By leveraging the PAB, our network becomes more adept at capturing relevant features for road crack segmentation. This means the PAB is vital in enabling the network to effectively understand the spatial relationships and dependencies within the road crack images, facilitating accurate and precise segmentation.

Multi-Scale Fusion Attention Block (MFAB) Our model incorporates the Multi-MFAB from [21] to capture interdependencies among feature channels. Inspired by the human visual system, MFAB selects salient information for accurate crack segmentation without introducing spatial dimensions. This mechanism enhances useful feature maps while suppressing less relevant ones, improving segmentation performance. By employ-

ing MFAB, our model effectively captures contextual dependencies, facilitating precise crack segmentation. The Multi-MFAB is vital in identifying and prioritizing relevant information across scales, leading to reliable and accurate crack segmentation results.

Cascade 1 comprises four encoder blocks from ResNet-50 and four MFAB blocks in the decoder. The MFAB blocks receive feature maps from previous MFAB blocks, PAB blocks, or corresponding ResNet encoder layers. These feature maps include low-level features from skip connections and high-level features from the previous MFAB or PAB blocks. High-level feature maps undergo 3x3 and 1x1 convolutions, Batch Normalization (BatchNorm2D), and ReLU activation. The resulting feature maps are upsampled and passed through a SE-Block to capture high-level attention [24]. The SE-Block includes Adaptive Average Pooling, Convolutional layers, ReLU, Convolutional with a 1x1 kernel, and Sigmoid activation. This combination of layers and blocks in Cascade 1 effectively integrates low-level and high-level features, captures spatial dependencies, and utilizes attention mechanisms for refined and improved feature maps, leading to accurate and robust road crack segmentation.

Additionally, we incorporate another SE-Block to process the low-level feature maps and obtain the low-level attention feature maps. Following this, we perform an element-wise addition of the attention feature maps produced by the SE-Blocks. The resulting feature map is then multiplied by the high-level up-sampled feature maps [21] and concatenated with the low-level feature maps obtained from the Skip-Connection. We apply two successive 3x3 convolutional layers, each followed by Batch Normalization (BatchNorm2D) and Rectified Linear Unit (ReLU) activation, to refine further and enhance the feature maps. These layers can capture meaningful and semantically rich information, allowing for the extraction of more discriminative features.

By incorporating these operations within the MFAB block, our model effectively integrates attention-based mechanisms for low- and high-level feature maps. This enables the network to emphasize relevant information, refine feature representations, and generate more precise crack segmentation results.

2.2.2. ***Cascade 2:*** *The Deeper-UNet Network*

The proposed cascaded model leverages the Deeper-UNet architecture as Cascade 2, as shown in Fig 3-(b). Deeper-UNet is a popular DL model for image segmentation tasks. It is an extension of the original UNet model [2] consisting of a deep encoder network followed by a decoder network with skip connections between corresponding layers.

The Deeper U-Net architecture consists of an input layer followed by a series of convolutional layers, each responsible for extracting increasingly complex features from the input image. The output of each convolutional layer is then passed to a pooling layer, which reduces the spatial dimension of the feature maps. This is followed by convolutional layers that upsample the feature maps and restore the original spatial dimension. The output of these convolutional layers is then concatenated with the corresponding feature maps from the encoder network to form the skip connections.

In our model, the Deeper-UNet network accepts a 4-channel image as input. We create the 4-channel image by concatenating the logit output of Cascade 1 with the original image. This concatenated image serves as the input for the Deeper-UNet model. The input passes through convolutional layers, doubling the number of channels at each layer until reaching 512 channels. Skip connections are used by upsampling and concatenating feature maps from the encoder with corresponding decoder feature maps. These

connections integrate local and global information for more accurate segmentation. The model's output is obtained through a 1x1 convolutional layer, reducing channels to 2 classes (background and defect) and generating the segmentation map for road cracks.

3. Experiments and Results

This section describes the experiments conducted to assess the efficacy of the proposed model. It includes details of the experimental setup, the evaluation metrics employed, and a comprehensive analysis of the resulting outcomes.

We used our constructed dataset for our experiments, which consisted of 6,246 images. The training set comprised 5,041 images, each with a size of 128 x 128, and the testing set contained 1,205 images. In the training pipeline, we utilized augmentation techniques such as horizontal and vertical flipping, rotation, and random brightness contrast for both Cascade 1 and Cascade 2 models. These techniques increased data diversity and improved the models' generalization ability. The models were trained with the Adam optimizer, using a batch size 32 and learning rates of 0.0001 and 0.001 for Cascade 1 and 2, respectively. Cascade 1 employed the binary cross-entropy loss function, while the cross-entropy loss function was used for Cascade 2. These training measures aimed to ensure accurate segmentation while maintaining robust and practical training for the models. The proposed model was evaluated in this work using various evaluation metrics, which included Intersection over Union (IoU), F1-score, Accuracy, and Precision.

In this subsection, we present the experimental results of our road crack segmentation models. First, we compared our proposed models with the baseline models, including DeepLabv3 [13], DeepLabv3+ [14], UNet [2], UNet++ [25], FPN [26], LinkNet [27], PAN [28], PSPNet [29], MANet [21], and Deeper-UNet. All networks used ResNet-50 as a backbone, except Deeper-UNet.

Table 1. Performance of the baseline segmentation models on the curated dataset.

Method	Backbone	IoU	F1-score	Accuracy	Precision
DeepLabv3	ResNet50	0.77640	0.86835	0.91499	0.80596
DeepLabv3+	ResNet50	0.78903	0.87605	0.91896	0.82091
UNet	ResNet50	0.79403	0.87949	0.91996	0.83489
UNet++	ResNet50	0.78430	0.87335	0.91743	0.81934
FPN	ResNet50	0.78732	0.87462	0.91630	0.82627
LinkNet	ResNet50	0.7921	0.87836	0.91680	0.83809
PAN	ResNet50	0.78286	0.87174	0.91614	0.82293
PSPNet	ResNet50	0.77707	0.86867	0.91285	0.81
MANet	ResNet50	**0.80252**	**0.88537**	0.9222	**0.84377**
Deeper-UNet	CNN	0.79639	0.88107	**0.93073**	0.83237

Table 1 shows the results of the baseline models as Cascade 1, where we evaluated the performance of each model in terms of IOU score, F1-score, Accuracy, and Precision. As shown in Table 1, MANet achieved the highest IOU score of 0.802, the highest F1-score of 0.885, and the highest Precision of 0.844 among the ten baselines tested. While Deeper-UNet achieved the highest Accuracy of 0.931. Additionally, we observed that some models, such as LinkNet and PSPNet, performed relatively poorly compared to

other models. The results showed the effectiveness of MANet in accurately capturing the extent of road cracks.

In the second experiment, we froze the weights of Cascade 1 and added Cascade 2 as the Deeper-UNet network to train the stacked network. The Deeper-UNet architecture incorporates an expanded network depth compared to the traditional UNet model with multiple stacked convolutional layers that capture more detailed and high-level representations, enhancing the overall segmentation performance. Table 2 shows the results of the proposed models combined with the nine networks of Cascade 1 mentioned above and the Deeper-UNet as a Cascade 2. As shown in Table 2, the proposed model with MANet, combined with Deeper-UNet, achieved the highest IOU score of 0.817, F1-score of 0.895, and Precision of 0.888. However, DeepLabv3 with Deeper-UNet achieved the highest Accuracy of 0.939. The cascaded models yielded more improvement in the four evaluation metrics than the one-stage segmentation models of the first experiment, with an improvement of 1.5% in terms of IOU.

Table 2. The results of the proposed models combined with the nine networks of Cascade 1.

Method1	Method2	IoU	F1-score	Accuracy	Precision
DeepLabv3	**Deeper-UNet**	0.80541	0.88784	**0.93935**	0.86173
DeepLabv3+	**Deeper-UNet**	0.81255	0.89165	0.93352	0.87491
UNet	**Deeper-UNet**	0.81543	0.89359	0.92403	0.88769
UNet++	**Deeper-UNet**	0.81358	0.89259	0.93171	0.87767
FPN	**Deeper-UNet**	0.80928	0.88926	0.93057	0.87473
LinkNet	**Deeper-UNet**	0.80779	0.88892	0.92753	0.87583
PAN	**Deeper-UNet**	0.80443	0.88615	0.93142	0.86879
PSPNet	**Deeper-UNet**	0.80180	0.88569	0.91994	0.87651
MANet	**Deeper-UNet**	**0.81709**	**0.89538**	0.92480	**0.88818**

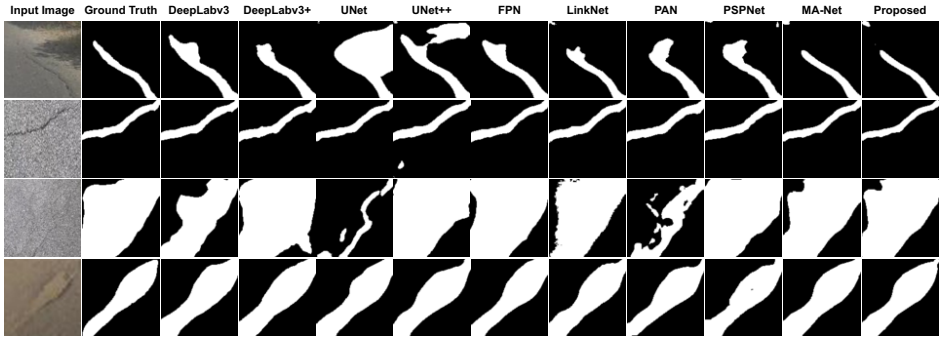


Figure 4. Segmentation results of different models. Each row presents a different crack—row1: Longitudinal Crack, row2: Transverse Crack, row3: Alligator Crack, and row4: Pothole Damage.

The experimental results depicted in Figure 4 demonstrate the effectiveness of our proposed cascaded models for road crack segmentation. The cascaded models consistently outperformed the one-stage models, showcasing their ability to accurately identify and segment road cracks. A quantitative evaluation of the results further supports this claim. Figure 4 visually compares the cascaded models with the ground truth, reveal-

ing that our approach generates segmentation results closely resembling the annotated ground truth. This indicates that our cascaded models capture fine details, handle complex crack structures, and produce accurate segmentation maps. Overall, these findings underscore the significance of cascading the MANet and Deeper-UNet networks, as it enables the integration of complementary features and contextual information from multiple stages, leading to improved segmentation performance. By leveraging the strengths of each model and incorporating them in a cascaded manner, our approach achieves superior results in road crack segmentation tasks.

4. Conclusion and Future work

In this work, we introduced a novel road crack segmentation dataset and developed a cascade network architecture designed explicitly for road crack segmentation. We conducted comprehensive experiments using our newly constructed dataset and compared our results with state-of-the-art methods in the field. Our proposed approach outperformed existing methods in terms of various evaluation metrics, including IoU, F1-score, Accuracy, and Precision, with an improvement of 1.5%, 1%, 0.8%, and 4.5%, respectively. These results provide that the cascade network architecture can achieve notable advancements in road crack segmentation performance. Future work will develop a comprehensive road crack detection merging crack multi-class identification and segmentation.

Acknowledgements

Diputació de Tarragona and Universitat Rovira i Virgili partly supported this study through Projecte d'Especialització i Competitivitat Territorial (PECT).

References

[1] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[3] Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, Hiroshi Omata, Takehiro Kashiyama, and Yoshihide Sekimoto. Crowdsensing-based road damage detection challenge (crddc-2022). *arXiv preprint arXiv:2211.11362*, 2022.

[4] Dongjun Jeong and Jua Kim. Road damage detection using yolo with image tiling about multi-source images. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6401–6406. IEEE, 2022.

[5] Ammar M Okran, Mohamed Abdel-Nasser, Hatem A Rashwan, and Domenec Puig. Effective deep learning-based ensemble model for road crack detection. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6407–6415. IEEE, 2022.

[6] Wenchao Ding, Xu Zhao, Bingke Zhu, Yinglong Du, Guibo Zhu, Tao Yu, Lei Li, and Jinqiao Wang. An ensemble of one-stage and two-stage detectors approach for road damage detection. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6395–6400. IEEE, 2022.

[7] Wadea Sindi and Bismark Agbelie. Assignments of pavement treatment options: Genetic algorithms versus mixed-integer programming. *Journal of Transportation Engineering, Part B: Pavements*, 146(2):04020008, 2020.

[8] S Jana, S Thangam, Anem Kishore, Venkata Sai Kumar, and Saddapalli Vandana. Transfer learning based deep convolutional neural network model for pavement crack detection from images. *International Journal of Nonlinear Analysis and Applications*, 13(1):1209–1223, 2022.

[9] Yahui Liu, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing*, 338:139–153, 2019.

[10] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[11] M Jabreel and M Abdel-Nasser. Promising crack segmentation method based on gated skip connection. *Electronics Letters*, 56(10):493–495, 2020.

[12] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2359–2367, 2017.

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[15] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[16] Li Deng, Xiaodong He, and Ji Gao. Deep stacking networks for information retrieval. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3153–3157. IEEE, 2013.

[17] Yu-Cheng Liu, Mohammad Shahid, Wannaporn Sarapugdi, Yong-Xiang Lin, Jyh-Cheng Chen, and Kai-Lung Hua. Cascaded atrous dual attention u-net for tumor segmentation. *Multimedia tools and applications*, 80:30007–30031, 2021.

[18] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[19] Ammar M Okran, Mohamed Abdel-Nasser, Hatem A Rashwan, and Domenec Puig. A curated dataset for crack image analysis: Experimental verification and future perspectives. 2022.

[20] D Arya, Hiroya Maeda, Yoshihide Sekimoto, Hiroshi Omata, Sanjay Kumar Ghosh, Durga Toshniwal, Madhavendra Sharma, Van Vung Pham, Jingtao Zhong, Muneer Al-Hammadi, Mamoona Birkhez Shami, Du Nguyen, Hanglin Cheng, Jing Zhang, Alex Klein-Paste, Helge Mork, Frank Lindseth, Toshikazu Seto, Alexander Mraz, and Takehiro Kashiyama. Rdd2022 - the multi-national road damage dataset released through crddc'2022, 2022.

[21] Tongle Fan, Guanglei Wang, Yan Li, and Hongrui Wang. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8:179656–179665, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE CVPR*, pages 770–778, 2016.

[23] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3141–3149, 2019.

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[25] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[26] Alexander Kirillov, Kaiming He, Ross Girshick, and Piotr Dollár. A unified architecture for instance and semantic segmentation, 2017.

[27] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017.

[28] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[29] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230658

Internal Representation and Shepard's Law in an Artificial Haptic System

Claudia SERRANO [1] and David MIRALLES
Human-Environment Research (HER), La Salle - Universitat Ramon Llull, Barcelona, Catalonia, Spain
ORCiD ID: Claudia Serrano https://orcid.org/0000-0003-4798-4142, David Miralles https://orcid.org/0000-0001-7127-199X

Abstract. In nature, the generalization in object recognition is based not in the stimuli space but the internal representation as Shepard proposed in his Universal Law of Generalization. The fact that this law applies to many biological system gives it its universal character. The extension of this universality to artificial systems remains to be studied. In this paper, we present how an artificial agent generates the internal representation of a collection of stimuli from a previous experience. Moreover, from the internal representation and the classification of the artificial agent, we are already able to verify Shepard's law. It should be noted that the presented methodology is independent of the classifier. In our case, we have applied it to an artificial system that captures haptic information from a collection of stimuli. We have verified that the relationship between the perceived distances in the internal representation and the probability of confusion between stimuli follows Shepard's law. Verifying compliance with this law in artificial systems and studying its implications can be relevant to understanding generalization in learning.

Keywords.
Artificial haptic system, Internal representation, Sheppard's Law, Generalization

1. Introduction

The intersection of AI and psychology is significant because understanding how humans think and learn can be inspirational for reproducing human-like behaviour. Generalization, the ability to apply learned knowledge and skills in new and different contexts, is an essential aspect of human learning. In this context, the cognitive psychologist Roger Shepard developed a mathematical model to describe the generalization of learning in 1987, now known as Shepard's Universal Law of Generalization [1].

To generalize, the model proposed by Shepard relates the probability of confusion between two stimuli to the distance between them, which he called the psychological space. Formally, given two items A and B, $P(R_B|S_A)$ is the probability of responding B (R_B) when the stimulus A (S_A) is exposed, and $d(A,B)$ is the distance in psychological space [2]. This psychological space, also known as an internal representation, is a mental construction that allows individuals to represent the external world and to use this repre-

[1]Corresponding Author: Claudia Serrano, c.serrano@salle.url.edu

sentation to guide their behaviour. The result he obtained relating these two characteristics was:

$$P(R_B|S_A) \propto e^{-d(A,B)} \tag{1}$$

This law was tested in numerous experiments on humans and animals , proving its universality (see Related Work). In these experiments, Shepard also tested the independence of the law in relation to the properties of the stimuli given for generalization. He found that the law holds regardless of the specific characteristics tested, suggesting that it holds in different contexts and domains. Surprisingly, it has recently been shown [3] that the origin of this law lies in an internal measurement invariance.

The aim of this paper is to create an internal representation and test the Universal Law of Generalization for an artificial agent. Our proposal is similar to the way humans and animals have their own internal representations through prior learning experiences. Following Shepard's approach and using a baseline as a reference, we investigate whether artificial agents are able to generalize in the same way as humans and whether they follow Shepard's Law. By exploring generalization in AI, we can gain a better understanding of how artificial agents learn and how they can be developed to behave more like humans.

2. Related Work

As we have already explained in the introduction, the generalisation law is mainly based on two concepts: a metric distance between stimuli and the probability of confusion between them.

The concept of psychological space or internal representation, first introduced by Shepard in [4], is crucial to understanding the similarity distance between stimuli. Shepard argued that this representation is not a passive copy of the external world, but an active, resonant and dynamic process influenced by the ecological constraints of the environment and the organism's interactions with it. Shepard used confusion matrices previously obtained to implement his first approaches to the generalization model [5]. For example, in a perceptual study [6] undergraduate students were presented with nine circles of different diameters for four seconds. Subsequently, given each stimulus, all possible confusion probabilities were determined.

Additionally, Shepard conducted numerous experiments to test his law of generalization on both humans and animals, varying the characteristics on which participants focused. For humans, these experiments included testing generalization over differences in size and lightness of geometric figures [6,8,9], spectral hues [10], vowel and consonant phonemes [11,12], and Morse code signals [13]. Meanwhile, experiments on animals focused on generalization from differences in spectral hues in pigeons were also tested [14,15].

While editing this paper, we came across another study that had similar goals of testing whether an artificial system follows the generalization law proposed by Shepard [7]. This study used a image data set and a neural network classification to obtain a confusion matrix, which was then used to obtain an internal representation through dimensionality reduction. Their results indicate that with this approach, the relationship between probability and distance between images follows a decreasing monotonic function.

After an extensive review of the literature on perception and generalization in artificial systems, the results of our proposed method show for the first time that the relationship between probability and distance not only decreases monotonically, but in particular follows an exponential decay, which is consistent with the Universal Law of Generalization proposed by Shepard.

3. Haptic System and Stimuli

In this study, we use an artificial haptic capture system that captures raw curvatures from different 3D stimuli in a virtual environment, described in [16]. To capture local curvature information, we employ three equally spaced end-effectors that collide with the surface of the stimulus, obtaining three relative time values with respect to the first capture. We obtain a local curvature value by approximating the second derivative of these time values. In this study, our focus is on object generalization rather than object recognition. In this way, we want to associate the captured curvatures with features that describe our stimuli.

Our stimuli are 3D virtual figures, specifically hexahedrons with variable curvatures ranging from concave to convex through the flat surface. These figures were generated using Gielis' parametric model, which enables the creation of 3D objects by adjusting different geometric parameters to its Superformula [17]. This system has been used in previous studies with similar aims [18]. The resulting objects are also known as Supershapes.

In order to study object generalization, we have developed a set of 21 stimuli which we refer to as Lats. These stimuli are divided into two subsets (see Figure 1). The first set, called Baseline (BL), serves as a reference for our perception system. This BL is the set of stimuli that are part of the system's prior perceptual experience and allows to represent the reality it will interact with. The second set, called Unknown (UK), includes unknown figures that we aim to generate an internal representation of. By analyzing the relationships between stimuli in both sets, we aim to generate an internal representation of the unknown figures (UK) based on their relationships with the stimuli in the baseline set (BL). Specifically, we examine how the unknown stimuli relate to each other through their relationship with the baseline stimuli.

4. Classification and Confusion

In Section 1, we discuss the importance of considering both probability and similarity between objects when studying generalization. In this section, we focus on obtaining these probabilities by calculating the probability of response R_B given a stimulus S_A, i.e. $P(R_B|S_A)$, for any two stimuli A and B from the UK figures set. These probabilities are commonly organized in a confusion matrix. To construct this confusion matrix, we first collect a dataset of raw curvatures from our stimuli using our haptic capture system. Next, we train and test an XGBoost classifier for recognition purposes. For a detailed explanation of our XGBoost model, including the parameters used for training, please refer to the following paper [16]. To gain a deeper understanding of the relationships between stimuli, we intentionally limited the number of test samples to just 5 haptic samples, in-

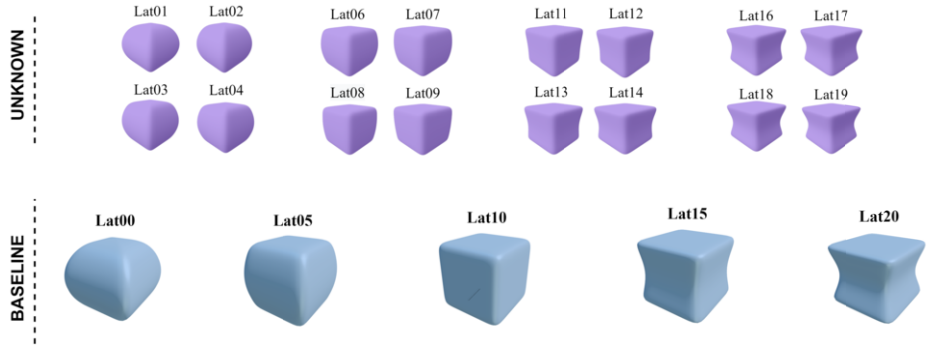


Figure 1. Complete set of 21 stimuli used in our study, which we refer to collectively as Lats. These stimuli were generated using Gielis' parametric model which allows us to adjust various geometric parameters to create 3D objects. The stimuli are split into two sets: a reference set, or baseline (BL), consisting of five stimuli (Lat00, Lat05, Lat10, Lat15, and Lat20) shown at the top of the figure in blue, and an unknown set (UK), consisting of 16 stimuli (Lat01, Lat02, Lat03, and so on, up to Lat19) shown at the bottom of the figure. Each stimulus is named Lat (from latitude) followed by its latitude parameter value in the Superformula (e.g., Lat05 has a parameter value of 5).

stead of using a large test set to achieve high-accuracy recognition results. This deliberate choice allowed us to obtain a confusion matrix that provides richer information about how stimuli are related to each other, rather than just high values for each stimulus with itself and less information regarding the others.

The confusion matrix obtained is not symmetric. Since our study involves examining the relationship between probability and distance, it is important to consider that given stimulus A and B, distances measured in a perceptual space from A to B are equal to distances from B to A. Following this argument, we symmetrize the confusion matrix. We refer to this new matrix as the probability matrix (see Figure 2). It is worth noting that Shepard also recognized this issue and addressed it by symmetrizing his probability matrices in his original work.

5. Internal Representation and Similarity Matrix obtention

Internal representation refers to the mental model or simulation of external objects or events that is encoded and stored in the brain [4]. These representations are actively constructed by the brain based on available sensory information, past experiences, and knowledge, and play a vital role in perception, interpretation, and interaction with the environment. To accurately measure similarities between stimuli in our artificial system, it is necessary to establish an internal representation of the stimuli.

As we mentioned earlier, the Universal Law of Generalization relates probability to similarity. This section focuses on the challenge of measuring similarities between stimuli from our artificial agent. Shepard's approach to measuring similarity between stimuli was initially tested with humans and animals, they could rate the similarity between pairs of unknown objects based on their prior experiences and knowledge. Extending this concept to artificial systems, we establish an internal representation of the stimuli based on some previous experience. In order to provide our agent with this prior knowledge, we include certain figures, known as baseline, as a reference (see Figure 1). These figures

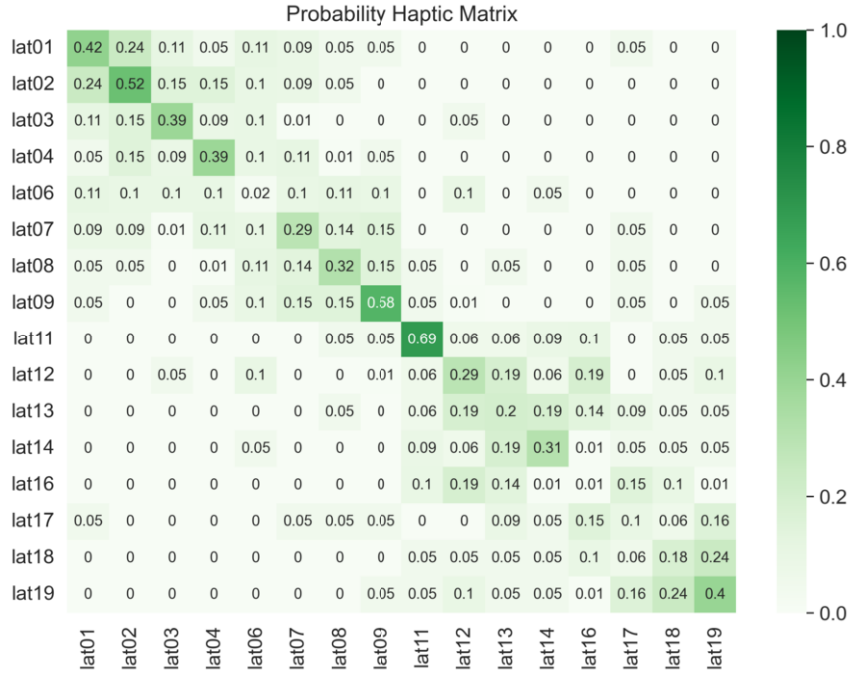


Figure 2. Symmetrized probability matrix from the confusion matrix. Obtained by $P_{ij} = \frac{C_{ij}+C_{ji}}{2}$, where P_{ij} is the probability of perceiving stimulus i as similar to stimulus j. This probability matrix provides a more detailed understanding of the relationships between stimuli than the confusion matrix, as all probabilities for a given pair of stimuli are now equal.

share some common characteristics with the unknown stimuli, making them suitable for use as a benchmark. The primary purpose is to train a classifier using these baseline figures and then use the model to obtain descriptors of the unknown figures. These descriptors are used to construct the internal representation of stimuli in the artificial system. Figure 3 displays the resulting descriptors of all unknown stimuli, obtained by training an XGBoost classifier using data from the baseline and testing it with the unknown figures.

The resulting descriptors obtained from the XGBoost classifier provide us with the system's perception of the new figures based on its previous experience with the baseline figures, without any prior knowledge of the unknown figures. We can observe that the concave unknown figures tend to have weights between lat00 and lat05, while convex figures fall more on lat15 and lat20. This observation makes sense because lat00 and lat05 correspond to concave shapes, and lat15 and lat20 correspond to convex shapes. Furthermore, this effectiveness of the descriptors in capturing even small differences between the unknown figures can be attributed to the selection of appropriate baseline figures. If the baseline is too generic, then the obtained descriptors may be too similar. In this case, as a result, the system is likely to perceive the unknown figures equally, leading to a loss of information regarding the individual differences between them. On the other hand, if the baseline is too specific, the obtained descriptors of the unknown figures would not share any qualities between them, making it difficult to compare them and find similarities. Therefore, selecting a suitable baseline with common characteristics to the

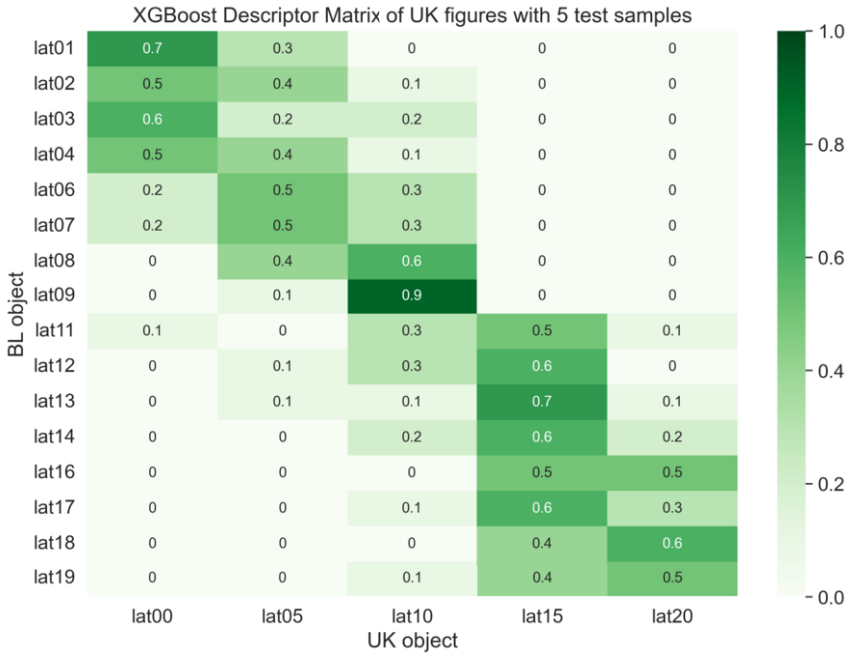


Figure 3. Descriptor Matrix for Unknown Stimuli obtained from XGBoost classifier. This figure displays the descriptor matrix generated by an XGBoost classifier for a set of unknown stimuli. The rows of the matrix correspond to the unknown stimuli, and the columns represent the degree of resemblance between each stimulus and each baseline figure. The matrix represents the output of the classifier, which is trained with the baseline figures and tested with the unknown stimuli. The matrix size is n x m, where n is the number of unknown stimuli and m is the number of baseline figures used for training the classifier. The effective descriptors generated by the XGBoost classifier can capture even subtle differences between the unknown stimuli, allowing for a reliable differentiation between them.

unknown figures is critical to obtain effective descriptors that can capture the essential features of each figure, allowing for a reliable distinction between them.

After obtaining the descriptors for each unknown figure, we can use them to calculate the similarity between each pair of unknown stimuli. To measure the similarity between two descriptors, we use the dot product between them (after their normalization). We compute the dot product of each unknown figure's descriptor with all the others, resulting in a square matrix of size n x n, where n represents the number of unknown figures that contains the similarities between these stimuli. The resulting values range from 0, indicating no similarity, to 1, indicating that the two stimuli are perceived equally. This similarity matrix enables us to identify which pairs of stimuli are more or less similar to each other creating an internal representation learned by the artificial agent.

6. Generalization Analysis for the Unknown Stimuli Set

Building on the similarity and probability values obtained in the previous sections 4 and 5, using our haptic capture system and XGBoost classifier, this section investigates the relationship between the probabilities and similarities perceived by our system for the

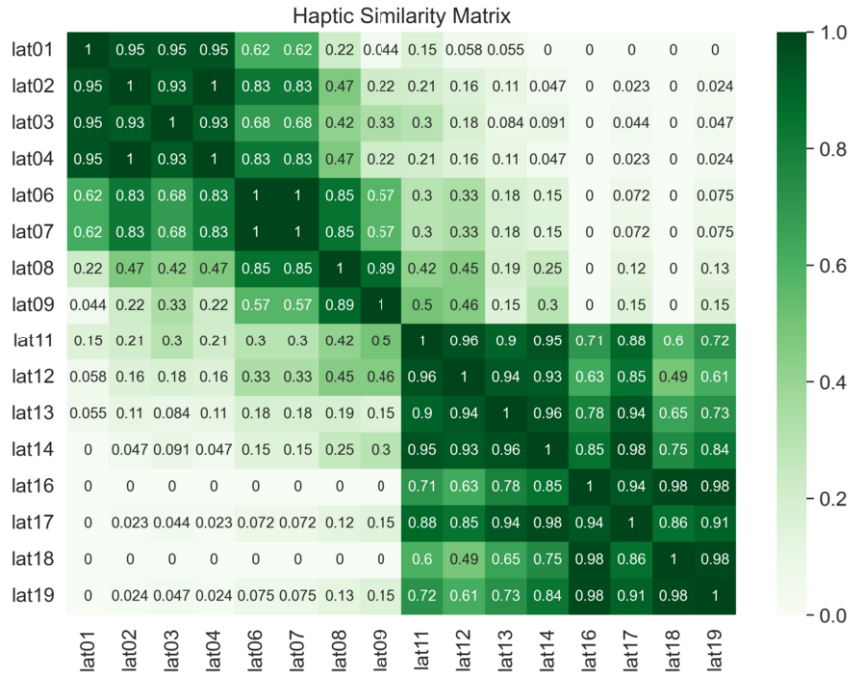


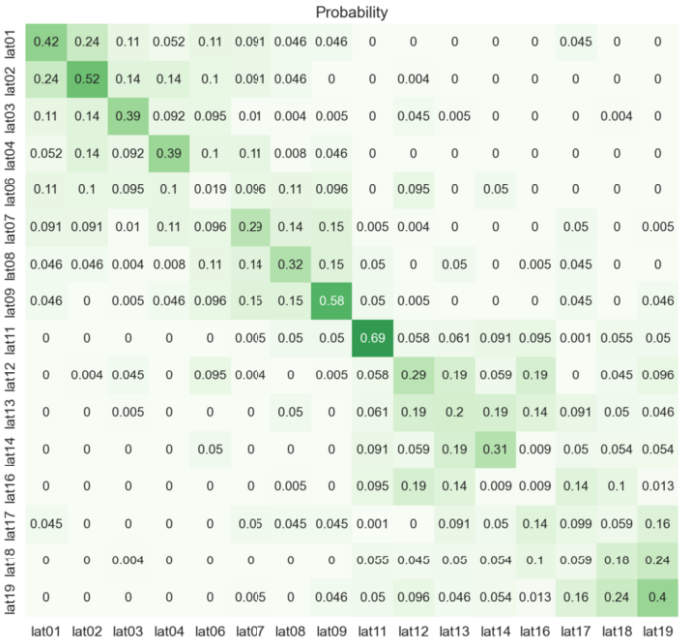
Figure 4. Similarity Matrix for Unknown Stimuli Based on Descriptor Similarities. This figure shows the similarity matrix for a set of unknown stimuli, computed based on their descriptor similarities. Each cell in the matrix represents the degree of similarity between a pair of stimuli, with higher values indicating greater similarity. The matrix is square and of size n x n, where n is the number of unknown stimuli. The similarity matrix provides a visual representation of the pairwise relationships between the unknown stimuli and can be used to identify clusters or patterns in the data.

UK stimulus set. The aim is to determine whether they follow the pattern of generalization observed by Shepard in his classic studies to humans and animals, which is described by a negative exponential or a Gaussian.

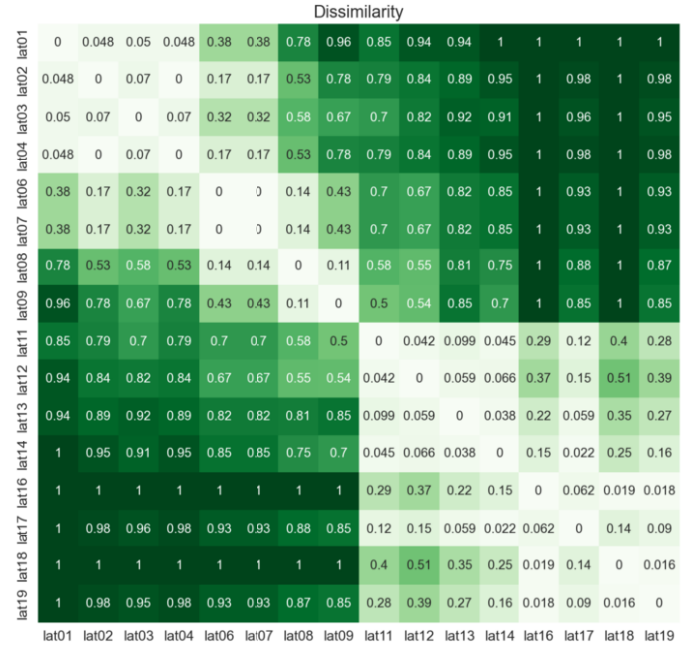
It is important to note that Shepard associated probability with the Euclidean distance obtained from a multi-dimensional scaling (MDS) technique of their psychological space. However, in our analysis, we have used the original similarity space between figures and the metric distance is based on the dissimilarity space.

It is noteworthy that this analysis will be conducted on each of the stimuli in the set separately, allowing us to examine the relationships between probabilities and distances for each figure individually. To better visualize the data used, we present Figure 5 where (a) displays the probabilities of each stimuli between them, while (b) shows the dissimilarities/distance between them.

We further analyze the relationship between probabilities and distances for each unknown figure in the stimulus set by creating a separate graph for each figure. Each graph represents an individual unknown figure, and each point on the graph corresponds to the relationship between the probability and distance of each other stimulus to that unknown figure. These graphs allow us to examine the generalization pattern for each individual stimulus and determine whether it follows the negative exponential or Gaussian



(a)



(b)

Figure 5. This figure presents the probability matrix (a) and dissimilarity matrix (b) for the UK stimulus set. The probability matrix displays the probability of perceiving stimulus *i* as similar to stimulus *j*, while the dissimilarity matrix shows the distance in the internal representation between stimulus *i* and stimulus *j*. The analysis aims to determine if the data follows the Universal Generalization Law.

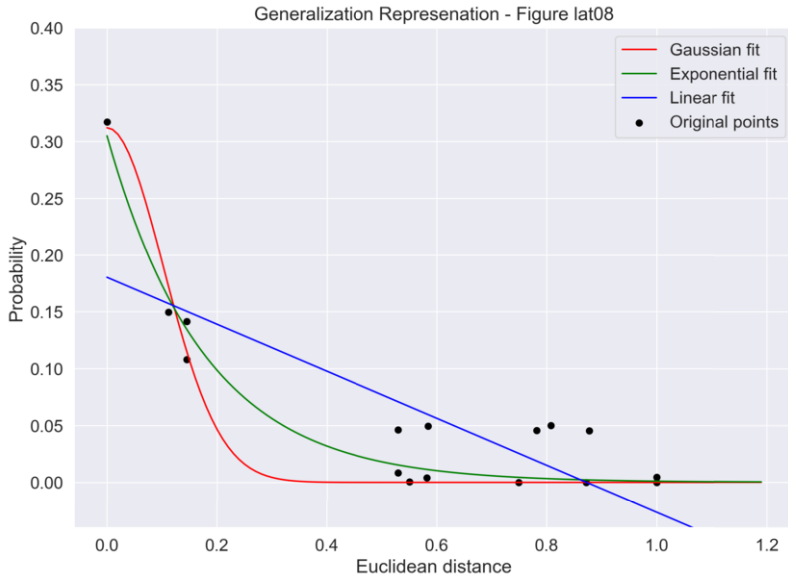


Figure 6. Graph representation of all the points corresponding to the unknown figure Lat08 with respect to all other stimuli in the UK set. Each point corresponds to the relationship between the probability and distance of the unknown figure to one of the other stimuli. Concretely it illustrates how generalization is made for the specific figure Lat08, where the green line represents the fitted curve for the negative exponential function, the red line represents the fitted curve for the Gaussian function and the blue line represents the fitted linear function. The RMSE of the negative exponential fit is 0.024, the RMSE of the Gaussian fit is 0.028 and the RMSE of the linear fit is 0.049.

pattern observed by Shepard. Figure 6 shows the graphical representation of all points corresponding to the unknown figure Lat08.

To evaluate the goodness of fit, we fitted the data for each figure into both the negative exponential and Gaussian functions and compute the root mean square error (RMSE) of the original points for each fit. In addition, we tested a decreasing linear function, which was not provided by Shepard, but was tested in a previous study with similar objectives to ours but different methodology [7].

The resulting RMSE values are presented in Table 1, which is divided into three sections: concave figures, convex figures, and the mean result of both. The RMSE values indicate that the negative exponential fit achieves a lower RMSE than the Gaussian fit, while the decreasing linear function has the highest RMSE. Furthermore, the convex figures tend to fit slightly better than the concave ones, suggesting that the system may have a better internal representation of concave figures than convex ones. This could be due to factors such as the dataset or the perceived curvatures of the figures.

Table 1. RMSE values for different fits and figure types.

	Concave figures	**Convex figures**	**Mean**
Exponential fit	0.0477	0.0465	0.0471
Gaussian fit	0.0538	0.0492	0.0515
Linear fit	0.0.648	0.0649	0.0649

7. Conclusions

In summary, our study provides evidence that Shepard's Universal Law of Generalization is applicable to artificial haptic systems. Specifically, we have shown that the relationship between perceived distances in the internal representation and the probability of confusion between stimuli follows Shepard's law, which is modelled by a negative exponential function. This finding has significant implications for understanding generalization in learning and for the development of human-like artificial agents. Moreover, our research has confirmed that an artificial agent can generate an internal representation of stimuli and use it for object generalization. This is a crucial step towards the development of advanced artificial agents that can interact with their environment in a human-like manner. Importantly, this means that any new figure or stimulus that fits into our baseline can be generalized using the law. However, we acknowledge that our study has limitations, such as using a limited number of stimuli and focusing on a specific sensory modality. Therefore, future research could investigate generalisation to other types of stimuli and other sensory modalities.

8. Acknowledgments

The work has been partially supported by the CMES project (grant No. PID2020-119725GA-I00) funded by the Spanish Ministry of Science and Innovation.

References

[1] Shepard RN. Toward a Universal Law of Generalization for Psychological Science. Science. 1987;237(4820):1317-1323.
[2] Chater N, Vitányi PMB. The Generalized Universal Law of Generalization. Cognition. 2003;87(2):B23-B34.
[3] Frank SA. Measurement invariance explains the universal law of generalization for psychological perception. Proceedings of the National Academy of Sciences. 2018 Sep 25;115(39):9803-6.
[4] Shepard RN. Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. Psychol Rev. 1984;91(4):417-447. doi: 10.1037/0033-295X.91.4.417.
[5] Shepard RN. Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. J Exp Psychol. 1958 Dec;55(6):509-23. doi: 10.1037/h0042354. PMID: 13602040.
[6] McGuire WJ. A multiprocess model for paired-associate learning. J Exp Psychol. 1961 Oct;62(4):335-47. doi: 10.1037/h0042676. PMID: 13893927.
[7] Rustom FB, Öğmen H, Yazdanbakhsh A. Object Detection, Recognition, Deep Learning, and the Universal Law of Generalization [Preprint]. arXiv:2206.05365. Published June 2022.
[8] Attneave, F. (1950). Dimensions of Similarity. The American Journal of Psychology, 63(4), 516. doi:10.2307/1418869.
[9] Shepard, R. N., & Cermak, G. W. (1973). Perceptual-cognitive explorations of a toroidal set of free-form stimuli. Cognitive Psychology, 4(3), 351–377. doi: 10.1016/0010-0285(73)90018-2.
[10] Ekman, G. (1954). Dimensions of Color Vision. The Journal of Psychology, 38(2), 467–474. doi: 10.1080/00223980.1954.9712953.
[11] Shepard, R. N. (1972). of Speech Sounds. Human communication: A unified view, (15), 67.
[12] Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. The Journal of the Acoustical Society of America, 24(2), 175–184. doi:10.1121/1.1906875
[13] Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. Journal of Experimental Psychology, 53(2), 94–101. doi:10.1037/h0041867

[14] Blough, D. S. (1961). The shape of some wavelength generalization gradients. Journal of Experimental Analysis of Behavior, 4(1), 31-40. doi: 10.1901/jeab.1961.4-31.

[15] Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. Journal of Experimental Psychology, 51(1), 79–88. doi:10.1037/h0046219

[16] Garrofé G, Parés C, Gutiérrez A, Ruiz C, Serra G, Miralles D. Virtual haptic system for shape recognition based on local curvatures. In: Advances in Computer Graphics: 38th Computer Graphics International Conference, CGI 2021, Virtual Event, September 6–10, 2021, Proceedings 38 2021 (pp. 41-53). Springer International Publishing.

[17] Gielis J. A generic geometric transformation that unifies a wide range of natural and abstract shapes. American journal of botany. 2003 Mar;90(3):333-8.

[18] Fougerolle YD, Gribok A, Foufou S, Truchetet F, Abidi MA. Rational supershapes for surface reconstruction. InEighth International Conference on Quality Control by Artificial Vision 2007 May 29 (Vol. 6356, pp. 206-215). SPIE.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230659

Expert Finding for Citizen Science

Bjoern KOMANDER [a,c,1], Jesus CERQUIDES [a], Jaume PIERA [b], Jeffrey CHAN [c] and Azadeh ALAVI [c]

[a] *IIIA-CSIC, Campus UAB, 08193 Cerdanyola, Spain*
[b] *EMBIMOS, ICM-CSIC, Barcelona, Spain*
[c] *School of Computing Technologies, RMIT University*

Abstract. Citizen Science brings together scientists and public participants to collaborate on a wide range of applications and fields. With this approach, Citizen Science advances scientific research and communication while accounting for various stakeholders. Across many Citizen Science projects, digital discussion platforms play an essential role for self-governance and self-organisation. In order to increase the quality of the discussions held on these platforms, we propose a model that recommends users to new discussions in which they are likely to contribute meaningful content. Our model learns relevant user representations based on the quality of past interactions between users and discussion threads, as well as the text content of questions, using a ranking loss function, an approximation of the NDCG metric, and matrix factorization. We demonstrate that our approach is able to predict potential experts on unseen discussion threads and outperforms several baselines. Compared to state-of-the-art expert finding techniques, the architecture of our model is significantly less complex, while focusing on a mostly overlooked ranking loss function.

Keywords. Expert Finding, Citizen Science, Learning-To-Rank, Online Discussion, Natural Language Processing

1. Introduction

Citizen Science (CS) [14] is scientific research conducted with participation from the public (who are sometimes referred to as amateur/nonprofessional scientists). An extension of this, Extreme Citizen Science (ExCiteS) [3], is a situated, bottom-up practice that takes into account local needs, practices and culture and works with broad networks of people to design and build new devices and knowledge creation processes to advance science and potentially help the community. In the quest for ExCiteS, digital tools which allow for discussion and deliberation among citizens are essential. Rich deliberation will enable proper self-governance for the CS project and will potentially increase the impact of the project on future policy[12].

Nowadays, the state-of-the-art in digital debate tools in CS projects range from non-existent for many small projects, to discourse[2] forums for some of the largest projects, such as iNaturalist[13][3] or Sensor.Community [16].

[1] Corresponding Author: Bjoern Komander, bkomander@iiia.csic.es
[2] https://discourse.org/
[3] https://www.inaturalist.org/

These discussion forums provide a tool to organise discussions and questions, where users can post a question or start a discussion in a *topic* across several *categories*. Other users can contribute to discussions or answer questions by posting within topics and yet other users can signal the usefulness of posts by awarding likes. Such discussion platforms play a vital role in the inclusive participation of all users in the Citizen Science process.

However, with increasing usage and content, these platforms become more and more cluttered, which can lead to several inefficiencies. As the amount of content continues to increase, potential answerers - i.e., participants who can contribute relevant answers - must filter through a larger volume of material to locate the post that aligns with their expertise. At the same time, question askers will have to wait for a longer time to receive the best (right) answer for their question. Tackling this problem by directing a new question to the right users and as quickly as possible is a prominent research area, called *expert finding* or *question routing*. While expert finding is a common research setting for online platforms like Stackoverflow and Yahoo! Answers, it has not yet been applied to Citizen Science discussion platforms. Particularly in these platforms, we can expect a very heterogenous distribution of knowledge across users with varying levels of expertise. An efficient allocation of users knowledge to the right content can therefore enhance the overall interaction with the platform and the inclusive participation of all users.

In this paper, we propose a model to enhance user engagement in Citizen Science discussion fora, by identifying potential experts for newly asked questions. Our approach combines multiple state-of-the-art concepts from expert finding to provide a ranking of potential experts for a new question. We utilize the question content, represented by Sentence-Bert embeddings [10], and past user-question interaction data to estimate the expertise of users and their likelihood of contributing valuable insights. Our model learns user embeddings by factorizing the interaction matrix using a ranking loss function. Unlike other expert finding methods, our model is computationally simple, yet outperforms several baselines. Moreover, our model can recommend experts for questions asked by unseen users, enabling it to generalize to new situations.

The remainder of this paper is structured as follows. Section 2 discusses related papers from the field of expert finding. Section 3 presents our concise problem and discusses the methodology of our approach. Section 4 introduces the data we obtained from several Citizen Science discussion platforms and section 5 presents our results. We conclude this work with a short discussion in section 6.

2. Related Work

To the best of our knowledge, there is only a small number of papers that address expert finding in the Citizen Science context. However these papers propose methods to find potential experts for the specific task of the Citizen Science platform, like the identification of insects in Spipoll [11]. We deviate from this problem as we seek to find experts for the respective Citizen Science discussion platform. This is more related to the task of expert finding/ question routing in community question answering, a sub-field of recommender systems. On community question answering websites such as StackOverflow and Yahoo! Answers, expert finding aims to identify user(s) with the adequate knowledge to answer a new question. The proposed methods in this field differ by the considered signals, i.e.,

the way they model questions and their content, as well as the modeling of users, the interactions between different users and the interaction between potential experts and questions. Moreover, these works propose a range of methods to make use of the different signals, from purely ranking and no learning algorithms such as PageRank to deep learning applications such as Convolutional Neural Networks. We categorize the different works into approaches that uses graphs to reflect the expertise of users, approaches primarily focused on the ranking objective, and deep learning approaches.

For the class of proposed methods which represent the interaction signals between questions and users, and among users in the form of graphs, Zhang et al. [17] note that these graphs differ from other online community graph structures, such as social networks like Twitter, as they are not based on social relationships but rather on the shared interests of the users. For instance Zhang et al. [17] and Wang et al. [15] construct a graph by connecting users with a directed edge from the author of a post to the answerer. Others [4,7] connect users if they answered the same question.

To make appropriate use of the graph structure, Wang et al. [15] apply an adapted version of PageRank, where they model a users expertise authority by the links pointing towards the user. Then, they combine the expert authority with a measure of expertise relevance by using text similarity based on TF-IDF embeddings. For each expert, they create an embedding for all past answered questions and then compare this to the new question through a distance measure. While they find their approach to be effective, they do not provide comparisons to other baselines or other discussion forums.

Similarly, Ghasemi et al. [4] use text similarity between a new question and all answers of existing questions in combination with user similarity between question asker and potential experts. They use Node2Vec to build user embeddings for all users in their graph and propose to use Word2Vec or TF-IDF for text embeddings. Then, they use a weighted sum to combine these two measures to create a ranking across potential experts for a new questions. They find that including the graph embeddings approach significantly increases the model's performance across all metrics. However, potential improvements of graph embeddings depend on the properties of the considered graph, as they find larger improvements for graphs with more links and nodes. Additionally, this approach requires sufficient information about all users for the user embeddings, and fails if an unseen user asks a new question.

With a learning-to-rank (LTR) approach, Ji et al. [5] apply RankingSVM, in a pairwise ranking approach, on user specific features, such as their activity and sum of best answers, and question specific features, such as text content. This approach is outperformed by Cheng et al. [2], who make use of LambdaMart, a popular Learning-To-Rank algorithm. They model expert finding as a multi-objective ranking problem, where they optimize answer quality and answer possibility of a potential expert at the same time. Their results suggest a very strong increase in performance over previous models.

Recently the expert finding task received increased attention in researchers that employ deep learning models: Li et al. [6] use a metapath guided network embedding to jointly embed information about the question raiser, question answerer and the question content. They then apply a Convolutional Neural Network ranking model to predict the

expertise across users for a new question. Again, as their model relies on data of the question raiser, it does not work for unseen or new question askers.

Liu et al. [8] aimed to improve the quality of text representations by retraining BERT embeddings with a label augmented masked language modelling task. To achieve this, they concatenated a user's past answers to a new question and used the binary label "expert" or "no expert" as the special CLS token.. They then fine tune the model on a binary expert / no expert classification task and report the best performance compared to a wide range of other models. For instance their model outperforms the approach by Li et al. [6] and Ghasemi et al. [4]. However, it is important to note that the pre-training stage of the model is computationally expensive. Additionally, the use of a binary classification in both pretraining and finetuning might limit the models ability to learn more meaningful text representations. When only considering one expert per question, the potentially helpful information of other answers is ignored. In general, it remains an open question if a ranking over more potential experts outperforms a binary classification for expert finding.

The existing literature in expert finding lacks a clear consensus on the optimal approach for representing questions, users, and their interactions. There is no evidence that graph based approaches outperform models that focus more on the text content and vice versa. Developing meaningful user representations that reflect their expertise based on their past interactions with questions is a major challenge in the expert finding literature. While previous studies have relied on graph-based methods to generate user representations, we propose a novel model that directly learns user embeddings by taking into account the quality of past interactions and the text representation of past questions. Our approach offers a simple and flexible way of learning meaningful user representations that help to recommend experts for unseen questions. Importantly, in comparison to graph-based user embeddings, our method does not fail to predict for unseen users. Furthermore, we integrate a ranking objective directly into the learning of user embeddings. This allows us to consider multiple potential experts for each question while learning user embeddings, through which we can better capture the complexity of the expert finding task.

3. Methodology

3.1. Problem Statement

Consider a (potentially non-finite) set of possible questions $\mathcal{Q}$, and a probability distribution P over $\mathcal{Q}$. Furthermore, consider a set of users $U = \{u_1, u_2, \ldots, u_n\}$, and a set of already answered questions $Q = \{q_1, q_2, \ldots, q_m\}$ which is a sample of P of size m. Let X_{ij} denote the matrix that reflects the expertise of user i for existing question j. For a new question q_{new}, we want to identify the user(s) with the adequate knowledge to answer q_{new}. We therefore define a function *f(text)* that, given the text features of q_{new} and the observed X_{ij}, creates a ranking across the set of users U, while maximizing the expected quality of the produced ranking:

$$\arg\max_{f} \mathop{\mathbb{E}}_{q\sim P} \kappa(f(q)), \tag{1}$$

where κ is a function to assert the quality of the produced ranking.

For the purposes of this study, we use the term "question" to refer to all types of initial posts in online discussion fora, including statements of opinion or requests for features. While for Community Question Answering fora (CQA), initial post are typically phrased as questions, in the Citizen Science discussion platforms, they may take on various forms. In this work we refer to all types of initial posts as questions, as we want to find the user with the proper knowledge to contribute to a post/discussion, regardless of the type. Moreover there might be more than one user with the appropriate knowledge for a given question. While many existing approaches view expert finding as a binary classification problem due to the prevalence of accepted answers on CQA platforms, our focus is on ranking potential experts based on their likelihood of providing a high-quality answer to a new question. By adopting this approach, we can identify multiple users who are well-suited to the task of expert finding, which is a critical step in encouraging greater participation among platform users.

In this paper, we define the knowledge adequacy $\tilde{X}_{ij}$ of user i for question j as

$$\tilde{X}_{ij} = \begin{cases} 0 & \text{if user did not engage} \\ \frac{\sum_{p\in R_{ij}} \text{likes(p)}+1}{\sum_{p\in R_j q(p)=j} \text{likes}(p)+1} & \text{otherwise,} \end{cases} \tag{2}$$

where R_{ij} is a set containing the responses to question j by user i and $R_j = \bigcup_{i\neq a(j)} R_{ij}$ is a set containing every response p provided to question j except those coming from the author. To model the difference of users who posted to a question but did not receive any likes from users that did not engage at all, we add one like for each user composing a response. We then normalize the adequacy to arrive at expertise matrix X_{ij} of user i for existing question j.

$$X_{ij} = \frac{\tilde{X}_{ij}}{\max_{i\neq a(j)} \tilde{X}_{ij}}. \tag{3}$$

Furthermore, in this paper we evaluate the quality of a produced ranking κ with the Normalized Discounted Cumulative Gain (NDCG), which can be defined as

$$NDCG(X_j, \hat{X}_j) = \frac{DCG(X_j, \hat{X}_j)}{DCG(X_j, X_j)}, \tag{4}$$

where DCG is defined as

$$DCG(X_j, \hat{X}_j) = \sum_{i=1}^{n} \frac{2^{X_{ij}} - 1}{\log_2(1+\pi_i)}. \tag{5}$$

To make the DCG measure comparable across different questions, each DCG measure is normalized by its ideal DCG, to arrive at NDCG. This measure reflects the quality of a predicted ranking by considering both the discounted rank of the prediction, π_i, as well as the relevance scores, in our case the adequacy scores, in the original ranking as $2^{X_{ij}}$.

3.2. Proposed Model

In this paper we propose a function that produces a ranking across potential experts in the form of

$$f(q,W) = \text{rank}(W \times H_q), \tag{6}$$

where W denotes a user embedding matrix and H_q denotes the encoded text representation of question q. As we want to avoid the shortcomings of methods that require exhaustive information about the question asker, and therefore fail for unseen users, we concentrate on information available for all users.

To model the question content H, we use Sentence-BERT, a computationally efficient adaptation of BERT that extracts semantically meaningful document representations [10]. The resulting question representation matrix has dimensions of *questions x 768*. We append an additional column of 1's to the question representation matrix, which serves as a constant term or bias for the subsequent learning process.

The main challenge is to develop a representation of the potential experts that reflects the relationship between users and questions, so that we can adequately predict their expertise for a new question. While previous studies have often modeled users using node embeddings in graph-based approaches, we propose factorizing the adequacy matrix X using a user embedding W and the encoded text representation H:

$$X = WH^T. \tag{7}$$

Therefore we can learn a user representation $\hat{W}$, which minimizes a loss $\mathscr{L}$ between X and the resulting $\hat{X}$. As our objective in predicting experts is to maximize the quality of the produced ranking in $\hat{X}$ denoted by κ, we can rewrite the objective in estimating $\hat{W}$ as

$$\hat{W} = \min \sum_{q}^{Q_t} \kappa(\hat{W}H^T, X). \tag{8}$$

Essentially we want to estimate $\hat{W}$ that maximizes the ranking quality κ of $\hat{X}$ with respect to X across all training questions $q \in Q_t$.

3.3. Loss Function

Previously, we defined κ as the NDCG. However since this metric, like other ranking metrics, is not differentiable, it cannot be used for direct optimization. As a result, there have been several approaches proposed in recent years to overcome this problem in listwise Learning-To-Rank (LTR) algorithms. Notably, algorithms like LambdaMART and LambdaRank, both based on Ranknet, work with gradient boosting or gradient optimization to optimize listwise ranking and related metrics, like the NDCG. For instance, LambdaMart uses a decision tree ensemble and gradient boosting to optimize pairwise differences in relevance scores between pairs of items. Other approaches aim to approximate the ranking metrics by a smooth and differentiable function. Recently, an older approach proposed by Qin et al.[9] gained more attention from researchers in the field,

including [1]. Qin et al.[9] propose an approximation of NDCG, that can be applied to other ranking metrics, where the rank π in the DCG metric is approximated by

$$\hat{\pi}_i = 1 + \sum_{k \neq i} \frac{1}{1 + \exp^{-\alpha(t-s)}}. \tag{9}$$

This approximation is now differentiable and thus can be optimized using (stochastic) gradient descent. We therefore define our loss function as

$$\mathscr{L}(X, \hat{W}H^T) = \min \sum_{q}^{Q_t} \hat{\kappa}(X, \hat{W}H^T), \tag{10}$$

where $\hat{\kappa}$ is the approximation of the NDCG. At the question level, the loss is defined as

$$\mathscr{L}(X_j, \hat{X_j}) = -\frac{1}{DCG(X_j, X_j)} \sum_{i=1}^{n} \frac{2^{X_{ij}} - 1}{\log_2(1 + \hat{\pi}_i)}. \tag{11}$$

3.4. Evaluation Metrics

The proposed approach is evaluated using several metrics including the NDCG at various cutoffs, namely for the top 5, 10, and 20. We also evaluate the performance using Mean Average Precision at k (MAP@k), where we compute the mean the average precision at k for all questions, as the intersection of ranking predictions and actual ranking of users at the cutoff k. Moreover, we use the Mean Recipropal Rank (MRR) as

$$MRR = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{\pi_i}, \tag{12}$$

where π_i denotes the predicted rank of the actual top ranked experts, across questions $j \in \{j_1, j_2, \ldots, J\}$

3.5. Baselines

We compare the model's performance to several baselines. The *CONSTANT* baselines predict the same ranking for all questions, based on the sum of training adequacy scores. Additionally we include an adaptation of the baseline *score*, where for a new question q_{new}, we weight the adequacy score for each training question by a similarity function $s()$ between the new the question and the respective text question

$$f(q) = \text{rank}(\sum_{j=k}^{J} s(q_{new}, q_i) \times X_{.i}). \tag{13}$$

We implement two different similarity functions, *EUC*

$$s_1(q_i, q_j) = e^{-d(H_{.i}, H_{.j})}, \tag{14}$$

and *COS*, where s_2 is defined as the cosine similarity between q_j and q_i.

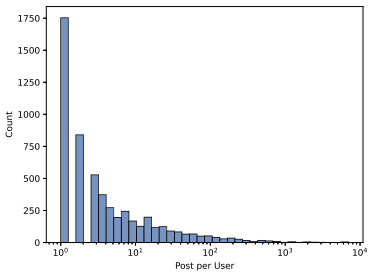
Name	Language	Active Users	Total Posts	Total Questions	Categories
iNaturalist	English	5568	161829	11886	9

Table 1. Characteristic of iNaturalist Discussion Forum.

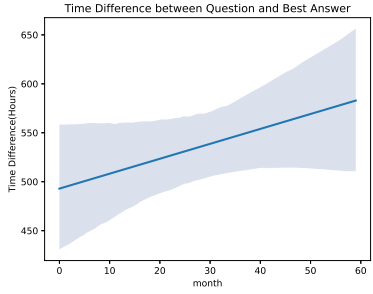
4. Data

We crawled several citizen science discussion fora, based on Discourse. We choose the largest forum, iNaturalist for training our model and evaluating our approach. Here users can post and discuss across several topics, such as *nature talk* and **feature requests**. An overview of the basic characteristics of iNaturalist is presented in table 1.

Similar to other online community discussion / question answering platforms, the iNaturalist discourse platforms consist of a small number of very active users as visible in figure 1a.



(a) Distribution of posts per user.



(b) Time difference between time of question and best answer in hours.

Figure 1.

The number of posts made by users can vary significantly, suggesting a difference in expertise levels. As shown in Figure 1b, we also observe a noticeable increase in the average wait time for obtaining a *best answer*, which is defined as the answer receiving the most likes, and it is noteworthy that this average time is already relatively high. To address this inefficiency, our model aims to route new questions to users who are most likely to provide meaningful answers, thus optimizing the process and reducing wait times. This approach could be implemented in the iNaturalist forum and other CS-fora by notifying users who our model identifies as top experts for a new question.

4.1. Experiments

After removing all questions with no answers, as well as any potential expert, where the sum of adequacy scores across all questions is below 1, we arrive at 11652 questions and 1625 potential experts. We split the questions into 80% training 10% validation and 10% testing using 5 folded cross validation.

We train the proposed model using stochastic gradient descent. Specifically we train each split for 150 epochs, with a batch size of 64 and a learning rate of 0.5. Our experiments suggest that training the model for more than 150 epochs will result in overfitting.

Model	MRR	MAP@1	MAP@3	MAP@5	MAP@10	MAP@20	NDCG	NDCG@5	NDCG@10	NDCG@20
MF	**0.25***	**0.129***	**0.225****	**0.227**	**0.184****	**0.132****	**0.5****	**0.325****	**0.368****	**0.407****
	(0.013)	(0.011)	(0.004)	(0.003)	(0.002)	(0.005)	(0.008)	(0.007)	(0.007)	0.007
COS	0.234	0.117	0.211	0.224	0.177	0.129	0.474	0.296	0.336	0.375
	(0.01)	(0.009)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.006)	(0.005)	(0.004)
EUC	0.241	0.124	0.214	0.219	0.179	0.129	0.486	0.307	0.35	0.39
	(0.01)	(0.009)	(0.003)	(0.004)	(0.003)	(0.002)	(0.003)	(0.007)	(0.004)	(0.005)
CONSTANT	0.226	0.105	0.2	0.224	0.177	0.128	0.468	0.291	0.331	0.369
	(0.009)	(0.005)	(0.017)	(0.002)	(0.003)	(0.002)	(0.004)	(0.007)	(0.006)	(0.005)

Table 2. Performance of several Models across Ranking Metrics with Standard Deviation in parenthesis. Best model for each score highlighted in bold. Stars indicate significance level of improvement over baselines at $p < 0.05$. * for baseline *CONSTANT*, ** for baseline *CONSTANT* & best baseline

5. Results

To assess the performance of our models we evaluate them on the previously discussed ranking metrics, presented in table 2. For each model we report the mean and standard deviation across the cross validation splits. We refer to our proposed model, the matrix factorization model as *MF*. Furthermore for the proposed baselines based on the distance functions, *EUC* denotes the baseline utilising the euclidean distance and *COS* the cosine similarity.

As it can be observed from the above table, our proposed model outperforms all the baselines across all metrics. For example, on average the most relevant expert is predicted at fourth position. However it should be noted that the other baselines already work reasonably well without any learning. All the baselines of comparison heavily rely on the sum of the adequacy score, which results in favoring the users that post more. While post frequency could be useful in indicating overall knowledgeability, we have found that incorporating question title embeddings into the *EUC* and *COS* models yields significant improvements over the *BASE* model. By further enhancing the signals from question titles, our approach has the potential to learn more meaningful user representations.

5.1. Quality of Predictions

We can further asses the quality of predictions across several models and several structural properties of iNaturalist. As figure 2 illustrates, the prediction quality for test questions differs across iNaturalist categories. For example all models perform relatively better for the categories *Bug Reports* and *Feature Requests*, categories where we would expect a well defined distribution of knowledge across users. We assume that these categories are more similar to community question answering fora such as Stackoverflow, where there are unique best answers to questions. However for the categories *Nature Talk*, *Curators* and *Tutorials* all models perform relatively bad. The knowledge for questions within these topics might be more evenly distributed, as these questions are usually more open. Interestingly, our model *MF* significantly outperforms the baselines in the categories *Nature Talk, News and Updates* and *Curators*.

6. Discussion

With this paper we present a model that can predict users with the adequate knowledge to contribute to new questions or posts in Citizen Science discussion platforms. We use

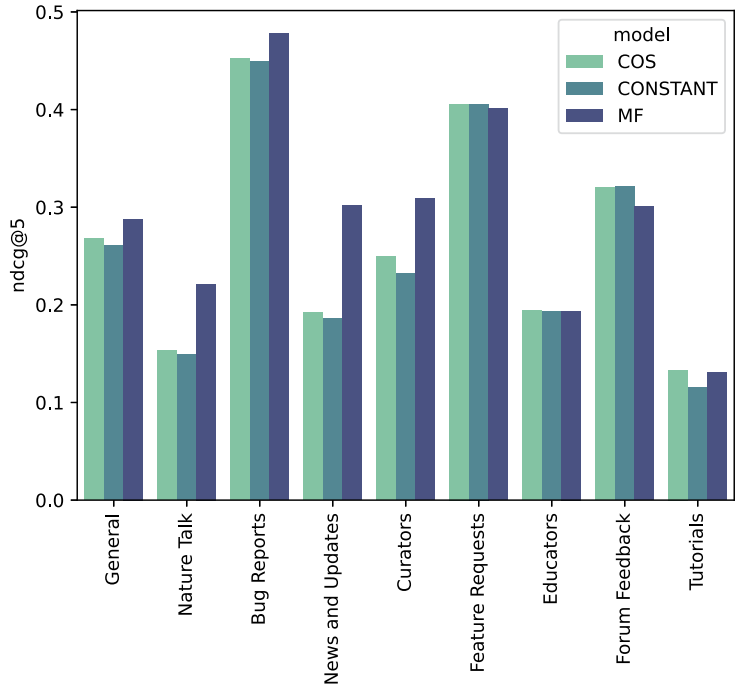


Figure 2. Differences in NDCG@5 across several categories and models

state-of-the-art techniques from the Expert Finding literature, including a novel Ranking Loss function and Sentence-BERT embedding with the goal to increase the usability of these discussion platform. Our proposed model outperforms several baselines while being significantly less computationally complex than state-of-the-art models. Moreover our model performs well on unseen question askers and is able to predict several potential experts for new questions. We believe that expert finding for online discussion can be a helpful tool for improving the quality and utility of these discussions. Our proposed model is adaptable and can be readily implemented on various online discussion platforms. However, our model can be improved by creating more meaningful question content representations and including more structural signals from the discussion fora.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement Nº 101034328, HumanE-AI-Net project N. 952026, and from the Horizon Europe programme project GUARDEN 101060693.

References

[1] Sebastian Bruch, Masrour Zoghi, Michael Bendersky, and Marc Najork. Revisiting Approximate Metric Optimization in the Age of Deep Neural Networks. In *Proceedings of the 42nd International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR'19, pages 1241–1244, New York, NY, USA, July 2019. Association for Computing Machinery.

[2] Xiang Cheng, Shuguang Zhu, Sen Su, and Gang Chen. A Multi-Objective Optimization Approach for Question Routing in Community Question Answering Services. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1779–1792, September 2017. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[3] Rafael Morais Chiaravalloti, Artemis Skarlatidou, Simon Hoyte, Marcos Moreu Badia, Muki Haklay, and Jerome Lewis. Extreme citizen science: Lessons learned from initiatives around the globe. *Conservation Science and Practice*, 4(2):e577, 2022.

[4] Negin Ghasemi, Ramin Fatourechi, and Saeedeh Momtazi. User Embedding for Expert Finding in Community Question Answering. *ACM Transactions on Knowledge Discovery from Data*, 15(4):70:1–70:16, March 2021.

[5] Zongcheng Ji and Bin Wang. Learning to rank for question routing in community question answering. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, CIKM '13, pages 2363–2368, New York, NY, USA, October 2013. Association for Computing Machinery.

[6] Zeyu Li, Jyun-Yu Jiang, Yizhou Sun, and Wei Wang. Personalized Question Routing via Heterogeneous Network Embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):192–199, July 2019. Number: 01.

[7] Chen Liu, Yuchen Hao, Wei Shan, and Zhihong Dai. Identifying Experts in Community Question Answering Website Based on Graph Convolutional Neural Network. *IEEE Access*, 8:137799–137811, 2020. Conference Name: IEEE Access.

[8] Hongtao Liu, Zhepeng Lv, Qing Yang, Dongliang Xu, and Qiyao Peng. ExpertBert: Pretraining Expert Finding. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pages 4244–4248, New York, NY, USA, October 2022. Association for Computing Machinery.

[9] Tao Qin, Tie-Yan Liu, and Hang Li. A general approximation framework for direct optimization of information retrieval measures. *Information Retrieval*, 13(4):375–397, August 2010.

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019. arXiv:1908.10084 [cs].

[11] Zakaria Saoud and Colin Fontaine. Expert finding in Citizen Science platform via weighted PageRank algorithm. October 2018.

[12] Lea A. Shanley, Alison Parker, Sven Schade, and Aletta Bonn. Policy Perspectives on Citizen Science and Crowdsourcing. *Citizen Science: Theory and Practice*, 4(1):30, December 2019. Number: 1 Publisher: Ubiquity Press.

[13] Shem Unger, Mark Rollins, Allison Tietz, and Hailey Dumais. iNaturalist as an engaging tool for identifying organisms in outdoor activities. *Journal of Biological Education*, 55(5):537–547, 2021. Publisher: Taylor & Francis.

[14] Katrin Vohland, Anne Land-Zandstra, Luigi Ceccaroni, Rob Lemmens, Josep Perelló, Marisa Ponti, Roeland Samson, and Katherin Wagenknecht. *The science of citizen science*. Springer Nature, 2021.

[15] G. Alan Wang, Jian Jiao, Alan S. Abrahams, Weiguo Fan, and Zhongju Zhang. ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, 54(3):1442–1451, February 2013.

[16] Michał Wróblewski, Joanna Suchomska, and Katarzyna Tamborska. Citizens or Consumers? Air Quality Sensor Users and Their Involvement in Sensor.Community. Results from Qualitative Case Study. *Sustainability*, 13(20):11406, January 2021. Number: 20 Publisher: Multidisciplinary Digital Publishing Institute.

[17] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 221–230, New York, NY, USA, May 2007. Association for Computing Machinery.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230660

Clustering of Qualitative Time Series for 3D Printing Quality Management

Ashutosh KARNA [a,1], Karina GIBERT [a]

[a] *Knowledge Engineering and Machine Learning Group at Intelligent Data Science and Artificial Intelligence Research Centre, Universitat Politècnica de Catalunya, Barcelona, Spain*

ORCiD ID: Ashutosh Karna https://orcid.org/0000-0002-9059-2796, Karina Gibert https://orcid.org/0000-0002-8542-3509

Abstract. In this research, we are attempting to extract relevant knowledge from sensor data of 3D printers. This is in continuation of our previous work to characterize the operational status of 3D printers using *Bootstrap-CURE* technique, including the post-processing techniques oriented to understand and conceptualize the clusters. Four clusters were identified regarding normal successful printing and abnormal situations due to different reasons like insufficient operating conditions, sensor failure, imbalance in internal temperature, etc. In the current work, the representation of data moves from punctual sensor readings to a vectorial description of a printing job in terms of the sequence of operational status over time. By representing a job by its sequence of clusters, it can be analyzed which sequential patterns are associated with successful or failing jobs. This opens the door to identifying trends and anticipating failures that are crucial to improving 3D printers' control and management. To cluster qualitative time series, specific qualitative distances are required, like χ^2-distance. One of the main challenges in clustering a time series dataset is to similarities among various time series rows. However, when the irregular length of the series and highly skewed series appear, special representation methods are required to avoid biases in results and the curse of dimensionality effects. The paper proposes to overcome these limitations.

Keywords. Clustering, time series, 3D printers, Pattern discovery, bootstrap clustering, decision support system

1. Introduction

In modern industrial applications, data flows continuously over time and is captured at short and fixed intervals. Any data science model to be used in industrial applications must adapt to near-real-time processing. That is, they should condense the data stream without losing the potential information.

While modern deep learning techniques are quite impressive in handling large volumes of data streams, however, they have very limited applications to clustering, which

[1]Corresponding Author: Ashutosh Karna, Knowledge Engineering and Machine Learning Group at Intelligent Data Science and Artificial Intelligence Research Centre, Universitat Politècnica de Catalunya, Catalonia, Spain; E-mail: ashutosh.karna@upc.edu

is often a good practice to understand the underlying patterns. The hierarchical clustering techniques fit quite appropriately due to their non-parametric approach, and can also explain the number of clusters very well using dendrograms. In this research, we are presenting a methodology to analyze large-scale data streams consisting of categorical values from a batch of real 3D printers.

Hence, the rest of the paper is structured as follows. A brief overview of related works in the literature is presented in section 2. Followed by that, the existing challenges that motivate this research and our proposals to deal with them are discussed in Section 3, A detailed methodology is presented in Section 4. An application to the proposal is discussed with the real-life data in Section 5. The results of our analysis are discussed briefly in Section 6 and finally, we conclude the paper with our insights and future scope of research in Section 7.

2. Literature Survey

Clustering is a nice strategy to derive meaningful classes from a large amount of data. Karna *et al.* in [1,2] uses hierarchical clustering-based methods to identify printing profiles from 3D printing data, which is scaled up to large datasets using a bootstrap approach as discussed in [3,4]. Since the long-term objective of our research is to build an intelligent decision support system, a methodology is required to synthesize long sequences of cluster labels over time, where a cluster label describes the operational status of a job at a particular time interval.

Aghabozorgi *et al.* [5] reviews several algorithms related to time-series clustering in a range of applications from finance to biology. The paper also highlighted the lack of focus on multivariate time series data with uneven length as well as discrete value time series with regard to clustering. In [6], Wang *et al.* proposes time-series clustering using the structural characteristics of the data, which eliminates distance computation. This approach extracts global features from the data and then uses clustering. Although this method reduces dimensionality, there is no universal set of features that can be applied to all the cases and hence this approach has limited usage in industrial applications where human intervention is avoided. In the current research work, we are analyzing 147 full printing jobs across 8 anonymous 3D printers, amounting to 562,777 records in total. A K-Nearest-Neighbour model has been built to predict the cluster labels (based on the clustering results obtained in precedent work [1]). This produces a qualitative time series associated with each printing job, where the pattern of the sensor instead of punctual measurement is considered. However, as the duration of these printing jobs varies, this results in long uneven lengths of cluster-label time series.

Determining the number of clusters in a set of time series is quite challenging since traditional numerical distances are not available and variable lengths of the series are also a drawback. In [7,8], Papadimitriou *et al.* discussed statistical measures to compute lag correlations among time series to find patterns. Another popular technique used in clustering the multivariate time series is *Dynamic Time Warping* (DTW) [9] which computes the warping path of the minimum distance between two different time series. Several applications of DTW in clustering are found in literature [10,11]. However, the distance in DTW is based on *Euclidean* distance which makes it incompatible to apply on categorical time series data. On the other hand, the clustering of categorical data is

often handled by considering the density of subspace without considering the ordering of attributes. The notion of distance in the case of categorical data is entirely different as the distance cannot be compared on a numerical scale. In [12], Zhang *et al.* proposes a dynamic system approach to cluster numerical data by updating the configuration mapping algorithm which improves another categorical data clustering proposal by Aggarwal *et al.* [13].

In terms of finding the distance between two arrays of categorical data, χ^2-distance can be used since it is non-parametric and can become applicable to compare frequency distributions across two or more columnar data. The frequency distribution from the original categorical data can be subject to Pearson's χ^2 test of homogeneity to check whether the difference between the expected and observed data is significant and the χ^2 statistic can represent the similarity between the two arrays.

This research involves a specific case of high dimensional data clustering which has its own challenges. In [14], Steinbach *et al.* discusses some of the key challenges of clustering in high dimensional data. In general, due to the high dimensionality, the data points get more sparse. Hence, the distance or similarity between two objects in such data becomes relatively uniform and the concept of distance of an object from its nearest and farthest point becomes complex. Beyer *et al.* in [15] provides a detailed description of nearest neighbors in high dimensions. Some other common approaches to deal with high dimensional data clustering include techniques like *Subspace-Clustering* [16,17] which finds clusters in different subspaces of the same data; *Projected-Clustering* [18], which projects data into lower dimensional spaces and then cluster it.

Chetouane et al. [19,20] discusses using a K-means clustering application to extract distinct scenarios from vehicle driving data, by building a cluster-episode graph to represent the chronological sequence of clusters.

However, as the application of this research directly applies to the real-life operation of 3D printers, any technique to deal with long sequences in clustering must be rather quick to enable faster decisions. Keeping this in mind, the authors propose splitting a long chain of qualitative values into a smaller mix of qualitative and quantitative values that can suitably be used in a clustering framework.

3. Research Challenges and Proposal

The data for this research was collected from 147 printing jobs consisting of 41 sensors amounting to 562,777 records. A K-Nearest-Neighbour model has been built to predict the cluster labels (based on the work in [1]) for all sensor records. Identifying the pattern of sensors, described by the cluster label, requires analyzing all print jobs in sequence. However, the duration of the print jobs varies significantly, which causes an imbalance in the data. This makes it difficult to place all cluster-label sequences together and apply further clustering due to the following challenges:

1. Imbalance in the length of the cluster label sequence
2. Categorical cluster labels do not allow Euclidean distance nor DTW as the appropriate distance measure
3. Curse of dimensionality appears due to long sequences of categories to be analyzed and this affects the quality of overall clustering results

Following are the proposed approaches to deal with them in order.

1. *Dealing with imbalanced sequences*: The variability in the length of sequences can be fixed by introducing a dummy label that indicates the *termination* of a printing job (represented by a sequence of cluster labels). The dummy variable can be appended at the end of each row as many times as required to complete the length of a longer job.
2. *Similarity measure in categorical data*: We propose using the χ^2 statistic to measure similarity among the printing jobs when the rows are purely categorical in nature. However, the long sequences of categorical data can adversely affect both the processing speed as well as the quality of clustering since the spatial distance diminishes as the dimensions keep increasing.
3. *Curse of high dimensionality on clustering*: In [21], Kriegel *et al.* discussed the challenges of clustering in a high dimensional space and recommends extracting local features in such cases. In particular, a long sequence can reasonably be transformed into a mix of categorical and numerical data, where the most important attributes are preserved as it is, while a long chain of categories can be collapsed into features such as frequencies of each individual feature. However, this converts categorical data into mixed data where the usual similarity distance such as χ^2 cannot be used, and more advanced mixed metrics such as *Gibert* mixed metric [22] can be applied. In this job, the proposal is to split the series into two blocks according to a parameter l^* that splits all series between time stamp $[1, l^*]$ and $[l^*, maxLenght]$ so that the lengths of initial split becomes equal for all rows. such that the main information is retained in $[0, l^*]$ and the remaining part of the series can be translated to a smaller set of relevant variables.

4. Methodology

This research attempts to identify patterns leading to various operation modes of a real-life multi-jet fusion 3D printer. Mathematically, a print job, denoted by P_i can be defined as a function of various sensor signals, represented by $s_1^t, s_2^t, ..., s_n^t$ at time t. The outcome of a job is defined both in terms of the part quality (physical properties of the items printed) as well as job status, denoted by $J \in (Success, Failure, Canceled, Error)$. In [1], the authors have identified four distinct clusters corresponding to different operational modes of the 3D printers. Hence, to scale the cluster labels to the target new sample, the following methodology is proposed.

1. Build a K-nearest-neighbor model to predict the cluster label individually for the entire sequence of all print jobs in this study.
2. A print job P_i of length l_i (number of time intervals) can then be expressed in terms of a sequence of cluster labels as :

$$P_i = (c_j^t; j \in (0,1,2,3)), t \in (0 : l_i) \tag{1}$$

   l_i is the length of i^{th} print job.
3. Introduce a dummy label, say (c_4) to denote the termination of a print job and append this at the end of all job sequences such that the length of all jobs becomes equal. If the length is already the same across all jobs, skip this step and proceed with the next step.

4. The resultant collection of sequences, can now be expressed in a matrix format where each column represents a cluster label in order and each row represents a whole print job.
5. Use suitable distance measures like χ^2 to compute the similarity among rows.
6. Compute pairwise χ^2 distance among all print jobs, apply hierarchical clustering, and interpret the meta clusters obtained from the sequence of cluster labels.

5. Application

A brief description of sensors and the 3D printer has been provided in [1]. The data contains a mixed pool of different job outcomes, namely, *success*, *failure*, *error*, and *canceled*. Data analysis is conducted using Python and R programming languages.

Earlier research in the same direction has already established 4 unique sensor profiles denoted by c_0, c_1, c_2, c_3 and leading to different operational modes. Following the methodology in the previous section, the analysis in this research can be split into three phases:

5.1. Phase 1: Prediction cluster labels

A KNN model has been developed to predict the cluster labels of unknown sensor records by training the model on the original set of sensor records and hierarchical clustering-based labels as discussed in [1]. A KNN model with 20 neighbors quite accurately predicted the cluster labels and hence this is applied to all the sensor records. A plot between miss-classification and the number of neighbors is shown in **Figure 1**.

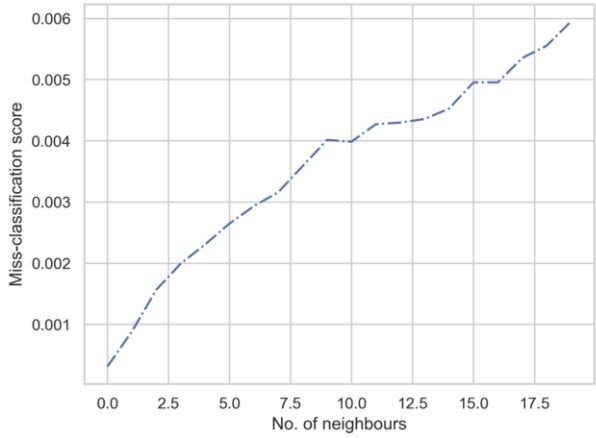


Figure 1. Miss-classification rate of KNN model against a number of neighbors

5.2. Phase II: Transforming multivariate data into sequence data

This phase deals with compressing the large-scale data into a sequence of cluster labels. In the previous phase, the cluster labels of all 562,777 records have been predicted. Thus, a row of 41 sensors is compressed into a single cell as a cluster label. Consequently, each printing job of dimension (n,m) (n records and m sensors) can now be represented by a vector of size l. **Figure 2** shows a sample time plot of a printing job along time.

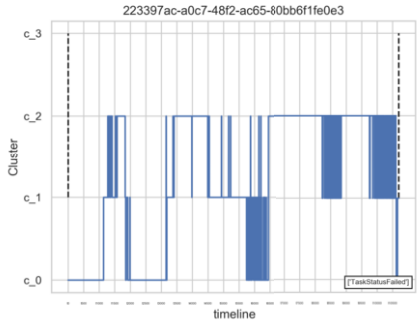


Figure 2. Time plot of sample print job

5.3. Meta-clustering of sequence data

Once transformed, the print jobs are converted into sequences of cluster labels, of unequal lengths. Hence, in order to bring them to the same scale, a dummy label, denoted by c_4 is introduced to all the jobs. The pairwise distance among print job sequences is calculated using χ^2 distance due to the categorical nature of the data. The χ^2 distance (analogous to the χ^2 test of homogeneity)[23] between two different sequences of values is calculated as shown in the **eq.2**.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \tag{2}$$

where $O_{i,j}$ and $E_{i,j}$ can be considered as two different sequences, corresponding to i^{th} row and j^{th} column of the data matrix obtained in phase II of the study. The aim of calculating χ^2 distance is to convert the data matrix consisting of categorical values into numerical distances that can directly be used in hierarchical clustering. Using *Ward's* method, the dendrogram revealed four meta-clusters, as shown in the **fig 3**.

However, on further inspection of the frequency distribution of no. of records across the job, the imbalance among the job length is quite evident as a vast majority of jobs terminate within the first 4000 time intervals, as shown in **Figure 4** while only a few jobs last for more than 8000 time-intervals, with the longest sequence being of 15,038. This implies that the long tail of the job sequence (column 4000 onwards) affects the clustering solution adversely and they need to be compressed and incorporated into the model without much loss of information. The hierarchical clustering obtained from χ^2 based distance represents 4 clusters where most of the objects fall under a big cluster, while 3 sporadic clusters are formed with fewer objects inside.

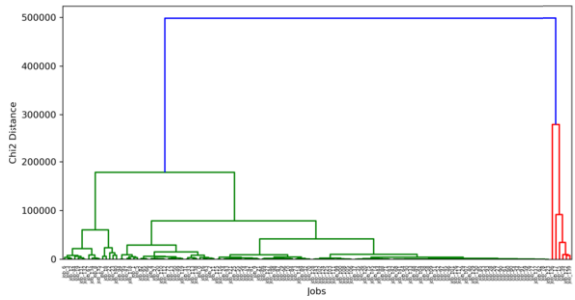


Figure 3. Dendrogram using χ^2 distance

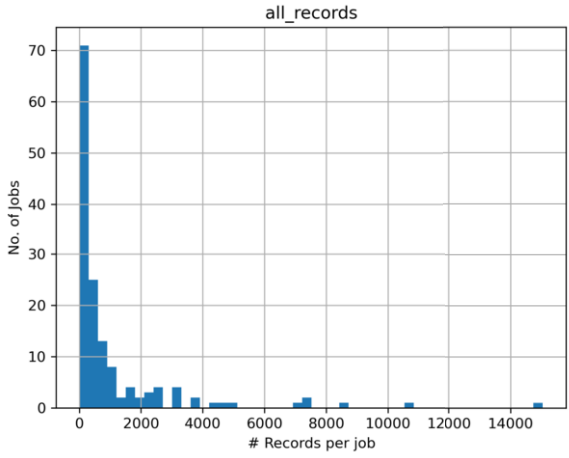


Figure 4. Histogram print-job length

5.4. Clustering based on Gibert mixed metric

This research encounters a specific case of high dimensionality with qualitaive series of lentgh= 15038. Clustering high-dimensional data brings several challenges. Consider a data matrix with n rows and p columns, where each row is a chain of qualitative values. Following the methodology as described in section 4, one can transform a collection of unequal-length sequences into a fixed-length sequence by introducing a dummy value in shorter jobs to complete the lentgh of the series. Let $P_i, i \in (0,1,2,...n)$ be a sequence of qualitative values with fixed length l and let the qualitative values in it be denoted by $c_j, j \in (0,1,2...p)$ and the corresponding frequency be denoted by m_j. Also, assume that N is the total number of elements in the matrix.

$$P_i = (c_j^t; j \in (0,1,2,...p)), t \in (0 : l_i) \tag{3}$$

$$\sum_j m_j = l_i; i \in (0,1,2,....n) \tag{4}$$

$$n * l = N \tag{5}$$

Following is the proposed approach to deal with long sequences in this research:

1. Inspect visually the histogram or the frequency distribution of the qualitative values across the data matrix.
2. Identify the length l^* such that $l^* << l$, which makes the distance computation feasible to manage.
3. For the remaining, $l - l^*$ columns, transform the qualitative chain into quantitative features synthesizing the pattern of the queue of the distributions by replacing them with their corresponding counts.
4. Choose an appropriate distance measure as now the object contains a mix of numeric and categorical features.
5. Used mixed metric distance matrix for hierarchical clustering.

Per our domain knowledge of 3D printing, we know that the data from the first 4000 time intervals (or columns in this case) are crucial in determining the quality and output of the job, it is imperative to freeze the initial 4000 records and the remaining queue of 11038 categorical columns can be squeezed into 5 columns each representing the counts of categories. In addition, two new numerical variables were augmented that capture the duration of jobs in minutes and the length of jobs in time intervals respectively. The *Gibert* mixed metric combines these qualitative and quantitative parts of a dataset so as to balance the influence given by each group of variables. Gibert's mixed metrics is a family of distances depending on two parameters α and β affecting the quantitative and qualitative part of the final distance, so that both components are balanced in proportion of the quantity of information represented. Following the approach in [22], the values of α_0 and β_0 are set to 0.7 and 0.3 respectively.

The dendrogram based on *Gibert* mixed metric is shown in **Figure** 5 and states for 4 clusters.

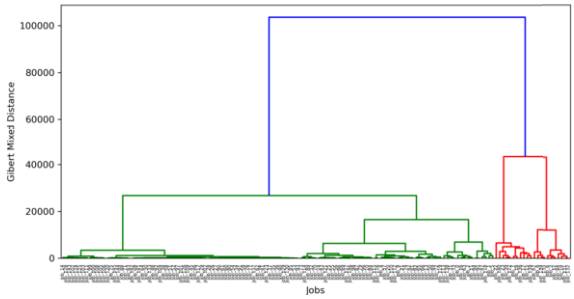


Figure 5. Dendrogram using *Gibert* mixed metric

6. Results

This section compares the results obtained from two different approaches - one using the original set of 15,038 categorical columns for 147 rows using χ^2 distance and an-

Table 1. Cross table between *Gibert* and χ^2 metric based meta-clusters

		χ^2 metric based clusters			
		k_0	k_1	k_2	k_3
Gibert mixed metric based clusters	k'_0	0	69	0	0
	k'_1	0	56	0	0
	k'_2	5	0	5	1
	k'_3	11	0	0	0

other with mixed metrics of 4000 categorical and 7 numerical columns synthesizing the information of the queue of 11038 columns and two additional variables using *Gibert* mixed metric. It is visually evident from the dendrograms that the one obtained using *Gibert* metric provides a more distributed set of clusters whereas the one obtained with χ^2 concentrates about a 90% of the observations as equal and belonging to a single class. A cross table between the class associations of these two methods is shown in **Table 1,** where $k'_i, i \in (0,1,2,3)$ denotes meta-clusters obtained using *Gibert* mixed metric and $k_j, j \in (0,1,2,3)$ denote those obtained using χ^2 distance measure.

7. Conclusions

This research analyzes a large dataset from 3D printing jobs and compares two different scenarios. Although the χ^2 metric can be used alone to measure pairwise similarity among all categorical data, the meta-clusters obtained therefrom are highly biased towards a big cluster due to the abundance of a single category in the tail end of the jobs. This is because the data points keep getting more sparse and the relevance of similarity between objects in higher dimensions reduces significantly. It is also evident from the meta-clustering results using χ^2 distance, where a large portion of jobs (125 out of 147) fall under 1 cluster whereas the other clusters contain too few objects (16, 5, and 1 jobs rest of the clusters respectively). Our proposal reduces this effect as the long tail of attributes in each job is collapsed into fewer features. This yields commendable improvement in the overall cluster quality. The *Gibert* mixed metric approach, however, is governed by the factors α and β that control the contributions of the qualitative part (fixed, important attributes) and the quantitative part (features from long-tail). In the current research, we have divided the dataset into qualitative and quantitative parts heuristically, based on the expert knowledge from the 3D printing domain and setting parameter $l = 4000$; however, a detailed analysis is needed to find a better way to calculate the best l. In the future lines of our research, we aim to explore this space and come up with a strategy to select the important qualitative attributes to retain in the analysis.

References

[1] Karna A, Gibert K. Using Hierarchical Clustering to Understand Behavior of 3D Printer Sensors. In: International Workshop on Self-Organizing Maps. Springer; 2019. p. 150-9.

[2] Karna A, Gibert K. Automatic identification of the number of clusters in hierarchical clustering. Neural Computing and Applications. 2021:1-16.

[3] SUMAN S, KARNA A, GIBERT K. Towards Expert-nspired Automatic Criterion to Cut a Dendrogram for Real-Industrial Applications. ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT. 2021:235.
[4] Suman S, Karna A, Gibert K. Bootstrap–CURE: A Novel Clustering Approach for Sensor Data—An Application to 3D Printing Industry. Applied Sciences. 2022;12(4):2191.
[5] Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering–a decade review. Information systems. 2015;53:16-38.
[6] Wang X, Smith K, Hyndman R. Characteristic-based clustering for time series data. Data mining and knowledge Discovery. 2006;13:335-64.
[7] Papadimitriou S, Sun J, Faloutsos C. Streaming pattern discovery in multiple time-series. In: Proceedings of the 31st international conference on Very large data bases; 2005. p. 697-708.
[8] Sakurai Y, Papadimitriou S, Faloutsos C. Braid: Stream mining through group lag correlations. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data; 2005. p. 599-610.
[9] Müller M. Dynamic time warping. Information retrieval for music and motion. 2007:69-84.
[10] Niennattrakul V, Ratanamahatana CA. On clustering multimedia time series data using k-means and dynamic time warping. In: 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07). IEEE; 2007. p. 733-8.
[11] Łuczak M. Hierarchical clustering of time series data with parametric derivative dynamic time warping. Expert Systems with Applications. 2016;62:116-30.
[12] Zhang Y, Fu A, Cai C, Heng P. Clustering categorical data. In: Proceedings of the International Conference on Data Engineering. Citeseer; 2000. p. 305-5.
[13] Aggarwal CC, Wolf JL, Yu PS. A new method for similarity indexing of market basket data. In: Proceedings of the 1999 ACM SIGMOD international conference on Management of data; 1999. p. 407-18.
[14] Steinbach M, Ertöz L, Kumar V. The challenges of clustering high dimensional data. New directions in statistical physics: econophysics, bioinformatics, and pattern recognition. 2004:273-309.
[15] Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful? In: Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7. Springer; 1999. p. 217-35.
[16] Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: a review. Acm sigkdd explorations newsletter. 2004;6(1):90-105.
[17] Elhamifar E, Vidal R. Sparse subspace clustering: Algorithm, theory, and applications. IEEE transactions on pattern analysis and machine intelligence. 2013;35(11):2765-81.
[18] Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, Park JS. Fast algorithms for projected clustering. ACM SIGMoD record. 1999;28(2):61-72.
[19] Chetouane N, Wotawa F. On the application of clustering for extracting driving scenarios from vehicle data. Machine Learning with Applications. 2022;9:100377.
[20] Chetouane N, Wotawa F. Extracting temporal models from data episodes. In: 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS). IEEE; 2022. p. 721-31.
[21] Kriegel HP, Kröger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. Acm transactions on knowledge discovery from data (tkdd). 2009;3(1):1-58.
[22] Gibert K, Cortés García CU. Weighting quantitative and qualitative variables in clustering methods. Mathware & soft computing 1997 Vol 4 Núm 3. 1997.
[23] Bolboacă SD, Jäntschi L, Sestraş AF, Sestraş RE, Pamfil DC. Pearson-Fisher chi-square statistic revisited. Information. 2011;2(3):528-45.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230661

Cognitive Similarity Through Bibliometric Analysis

Lucas MARCO-TORDERA [a], Daniel GARCÍA-COSTA [a] and Francisco GRIMALDO [a,1]

[a] *Departament d'Informàtica, Universitat de València, Spain*

Abstract. Cognitive similarity is a fundamental concept in the peer review process and the research funding programs. It can help to find qualified experts in a field in the refereeing or evaluation process. The main objective of this work is to find an automatic way to determine cognitive similarity based on bibliographic information retrieval using the Scopus database. In this paper, we will compare different measures of cognitive similarity between pairs of authors through the whole publication portfolio of both. These measures are based on the authors' publications and citations, at different levels of depth of authorship and citation networks. We have applied bibliographic coupling and text-based techniques. We have expressed these measures as cosine similarity. For this work, we use a small empirical case study with authors in the fields related to artificial intelligence to compare the measures of cognitive similarity. The results of this study show that the use of combined measures using bibliographic coupling and text-based can provide more information on the cognitive similarity between a pair of authors.

Keywords. cognitive similarity, bibliographic coupling, bibliometric analysis, refereeing, peer-review, citation networks, authorship, cosine similarity, author bibliographic coupling, stop words

1. Introduction

In the epistemologies of Piaget, categories of thought are developed in interaction with physical and cultural environment [1]. Cognitive similarity quantification can be complex since it considers experiential or ambient factors [2]. In the field of research, we can simplify this broad concept by reducing it to the research lines' similarity of each author. In this way, two authors will have greater cognitive proximity the more their researches are similar. Scientific production in specialized journals allows us to quantify cognitive similarity by comparing the bibliography of each author. Cognitive similarity has direct applications such as determining the suitability of experts in the peer review process and the research funding programs in the refereeing or the evaluation process [3]. The measurement of cognitive similarity can be approached through different techniques such as authorship networks and citation networks based on establishing coincident articles or authors, as well as the analysis of the text of the articles in these networks [4]. Other works have studied document-document similarity approaches using bibliographic cou-

[1]Corresponding Author: Francisco Grimaldo, e-mail: francisco.grimaldo@uv.es

pling and a text approach based using a cluster classification [5], combined bibliometric or link information with textual content for clustering classifier [6], identified research topics related to Big Data research through citation networks [7][8].

In this study, we use the author bibliographic coupling ABC [9], which is an extension of the bibliographic coupling [10], in which we use the portfolio of each author through their citation networks at different levels of network depth. We have introduced a cognitive similarity indicator based on an author citation network extended to the portfolio's author and compared it with the other similarity indicators based on co-authorship networks, citation networks, and similarities between the keywords of the articles.

When collecting the entire portfolio of an author, we find a temporary gap between the different publications. This effect is considered a bias because older papers may have more citations than recent ones, not necessarily because of their merits, but because they stayed in the literature longer and had more time to be cited [11] and in the case of authors who changed their research line at some point in their careers. Other studies address this problem by classifying the data according to the date and the journal of publication or by creating dynamic slices of the database containing the same number of citations [12]. In our study, temporality has not been taken into account, but a decay function depending on the publication date can be introduced.

Schematically, we can represent the bibliographic network at different depth levels for the study in Figure 1.

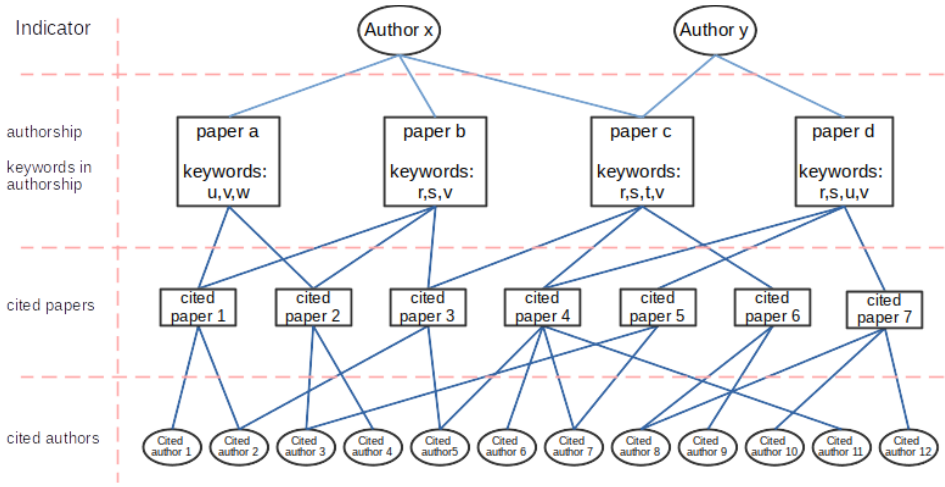


Figure 1. Scheme of cognitive network.

Using the whole publication portfolio of both authors, we compare 4 different measures of cognitive similarity between pairs of authors for establishes:

$SIM_{authorship}$: the level of co-authorship in the scientific production of two authors in the first level of the network's depth.

SIM_{cited_papers}: the level of coincidences between the articles referenced by each author in the scientific production of two authors in the second level of the network's depth.

$SIM_{cited_authors}$: the level of coincidences between the authors referenced in the scientific production of two authors in the third level of the network's depth.

$SIM_{keywords}$: the level of keyword matches of the articles made by each author in the first level of the network's depth using text-based techniques.

We build a weighted graph [13] from the cognitive similarity between each pair of authors. The weights of each link are the cognitive similarity values. We show the cognitive community structure created as a division of network nodes into densely connected subgroups [14].

2. Methods

The bibliographic data for each author was retrieved from the Scopus database using the scheme in Figure 1. Starting from an author, we have recovered all its publications. These publications belong to the first level of the network's depth. We have obtained the bibliographic information of the cited articles from the references or citations of these publications. These referenced publications belong to the second level of the network's depth. We have obtained the set of co-authors for each referenced publication. This set of authors belongs to the third level of the network's depth. For the case of the study, we have used the people who were part of the Scientific Committee of the Proceedings of the 24th International Conference of the Catalan Association for Artificial Intelligence (CCIA), held in Sitges, in October 2022. For the 56 members of the Scientific Committee, we have recovered 5,414 papers quoting 63,035 articles by 105,899 authors.

From the articles of the members of the Scientific Committee, we have recovered 41,065 keywords that Scopus provides per each paper, which we have processed by removing stop words to avoid false matches between pairs.

With the information retrieved from Scopus, we calculated four different measures of cognitive similarity between pairs of authors.

authorship network			**cited papers**			**cited authors**			**keywords in authorship network**		
paper	author x	author y	cited paper	author x	author y	cited author	author x	author y	keyword	author x	author y
a	1	0	1	2	0	1	1	0	r	2	2
b	1	0	2	2	0	2	2	1	s	2	2
c	1	1	3	2	1	3	1	1	t	1	1
d	0	1	4	0	2	4	1	0	u	1	1
			5	0	1	5	2	2	v	3	2
			6	0	1	6	1	1	w	1	0
			7	0	1	7	1	2			
						8	1	2			
						9	1	1			
						10	0	1			
						11	1	1			
						12	0	1			

Table 1. Frequencies for the different studied indicators corresponding to Figure 1.

For $SIM_{authorship}$ we have started from the publications that appear in the Scopus database for a pair of authors. Once the information has been compiled, based on the identifier that Scopus assigns to each publication, we create a frequency table. The frequencies are binary (1 if he or she is the publication's author and 0 otherwise) (Table 1).

For SIM_{cited_papers} we have also used the Scopus identifier for the publications, but this time these are the articles referenced by each author. To generate the frequencies,

we count the number of times that each author has cited each publication (Table 1). The frequencies not include self-citations.

For $SIM_{cited_authors}$ we used the Scopus identifier for the authors of the articles cited by each author. To generate the frequencies, we count the number of times an author references another. Each cited publication is associated with one or more authors (Table 1). The frequencies not include self-citations.

For $SIM_{keywords}$ (text-based), we used the keywords that Scopus includes with each article. We have proceeded similarly to build the frequency table where we count the number of times a keyword appears in all the authors' articles (Table 1).

For each of the 4 measures of cognitive similarity, we have calculated the unweighted cosine similarity [15], as a first approximation, from the frequency vectors of each pair of authors.

The process to calculate the cosine similarity was repeated for the 1540 combinations of the 56 members of the Scientific Committee.

We have created a complete graph K_{56} = (V, E) (simple and undirected). In this graph the vertices V are the authors of the study $v_1, v_2 \ldots v_{56} \in V$, and the edges E represent a relationship between each pair of vertices $\forall i \neq j | i, j \in \{1...56\}, \{v_i, v_j\} \in E$. The number of edges of this graph is $\binom{56}{2} = 1540$. We have transformed the created graph into a weighted graph, associating a positive real number called weight to each edge. The weight used has been the indicators of cognitive similarity. The adjacency matrix of a weighted graph is a M=V × V matrix where each element $M_{i,j}$ contains the weight $w_{i,j}$ assigned to the edge $\{v_i, v_j\}$. The non-zero values of the weight $w_{i,j}$ imply that the vertices v_i and v_j are adjacent and the zero values that they are not. The members of the Scientific Committee have grouped into communities. The communities are classified using the cluster_fast_greedy algorithm [14] from the igraph [16] R [17] library.

3. Results

The results for the measure of similarity $SIM_{authorship}$ show a low level of co-authorship. There are only 96 pairs of authors with collaboration out of the total of 1540, which represents 6%. Table 2 shows the 10 authors with the highest co-authorships. A high value indicates similar lines of research and a priori an appropriate choice in a panel of experts or for the peer-review process and also alerts us to possible cognitive biases or homophily.

The SIM_{cited_papers} measure generally shows low similarity values. This occurs because the same line of research can give rise to more than one article or several types of publications such as non-peer-reviewed articles, peer-reviewed articles, and conference proceedings. This multiplicity of types of publications for the same line of research makes it more difficult to get matches in cited articles. However, the number of pairs that have a value greater than zero is 348. In the case that $SIM_{authorship}$ is high SIM_{cited_papers} increases.

The measure $SIM_{cited_authors}$ is less selective than SIM_{cited_papers}, since it does not look for the exact match of the cited article but of the authors. The number of pairs that have a value greater than 0 is 971. We highlight that authors without collaboration and with relatively low values of SIM_{cited_papers} have relatively high values of this measure in Table 3. These values are comparable with other authors that have collaborated.

Author 1	Author 2	$SIM_{authorship}$
Lledó Museros (UJI)	Zoe Falomir (UJI)	0.48
Javier Vázquez-Salceda (UPC)	Sergio Alvarez-Napagao (BSC)	0.46
Emilio García (UIB)	Francisco Bonnín (SRV-UIB)	0.44
Aïda Valls (URV)	Antonio Moreno (URV)	0.36
Jesús Cerquides (IIIA-CSIC)	Juan Antonio Rodríguez (IIIA-CSIC)	0.36
Javier Béjar (UPC)	Ulises Cortés (UPC)	0.25
Miquel Sànchez (UPC)	Ulises Cortés (UPC)	0.24
Dario Garcia-Gasulla (BSC)	Sergio Alvarez-Napagao (BSC)	0.22
Dario Garcia-Gasulla (BSC)	Ulises Cortés (UPC)	0.22
Enric Plaza (IIIA-CSIC)	Eva Armengol (IIIA-CSIC)	0.19

Table 2. Top-10 most similar researchers using $SIM_{authorship}$.

Author 1	Author 2	$SIM_{authorship}$	SIM_{cited_papers}	$SIM_{cited_authors}$
Anders Jonsson (UPF)	Héctor Geffner (ICREA-UPF)	0	0.16	0.47
Jordi Sabater (IIIA-CSIC)	Vicente Julián (UPV)	0	0.04	0.41
Javier Vázquez-Salceda (UPC)	Juan Antonio Rodríguez (IIIA-CSIC)	0	0.05	0.37
Jordi Coll (Aix-Marseille Université)	Jordi Levy (IIIA-CSIC)	0	0.07	0.35
Enric Plaza (IIIA-CSIC)	Eva Armengol (IIIA-CSIC)	0.19	0.14	0.34
Enric Plaza (IIIA-CSIC)	Marco Schorlemmer (IIIA-CSIC)	0.14	0.21	0.33
Isabel Aguiló (UIB)	Manuel González-Hidalgo (UIB)	0.06	0.03	0.33
Elisabet Golobardes (URL)	Eva Armengol (IIIA-CSIC)	0.09	0.19	0.32
Miquel Sànchez (UPC)	Ulises Cortés (UPC)	0.24	0.14	0.32

Table 3. High values of $SIM_{cited_authors}$ with low values of $SIM_{authorship}$ and SIM_{cited_papers}.

The communities in the graph (Figure 2) are described by the most frequent keywords in each group (Table 7). These keywords show us the profile of the people that make up each community and their specialization within the field of artificial intelligence.

The number of pairs with a greater value than 0 is 1385 for the $SIM_{keywords}$ measure. We observed a Pearson correlation coefficient of 0.81 between $SIM_{cited_authors}$ and $SIM_{keywords}$. We find significant values of $SIM_{keywords}$ for low values of $SIM_{authorship}$ and SIM_{cited_papers} (Table 4) due to the authors sharing keywords in their articles even if they do not collaborate or do not cite the same paper. There are null values of $SIM_{authorship}$ and low values of SIM_{cited_papers} and $SIM_{cited_authors}$ that show high values of $SIM_{keywords}$. This case could be associated with portfolios of authors with similar research but with different articles' references.

The communities in the graph (Figure 3) are described by the most frequent keywords in each group (Table 6). These keywords show us the profile of the people that make up each community and their specialization within the field of artificial intelligence.

We labeled the communities obtained with $SIM_{cited_authors}$ in Figure 2 as A, B, C, D, and E. We also labeled the communities obtained with $SIM_{keywords}$ in Figure 3 as 1, 2, 3, 4, 5, and 6. Table 5 shows the correspondence between the communities obtained with the different measures of cognitive similarity. The correspondence between these four groups maximizes the number of common members classified by the two measures

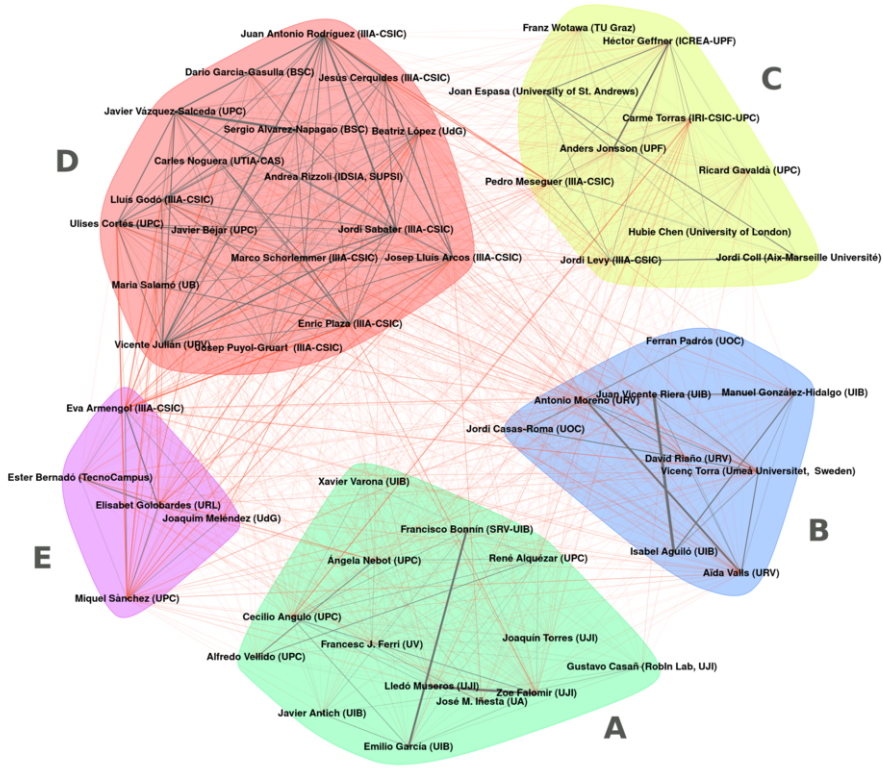


Figure 2. Communities in cognitive network $SIM_{cited_authors}$

Author 1	Author 2	$SIM_{authorship}$	SIM_{cited_papers}	$SIM_{cited_authors}$	$SIM_{keywords}$
Anders Jonsson (UPF)	Héctor Geffner (ICREA-UPF)	0	0.16	0.47	0.54
Javier Vázquez-Salceda (UPC)	Juan Antonio Rodríguez (IIIA-CSIC)	0	0.05	0.37	0.35
Isabel Aguiló (UIB)	Manuel González-Hidalgo (UIB)	0.06	0.03	0.33	0.32
Elisabet Golobardes (URL)	Miquel Sànchez (UPC)	0	0.07	0.21	0.31
Jordi Sabater (IIIA-CSIC)	Vicente Julián (UPV)	0	0.04	0.41	0.3
Josep Puyol-Gruart (IIIA-CSIC)	Vicente Julián (UPV)	0	0	0.11	0.21
Eva Armengol (IIIA-CSIC)	Ricard Gavaldà (UPC)	0	0	0.02	0.21
Javier Béjar (UPC)	Ricard Gavaldà (UPC)	0	0	0.01	0.21
Dario Garcia-Gasulla (BSC)	Ferran Padrós (UOC)	0	0	0	0.2
Jesús Cerquides (IIIA-CSIC)	Ricard Gavaldà (UPC)	0	0	0.03	0.19

Table 4. High values of $SIM_{keywords}$ with low $SIM_{authorship}$ and SIM_{cited_papers}

$SIM_{cited_authors}$ and $SIM_{keywords}$. The maximum number of matching members in the two graphs is 40 out of a total of 56, which represents a 71% match.

Comparing the communities' keywords of the $SIM_{cited_authors}$ and $SIM_{keywords}$ measures from Table 6 and 7, we see that the most frequent keywords are maintained for the groups described in Table 5.

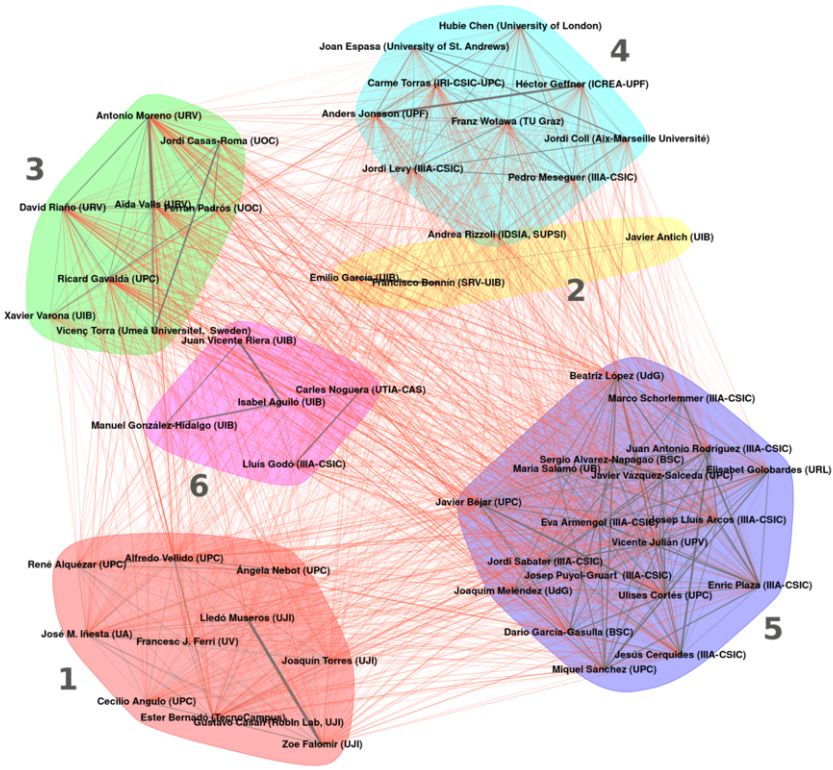


Figure 3. Communities in cognitive network $\text{SIM}_{keywords}$

$\text{SIM}_{cited_authors}$	$\text{SIM}_{keywords}$	coincident authors
A	1	10
B	3	6
C	4	9
D	5	15

Table 5. Similar communities between $\text{SIM}_{cited_authors}$ and $\text{SIM}_{keywords}$.

4. Conclusion

The results for the similarity measure $\text{SIM}_{authorship}$ show a low level of co-authorship in the Scientific Committee of CCIA. Using this measure as the only indicator of cognitive distance is not adequate since it does not reveal many of the relationships shown by the other similarity measures presented in this study. Low values in this measure of cognitive similarity would be desirable among the members' panel and the applicants in funding programs.

The measure of similarity although $\text{SIM}_{cited_papers}$ triples the number of relationships between pairs of authors for $\text{SIM}_{authorship}$ but is more selective than expected. It is logical to think that when a researcher publishes several articles within the same line

of research, he may be referenced by another author in one of them for sharing ideas or methodologies. This article multiplicity increased when each article can have several types of publications such as non-peer-reviewed articles, peer-reviewed articles, and conference proceedings. This cause may justify the low levels in this measure of cognitive similarity since it makes it difficult for two authors to share the same publication.

The $SIM_{cited_authors}$ similarity measure complements the previous measures since it shows the citation relationships at the author level and not at the publication level. This indicator reveals the relationships between authors who have not collaborated but whose research papers are based on the work of other researchers that they have in common through their citations.

The measure of similarity $SIM_{keywords}$ shows values for practically all the members of the Scientific Committee of CCIA. Given the committee's specialization in artificial intelligence, this fact seems logical.

The communities formed by $SIM_{cited_authors}$ and $SIM_{keywords}$ show similarities explained by the correlation. Although they are correlated, we cannot opt for one of the two measures since their joint analysis allows us to appreciate cases where there may be a high similarity in the keywords but a low similarity in the cited authors.

The proposed measures of cognitive similarity are complementary and show different aspects of the authors' portfolios that help us understand their relationships, reference authors, and research fields.

These dimensions of cognitive similarity can give us a better image of forming different panels of experts. In the same way, this methodology could automate the choice of a referee in the peer review process.

keyword	freq	community	keyword	freq	community
fuzzy inductive reasoning	67	1	model based diagnosis	82	4
indoor positioning	62	1	planning	59	4
machine learning	62	1	constraint satisfaction	53	4
protein coupled receptors	46	1	fault localization	52	4
generative topographic mapping	38	1	model based testing	49	4
micro aerial vehicle	24	2	based reasoning	135	5
control architecture	18	2	multi agent systems	129	5
supervised autonomy	17	2	machine learning	99	5
vessel inspection	16	2	multi agent	88	5
localization	15	2	artificial intelligence	62	5
humans	115	3	norm	60	6
multiple sclerosis	90	3	mathematical fuzzy logic	45	6
diabetic retinopathy	75	3	discrete fuzzy	41	6
magnetic resonance imaging	74	3	uninorms	34	6
anonymity	67	3	aggregation functions	30	6

Table 6. Top-5 communities keywords in cognitive network $SIM_{keywords}$

keyword	freq	community	keyword	freq	community
fuzzy inductive reasoning	67	A	multi agent systems	123	D
indoor positioning	62	A	multi agent	90	D
machine learning	55	A	machine learning	78	D
protein coupled receptors	46	A	based reasoning	73	D
generative topographic mapping	38	A	electronic institutions	60	D
humans	108	B	based reasoning	65	E
multiple sclerosis	88	B	machine learning	33	E
diabetic retinopathy	75	B	clustering	29	E
magnetic resonance imaging	74	B	data mining	28	E
anonymity	67	B	classification	19	E
model based diagnosis	82	C			
planning	59	C			
constraint satisfaction	54	C			
fault localization	52	C			
model based testing	49	C			

Table 7. Top-5 communities keywords in cognitive network $\text{SIM}_{cited_authors}$

References

[1] Nooteboom B. Towards a dynamic theory of transactions. Journal of Evolutionary Economics. 1992;2(4):281.

[2] Nooteboom B, Van Haverbeke W, Duysters G, Gilsing V, van den Oord A. Optimal cognitive distance and absorptive capacity. Research Policy. 2007 Sep 1;36(7):1016–34.

[3] Wang Q, Sandström U. Defining the role of cognitive distance in the peer review process with an explorative study of a grant scheme in infection biology. Research Evaluation. 2015 Jul;24(3):271–81.

[4] Ahlgren P, Colliander C. Document–document similarity approaches and science mapping: Experimental comparison of five approaches. Journal of Informetrics. 2009 Jan;3(1):49–63.

[5] Ahlgren P, Jarneving B. Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping. scientometrics. 2008 Aug 1;76(2):273–90.

[6] Janssens F, Quoc VT, Glänzel W, Moor BD. Integration of textual content and link information for clustering of science fields. :6.

[7] Lyu X, Costas R. Studying the cognitive relatedness between topics in the global science landscape: The case of Big Data research. Journal of Information Science. 2022 Sep 18;01655515221121970.

[8] Janssens F, Glänzel W, De Moor B. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. Katholieke Universiteit Leuven, Open Access publications from Katholieke Universiteit Leuven. 2007 Jan 1;

[9] Zhao D, Strotmann A. Author Bibliographic Coupling: Another approach to citation-based author knowledge network analysis. Proc Am Soc Info Sci Tech . 2008 ;45(1):1–10.

[10] Kessler MM. Bibliographic coupling between scientific papers. American Documentation. 1963; 14(1):10–25.

[11] Radicchi F, Fortunato S, Vespignani A. Citation Networks. In: Scharnhorst A, Börner K, van den Besselaar P, editors. Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences [Internet]. Berlin, Heidelberg: Springer; 2012 [cited 2023 Jul 16]. p. 233–57. (Understanding Complex Systems). Available from: https://doi.org/10.1007/978-3-642-23068-4_7

[12] Radicchi F, Fortunato S, Markines B, Vespignani A. Diffusion of scientific credits and the ranking of scientists. Phys Rev E. 2009 Nov 11;80(5):056103.
[13] Newman MEJ. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E. 2001 Jun 28;64(1):016132.
[14] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004 Feb 26;69(2):026113.
[15] Baeza-Yates R, Ribeiro-neto B, Mills D, Bonn O, Juan S, Mexico M, et al. Modern Information Retrieval. 1999 Jul 17; (Chapter 2)
[16] Csardi G, Nepusz T (2006). "The igraph software package for complex network research". InterJournal, Complex Systems, 1695.
[17] R Core Team. R: A Language and Environment for Statistical Computing, 2023.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230662

Full-Range Yaw Prediction: A Multi-View Approach for 3D Head Model Pose Estimation Using Convolutional Neural Networks

Álvaro HEREDIA-LIDÓN [a,1], Neus MARTÍNEZ-ABADÍAS [b], and Xavier SEVILLANO [a]

[a] *HER - Human-Environment Research Group, La Salle - Universitat Ramon Llull, Barcelona, Spain*

[b] *Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals (BEECA), Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain*

ORCiD ID: Álvaro Heredia-Lidón https://orcid.org/0000-0001-6325-8744, Neus Martínez-Abadías https://orcid.org/0000-0003-3061-2123, Xavier Sevillano https://orcid.org/0000-0002-6209-3033

Abstract. Head pose estimation, a crucial task in computer vision, involves determining the orientation of a person's head in 3D space through yaw, pitch, and roll angles. While recent techniques present excellent results in estimating head pose from a single 2D RGB image when the head faces the camera directly, few methods exist for pose estimation from arbitrary viewpoints. This problem is emphasised when the input data is in 3D, such as heads reconstructed models from magnetic resonances, where an accurate estimation of the pose is necessary for diagnostic purposes. To overcome these limitations, we make a first step by proposing a method for fine-grained head pose estimation across the full-range of yaw angles using 3D head synthetic models. Our approach involves transforming the 3D pose estimation problem into a multi-class 2D image classification problem by representing 3D head models as multi-view projection images. Leveraging a fine-tuned ResNet50 convolutional neural network, we tackle the task of head pose estimation with fine granularity of 5°, effectively discretizing the 360° yaw orientations. For the evaluation of our proposal, we train and test our models with the publicly available FaceScape and 3D BIWI datasets obtaining promising results.

Keywords. Head pose estimation, 3D orientation, Multi-view projections, Yaw, Wide range

1. Introduction

Head pose estimation (HPE) refers to a set of techniques within the field of computer vision that aim at determining the orientation of a person's head in the 3D space through

[1] Corresponding Author: Álvaro Heredia-Lidón, alvaro.heredia@salle.url.edu

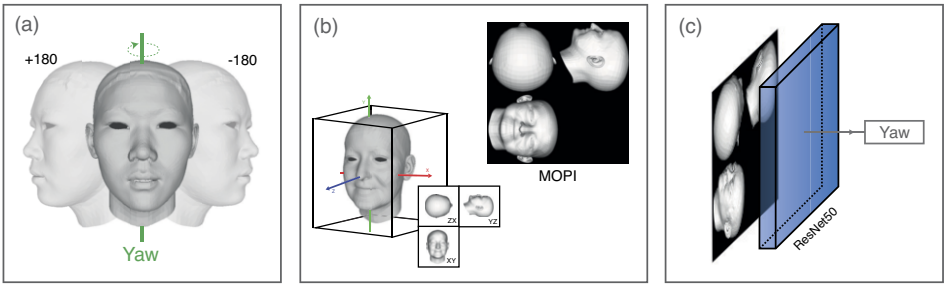


Figure 1. (a) Applied (±180°) rotation angles in yaw. (b) MOPI generation from tree orthogonal 2D projections (i.e. XY, YZ, ZX plane projections) images of the 3D head model and combining them into a single RGB image following the first dihedral European projection method as presented in [7] (c). Yaw angle prediction from a MOPI based on a fine-tuned ResNet50 [8] backbone.

three angles: yaw, pitch and roll. For an extensive survey of HPE methods and applications see [1].

The vast majority of HPE methods usually work with 2D data, such as RGB images, depth images or videos [1]. These techniques have demonstrated high performance when the head is frontal to the camera and the range of angles to be predicted is narrow (i.e. ±90° in yaw, ±30° pitch and ±45° in roll).

However, few methods exist that can estimate the head pose from any viewpoint, what is referred to as *'full-range'* techniques [2]. Very recent proposals are WHENet [2] and 6DRepNet [3], both based on a convolutional neuronal network (CNN) backbone.

These methods present excellent results from 2D data. Nevertheless, there are no proposals in the literature capable of estimating full-range head orientation from 3D data, which might be useful in accurately orienting and aligning reconstructed 3D head models from structural magnetic resonance imaging (MRI) for diagnostic purposes [7].

In this paper, we present an ongoing work that seeks to overcome these limitations. Due to the lack of ground truth 3D models reconstructed from MRI for the evaluation, our initial proposal estimates the orientation of 3D head synthetic models across the full-range of possible yaw angles (±180°) (see Figure 1a). To do so, we take 2D projection images of the 3D head models and fine-tune a ResNet50 CNN [8] to solve the HPE problem as a multi-class image classification challenge.

For the evaluation of our proposal, we trained our method with 2D multi-view projections of the 3D head models of the FaceScape [5] dataset and tested with 2D multi-view projections of the 3D head models of BIWI [6] (3D BIWI) dataset combining yaw, pitch and roll rotations. The preliminary results obtained provide us with an accuracy higher than 98%, auguring promising results for our work in progress.

2. Our proposal

The idea of our proposal is to estimate each of the three angles separately. In this case, we make a first step by estimating the full-range of yaw angles. The main feature of our proposal involves converting the 3D head pose estimation problem into a multi-class 2D image classification problem. In that sense, each 3D head model is represented with

Table 1. Evaluation of our proposal in terms of accuracy combining different yaw (Y), pitch (P) and roll (R) rotations, and datasets.

	Rotations	Train	Test	Yaw prediction
Accuracy (%)	Y=±180° P=0° R=0°	FaceScape (70%)	FaceScape (30%)	99.54
Accuracy (%)		FaceScape (70%)	3D BIWI	64.47
Accuracy (%)	Y=±180° P=±30° R=±45°	FaceScape (70%)	FaceScape (30%)	98.55
Accuracy (%)		FaceScape (70%)	3D BIWI	98.05

multi-view orthographic projection images (MOPI) [7], where multiple orthogonal 2D views of the 3D head are combined into a single RGB image (see Figure 1b).

For yaw prediction, we select a multi-class classification convolutional neural network as ResNet50 [8] pre-trained on ImageNet [9] and perform a fine-tunning to estimate the 72 classes of our training set, which represents the ±180° yaw angles orientations with a granularity of 5° (see Figure 1c).

3. Experiments and results

For the evaluation of our proposal, we present two experiments[2]: *i)* training and testing the model only with yaw rotations (in the ±180° range) and *ii)* training and testing the model with yaw rotations combining random pitch (in the ±30° range) and roll (in the ±45° range) rotations. In terms of training hyperparameters, we initialised the learning rate at 0.003, set the batch size to 16, and capped the maximum number of epochs at 6, as no further loss reduction was obtained if trained for longer.

In terms of data used, we have employed two different publicly available datasets: the 846 3D head models of the FaceScape [5] dataset and the 24 3D head models of the BIWI [6] dataset (3D BIWI), all of them with a neutral expression. For training and testing, the data have been randomly split and there were no data shared between the two groups. The metric used for evaluation is accuracy.

In the first experiment (see two top rows of Table 1), we rotated each 3D head of FaceScape dataset in 72 orientations, covering the ±180° yaw angles with a granularity of 5°. This generated 846 MOPIs per class. For testing, the same process was applied with 3D BIWI, generating 24 MOPIs per class. As can be seen in Table 1, in the split of train and test of FaceScape the accuracy is quite good, but when testing the model with 3D BIWI, accuracy decreases. This major error was probably because there were not enough training images per class and the model is not able to generalise enough.

For the second experiment (see two bottom rows of Table 1), each 3D head of FaceScape dataset has been rotated at each of the possible yaw angles (maintaining 5° in granularity) and for each of them, 12 MOPIs have been generated where pitch (in the ±30° range) and roll (in the ±45° range) angles are randomly combined to train the model. This generates a total of 10152 MOPIs per class (846 3D models x 12 MOPI). The same protocol has been replicated with 3D BIWI, generating 288 MOPIs per class for testing. In this case, we have observed that the accuracy decreases in the train and test split of FaceScape with respect to the previous experiment. However, when testing

[2]Run on a PC with i9-10980XE CPU @ 3.00GHz × 36 cores and NVIDIA GeForce RTX 2080 Ti GPU

with 3D BIWI, we observed that the accuracy increases considerably, which shows that having a larger volume of training images per class and more variability within the data (combination of yaw, pitch and roll rotations) allows us to obtain a more robust model, offering good results with datasets different from the training set.

4. Conclusions and future work

In this paper, we present a first step in the investigation of a method for the 360° 3D head pose estimation, by designing a method for the fine-grained estimation of full-range yaw angles. The key characteristic of our approach lies in the use of multi-view projection images, which enables the generation of a unique 2D representation of the pose of the 3D heads. This allows us to address the 3D orientation problem as a multi-class 2D image classification problem that we solve using a ResNet50 backbone.

Our preliminary results indicate that the proposed approach is capable of determining the yaw angle with a very high accuracy and a fine granularity of 5°. This validates the idea of generalising our previous proposal, which was limited to a few orthogonal orientations in the context of magnetic resonance imaging [7], to any head pose.

In this sense, we will continue working in several directions, such as *i)* completing our HPE proposal by including a cascade of classifiers that estimate the pitch and roll angles, *ii)* studying the possibility of decreasing the current angular estimation granularity of 5° and including a regression layer to estimate a continuous range of angles, and *iii)* comparing our proposal with state-of-the-art HPE methods.

Acknowledgements The research in this paper was supported by the Joan Oró grant (FI 2022) from the Recerca i Universitats Departament (DRU) of the Generalitat de Catalunya with grant identification 2023 FI-2 00160 and the European Social Fund. The authors would also like to thank the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya (2021 SGR01396, 2021 SGR00706).

References

[1] Abate AF, Bisogni C, Castiglione A, Nappi M. Head pose estimation: An extensive survey on recent techniques and applications. Pattern Recognition. 2022 Jan;127:108591.
[2] Zhou Y, Gregson J. WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose. British Machine Vision Virtual Conference. 2020.
[3] Hempel T, Abdelrahman AA, Al-Hamadi A. 6D Rotation Representation For Unconstrained Head Pose Estimation. IEEE International Conference on Image Processing (ICIP). 2022. p. 2496-500.
[4] Voit M, Nickel K, Stiefelhagen R. Multi-view head pose estimation using neural networks. The 2nd Canadian Conference on Computer and Robot Vision (CRV'05). 2005. p. 347-52.
[5] Yang H, Zhu H, Wang Y, Huang M, Shen Q, Yang R, et al. FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. p. 598-607.
[6] Fanelli G, Dantone M, Gall J, Fossati A, Van Gool L. Random Forests for Real Time 3D Face Analysis. Int J Comput Vis. 2013 February;101(3):437-58.
[7] Heredia-Lidón A, et al. Automated Orientation Detection of 3D Head Reconstructions from sMRI using Multiview Orthographic Projections: An Image Classification-Based Approach. 11th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA). 2023.
[8] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016. p. 770-8.
[9] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84-90.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230663

A Deep Learning-Based Pipeline for Computing Brain Biomarkers from MRI

Jordi MALÉ [a,1], Yann HEUZÉ [b], Juan FORTEA [c], Neus MARTÍNEZ-ABADÍAS [d] and Xavier SEVILLANO [a]

[a] *HER - Human-Environment Research Group, La Salle - Universitat Ramon Llull, Barcelona, Spain*
[b] *CNRS, Université de Bordeaux, France*
[c] *Hospital of Sant Pau, Barcelona, Spain*
[d] *Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals (BEECA), Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain*
ORCiD ID: Jordi Malé https://orcid.org/0000-0003-4566-1921, Yann Heuzé https://orcid.org/0000-0002-0660-9613, Juan Fortea https://orcid.org/0000-0002-1340-638X, Neus Martínez-Abadías https://orcid.org/0000-0003-3061-2123, Xavier Sevillano https://orcid.org/0000-0002-6209-3033

Abstract. This ongoing work outlines a computer vision and deep learning-based pipeline to identify and detect brain biomarkers of diagnostic potential from magnetic resonance imaging (MRI) scans. In this context, this paper describes and analyses two strategies for brain landmark detection, which is a key step in brain biomarker identification: one based on a single Deep Convolutional Neural Network (DCNN) that detects multiple landmarks, and the other based on an ensemble of DCNNs trained to detect one landmark each. Based on our evaluation using two distinct datasets, our preliminary findings demonstrate that the ensemble of DCNNs achieves superior accuracy in landmarking. Specifically, it successfully detects 84% of the landmarks within a 3mm proximity to their actual locations, with an average error of less than 2mm. In contrast, a single DCNN exhibits an average error of approximately 3mm and locates only 59% of the landmarks within a 3mm distance from their true positions.

Keywords. brain landmarking, magnetic resonance imaging, deep learning

1. Introduction

The analysis of brain magnetic resonance imaging (MRI) is used for the diagnosis and prognosis of multiple genetic and developmental disorders, such as schizophrenia or Alzheimer's disease, among others. Automatic definition [1] and detection of brain landmarks [2] is crucial for this analysis, as landmark information is used for segmentation [3], registration [4] and diagnosis [5]. In this context, the obtainment of disorder-specific brain biomarkers paves the way towards personalized medicine.

[1] Corresponding Author: Jordi Malé, jordi.male@salle.url.edu.

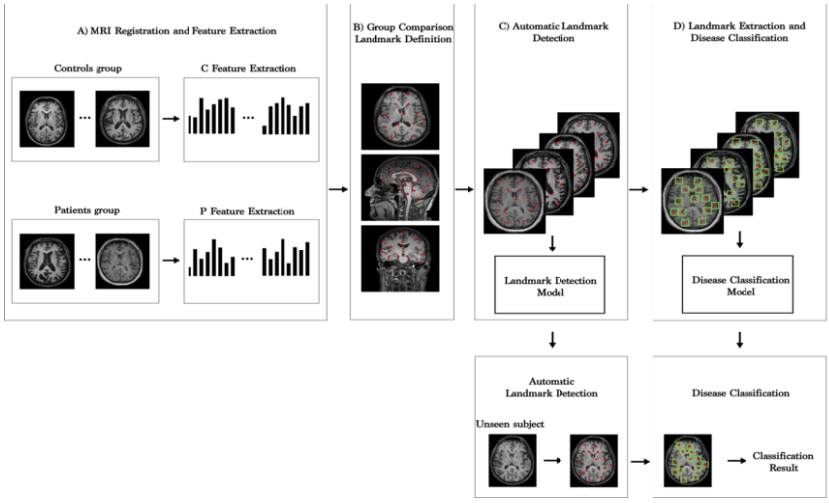


Figure 1. Proposed landmark-based disease identification framework, including (1) MRI registration and feature extraction, (2) statistical group comparison and data-driven landmark definition based on the identified discriminative voxels, (3) automatic landmark detection using DCNN, (4) landmark information extraction and (5) model-based disease classification based on the features of each input landmark

This paper presents an end-to-end pipeline –see Figure 1– that, leveraging on large datasets of brain MRI scans of healthy subjects and patients, *i)* identifies statistically significant brain anatomy differences between groups, *ii)* defines brain biomarkers of diagnostic potential based on those anatomical differences, *iii)* defines a set of brain landmarks related to the identified biomarkers, *iv)* detects those landmarks on brain MRI scans of unseen subjects, and *v)* uses the detected biomarkers to diagnose new subjects.

In the context of the proposed pipeline, the goal of this paper is to present our first results in the automatic brain landmark detection task. To that end, we trained two different Deep Convolutional Neural Network (DCNN) architectures: firstly, Multi-Landmark (ML) DCNN that predicts N landmarks, and secondly, an ensemble of N individual Single-Landmark (SL) DCNNs that predict only one landmark each.

Our preliminary results indicate that the ensemble of SL models is much more accurate than the ML model, placing landmarks with an average error below 2mm, and detecting nearly 3 times more landmarks less than 1mm away from its true location.

2. Automatic Landmark Detection

This section describes the detection of 2-dimensional landmarks on the Mid-Sagittal Plane (MSP[2]) of brain MRI scans. The ground truth coordinates of $N = 8$ landmarks (identified as diagnostic potential landmarks by our experts) were obtained by [6].

We compared two architectures of DCNN-based models: a single DCNN trained to detect the 8 landmarks, (multi-landmark model, or ML), and an ensemble of 8 DCNNs trained to detect one landmark each (single-landmark models, or SL).

[2]The MSP is the MRI slice that separates the brain into two almost-identical hemispheres.

All the DCNNs in both architectures are fed with the MSP of the brain MRI scans, and comprise: four CNN blocks (a convolutional layer, batch normalization layer, ReLU activation function, and a max-pooling layer), followed by three fully connected layers, ReLU activation functions and dropout layers to avoid overfitting. The loss function of the ML model is the average of the 8 Euclidean distances between the predicted and the ground truth landmarks. Whereas, each DCNNs in the SL model ensemble used the distance between the predicted and ground truth landmark coordinates as loss function.

3. Experimental setup

The DCNN models were trained[3] with 1837 brain MRI scans obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI-1) database. To evaluate the implemented models, we tested them with a dataset of 141 brain structural T1 MRI scans provided by Hospital Sant Pau Memory Unit (Barcelona, Spain). Different datasets were employed to evaluate the model's ability to generalize across different sources of data.

The accuracy of the landmarking results were measured in terms of: *i)* landmarking error, i.e. Euclidean distance (in mm) between the coordinates of the ground truth and the predicted landmark coordinates, and *ii)* Successful Detection Rate (SDR), defined as the percentage of predicted landmarks to be within a certain radius of their real position. We evaluated SDR at increasing radiuses from 1 to 5 mm.

4. Results, conclusions and further work

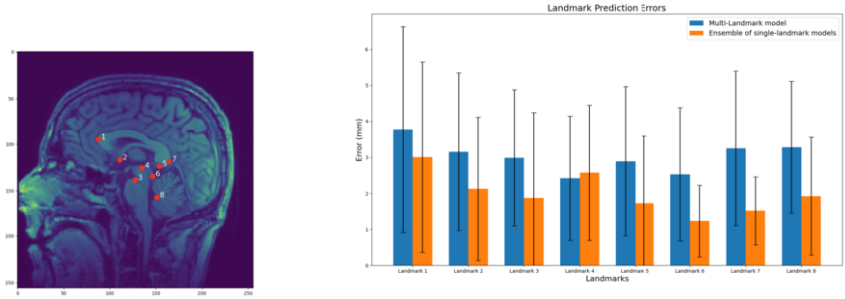
Table 1 presents the landmarking error (mean±standard deviation), SDR for the ML and SL models, and testing time (time needed to make the predictions) per image. Moreover, Figure 2b depicts the landmarking error for each specific landmark for both models.

Table 1. Landmarking error (in mm) of all landmarks for each model, and SDR of both models.

Model	Error (mm)	SDR (%)					Testing time (ms)
		1mm	2mm	3mm	4mm	5mm	
Multi-Landmark	3.03 ± 2.13	12.32	36.52	58.86	75.79	87.06	11.88
Single-Landmark	1.99 ± 1.94	29.52	66.31	83.86	90.51	93.52	92.15

As our preliminary results indicate, the ensemble of SL models outperforms the ML model both in terms of SDR and landmarking error. It is to note that the ensemble of 8 SL models places nearly 84% of the landmarks less than 3 mm away from their true location. Moreover, this higher accuracy is found for each of the 8 landmarks. This leads to the conclusion that having a model that focuses only on the loss function of the Euclidean distance of a single landmark is preferable than having a model that focuses on the average loss functions of 8 landmarks and the relative position between them.

[3] All models were implemented using PyTorch. The learning rate was initially set to 0.005 was gradually reduced to a final value of 0.0005 by applying a gamma factor of 0.95. An Adam optimizer was used with a batch size of 256. The SL models were trained for 5,000 epochs, and the ML model was trained for 10,000 epochs. All experiments were conducted on a PC with Intel Core™ i9-10980XE CPU @ 3.00GHz × 36 cores and NVIDIA GeForce RTX 2080 Ti GPU.



(a) (b)

Figure 2. (a) Set of N landmarks located on the MSP, and (b) Landmarking error (in mm) using the ensemble of SL models (orange bars) and the ML model (blue bars) (average value and standard deviation per landmark).

This paper has presented a deep learning-based pipeline to identify discriminative biomarkers for diagnosing psychotic disorders. A first approach to automatically detect landmarks has been implemented and validated with different datasets, obtaining promising results. Future research will include *i)* training the models using data augmentation techniques, *ii)* 3D landmarking, and *iii)* implementing the pipeline depicted in Figure 1.

Acknowledgements

This work was partly supported by Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya (2021 SGR01396, 2021 SGR00706), the Spanish Ministry of Science, Innovation, and Universities under grant PID2020-113609RB-C21, the Fondation Jerome Lejeune under grant 2020b cycle-Project No.2001, and the Joan Oró grant (FI 2023) from the DRU of the Generalitat de Catalunya and the European Social Fund (2023 FI-1 00014). Data used in this work were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[4].

References

[1] Zhang J, Gao Y, Gao Y, Munsell B. Detecting Anatomical Landmarks for Fast Alzheimer's Disease Diagnosis. IEEE Transactions on Medical Imaging. 2016 06;35:1-1.

[2] Yang X, Tang WT, Tjio G, Yeo SY, Su Y. Automatic detection of anatomical landmarks in brain MR scanning using multi-task deep neural networks. Neurocomputing. 2020;396:514-21.

[3] Bokde ALWea. A new rapid landmark-based regional MRI segmentation method of the brain. Journal of the Neurological Sciences. 2002;194(1):35-40.

[4] Han D, Gao Y, Wu G, Yap PT, Shen D. Robust Anatomical Landmark Detection for MR Brain Image Registration. vol. 17; 2014. p. 186-93.

[5] Liu M, Zhang J, Adeli E. Landmark-based Deep Multi-Instance Learning for Brain Disease Diagnosis. Medical Image Analysis. 2017 10;43.

[6] Asaei A. Brain MRI landmark identification and detection; 2015. Available from: http://hdl.handle.net/20.500.12648/650.

[4]The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report (see http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf).

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230664

Identifying Patterns Between Acoustic Environment and Visual Landscape Through Semantic Segmentation Based on Deep Learning

Riccardo Giacalone[a,1], Carlos Guerrero-Mosquera[a,1] and Xavier Sevillano[a,1]
[a] *HER - Human-Environment Research Group, La Salle - Universitat Ramon Llull, Barcelona, Spain, Quatre Camins, 30, 08022 Barcelona, Spain*
ORCiD ID:
Riccardo Giacalone https://orcid.org/0009-0008-0581-4990
Carlos Guerrero-Mosquera https://orcid.org/0000-0001-8265-3651
Xavier Sevillano https://orcid.org/0000-0002-6209-3033

Abstract. This work is part of the research project "Sons al Balcó" conducted by La Salle - Universitat Ramon Llull, which examines the impacts of noise pollution on human perception and mental health, specifically focusing on the perception of noise in Catalonia during the lockdown in 2020 and the return to normalcy in 2021. The purpose of this research is to identify patterns between the soundscape and the visual landscape of participants' environments. To achieve this, we have developed a pipeline to automatically analyse the visual landscape of participants' environments by semantically segmenting the keyframes of their videos using deep neural networks. Specifically, we use the SegFormer model, a Transformer-based framework for semantic segmentation that integrates Transformers with lightweight MLP decoders. This pipeline facilitates the efficient and accurate identification of different objects, to understand the complex relationships among the acoustic environment, visual landscape, and human perception. We expect that our findings will offer insights into the design of urban and suburban areas that promote well-being and quality of life.

Keywords. Artificial Neural Networks, Artificial Vision, Image Processing, Machine Learning, Deep Learning, Transformers, Semantic Segmentation

1. Introduction

In recent years, there has been a rise in interest in the effects of noise pollution on human perception and mental health. The project "Sons al Balcó" conducted by La Salle - Universitat Ramon Llull seeks to investigate the effects of the lockdown during the COVID-19 pandemic on human perception of the acoustic environment in Catalonia [1]. This paper aims to understand the impact of COVID-19 lockdowns on noise perception

[1] Corresponding Author:
Riccardo Giacalone; E-mail: ricardo.giacalone@students.salle.url.edu
Carlos Guerrero-Mosquera; E-mail: carlos.guerrero@salle.url.edu
Xavier Sevillano; E-mail: xavier.sevillano@salle.url.edu

due to changes in human activity and mobility patterns and analyzes the relationship between participants' perception of the acoustic environment and the landscape of their surroundings using semantic image segmentation.

2. Datasets

Two databases were used in this study: Cityscapes and Sons al Balcó. Cityscapes is a benchmark dataset for semantic segmentation [2], which comprises high-resolution images of urban environments captured across 50 cities in Germany. This dataset includes pixel-level annotations for 30 distinct object classes [3]. For this study, we used the 19 classes and two files which contain RGB images and ground truth annotations. To prepare the data and convert the annotations the label images, we assign a unique grayscale value to each class label and create grayscale label images where each pixel corresponds to a specific class label.

The Sons al Balcó database is a comprehensive dataset that was collected during the COVID-19 lockdown in Catalonia. The data was obtained through a socio-acoustic digital participatory survey, which was made available to Catalonian citizens via social media and the press. The survey included a series of questions regarding socio-demographics, residential soundscape quality, and individual positive and negative perceptions prior to and during the lockdown. Participants were also given the option to contribute videos of their residential soundscapes.

3. Experimental setup

To ensure the quality of the data, each video was manually reviewed before analysis to eliminate those with insufficient lighting, private information, or a singular focus on an object. After this filtering process, 188 videos from 2020 and 165 videos from 2021 were selected for further analysis.

Next, we evaluated the performance of five semantic segmentation models on the Cityscapes dataset using the MMSegmentation framework [4]. To develop the models, we built on top of the MMSegmentation library, which offered the training and testing pipelines needed for this task and was commonly used in the development of these models. The models were evaluated based on the mean Intersection over Union (mIoU) metric, a commonly used metric in the evaluation of segmentation models.

After evaluating the models, we selected the best performing semantic segmentation model to perform inference on videos from both the 2020 and 2021 Son al Balcó datasets. For each video, the average presence of every class was computed by averaging the predicted classes from the segmented frames. Then, we created mappings between video files and segmentation results to process the data.

Finally, the survey data was merged with the segmentation data for each video using a Python Pandas DataFrame, and Pearson and Spearman correlations were plotted to examine the relationships between the variables. Finally, an attempt was made to interpret the correlations using a logistic regression model and examining its coefficients.

4. Results and Discussion

To compare the selected Cityscapes models, they were all evaluated on the 500 validation images set and ranked by the mean Intersection over Union metric.

Table 1. Comparison of Cityscapes Models on Validation Set - Comparison of selected Cityscapes models based on Mean Intersection over Union (mIoU) scores on the 500-image validation set. Models are ranked in descending order of mIoU scores: SegFormer (82.25), OCRNet (81.35), HRNet (80.67), DeepLabv3 (80.48), and DeepLabv3+ (80.46).

Model name	mIoU
SegFormer	82.25
OCRNet	81.35
HRNet	80.67
DeepLabv3	80.48
DeepLabv3+	80.46

The table shows that SegFormer is the clear winner in terms of performance with an mIoU of 82.25%.

The next result shows the Spearman correlation between the noise perception reported in the survey and the percentage of presence of the class labels in the recorded videos. The first figure corresponds to the surveys taken in 2020, while the second in 2021.

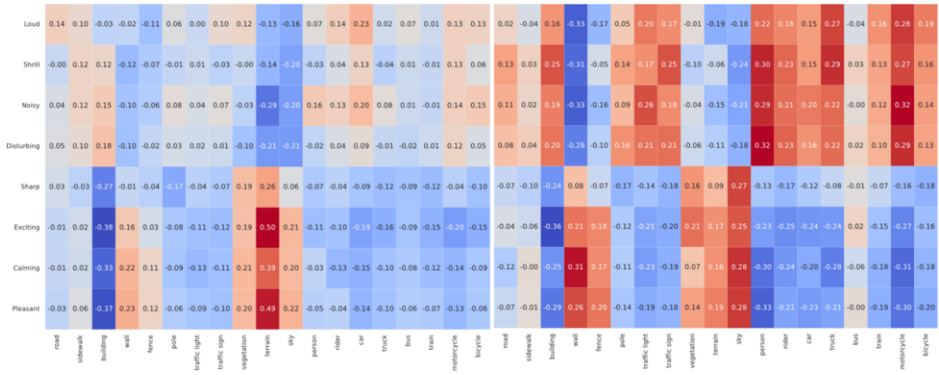


Figure 1. Spearman correlation between the 2020 (left) and 2021 (right) segmentation classes and perception adjectives. Note that the Spearman correlations do not show strong relationships; however, it can be noticed that the relationships differ significantly between the 2020 and 2021 surveys. In 2020 the strongest positive correlations are between positive attributes and "terrain". In 2021 the highest correlations occur with the class "person" and with vehicles and they are all with negative attributes.

To introduce more insights into the analysis with significant correlations, another approach was performed by training a logistic regression model to classify the class label data into a positive or negative perception.

The results indicate that SegFormer performs better in terms of performance metrics. However, in addition to its top-performance, this model offers other range of features that make it an interesting choice for this type of problem, including: (i) It is not a CNN based model, but it is based on an encoder-decoder architecture, (ii) it uses Mix-FFN layers as positional embeddings, addressing the drawbacks of resolution-dependence, redundancy, and inefficiency, (iii) uses outputs multiscale features due to hierarchically structured transformer encoder, (iv) avoids complex decoders combining both local attention and global attention to render powerful representations.

The step of fitting a logistic regression model and analyzing the coefficients was performed after observing a high correlation between the 4 negative attributes ("Loud",

"Shrill", "Noisy", and "Disturbing") and the four positive attributes ("Sharp", "Exciting", "Calming", "Pleasant"). A negative coefficient means that it influences negatively in the calculation of the perception. A positive coefficient is the opposite, it positively affects the overall perception. The results from the logistic regression model trained on the 2020 and 2021 datasets show that the coefficients for 2020 are different than the coefficients for 2021. This is because the class labels can include more than one object, like "rider", which could be a person riding a motorbike or a bicycle.

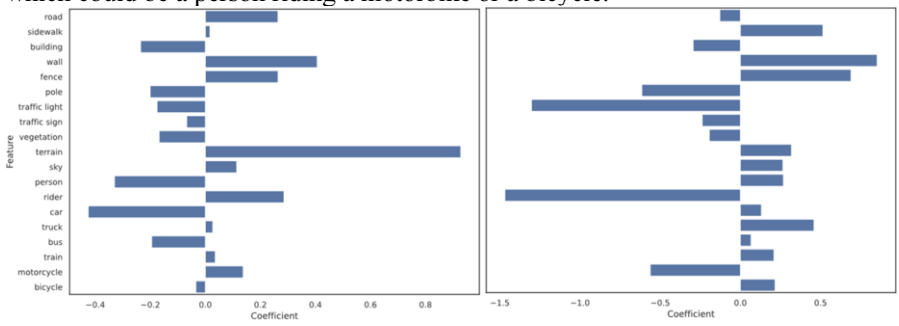


Figure 2. Logistic regression model coefficients for 2020 data (left) and 2021 data (right). Note that 2020 the feature with highest coefficient is "terrain", which contributes the most to a positive perception. About the negative perceptions, "car" is the feature with the lowest coefficient, meaning that negatively influences the overall perception of noise if present in the video. In 2021 the strongest coefficients are negative. "Traffic light" and "rider" are twice and thrice greater than the highest positive coefficients. Notice how rider in 2020 was a positive feature, while in 2021 it is the most negative coefficient; this is because rider includes several categories. In 2020 there were probably more bicycles and electric scooters than people riding a motorbike, due to the confinement. Also, in 2020 "car" is the most negative feature while in 2021 it is slightly positive. There could be several reasons for this divergence, for instance a car in 2020 could be heard from further since there were less cars and could impact the noise perception more than a car in 2021 as there were more cars on the road.

In conclusion, this study indicates that noise perception is not solely dependent on the visual features, as evidenced by the correlations and coefficients of the logistic regression model. Rather, it is also influenced by contextual factors. Specifically, our results suggest that the year 2020 played a significant role in shaping the importance of the various features. To improve the project, it would be interesting to create a model that predicts noise perception more reliably by taking into consideration the sound present in the videos.

Acknowledgment

The research in this paper was supported by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya (grant 2021 SGR01396) and the Secretaria d'Universitats i Recerca from the Departament d'Empresa i Coneixement (Generalitat de Catalunya) and Universitat Ramon Llull (grant 2020-URL-Proj-054).

References

[1] LaSalle. Ramon Llull University. SONS AL BALCÓ - Map of the Soundscape of the Lockdown in Catalonia. 20 May 2020. [Online]. Available: https://www.salleurl.edu/en/research/project/sons-al-balco.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. 6 Apr 2016. [Online]. Available: https://arxiv.org/abs/1604.01685.

[3] Cityscapes Dataset – Semantic Understanding of Urban Street Scenes. 17 Oct 2020. [Online]. Available: https://www.cityscapes-dataset.com/.

[4] M. Contributors. OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020. [Online]. Available: https://github.com/open-mmlab/mmsegmentation.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230665

OPNet: A One-Shot Image Similarity Algorithm for Production Systems

Karl Fabian Svensson[a,1] and Carlos Guerrero-Mosquera[a,2]
[a] *La Salle Ramon Llull University*
ORCiD ID: Carlos Guerrero-Mosquera https://orcid.org/0000-0001-8265-3651

Abstract. The appearance of new trends in the field of cognitive neuroscience, for example object persistence, has paved the way for the evolution of deep CNNs into Siamese Neural Network architectures such as OPNet. These networks allow for image recognition without the need for expensive labelled data. In this work, we apply this technology to a small Spanish tech e-commerce struggling with the production of their customizable products. Our goal was to automatically identify each product's order in the company's internal system by matching photos of the products taken by workers with system-generated images. After testing various architectures, we achieved 91% accuracy with a triplet loss model using deep CNN embedding networks. The algorithm was trained on a dataset of 9696 unique product images captured in the company's production department. The paper details the technical aspects of the Siamese Neural Network architecture, including the triplet loss and SoftMax distance function used to train it. Our results demonstrate the potential of these deep learning models to generate practical benefits for firms, since it reduces human errors, while improving the effectiveness and efficiency of the company's internal processes.

Keywords. SNN, Image-Similarity, Triplet-Loss, Artificial Vision, Image Processing, Artificial Neural Networks, AI Problem Solving

1. Introduction

Despite the success of the neural network when it comes to image classification and object detection, image similarity has traditionally proven to be a more daunting task.

Advanced statistical comparison techniques such as comparing ontology trees or latent semantic indexing perform poorly when the data is not organized or structured in a specific and predictable manner.

Extracting feature vectors is an excellent way of analyzing images, since the information therein lies in the interaction between the pixels. Arguably the most popular machine learning algorithm for extracting feature vectors is called Convolutional Neural Network (CNN). The ConvNets which constitute these are excellent at capturing the spatial dependencies in an image by applying certain filters[3].

[1] Karl Fabian Svensson, karlfabian.svensson@students.salle.url.edu
[2] Carlos Guerrero-Mosquera carlos.guerrero@salle.url.edu

2. Algorithm

SNNs are state-of-the-art in the field of image similarity. They contain two identical subnetworks which perform an identical transformation on some input data. The two subnetworks commonly have a CNN architecture which ends in a fully connected network that constitutes some feature vector or output.

Two major sub-architectural branches can be distinguished within SNNs depending on the loss function employed.

The traditional SNN (Contrastive Loss Function) relies on an activation function, commonly SoftMax, to evaluate whether two images are similar. Du and Fang at Stanford output a SoftMax cross-entropy loss which returns 1 if the inputs are similar and 0 if they are not[2]. DeepFace use a similar approach, exploring different similarity metrics such as the Kai squared formula[4].

The second branch, called Triplet Loss Function, was first introduced by Hoffer et al and trains on three images for each iteration. Its objective is to minimize the distance between a query image and a positive image, and to maximize the distance between the query image and the negative image (see *formula 1*).

$$\| f(x_i^a) - f(x_i^p) \|_2^2 + \propto < \| f(x_i^a) - f(x_i^n) \|_2^2, \sum_i^N [\| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \propto]$$

Formula 1. Triplet Loss Distance Function

2.1. One-Shot Learning

The purpose of few-shot learning is to bring artificial intelligence models closer to human conceptual knowledge.

While humans can learn from just a handful of examples, the highest performing deep learning models are very “data-hungry”. One-Shot learning consists of training models to correctly classify elements having seen only one example of its class.

3. Methods

OPNet was proposed to improve the efficiency of the production department of a Spanish company that offers personalized products. Because of the personalization, the dataset used to train and test OPNet consists entirely of unique classes, i.e., the model is both one-shot training and one-shot testing. The company’s main product is the die-cut sticker.

In order to attain the best result possible, different types of embedding models were experimented with. In total, 7 variations were created, 3 using a SoftMax activation function and 4 with a Triplet Loss function.

4. Data Setting

OPNet was trained using 9696 black and white images of dimension 320x180. An additional 1250 images were stored independently of the model to be used uniquely for specific validation purposes.

All these images were taken by the production department at OriginalPeople throughout a 3-month period ranging from May 2nd, 2022, to August 5th, 2022.

In order to ensure data consistency, a small wooden structure was built with a camera attached to the top. A piece of black vinyl was then placed on the area directly below the camera to provide a neutral background for all images.

One difficulty was choosing the height at which to place the camera. The customizable stickers vary in size from 15x25mm to 575x140mm.

Ultimately, to achieve the best trade-off between resolution and not cutting off too much of large stickers, a height of 48cm was chosen which covers an area 400 mm wide. The images produced are 1280x960 pixels and weigh around 70 – 80 KB.

4.1. Preprocessing

In order to avoid the overhead that comes with preprocessing images at execution time, all images are preprocessed when taken and stored both in their original format and their manipulated format in the OPNet database. Refer to *figure 1*.

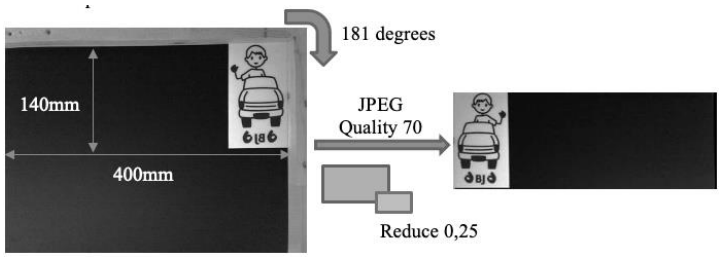


Figure 1. Preprocessing of images.

The resulting images have a dimension of 320x118 pixels and weigh around 1-2 KB. Both the original and the preprocessed images were stored to allow for future experimentation with pre-processing.

4.2. Sampling

Similar images are stored adjacent in memory to make it less costly to sample triplets. OPNet is trained using curriculum learning, which means that a sub-batch of semi-hard triplets are purposefully selected. Triplets are considered semi-hard if the distance between the negative and the query image is very similar to the distance between the latter and the positive image.

Based on tests performed, and in accordance with research by Appalaraju et al.[1], this speeds up training by about 20-30% and results in accuracy levels of a few percentage points higher for the same number of epochs, compared to if only randomly sampled triplets are used.

5. Results and Analysis

All models were evaluated using accuracy, by taking large batches of pairs and triplets and returning a percentage. Other visual functions were implemented to get a better understanding of why an error occurs.

Variations in complexity, depth and feature vector dimension were experimented with. See the summary of the results in *table 1* below:

Table 1. Results of all 7 models.

Model	Loss	Accuracy
SNN Iceberg	0,226	83,95%
SNN Gregory	0,35	88,15%
SNN Flowers	0,76	87,7%
Triplet Dense 1	-	77,6%
Triplet Dense 2	-	70,5%
Triplet CNN 1	-	90,7%
Triplet CNN 2	-	88,4%

5.1. Best Performing model: Triplet CNN 1

This model uses a CNN embedding network, same as the SNN models. The accuracy achieved was 90,7% and the AUC was 90% (see *figure 2)*.

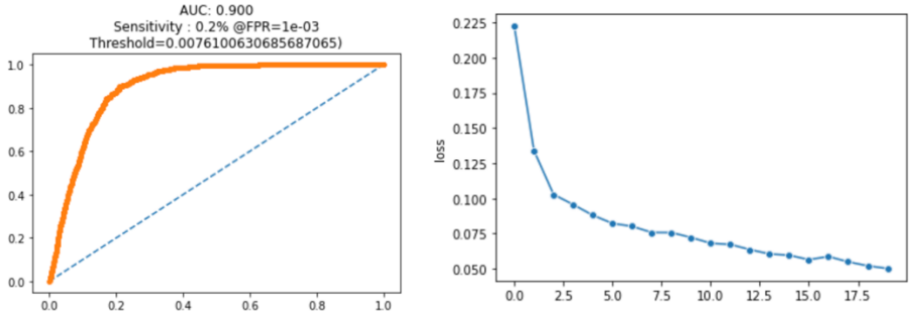


Figure 2: Triplet Model 3 – ROC and training of Triplet CNN 1

This is the best performing model so far, and the training does not converge after 20 epochs, meaning that better results should be attainable through further training.

6. References

[1] Appalaraju, S., Chaoji, V. (2017) *Image similarity using Deep CNN and Curriculum Learning*. ArXiv, abs/1709.08761.

[2] Du W, et al (2017) Siamese convolutional neural networks for authorship verification. Tech. rep. Stanford University

[3] Saha, S., 2018. *A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way*. [online] Medium. Available at: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 [Accessed 01 May 2022].

[4] Taigman Y, et al (2014) *DeepFace: closing the gap to human-level performance in face verification*. In: Proceedings of CVPR 2014 – the IEEE conference on computer vision and pattern recognition, pp 1701–1708

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230666

Analyzing Car Configurator Impact Through Genetic Algorithm from a Regional Perspective[1]

Juan Manuel GARCÍA-SÁNCHEZ [a,2], Xavier VILASÍS-CARDONA [a], Álvaro GARCÍA-PIQUER [a] and Alexandre LERMA-MARTÍN [b]

[a] *Research Group of Smart Society, La Salle-Ramon Llull University, 08022 Barcelona, Spain*

[b] *SEAT S.A., 08760 Martorell, Spain*

ORCiD ID: Juan Manuel García-Sánchez https://orcid.org/0000-0003-2569-6561, Xavier Vilasís-Cardona https://orcid.org/0000-0002-1915-9543, Álvaro García-Piquer https://orcid.org/0000-0002-6872-4262

Abstract. This study examine whether visits to the Car Configurator website from a specific area in Spain, referred to as "compound," have a similar impact compared to visits from other locations. The impact is measured by the correlation between clickstream data and sales records. To analyze this relationship, genetic algorithms are employed. The findings reveal that the genetic algorithm surpasses the benchmark values by more than 65 points, indicating its effectiveness. Moreover, the correlation achieved from locations outside the compound is found to be equivalent to the fitness obtained from the regions comprising these compounds. This suggests that the impact of website visits on sales is consistent across different geographic locations.

Keywords. Car Configurator, Genetic Algorithm, Correlation, Automotive Industry

1. Introduction

The Car Configurator (CC) website, offered by car manufacturers, allows customers to select from the company's range of models, estimate a purchase value and schedule a visit to the dealership [4]. This manuscript postulates the possibility of extracting customers' profile from this resource. We pursue to figure out if the visits to the webpage registered from one area of Spain, named compound, have the same impact than the ones received from another location. The impact is measured by to the correlation between clickstream and sales record. Data mining techniques employing genetic algorithms will be utilized. Results show that genetic algorithm outperforms in more than 65 points the benchmark values. Additionally, the correlation achieved from locations outside compound is equivalent to the fitness coming from the regions that make up these compounds.

[1] This work is partially funded by the Department de Recerca i Universitats of the Generalitat de Catalunya under the Industrial Doctorate Grant DI 2019-34.

[2] Corresponding Author: juanmanuel.g@salle.url.edu

The article is structured in the following way. Firstly, in Section 2, we present related works for the research topic. Hence, Section 3 describes the dataset provided by the automotive OEM source. Next, the methodology and results of the research are in Section 4 and Section 5, respectively. The discussion takes place in Section 6. Finally, Section 7 provides conclusions gained and future research paths.

2. Related Works

In the literature, there is a research study that investigates the optimization of compound delivery time distribution [2]. However, this study did not consider information from customers. Our approach to this task involves extracting consumers' purchasing behavior from the massive and noisy clickstream data. This can be thought of as a feature selection exercise. A comprehensive examination of the current advancements in methods for selecting relevant features can be found in [6]. Genetic algorithms deserve special mention, as they are a type of optimization algorithm. We focus on works that utilize correlation as an assessment metric of the genetic algorithm. This technique has demonstrated its effectiveness in various domains, including: (a) cancer research [5]; (b) build a data warehouse [7]; (c) feature selection in high-dimensional datasets [8]; and (d) identifying apple leaf disease in computer vision [1].

Unfortunately, we were unable to find research that specifically addresses user-generated data from non-transactional webpages. Online activity data can offer non-intrusive means to gain insight into consumers' purchasing behavior.

3. Dataset description

The dataset under analysis spans from April 2017 to January 2020, both included, and pertains exclusively to the Spanish national market. It contains what it has been called car variant, i.e., the user's geographical location (*GeoSeg*) and attributes of the configured car (TRIM, Engine and Color). Additionally, the users' connection day is saved (DOW). The user's geographical location is defined by the Spanish provinces. They are divided by the area of influence of the compound planted by the car manufacturer along the national territory. These compounds act as the company's warehouses. In total, there are six of them, which contain from 4 to 15 provinces each. Additionally, the OEM source has furnished sales records for identical time periods, car models, and geographic regions. Average values from both registers are on Table 1

Table 1. Average values of the weekly volume of configurations contained in clickstream data and sales per compound.

	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**	**Comp. 5**	**Comp. 6**
Avg. [config.]	501.83	8073.13	3249.97	2668.20	7056.11	3212.46
Avg. [sales]	43.94	145.03	130.31	96.07	181.74	126.42

4. Methodology

The genetic algorithm (GA) contains chromosomes made of the attributes defining car variant and DOW the user visited the webpage. The raw clickstream data is filtered using the information in the chromosome, which allows for the creation of CC filtered time

series and to compute the correlation with respect to the sales record, based on findings from note [3]. Nevertheless, we impose a delay of 8-week between both registers. It simulates the client exploration phase and it is based on the automobile manufacturing process. The lagging is executed monthly and results are averaged. If there are no users that meet the chromosome criteria, the fitness is penalized. New chromosomes are created by means of crossover and mutation. In addition, elitism is employed to prevent a decrease in fitness although if it remains constant for consecutive generations, an anti-stagnation systems is triggered.

For each one of the six compounds, GA is executed and the results are compared with baseline values. These baselines are obtained by computing the correlation between sales and online visits made from same compound. Subsequently, a detailed analysis of the best individuals is performed based on how provinces are segmented. Specifically, a fitness comparison is conducted between locations belonging to the compound and those ones out of the area of influence, including Kolmogorov-Smirnov test (KS test) to assess the equivalency between the two distributions.

5. Results

The parameters utilized in the study are: (a) the population size was 300 individuals, (b) each chromosome had 150 distinct rules, (c) there were 200 generations, (d) the tournament probability was set to 0.3 times the population size, (e) the crossover probability was set to 0.9, (f) the mutation probability was set to 1 divided by the population size, and (g) the theoretical maximum fitness value was set to 100.

Table 2 reflects the correlation obtained by the best candidate and the baseline for each of the compound regions. Furthermore, it is examined at province level per compound.

Table 2. Correlation obtained by the best GA solution,the baseline and fitness analysis per location within each compound. *GeoC* stands for provinces that belongs to the compound and *GeoA* signifies locations outside the compound

	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**	**Comp. 5**	**Comp. 6**
Benchmark [R2 Score]	31.08	29.46	30.48	32.08	35.35	29.92
GA Fitness [R2 Score]	96.92	97.22	98.32	98.00	99.25	96.15
Rate GeoC [%]	8.00	14.00	19.33	7.33	20.00	29.33
Rate GeoA [%]	92.00	86.00	80.67	92.67	80.00	70.67
Avg. GeoC	0.34 ±	1.70 ±	2.01 ±	2.24 ±	4.61 ±	1.89 ±
[R2 Score]	0.83	3.29	3.06	2.17	8.53	2.90
Avg. GeoA	2.26 ±	2.48 ±	2.00 ±	2.16 ±	3.00 ±	3.53 ±
[R2 Score]	3.15	4.15	3.45	4.19	4.61	5.72
KS test [p-value]	0.004	0.175	0.336	0.336	0.081	0.983

6. Discussion

In the present analysis from Table 2, all cases exhibited a considerable enhancement in terms of correlation when compared to the baseline values. Among the considered regions, Compound 2 displayed the most significant improvement, while Compound 5

obtained the highest correlation value. On average, the R2 Score was augmented by 66.25 $\pm$ 1.46.

On the other hand, it demonstrates that common locations are underrepresented in all analyzed scenarios. Compound 6 exhibits the highest frequency of common locations. Conversely, Compound 4 has the lowest frequency of common locations. In terms of the individual correlation of each category, KS test proves that fitness distributions are equivalent in each case, but Compound 1. In this last scenario, together with Compound 2 and Compound 6, the external locations provide more fitness that the common ones, as the average fitness is larger. The contrary occurs in Compound 4 and Compound 5. Finally, the difference between the two groups in Compound 3 is not significant.

7. Conclusions

The outcomes given by genetic algorithm outperforms the baseline values for all compounds under analysis. Furthermore, the analysis of the provinces revealed an underrepresentation of locations that define the compound's influence area. Nevertheless, the distribution of individual fitness scores for both groups of locations was found to be equivalent based on KS-test. The reasons behind this behavior cannot be determined with the available information. Therefore, we recommend exploring demographic data, including social and economic factors, as an external source that may shed light on this matter.

To conclude, we recommend that car manufacturers integrate this methodology into their commercial and logistics operations, as genetic algorithms are highly adaptable and customizable to various situations. Further steps involve leveraging online data sources to determine the optimal location for allocating manufactured cars, ensuring they are easily accessible to customers.

References

[1] Chuanlei, Z., Shanwen, Z., Jucheng, Y., Yancui, S., Jia, C.: Apple leaf disease identification using genetic algorithm and correlation based feature selection method. International Journal of Agricultural and Biological Engineering **10**, 74–83 (01 2017).

[2] García-Sánchez, J.M., Cardona, X., Martín, A.: Binary Delivery Time Classification and Vehicle's Reallocation Based on Car Variants. SEAT: A Case Study, pp. 147–150. Frontiers in Artificial Intelligence and Applications, IOS Press (10 2022).

[3] García Sánchez, J.M., Vilasís Cardona, X., Lerma Martín, A.: Influence of Car Configurator Webpage Data from Automotive Manufacturers on Car Sales by Means of Correlation and Forecasting. Forecasting **4**(3), 634–653 (2022). , https://www.mdpi.com/2571-9394/4/3/34

[4] Scholz, M., Dorner, V., Schryen, G., Benlian, A.: A configuration-based recommender system for supporting e-commerce decisions. European Journal of Operational Research **259** (3 2017).

[5] Shah, S., Kusiak, A.: Cancer gene search with data-mining and genetic algorithms. Computers in Biology and Medicine **37**(2), 251–261 (2007). , https://www.sciencedirect.com/science/article/pii/S0010482506000217

[6] Shroff, K.P., Maheta, H.H.: A comparative study of various feature selection techniques in high-dimensional data set to improve classification accuracy. In: 2015 International Conference on Computer Communication and Informatics (ICCCI). pp. 1–6 (2015).

[7] Tiwari, R., Singh, M.: Correlation-based attribute selection using genetic algorithm. International Journal of Computer Applications **4** (08 2010).

[8] Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03). pp. 856–863 (2003)

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230667

Navigating Black Swan Events in Algorithmic Trading: A Reinforcement Learning Perspective

Fernando VILLAMARÍN DÍAZ [a] Carlos GUERRERO-MOSQUERA [a]
[a] *HER - Human-Environment Research Group, La Salle - Universitat Ramon Llull, Barcelona, Spain, Quatre Camins, 30, 08022 Barcelona, Spain*
ORCiD ID: Fernando Villamarín Díaz https://orcid.org/0009-0000-1877-5902, Carlos Guerrero-Mosquera https://orcid.org/0000-0001-8265-3651

Abstract. This study evaluates Reinforcement Learning (RL) techniques for financial trading during unpredictable market conditions, such as black swan events. Three experiments were conducted: one where the algorithms were trained and tested over the same period; another where they were trained and tested over different periods; and the final one where they were trained over a certain period and then tested during a period that included a black swan event (the market crash of March 2020). Results show that RL methods outperform traditional strategies in the in-sample period, but struggle to adapt during the black swan event. The results show the potential of RL techniques in financial trading with the right approach.

Keywords. AI applications, reinforcement learning, financial trading, black swan events, stock market, algorithmic trading, portfolio management

1. Introduction

Financial markets, often assumed to be efficient, have experienced significant bubbles and crashes, revealing their vulnerability to rare, high-impact ”black swan” events. Over the past two decades, the financial industry has undergone a technological transformation, with algorithmic trading becoming increasingly prevalent. Simultaneously, advancements in **Machine Learning**, particularly **RL**, have shown promising results across various problem domains.

Despite growing interest in applying RL algorithms to trading, their performance during black swan events remains uncertain. This study aims to investigate the application of RL algorithms to trading by capturing the state of the market through technical indicators and exploring their performance during unexpected market downturns. Additionally, the study compares the performance of **RL-based trading strategies** with traditional man-made trading strategies to evaluate their efficacy and potential advantages.

2. Literature Review

Reinforcement learning (RL) is a learning paradigm where an agent learns to map situations to actions to maximize cumulative reward through trial and error [9] . Despite its increasing prominence, research on RL's applications in finance remains limited.

Few studies have investigated RL in financial settings, such as the application of inverse RL in order book dynamics [6] and the development of granular trading simulators for RL agents [1]. Deng et al. [3] stand out for applying **Deep Reinforcement Learning (DRL)** to financial trading, demonstrating the effectiveness of their proposed system in summarizing market conditions and learning optimal actions.

New approaches include extending the DRL framework to handle **multiple assets** and learn portfolio management strategies, as well as developing a method for intelligently selecting appropriate training periods to address the **non-stationary** nature of financial markets [3].

3. Methodology and Results

This study utilizes **daily stock price data** for JPMorgan Chase & Co. (ticker: JPM) from January 1, 2000, to January 1, 2023. This specific time period was selected due to its availability on Yahoo Finance [12]. The data is transformed into **technical indicators**—price-EMA ratio, %B, and momentum—and then discretized into n bins to capture the market's state. This state information is subsequently fed into the models, which use it to determine their actions: either buy, sell, or do nothing. The market simulator iteratively repeats this process for all the training days in the sample, drawing from historical data.

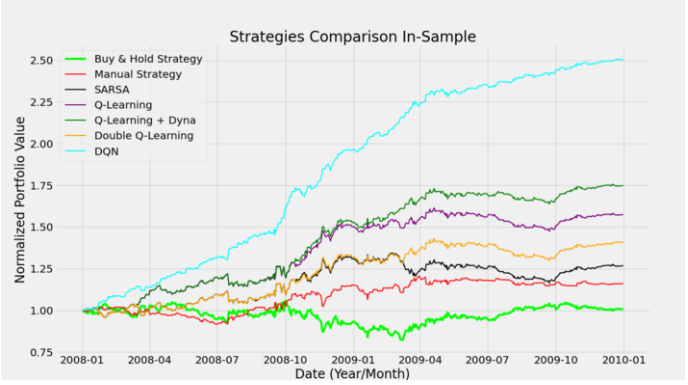


Figure 1. In-sample performance comparison of RL methods, manual, and buy-and-hold strategies. The DQN strategy outperforms all, with a Sharpe ratio of 4.75 and a 150% portfolio value increase. Other RL strategies also outperform the manual (Sharpe ratio: 0.604) and buy-and-hold (Sharpe ratio: 0.112) strategies. SARSA yields the lowest Sharpe ratio (0.810) among RL methods.

The authors have developed a **market simulator** that reads financial data from a CSV file to simulate trading over a specified period. The simulator calculates the daily portfolio value, which is the sum of stock holdings and cash. It also incorporates transaction costs such as a set commission fee per trade and accounts for slippage, representing the adverse price movement a trader might experience in a real market. The simulator restricts the range of stock holdings to between -1000 and 1000 shares. The reinforcement learning model integrates with the market simulator by transforming state elements (price-EMA ratio, %B, momentum, and holdings) into a single value using a **discretization formula**. Values have been previously discretized into n bins, for this article, $n = 10$ was used.

$$\text{State} = \text{Portfolio} \times n^3 + \%B \times n^2 + \text{Momentum} \times n + \frac{\text{Price}}{\text{EMA}} \tag{1}$$

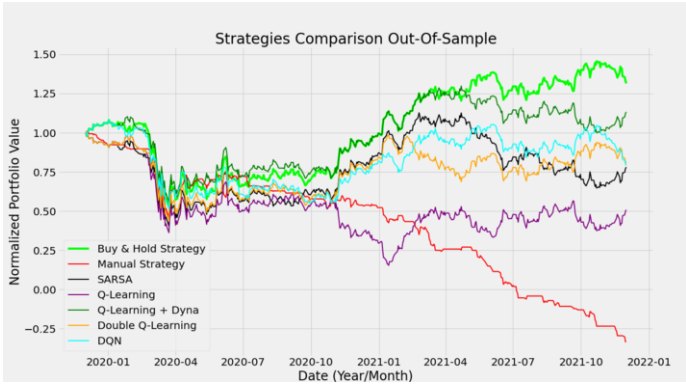


Figure 2. Out-of-sample performance comparison. DQN underperforms with a Sharpe ratio of -1.471, suggesting possible overfitting. Other RL methods continue to outperform manual and buy-and-hold strategies, though the hypothesis that RL models will outperform these strategies in the out-of-sample period is not fully validated.

The models used are **SARSA** [5], **Q-Learning** [11], **Dyna Q-Learning** [8], **Double Q-Learning** [10], and **Deep Q Learning (DQN)** [4].

The simulations use a straightforward **reward function** based on daily returns, allocating a positive reward of +1 when the algorithm generates a profit and -1 when it incurs a loss. This function guides the trading agent's learning process and decision-making within the simulated market environment. To evaluate the performance of the reinforcement learning methods, two benchmark strategies are employed: a **buy-and-hold strategy**, where the trader buys at the beginning of the time period and maintains the position, and a **manual strategy**, which combines technical indicators with simple logic to generate trading signals. These serve as comparison points for assessing algorithmic trading performance.

These experiments were designed to comprehensively evaluate the performance of the algorithms under a variety of conditions in the financial market. The **in-sample experiment** (Experiment 1) aimed to verify the learning capabilities of the RL algorithms in a controlled environment using known data. **Out-of-sample testing** (Experiment 2) was conducted to assess the generalizability of these methods to unseen data and their ability to avoid overfitting. Finally, the **black swan event experiment** (Experiment 3) was crafted to test the resilience and adaptability of these models during extreme market upheavals, such as those characterized by the **COVID-19 pandemic**. The timeframe for this last experiment is more restricted, focusing solely on the market crash of March 2020. To ensure a fair comparison, we optimized the hyperparameters of all models using a grid search approach, focusing solely on the in-sample data. Each experiment was run 100 times to account for variability. The results were then ranked based on the final Sharpe ratio. Graphs were created using data from the models that landed in the median rank for a more robust representation.

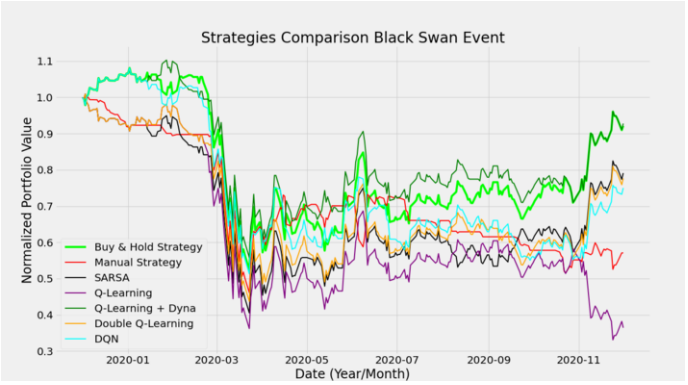


Figure 3. Performance comparison during the black swan event (COVID-19 pandemic). No RL methods outperform manual and buy-and-hold strategies. The buy-and-hold strategy tops with a Sharpe ratio of 0.224. Q-Learning performs the worst among RL methods (Sharpe ratio: -0.580). All strategies yield negative returns, indicating inability to predict or mitigate a black swan event.

4. Discussion and Conclusion

The purpose of this study was to evaluate the **effectiveness of RL techniques** for financial trading and to determine if these methods perform better than traditional strategies during unstable and unpredictable market conditions, such as black swan events. The study focused on experimenting with several value-based RL methods and compared them to a basic "buy-and-hold" strategy and a manual strategy using the same technical indicators.

In the second experiment, RL methods performed significantly better than the benchmark strategies during the in-sample period. However, in the third experiment, the **Deep Q Network (DQN)** method suffered from overfitting, leading to a decrease in performance. The other RL methods, namely the tabular methods, performed much better, indicating a certain level of generalizability.

In the fourth experiment, **none of the RL methods** were able to adapt their behavior or take preventive measures during the black swan event, the market crash in March 2020 due to the COVID-19 pandemic. Despite mixed outcomes, the results demonstrate the potential of these algorithms, showing promise and effectiveness with the right approach.

Despite the promising results of this study, there are several areas of exploration for future research. Extensive model tuning could potentially have enhanced the performance of the RL methods used in this study, but the complexity of financial trading makes this a challenging task. Particularly for the DQN model, future work could focus on implementing techniques to combat overfitting, such as L2 regularization [2] and neural network dropout [7], as proposed in prior literature.

References

[1] David Byrd, Maria Hybinette, and Tucker Hybinette Balch. ABIDES: Towards High-Fidelity Market Simulation for AI Research. 4 2019.

[2] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L 2 Regularization for Learning Kernels. Technical report, 2012.

[3] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep Direct Reinforcement Learning for Financial Signal Representation and Trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 3 2017.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. 12 2013.

[5] Mahesan Niranjan. On-Line Q-Learning Using Connectionist Systems Machine Learning in Fixed Income Markets View project Deep Learning based visual object Tracking View project. Technical report, 1994.

[6] Jacobo Roa-Vicens, Cyrine Chtourou, Angelos Filos, Francisco Rullan, Yarin Gal, and Ricardo Silva. Towards Inverse Reinforcement Learning for Limit Order Book Dynamics. 6 2019.

[7] Nitish Srivastava. Improving Neural Networks with Dropout. Technical report, 2013.

[8] Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 7 1991.

[9] Richard S Sutton and Andrew G Barto. Reinforcement Learning An Introduction second edition. Technical report, 2018.

[10] Hado Van Hasselt. Double Q-learning. Technical report, 2010.

[11] Christopher J C H Watkins and Peter Dayan. Technical Note Q,-Learning. Technical report, 1992.

[12] Yahoo Finance. JPMorgan Chase & Co. (JPM) Stock Price, News, Quote & History, 2023.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230668

Machine Learning in Particle Physics Experiments : The Case of LHCb

Sergi BERNET ANDRÉS [a],Míriam CALVO GÓMEZ [a], Àlvaro GARCÍA PIQUER [a], Elisabet GOLOBARDES RIBÉ [a], Núria VALLS CANUDAS [a]and Xavier VILASÍS-CARDONA [a,1]

[a] *Smart Society Research Group, La Salle-Universitat Ramon Llull, Sant Joan de La Salle 42, 08022 Barcelona, Spain*

ORCiD ID: Sergi Bernet Andrés https://orcid.org/0000-0002-4515-7541, Míriam Calvo Gómez https://orcid.org/0000-0001-5588-1448, Àlvaro García Piquer https://orcid.org/0000-0002-6872-4262, Elisabet Golobardes Ribé https://orcid.org/0000-0001-8080-0769, Núria Valls Canudas https://orcid.org/0000-0001-8748-8448, Xavier Vilasís-Cardona https://orcid.org/0000-0002-1915-9543

Abstract. Particle physics is a source of engineering challenges, also for Machine Learning techniques. We showcase three current uses of Machine Learning in the LHCb experiment, one of the four main experiments of the Large Hadron Collider (LHC) at CERN. Two are in the Real Time Analysis framework, which is in charge of processing the detector 4TB/s dataflow in real time: one to locate the points where particles issued from the accelerator collisions decay, and the other to ensure a smooth choice in the data to be stored. A third use is about speeding the detector simulation with generative techniques. In all three cases, computing speed is the key factor for using Machine Learning algorithms.

Keywords. Particle Detectors, LHCb, Machine Learning.

1. Introduction

High Energy Physics experiments have been for a long time good customers of Machine Learning (ML) applications. As soon as the late 80's and early 90's, Neural Network models where applied to solve several of the challenges met in the intepretation of data, as reference [1] shows. Later on, the dectectors for the Large Hadron Collider (LHC) at CERN continued to selectively use ML techniques at different stages of the data processing. In the case of the LHCb detector, neural networks where currently used for particle identification at several stages of the data processing flow and also Boosted Decision Trees where applied to select the collisions worth keeping for physics analysis [2]. After two very successful data taking periods, the LHC has undergone an upgrade which increases the average number of collisions per time unit, and so, increases the detector data rate. In the case of LHCb this amounts to processing 4TB/s to select what data is worth

[1]Corresponding Author: Xavier Vilasís-Cardona, xavier.vilasis@salle.url.edu

storing for physics analysis. In this note, we overview some examples of machine learning solutions applied in the current version of the detector. For this, we shall first explain through LHCb how a particle detector works and what are requirements in terms of data processing. Next we shall highlight some ML uses to finally elaborate on possibilities and drawbacks of these techniques in this framework. It is not the objective of this note to give an exhaustive review of all possible applications but rather to illustrate some uses and raise the awareness of a potential field of application with a number of challenges.

2. Particle Physics Experiments

Particle Physics Experiments generally use dectectors to observe the decay of particles coming from external sources or, as in the case of the LHC, issuing from the collision of accelerated particles. The physics analysis related to these observations consists in determining what are the products, the features and the rates of some decay modes with respect to others and compare those with theoretical models. Particle dectectors are divided in several subdetectors that measure different features of the particles issued from these decays. Roughly, subdetectors may have three functions; tracking that is determining the trajectories of the particles; identification, determining the type of particle and finally the measure of the energy or the momentum of the particles. Each subdetector has its own type of sensors. The data collected is put together and processed in order to translate the measurements of the sensors into the descriptors needed for physics analysis. One common issue in this type of experiment is the overabundance of data. The amounts of collected data cannot be possibly stored, while some some decays may not be of interest and can be discarded. The system in charge of keeping or discarding an event is called the Trigger.

3. LHCb

The LHCb experiment is one of the four main experiments of the LHC [3]. Its goal is to test the current standard model of particle physics and look for indirect evidence of new phenomena mainly by studying the decays particles containing b and c quarks. Such particles are produced in abundance in the proton-proton collisions provided by the LHC. Because of the kinematics of the decays, the detector just needs to cover a part of the solid angle around the collision point. In exchange, the search for indirect evidences requires precision measurements for which LHCb is designed and built. Figure 1 shows the layout of the detector [4]. Each chamber uses different sensing technology, being the total number of channels of around 1 million. The frequency of collisions provided by the LHC is 30MHz, so that the actual data flow in data taking conditions is around 4TB/s. In order to deal with this data flow, a system call Real Time Analysis has been put in place to select the events that are with relevant information for physics analyses purposes. The scheme of this system is depicted in Figure 2.

4. Machine Learning Opportunities

The operations performed on the detector data both to evaluate the interest of the observed decaying particle, and the posterior physics analysis, are suited to use machine

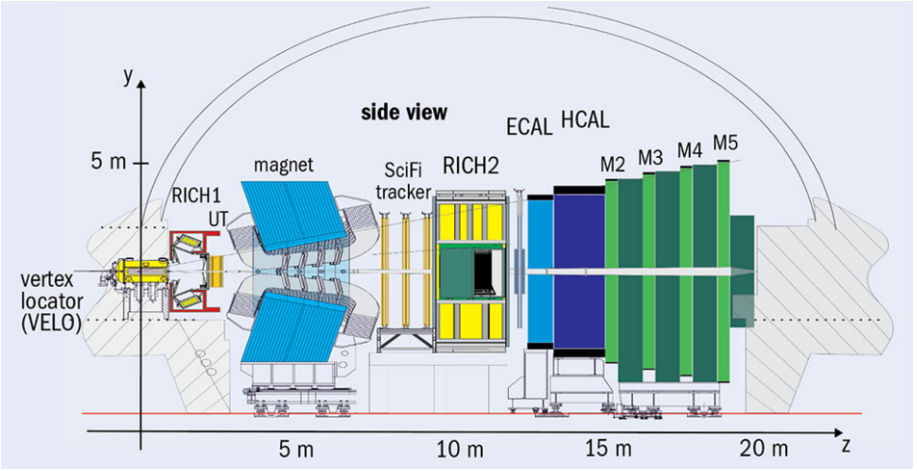


Figure 1. Layout of the LHCb detector. The detector elements, in order are the Vertex Locator, the first Cherenkov Radiation Imaging Chamber, the upstream tracker, the magnet, the scintillator fiber tracker, the second Cherenkov Radiation Imaging Chamber, the electromagnetic calorimeter, the hadronic calorimeter and the muon chambers.

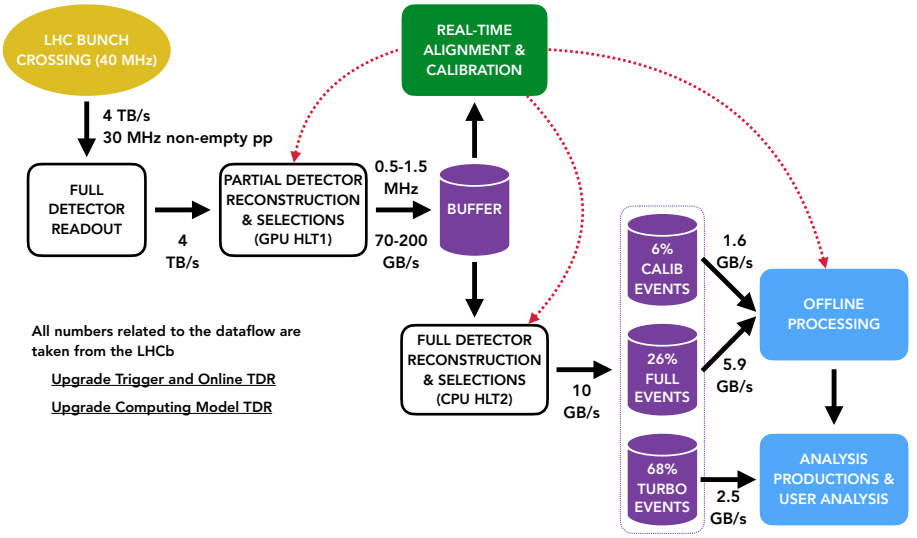


Figure 2. Scheme of the Real Time Analysis System. Data are first analysed in a GPU-based system called Allen, then buffered and finally reconstructed. The events with potentially interesting features are stored while the rest are discarded [5]

learning algorithms. Many require advanced pattern recognition or are based on classifiers. Moreover, another important aspect in the analysis processes are Monte Carlo simulations. A full model of the interaction and the detector response is used to evaluate the analysis procedures and test algorithms. These simulations are very detailed and thus time and resource consuming. We shall single out three interesting approaches among

the current investigations to apply ML techniques in LHCb.

4.1. Lipschitz Neural Newtorks for selection algorithms

The function providing the selection variable that shall determine whether the event under study is relevant for further physics analysis or not should meet two essential requirements: robustness and interpretability. By robustness is understood that it has mitigated sensitivity to experimental instabilites during the data taking and inacuracies in the simulation. Interpretability would be implemented through a built in inductive bias on some selected input variables. These two conditions can be met by designing neural network architectures monotonic with respect to a set of input features and by enforcing Lipschitz constraints. These requirements are embedded into the learning algorithm of a neural network. This strategy is successfuly applied to the so-called LHCb inclusive heavy-flavour trigger as shown in reference [6].

4.2. Deep Learning for identifying and locating primary vertices

One of the most important requirements to achieve LHCb's desired precision is an accurate location of the points where the particles produced in the LHC proton-proton collisions decay, the so-called Primary Vertices (PV). In reference [7], a refinement of the actual method, which is found to be very slow, based on comparing histograms is presented. It consists of using the VELO tracks parameters to feed a system of two consecutive neural networks, a fully connected Multi Layer Perceptron with a UNet that constructs the final predicting histogram. The efficiency in detecting the actual PVs, on simulation data, and the accuracy of the position is found to be below the bin width of the target histogram.

4.3. CaloGAN

One of the most time consuming parts of the simulation software is computing the energy deposit of particles in the calorimeters. This task makes heavy use a very accurate software describing the interaction of particles with matter called GEANT [8]. A strategy to reduce the simulation time would be to be able to obtain equaly accurate simulations on a faster bases. The approach of references [9] and [10] is to use Generative Techniques, both Generative Adversarial Networks and Variational Autoencoders, trained from actual GEANT simulations. These approaches are on the way of being incorporated in the simulation software framework of LHCb called Gaussino [11].

5. Discussion and Overview

The three cases described in the previous section are good examples of challenging problems for Machine Learning experts. Again, let us insist that the list is not exhaustive, but rather illustrative. Another example can be found in reference [12]. Actually, the field has a peculiarity. Unlike many other situations where Machine Learning is used for complex problems that require advanced solutions, here, many applications depend on the ability to speed up calculation of neural network parallel inference. Not only complexity but

also speed is a good reason to use ML. However, one of the limitations found in some applications is that inference sofwares do not deliver the required data throughput. This involves both the programming of the inference algorithms, but also the underlying hardware. A sample discussion on this topic can be found in [13]. The coming challenges to be faced in this field will require of sensible ML experts with deep understanding of training procedures, model optimisation and a sensitivity to find the simplest solution for each problem. In exchange, many interesting lessons may be learned and extrapolated to other fields.

Acknowledgment

This work is supported by MICINN under Grant PID2019-106448GB-C32 and by Generalitat de Catalunya under Grant 2021 SGR 01398.

References

[1] Denby, B. (1993). The use of neural networks in high-energy physics. Neural Computation, 5(4), 505-549.
[2] LHCb Collaboration. (2015). LHCb detector performance. International Journal of Modern Physics A, 30(07), 1530022.
[3] Alves Jr, A. A., et. al. (2008). The LHCb detector at the LHC. Journal of instrumentation, 3(08), S08005.
[4] Bediaga, I., et. al. (2012). Framework TDR for the LHCb upgrade: technical design report (No. CERN-LHCC-2012-007). LHCb-TDR-012.
[5] LHCb Collaboration (2020) LHCb-Figure-2020-016.
[6] Kitouni, O., Nolte, N., Williams, M. (2021, January). Robust and Provably Monotonic Networks. In 35th Conference on Neural Information Processing Systems. arXiv preprint arXiv:2112.00038.
[7] Akar, S., Peters, M., Schreiner, H., Sokoloff, M. D., Tepe, W. (2023). Comparing and improving hybrid deep learning algorithms for identifying and locating primary vertices. arXiv preprint arXiv:2304.02423.
[8] Agostinelli, S., Allison, J., Amako, K. A., Apostolakis, J., Araujo, H., Arce, P., ... Geant4 Collaboration. (2003). GEANT4—a simulation toolkit. Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 506(3), 250-303.
[9] Chekalina, V., Orlova, E., Ratnikov, F., Ulyanov, D., Ustyuzhanin, A., Zakharov, E. (2019). Generative models for fast calorimeter simulation: the LHCb case. In EPJ Web of Conferences (Vol. 214, p. 02034). EDP Sciences.
[10] Sergeev, F., Jain, N., Knunyants, I., Kostenkov, G., Trofimova, E. (2021). Fast simulation of the LHCb electromagnetic calorimeter response using VAEs and GANs. In Journal of Physics: Conference Series (Vol. 1740, No. 1, p. 012028). IOP Publishing.
[11] Mazurek, M., Corti, G., Muller, D. (2021). New simulation software technologies at the LHCb Experiment at CERN. arXiv preprint arXiv:2112.04789.
[12] Valls Canudas, N., Calvo Gómez, M., Golobardes Ribé, E., Vilasis-Cardona, X. (2021). Use of Deep Learning to Improve the Computational Complexity of Reconstruction Algorithms in High Energy Physics. Applied Sciences, 11(23), 11467.
[13] Valls Canudas, N. et. al. Reconstruction of the LHCb Calorimeter using Machine Learning: lessons learned, CCIA-2023.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230669

Petri-Nets-Based Controllers Generation Using Genetic Programming

Carlos A. GARCIA-SANCHEZ [a,1], Manel VELASCO [a], Cecilio ANGULO [b], Pau MARTÍ [a] and Antonio CAMACHO [a]

[a] *SEPIC, Universitat Politècnica de Catalunya, Vilanova i la Geltrú, Spain*
[b] *IDEAI, Universitat Politècnica de Catalunya, Barcelona, Spain*

ORCiD ID: Carlos A. Garcia-Sanchez https://orcid.org/ 0000-0002-5473-9308, Manel Velasco https://orcid.org/ 0000-0002-0764-3063, Cecilio Angulo https://orcid.org/ 0000-0001-9589-8199, Pau Martí https://orcid.org/ 0000-0002-5189-0782, Antonio Camacho https://orcid.org/ 0000-0002-0673-6452

Abstract. Existing research has shown the effectiveness of genetic strategies in generating Petrin-Net (PN)-based controllers, but limitations exist in the ease of controller generation due to the designer's ability and the system's complexity. In the case of automated controller generators based on genetic programming (GP), limitations arise from the static nature of their chromosome over the evolution process. In this short paper we introduce a first discrete PN-based controller designer that can accept systems modeled either continuously or discretely, making it more flexible in handling a wide range of systems. By utilizing genetic algorithms and PNs, the program can generate controllers tailored to the specific requirements of a given system, including the optimal size of the controller. This novel approach has the potential for far-reaching applications in various fields.

Keywords. petri nets, genetic programming, energy microgrids

1. Introduction

Petri Nets (PNs) are a powerful graphical and mathematical modeling tool for complex systems [1]. They consist of two node types: places (representing system states) and transitions (representing events/actions changing states). PNs' graphical representation aids in identifying problems and designing control structures, while the mathematical matrix representation allows formal analysis of properties like reachability, liveness, and deadlock-freeness.

Comparing PNs with alternatives, they excel in handling concurrency and synchronization, making them valuable for modeling parallelism (unlike Automata Theory). Additionally, PNs' formal foundation enables rigorous analysis and verification (unlike Dataflow Diagrams and Object-Oriented Modeling). PNs find extensive use in manufacturing, transportation, communication systems, and software engineering due to their

[1]Corresponding Author: Carlos A. Garcia-Sanchez, FME Building - Office 506. Pau Gargallo 14, 08034 Barcelona, Spain.
This research was funded by Siemens Energy under agreement Siemens-UPC (C-11193).

effectiveness and flexibility. Designing PN-based controllers poses challenges related to system complexity, specifying desired behavior, selecting control parameters, verifying correctness and performance, and considering conflicting criteria.

To address these challenges, advanced evolutionary techniques can be used. Genetic algorithms (GAs) and genetic programming (GP) are both based on the propagation of generations of individuals by selection through the criteria of "survival of the fitness". The individuals that are part of a generation are initially generated in a random way, and each one is evaluated and ranked using a fitness function related with the process goal. An individual in a GA corresponds to a set of values in a parameterized model to be optimized. In contrast, an individual in GP is a tree-based structure. Operations considered in the genetic algorithm include mutation, crossover, and replication. Mutation involves changing an individual's matrix representation by adding or removing rows or columns, or by changing the weight of specific transitions. Crossover involves combining the matrices of two parent individuals to create a child individual.

The use of evolutionary algorithms in PN controller design, as an automated design tool, has the potential to improve control performance, system efficiency, and reduce the time and cost associated with manual controller design. They can handle both discrete and continuous variables, making them suitable for PNs, which often involve both types of variables. Moreover, they can handle noisy or incomplete data, making them robust to uncertainties and variations in the system's environment. Finally, using evolutionary algorithms for designing PN-based controllers can reduce the time and effort required by the designer, as they provide a systematic and automated approach to designing control policies. In this short paper, we aim to explore their use for the automatic design of PN-based controllers for discrete systems. Specifically, this work focuses on the generation of a general PN-based controller designer which uses as input generic plants regardless of whether they are modeled continuously or discretely.

2. State of the Art

Petri-Nets are a powerful tool for modeling and analyzing discrete control structures in a wide range of systems. PNs in supervisory control of discrete-event systems [2] allows to model the behavior of a system and its supervisor. The supervisor ensures that the system always behaves in a safe and correct manner by controlling its transitions based on the current state of the system. In the reverse sense, structural analysis techniques, such as state-space analysis and reachability analysis, can be used to identify control structures in PNs [3]. PNs are also used to model and analyze workflows, such as business processes and software systems [4]. In these applications, PNs allow for the identification of bottlenecks and inefficiencies. A similar modeling and control alternative are Hybrid PNs [5]. They combine discrete and continuous PNs to model and control hybrid systems, such as manufacturing systems and chemical processes,in a single framework. An additional tool are Timed PNs [6], which extend traditional PNs with timing information, allowing for the modeling and analysis of timed systems, such as real-time embedded systems and communication protocols.

Thanks to the ease in which PNs can be mathematically represented using matrices, the automatic design and generation of PNs structures is highly compatible with evolutionary techniques, such as genetic algorithms (GAs) or genetic programming (GPs).

One of the earliest works in this area [7] proposes a method for designing PN controllers using a GA. A binary representation encodes the PN structure and firing sequences, then candidate solutions are evaluated by simulating the PN to measure its performanceuntil an optimal controller is found [8]. Later, several other works have built upon this approach and explored different variations of the GA algorithm. Some works have used multi-objective GAs to optimize multiple conflicting objectives, such as minimizing resource usage and maximizing system throughput [9]. Other works have used hybrid optimization algorithms that combine GAs with other optimization techniques, such as simulated annealing or tabu search. Hence, in [10] it is proposed a method for designing a PN controller for a manufacturing system using a hybrid optimization algorithm that combines a GA with a particle swarm optimization (PSO) algorithm. This method is able to find a near-optimal solution that minimizes the makespan of the system, while ensuring that all production orders are completed within their due dates.

3. Proposal

The literature demonstrates that using GAs to design PN controllers is a promising approach that has been shown to be effective in a wide range of applications. However, it is important to carefully choose the encoding, fitness function, and other GA parameters to ensure that the optimization problem is well-defined and that the algorithm can converge to an optimal solution. The main objective of this research proposal is to develop a discrete PN-based controller designer based in genetic algorithms. The designer tool is intended to be capable of accepting systems modeled either continuously or discretely, allowing for greater flexibility in the types of systems it can handle. The tool's input is a model of the system to be controlled, and its output is a PN-based controller that can regulate the system.

The development of this controller designer tool involves several challenges. One of the most significant is determining how to generate a generic link between the system model and the controller to allow not only simulation, but physical implementation of the controller. Another challenge to consider is the definition of a useful fitness function to evaluate the quality of individual controllers. This fitness function must be designed to reward controllers that effectively regulate the system and penalize those that fail in this task. Hence, it must be defined enclosing both, the desired behaviour for the controlled process and the control action. A control structure to be considered in the optimization porocess should meet the following five requirements:

R1. Small-sized expression for the controller;
R2. Control action is as smooth as possible;
R3. Control action is robust and can withstand variations in the system's dynamics;
R4. Process variable does not surpass a specified overshoot;
R5. Steady state error is minimal after a specified time.

The individuals (controllers) generated by the designer tool are encoded in a matrix form based on their input and output matrices. Specifically, the rows of the input matrix correspond to the places of the PN, while the columns correspond to the transitions. Similarly, the output matrix defines the PN's outputs, with each row corresponding to a place and each column corresponding to an output signal. The fitness function

is based on simulating the system and controller together to evaluate their performance. The simulation is run over a defined time period, and the fitness function evaluates the controller's performance based on the proportion of the simulation time that the system spends in designated desired system behavior.

Operations considered in the genetic algorithm include mutation, crossover, and replication. However, crossover represents a challenge in this tool because the size of the matrix representation can vary during the evolution process. As a result, a process must be defined to determine the size of the child individual and how to combine the parents' genes.

4. Conclusions

The tool's primary use case is in the design of controllers for complex systems. By utilizing genetic algorithms and PNs, the designer tool can generate controllers that are tailored to the specific requirements of a given system. Additionally, the program's flexibility in accepting continuous and discrete system models allows for greater versatility in the types of systems it can handle. This program's development is an exciting opportunity to use genetic algorithms and PNs to solve challenging control problems. While the program's development presents several challenges, including the need to create a generic link between the system model and the controller and designing an effective fitness function, the potential benefits of the program make these challenges worthwhile. The ability to generate custom controllers for complex systems can have far-reaching applications in a variety of fields, from manufacturing to healthcare to transportation.

References

[1] Murata, T. (1989). PNs: Properties, analysis and applications. Proceedings of the IEEE, 77(4), 541-580. https://doi.org/10.1109/5.24143

[2] Zhou, M. C., & Wonham, W. M. (1993). Supervisory control of discrete event systems using PNs. IEEE Transactions on Systems, Man, and Cybernetics, 23(2), 438-449. https://doi.org/10.1109/21.221084

[3] Jensen, K. (1997). Coloured PNs: Basic concepts, analysis methods and practical use. Vol. 1. Springer. https://doi.org/10.1007/978-3-642-60894-4

[4] van der Aalst, W. M. P., & van Hee, K. M. (2002). Workflow management: models, methods, and systems. MIT Press. https://doi.org/10.7551/mitpress/3434.001.0001

[5] David, R., Alla, H., & Salhi, H. (2007). Hybrid PNs for modeling and control of hybrid systems: A survey. Applied Mathematics and Computation, 190(1), 155-177. https://doi.org/10.1016/j.amc.2007.01.056

[6] Desel, J., & Reisig, W. (1998). A class of timed PNs for modeling and analyzing timed systems. Lecture Notes in Computer Science, 1420, 200-219. https://doi.org/10.1007/3-540-64574-8_21

[7] Lee, C. Y., & Lin, Y. J. (1995). Genetic algorithm-based approach to the design of PN controllers. IEEE Transactions on Systems, Man, and Cybernetics, 25(12), 1585-1597. https://doi.org/10.1109/21.477245

[8] He, J., & Yang, X. (2009). An optimization algorithm for PN controller design based on particle swarm optimization and simulated annealing. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 39(3), 666-677. https://doi.org/10.1109/TSMCB.2008.2011167

[9] Le, T. M., Yoon, W. Y., & Kim, Y. H. (2007). Multi-objective genetic algorithm for PN-based control system design. IEEE Transactions on Industrial Electronics, 54(3), 1653-1663. https://doi.org/10.1109/TIE.2007.892244

[10] Jiang, P., & Lu, Y. (2014). A hybrid optimization algorithm for PN controller design. IEEE Transactions on Industrial Informatics, 10(1), 563-571. https://doi.org/10.1109/TII.2013.2273711

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230670

Addressing Scale and Density Challenges in LiDAR Point Classification

Mariona CARÓS [a], Ariadna JUST [b] Santi SEGUÍ [a] and Jordi VITRIÀ [a]
[a] *Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB), Gran Via Corts Catalanes, 585, 08007 Barcelona, Spain*
[b] *Institut Cartogràfic i Geològic de Catalunya, Barcelona, Spain*

Abstract. Semantic segmentation of LiDAR point clouds has received significant attention due to its applications in autonomous driving, forestry, and urban planning. Despite their potential, accurately classifying three-dimensional points remains a significant challenge due to the irregular distribution of data and density variation. To address this, state-of-the-art approaches use various techniques, such as voxelization, point-based networks, and graph-based methods. However, these techniques have limitations regarding the point cloud size they can handle and can be computationally expensive. Therefore, in this work, we propose a method to process point clouds of different scales and densities for point classification.

Keywords. LiDAR, Point Clouds, Deep Learning, Segmentation, Remote Sensing

1. Introduction

Airborne LiDAR (Light Detection and Ranging) is a remote sensing technology that is used to generate highly precise three-dimensional representations of the Earth's surface by using near-infrared light. The number of points in a scene of 100×100 meters acquired by a LiDAR sensor is usually in the order of 10^6. A common practice in remote sensing is to apply ground-point filtering [1] as part of the processing pipeline. By eliminating ground points, a significant reduction in the number of points is achieved, producing point clouds of different scales. The size of the resulting point cloud is correlated with the quantity of vegetation and structures within the observed scene.

The objective of our study is to perform semantic segmentation on point clouds of arbitrary scale, enabling the accurate segmentation of varying size and density objects, while maintaining minimal computational demand. Current methods face a limitation in terms of the point cloud size they can handle, either being restricted to small-scale (10^3 points) or large-scale point clouds (10^6 points) [2]. They are unable to effectively address both small and large-scale point clouds in a unified manner, which can limit their practical utility and efficiency. Therefore, it is desirable to further investigate the problem of efficient segmentation of varying-size point clouds. Existing approaches such as PointNet[3], PointNet++[4], and R-CNN [5] address large-scale point cloud segmentation by partitioning the block into sub-blocks of a small number of points (< 4k) before being fed into the network. However, such an operation of partition degrades the spatial geometric contextual information among large objects within the scene. Recently, several

works have tackled the challenge of semantic segmentation of point clouds of millions of points. In SPG [6] the authors propose a representation of points as a collection of interconnected shapes using graphs. However, its approach of processing point clouds as super graphs prior to applying neural networks for learning local semantics results in a computationally intensive method. RandLA-Net [2] proposes a local feature aggregation module to capture geometric features. Although efficiency is improved by applying random sampling and shared MLPs, it is designed to work with a single scale, as points of small point clouds are replicated generating substantial noise in the resultant point cloud. Voxel-based techniques [7] provide effective network designs for large-scale 3D scene comprehension, nevertheless, they may compromise on detailed geometric patterns due to quantization artifacts. Zhao et al. [8] realized the success of transformers in natural language processing and designed a self-attention layer for point clouds to construct a Transformer network. While Point Transformer is a powerful tool for processing point clouds, its efficacy may be constrained when dealing with larger point clouds.

In order to learn the irregular distribution of points, the optimal strategy would be using all points present in the point cloud. However, the use of deep learning requires the parallelization of samples during training. For this reason, it is necessary to ensure that all training samples have the same size. To achieve this, a common approach involves downsampling larger point clouds and upsampling or padding smaller ones. Several sampling strategies, such as random sampling, Farthest Point Sampling (FPS) [4], and Constrained Sampling [9], are employed based on how points are selected. Nevertheless, in cases where there exists a significant disparity between point cloud sizes, undersampling strategies result in substantial point loss. To tackle this challenge, we propose an adaptive architecture that consists of encoding subsets of a point cloud scene and modeling its relation by using a recurrent neural network. Our proposed method not only accommodates diverse point cloud scales but also exhibits efficient computational performance due to its weight-sharing architecture.

2. Method

The sampling technique significantly impacts the model's performance as it alters the training data. In order to maintain the training performance during inference, it is crucial that any sampling strategy employed during training preserves the original point distribu-

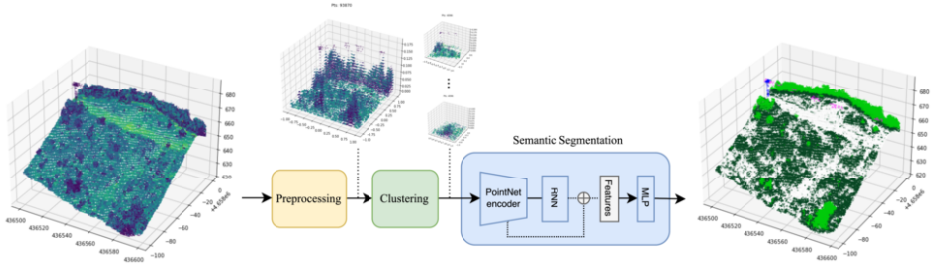


Figure 1. Overview of the proposed methodology. In the first stage, a large point cloud (100×100 m) is flattened and its ground surface is removed. Then, random sampling is applied, and the point cloud is partitioned into equal-size subsets. Finally, subsets of points are fed into the recurrent network, enabling semantic segmentation of the entire scene.

tion and avoid introducing bias. This requirement stems from the fact that the inference data is not subject to sampling, as every point must receive a predicted label. We propose a method that minimizes the sampling impact and handles raw point clouds with different sizes effectively. An overview is shown in Figure 1 and consists of: (I) Preprocessing, (II) Clustering, and (III) Recurrent Semantic Segmentation.

Preprocessing. To facilitate the segmentation of point clouds, the first step involves normalizing the orthonormal heights by subtracting the height of the ground surface[1]. Subsequently, points located at ground level (z=0) and those exceeding a height of 100 meters are filtered out, as we are not interested in their classification. This leads to large variations in the number of points across different point cloud scenes. Then, random sampling is applied to reduce the size of large point clouds down to a maximum size S_{max} so that they can be split into equal-size clusters. S_{max} is defined considering the number of clusters to be used and the number of points per cluster ($S_{max} = C \times S_{cluster}$). We experiment with different C sizes and set $S_{cluster}$ to 2048 points, which is the default number of points defined by PointNet authors.

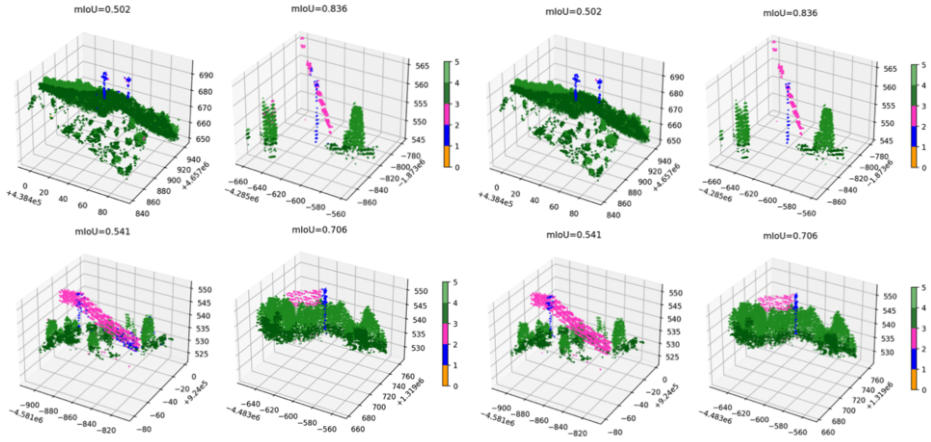
Clustering. We partition large point clouds into subgroups of points to handle point clouds with different sizes effectively. This approach facilitates the extraction of features from small objects and reduces computational complexity. Specifically, we convert each point cloud into a variable sequence of clustered points by using constrained K-means [10]. To ensure that the partitioning does not damage the geometry of the objects, we employ x, y, and *normalized vegetation index* [11] features. A maximum number of clusters is established and point clouds with fewer clusters are padded and masked during training. Finally, clusters are arranged by sorting them along the (x, y) plane.

Recurrent Semantic Segmentation. We implement a lighter version of PointNet to obtain a representation for each cluster. Then, we use a recurrent neural network on the cluster features obtained from the PointNet encoder to model the relationship between them. Specifically, we use a bidirectional GRU [12] with a hidden dimension set to 256. Then, we concatenate the recurrent representation with the local features obtained by PointNet and we feed them into a multi-layer perceptron to predict per-point category scores. This mechanism permits using information from all parts of the scene for predicting each point label.

3. Results and Discussion

The proposed method is evaluated on point clouds of different sizes, ranging from 2k to 32k points, producing promising results shown in Figure 2 and Table 1. While the method may face particular challenges associated with the generation of false positives from hallucinated points, it exhibits significant potential in effectively segmenting objects such as power lines of diverse sizes from LiDAR data. Moreover, the network's relatively small size enhances its computational efficiency for LiDAR data processing. Future work will involve evaluating the methodology with alternative encoding networks and expanding the scope of comparison to include additional state-of-the-art methods.

[1] provided by Institut Cartogràfic i Geològic de Catalunya



(a) Point cloud predictions ranging from 2k to 32k pts. **(b)** Ground truth

Figure 2. Semantic segmentation predictions where each color defines a semantic category.

Table 1. Quantitative comparison on our experimental dataset. The mean Intersection Over Union (mIoU) and IoU per class (%) are used for evaluation. Our method with 6 clusters achieves the highest mIoU, significantly improving the tower and power lines categories.

Methods	param.	mIoU	high veg.	low veg.	other infr.	tower	power lines
PointNet	3.5M	54.7	**83.9**	79.9	**36.4**	32.6	40.9
Ours (clusters=9)	2.7M	57.2	77.6	86.7	29.3	57.0	35.3
Ours (clusters=6)	2.7M	**60.0**	83.4	**86.9**	29.5	**64.4**	**41.1**

References

[1] Mandlburger, et al. Improved Topographic Models Via Concurrent Airborne LiDAR and Dense Image Matching. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences. 2017;4.
[2] Hu Q, et al. Randla-net: Efficient semantic segmentation of large-scale point clouds. In: CVPR; 2020. .
[3] Qi C, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE CVPR; 2017. p. 652-60.
[4] Qi C, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems. 2017;30.
[5] Chen Y, et al. Fast point r-cnn. In: Proceedings of the IEEE/CVF ICCV; 2019. .
[6] Landrieu L, et al. Large-scale point cloud semantic segmentation with superpoint graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4558-67.
[7] Zhao L, et al. SVASeg: Sparse Voxel-Based Attention for 3D LiDAR Point Cloud Semantic Segmentation. Remote Sensing. 2022;14.
[8] Zhao H, et al. Point transformer. In: Proceedings of the IEEE/CVF ICCV; 2021. .
[9] Caros M, et al. Object Segmentation of Cluttered Airborne LiDAR Point Clouds. In: Artificial Intelligence Research and Development. vol. 356. IOS Press; 2022. .
[10] Bradley P, et al. Constrained k-means clustering. Microsoft Research, Redmond. 2000;20.
[11] Pettorelli N. The normalized difference vegetation index. Oxford University Press; 2013.
[12] Chung J, et al.. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling; 2014.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230671

A Bayesian Network Framework to Study Class Noise: Exploring the Filtering of Completely Random Noise

David MARTÍNEZ-GALICIA [a,1], Alejandro GUERRA-HERNÁNDEZ [a], Xavier LIMÓN [b], Nicandro CRUZ-RAMÍREZ [a] and Francisco GRIMALDO [c]

[a] *Inst. de Invest. en Inteligencia Artificial, Universidad Veracruzana, Xalapa, México*
[b] *Facultad de Estadística e Informática, Universidad Veracruzana, Xalapa, México*
[c] *Departament d'Informàtica, Universitat de València, València, España*

Abstract. Although the negative consequences of noise during induction have been widely studied, previous work often lacks the use of validated data to measure its impact. We propose a framework based on Bayesian Networks for modeling class noise and generating synthetic data sets where the kind and amount of class noise are under control. The benefits of the proposed approach are illustrated evaluating the filtering of noise completely at random in class labels when inducing decision trees. Unexpectedly, this kind of noise showed a low effect on accuracy and a low occurrence on real datasets. The framework and the methodology developed here seem promising to study other kinds of noise in class labels.

Keywords. Noise modeling, Bayesian Networks, Data generation, Noise filtering

1. Introduction

In supervised learning, noise refers to anything that obscures the relationship between attributes and class labels [1]. However, previous studies often use not validated data to measure its impact. A common approach while studying class corruption is to assume that datasets from repositories are clean, which may produced biased results [2]. We propose a Bayesian Network (BN) framework to model domains affected by class noise and provide a controlled experimental setting. We illustrate its benefits with a case study evaluating the effect of filtering completely random noise when inducing decision trees.

2. Modeling Class Noise

Frénay and Verleysen [3] pioneered the probabilistic approach to model class noise. Their modeling represents complex relationships in noisy scenarios using a domain distribution and a noise mechanism. However, it requires a complete description, which can cause problems when the parameters representing the full joint distribution of the

[1] Corresponding Author: Martínez-Galicia David, davidgalicia@outlook.es

domain grow exponentially. We propose a framework that extends their approach using BNs to reduce the number of parameters and generate experimental data with sampling algorithms. Three models arise considering certain statistical relationships, as shown in Figure 1.

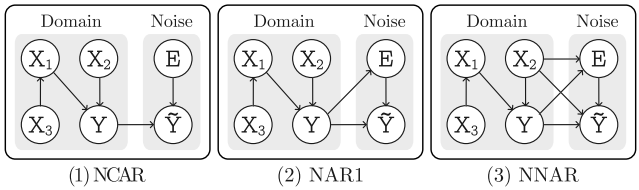


Figure 1. Models of class noise where X_i denotes the attributes, Y denotes the true class, E is a binary variable symbolizing the probability of a labeling error, and $\tilde{Y}$ denotes the observed class. *(1) Noisy Completely at Random:* errors are independent of other domain variables. *(2) Noisy at Random:* errors depends on the true class. *(3) Noisy Not at Random:* errors depends on a set of attributes and the true class.

Building the domain distribution is not trivial and requires prior knowledge. We recommend using reliable sources such as curated data or a domain expert to avoid spurious noise. Other suggestions for generating reliable data are: 1) defining the noise level to be less than 50% and, 2) in case the model was learned from curated data, restricting the generation of synthetic examples to those appearing in the original dataset.

3. Case Study

We illustrate the benefits of adopting our approach with a case study that evaluates the efficiency of a new filter for class noise completely random. Class noise poses a challenge to estimate the reliability of class labels and it has the most detrimental effect on learning [4]. The Inconsistency Deletion Filter (IDF) identifies sets of examples with the same attribute values but different class and keeps only the ones belonging to the majority class, which are arguably the most likely to be clean. The case study has two parts:

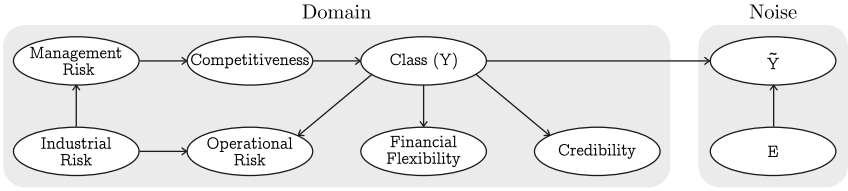


Figure 2. NCAR model for the QB domain.

1. Evaluating IDF on synthetic data. We define an NCAR model using an apparently clean domain, the Qualitative Bankruptcy (QB) dataset [5], see Figure 2. The model is fed with noise levels from 10% to 40% with increments of 10%. For each level, four datasets are generated with the following number of examples: 500, 2000, 5000, and 10000. A 10-fold cross-validation process is performed to obtain pairs of train and test sets. To evaluate the efficiency of IDF, three metrics are considered: the ratio of clean examples removed (ER1), the ratio of noisy examples not removed (ER2), and the per-

centage of conserved examples (CE) after filtering. ER1 and ER2 are known as errors in data cleansing [6]. To measure the impact on learning, decision trees are induced using J48 (with default parameters) [7] and the area under the ROC curve (AUC) is measured over clean and noisy test sets. Clean test sets are obtained by $P(E = True) = 0$, while noisy test sets are produced by a cross-validation process over noisy data.

2. Evaluating IDF on UCI data. We adopt seven UCI datasets, see Table 1. For continuous data, five methods from the discretization R package [8] were adopted: AMECA, CACC, CAIM, CHI2, and MDLP. Using various discretizers avoids their individual limitations for certain types of data or models and allows the description of their average performance. The methodology previously introduced is modified since the noisy examples in UCI datasets are unknown: 1) we use one type of test set obtained by a 10-folds cross-validation, which might be noisy; 2) ER1 and ER2 are not measured.

Table 1. Characteristics of the adopted datasets.

Dataset	Solar Flare	Nursery	Balance Scale	Breast Cancer	Diabetes	Ecoli	Semgent
Type	Disc.	Disc.	Disc.	Cont.	Cont.	Cont.	Cont.
Attributes	13	9	4	9	9	8	20
Class labels	5	5	3	2	2	8	7
Examples	323	12960	625	683	768	336	2310

4. Results and Discussion

Table 2 summarizes the efficiency of IDF on the synthetic datasets generated by our framework. Values of ER1 and ER2 closer to zero denote the best outcome, most of the clean data is conserved while noise is removed. Our filter seems to be quite effective in synthetic datasets with equal or less than 20% of noise. Increasing the noise level reduces its efficiency. However, this effect is mitigated in larger datasets. The remaining noisy examples in the filtered training set suggest that inconsistencies are a subset of noisy examples, i.e., there are noisy examples that do not produce inconsistent sets. Regarding the percentage of remaining examples, CE suggest that the proportion of deleted examples is closer to the noise level of a dataset, which is consistent with the low values of errors in data cleansing.

Table 2. Mean values of ER1, ER2 and CE. The size of the training sets represent 90% of the generated data.

Noise	10%			20%			30%			40%		
Size	ER1	ER2	CE	ER1	ER2	CE	ER1	ER2	CE	ER1	ER2	CE
450	1.6%	1.3%	89.9%	2.5%	2.2%	79.1%	7.0%	6.2%	66.6%	12.7%	12.3%	59.8%
1800	0.1%	0.0%	89.4%	0.4%	0.1%	79.3%	1.3%	0.9%	69.6%	6.8%	7.6%	60.2%
4500	0.0%	0.0%	89.6%	0.1%	0.0%	79.9%	1.1%	1.0%	70.1%	6.7%	5.6%	58.0%
9000	0.0%	0.0%	89.8%	0.0%	0.0%	79.9%	0.1%	0.0%	69.6%	1.6%	1.5%	59.4%

Accuracy results are not presented because of space limitations. However, in most cases, IDF preserves the levels of AUC. Experiments were performed on clean and noisy test sets. When facing clean test sets the induced decision trees seem robust to noise (AUCs between 1.0 and 0.97), perhaps because of the low complexity of the QB domain. Noisy tests sets exhibits a linear degradation, e.g., AUCs near 0.9 for 10% noise, decreasing to 0.6 in tests with 40% noise.

Table 3 shows the results over UCI datasets. Even though the kind and amount of noise affecting these datasets are unknown, it is possible to make assumptions about their noise type. Balance Scale and Nursery do not seem to include NCAR given the lack of data reductions. However, Balance Scale could probably have other types of noise that affect the performance of induced decision trees. Poor reductions and high AUCs on Breast Cancer and Segment suggest a low percentage of noise, possibly NCAR. Finally, Solar Flare, Diabetes, and Ecoli seem to include NCAR since our filter reduces up to 15% of data. Indeed, these observations imply NCAR noise is not frequent in UCI data.

Table 3. Mean values of conserved examples and AUC on UCI datasets. For continuous sets, results are calculated as an average of all discretized versions. J48 stands for the case where no filter was applied.

	CE	AUC			CE	AUC	
Data	IDF	J48	IDF	Data	IDF	J48	IDF
Balance Scale	100.0%	68.5	68.5	Breast Cancer	98.8%	96.5	96.4
Solar Flare	85.9%	71.9	71.5	Diabetes	84.4%	75.0	70.8
Nursery	100.0%	99.5	99.5	Ecoli	93.3%	79.4	78.5
				Segment	99.8%	93.1	93.1

5. Conclusions and Future Work

Our BN framework allows the compressed description of noisy domains and provides tools to build a controlled experimental setting. These benefits help to assess the efficiency of thetechniques proposed to cope with noise since the kind and amount of noisy examples are known. Our case study shows the significant filtering achieved by IDF, although it does not improve the accuracy of decision trees. These observations are relevant because they contradict some results from the literature. First, according to our experiments, class noise seems not as harmful as suggested when assessed on clean test sets. Second, filtering synthetic data does not lead to better predictive models, perhaps because of the low complexity of the QB domain. The last observation, derived from our experiments on real datasets, is that completely random class noise does not seem to be usual in the UCI repository. These observations would not be possible without using our framework. Future work includes investigating other types of class noise (NAR1, NAR2, and NNAR) and exploring the application of IDF to inductive algorithms that do not use noise-handling methods, e.g., naive Bayes.

References

[1] Hickey, R. Noise Modelling and Evaluating Learning from Examples. *Artif. Intell.*. **82** (1996).
[2] Zhu, X. & Wu, X. Class Noise vs. Attribute Noise: A Quantitative Study. *Artif. Intell. Rev.*. **22** (2004).
[3] Frénay, B. & Verleysen, M. Classification in the Presence of Label Noise: A Survey. *IEEE Trans. Neural Networks Learn. Syst.* **25** (2014).
[4] Quinlan, J. Learning from noisy data. *Proc. 1983 Int. Machine Learning Workshop*. (1983).
[5] Kim, M. & Han, I. The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Syst. Appl.*. **25** (2003).
[6] Brodley, C. & Friedl, M. Identifying Mislabeled Training Data. *CoRR*. (2011)
[7] Witten, I., Frank, E. & Hall, M. Data mining: practical machine learning tools and techniques, 3rd Ed. *Morgan Kaufmann, Elsevier* (2011).
[8] Kim, H. discretization: Data Preprocessing, Discretization for Classification. (2022), R package v. 1.1.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230672

EduSign: Real-Time Application for Spanish Sign Language Recognition

Víctor VALLÉS [a], Ester VIDAÑA-VILA [a,1]
[a] *Human-Environment Research, La Salle Campus Barcelona, Universitat Ramon Llull*
ORCiD ID: Víctor Vallés https://orcid.org/0009-0005-3795-4654, Ester Vidaña-Vila https://orcid.org/0000-0002-4265-2985

Abstract. According to a survey by the National Institute of Statistics in Spain, around 2.2% of the hearing impaired population, equivalent to 27 300 people, use Sign Language to communicate. This highlights the importance of ensuring accessibility to Sign Language as a fundamental resource to guarantee inclusion and equal opportunities for deaf and hearing impaired people. To achieve this goal, it is essential to promote and disseminate Sign Language and to remove barriers to its usage and access. Technology can be of great help in this field. Therefore, this work presents EduSign, a new AI-based application that classifies the alphabet of the Spanish Sign Language in real time aiming at facilitating the communication of people with hearing disabilities. Specifically, EduSign is a web application that runs a machine learning algorithm (i.e., a neural network) where users can learn and practice the Spanish Sign Language alphabet. The methodology that enabled the creation of EduSign consisted on capturing the coordinates of the gestures made in Sign Language using the MediaPipe tool. These coordinates were then used to train several machine learning classifiers (i.e., K-Nearest Neighbors, Ridge Regression, Gradient Boosting, Random Forest, Logistic Regression and a Neural Network). Out of these classifiers, the neural network offered the best performance. More specifically, using a custom dataset with the complete Spanish alphabet, it shown an accuracy of 99%. These results reinforce the idea that the application of artificial intelligence techniques can be of great help in improving communication for deaf and hearing impaired people. Hence, it is hoped that the system can be used effectively by the target audience.

Keywords. Sign language recognition, machine learning, AI real-time application, neural networks, Spanish Sign Language.

1. Introduction

Artificial intelligence (AI) has revolutionized communication and learning, making it more accessible for people with hearing impairment. EduSign is an AI-powered web application designed to promote inclusion and equal opportunities for individuals with hearing disabilities and those interested in learning sign language. EduSign allows real-time learning and practice of Spanish Sign Language (SSL) through its web interface, which can be accessed via this GitHub link: https://github.com/victorr97/

[1]Corresponding Author: Ester Vidaña-Vila, ester.vidana@salle.url.edu

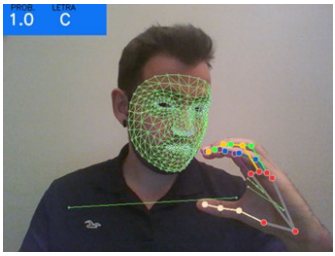


Figure 1. MediaPipe coordinate detection and letter classification.

`DetectionSignLanguageRealTime`. Its unique features, such as accurate sign classification and real-time practice, distinguish EduSign from other applications [1,2].

The EduSign web application consists of three main sections: data contribution, practice signs, and user evaluation. In the data contribution section, users add sign language examples to the dataset by selecting a letter and performing the corresponding sign. Correctly recognized signs are stored for future use. The practice signs section allows users to practice sign language by selecting a letter and receiving feedback on their performance. The user evaluation section is a game-like challenge where users aim to perform all alphabet letters in the shortest time possible. The MediaPipe library, like in other research projects [3], is used to detect coordinates of the face, body, and hands. In this project, coordinates from the right hand are stored with their corresponding sign in a CSV file. Figure 1 shows an example of the MediaPipe platform.

For the training of the automatic classification, the dataset used is a CSV file with sign coordinates labeled with corresponding signs. After splitting into train and test sets, multiple classifiers were trained, with the neural network (multilayer perceptron) achieving the best performance after hyperparameter optimization. Evaluation metrics like Accuracy, Confusion Matrix, Precision, Recall, and ROC curve were used, resulting in 99% accuracy for the neural network.

2. Conclusions

This work introduces EduSign, a real-time platform for learning SSL. It demonstrates that neural networks and MediaPipe coordinates achieve high performance in SSL classification. Future directions include incorporating movement gestures in sign language and the addition of other sign language alphabets such as the Catalan Sign Language.

References

[1] Duarte A, Palaskar S, Ventura L, Ghadiyaram D, DeHaan K, Metze F, ..., Giro-i-Nieto X. How2sign: a large-scale multimodal dataset for continuous american sign language. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021 (pp. 2735-2744).

[2] Forster J, Schmidt C, Koller O, Bellgardt M, Ney H. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In LREC; 2014 May (pp. 1911-1916).

[3] Harris M, Agoes AS. Applying hand gesture recognition for user guide application using MediaPipe. In 2nd International Seminar of Science and Applied Technology (ISSAT 2021); 2021 Nov. (pp. 101-108). Atlantis Press.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230673

Automated Cleanliness Scoring and Digestive Content Segmentation for Capsule Endoscopy

Pere GILABERT [a,1,2] Carolina MALAGELADA [b] Hagen WENZEK [c] Jordi VITRIÀ [a] Santi SEGUÍ [a]

[a] *Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona*
[b] *Digestive System Research Unit, University Hospital Vall d'Hebrón, Barcelona*
[c] *CorporateHealth International*

Abstract. Accurate evaluation of capsule endoscopy videos plays a crucial role in diagnosing gastrointestinal disorders. One important aspect is the assessment of the cleansing quality, which indicates the visibility of the gastrointestinal mucosa during the examination. In this study, we propose a novel method for assigning a cleansing score to capsule endoscopy videos. Moreover, our system proposes a segmentation masks for each frame, highlighting the regions of the image that are not visible. This additional information aids in the interpretation and analysis of the videos, enabling more accurate diagnoses and improved patient care.

Keywords. capsule endoscopy, cleansing, segmentation, digestive content, video quality

1. Introduction

Performing a capsule endoscopy involves several important steps to ensure a successful examination. The process begins with bowel preparation, which is crucial for optimal visualization of the gastrointestinal tract. Bowel preparation typically involves a restricted diet for a day or two before the procedure, which may include avoiding solid foods and consuming only clear liquids. Additionally, patients may be required to take laxatives or other medications to cleanse the bowels and remove any residual debris or stool [1]. In this work, we present a threefold approach to enhance the quality of annotation. Firstly, we aim to simplify the annotation process by converting the problem into a binary decision for annotators, focusing on specific regions within an image. Secondly, we propose an automated method for generating segmentation masks, which effectively highlight regions in an image that may be occluded by digestive content. Lastly, we introduce a cleanliness score that incorporates various factors, providing a comprehensive assessment of video cleanliness. This score considers multiple aspects and serves as a concise summary of the overall cleanliness level observed in the video. By implementing these

[1] pere.gilabert@ub.edu
[2] Funding declared: 2021 SGR 01104 grant, Innovate UK project 104633 and FPU20/01090 grant.

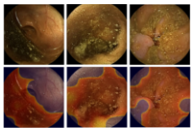
three strategies, we aim to streamline the annotation process, enhance annotation accuracy in occluded regions, and provide a comprehensive cleanliness evaluation metric for videos.

2. Methods

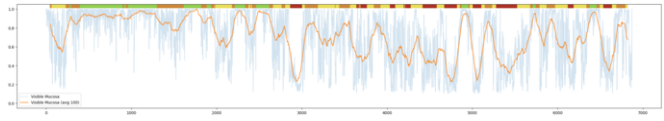
Instead of manually segmenting digestive content masks, a laborious task, we leveraged labels provided by physicians who annotated 2176 random patches of 64x64-pixel images as 1 (bad visibility) or 0 (good visibility) from 18 different videos. The labels were assigned based on whether the presence of digestive content would obstruct any potential pathology. Using this data, we trained a ResNet50 classifier with 14 videos (1696 patches) and we used 4 different videos (480 patches) for validation. To obtain the segmentation masks, we employed a sliding window approach with 32 pixels of overlap on each frame. This allowed us to compute the percentage of visible mucosa per frame using the segmentation masks. After applying this process to all images in a test video, we generated a plot illustrating the visibility level in each video segment. This plot provides valuable insights into the level of visible mucosa throughout the video.

3. Results

The classifier achieved an accuracy of 90.2% and an AUC of 95.7% on the classification task. In Figure 1(a), we present three examples of segmentation masks generated through the previously described procedure. Patches with predicted values exceeding 0.7 were classified as positive. Figure 1(b) illustrates the percentage of visible mucosa in blue, along with a moving average for 100 frames in orange. Additionally, we have overlaid a discretized scale from 0 to 3 on this plot, following score proposed by [2], to provide a clearer representation of the scores. This visualization helps in understanding the variations and trends in the scores throughout the video.



(a) Segmentation



(b) Video score

Figure 1. Output of the proposed system.

References

[1] Bucci C, Rotondano G, Hassan C, Rea M, Bianco MA, Cipolletta L, et al. Optimal bowel cleansing for colonoscopy: split the dose! A series of meta-analyses of controlled studies. Gastrointestinal Endoscopy. 2014;80(4):566-76.

[2] de Sousa Magalhães R, Arieira C, Carvalho PB, Rosa B, Moreira MJ, Cotter J. Colon Capsule CLEansing Assessment and Report (CC-CLEAR): a new approach for evaluation of the quality of bowel preparation in capsule colonoscopy. Gastrointestinal endoscopy. 2021;93(1):212-23.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230674

Development of an Automated CBCT Nerve and Teeth Segmentation Tool Based on Deep Learning

Vicent CASELLES [a], Wenceslao PIEDRA [b], Jordi CASAS [a], Xavier Paolo BURGOS [b] and Ferran PRADOS [a,c,d]

[a] *eHealth Center, Universitat Oberta de Catalunya, Barcelona, Spain*
[b] *Movumtech S.L., Madrid, Spain*
[c] *Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK*
[d] *Queen Square MS Centre, Department of Neuroinflammation, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, London, UK*
ORCiD ID: Vicent Caselles https://orcid.org/0000-0002-8697-3692, Jordi Casas https://orcid.org/0000-0002-0617-3303 , Xavier Paolo Burgos https://orcid.org/0000-0001-9681-5404, Ferran Prados https://orcid.org/0000-0002-7872-0142

Abstract. This long abstract details the translation work from research to industry that stems from the collaboration between a research group in Applied Data Sciences and a digital dentistry start-up, with the aim of developing a deep learning-based nerve and teeth segmentation tool for aiding dentist treatment decision-making and diagnosis.

Keywords. Deep Learning, CBCT Image Segmentation, Medical Imaging

1. Translation to develop an automated CBCT segmentation tool

3D Imaging techniques such as Cone-Beam Computed Tomography (CBCTs) and Intra-Oral scans (ICs) have become pivotal in diagnosis and treatment planning in orthodoncy, since they offer the possibility of extracting invaluable volumetric information. In this field, one of the main challenges that stem from implant dentistry and maxillofacial surgery is related to neural and dental damage during the procedure, with the Inferior Alveolar Nerve (IAN) being the most commonly injured nerve [1].

However, a key bottleneck in this workflow is the processing of the DICOM images produced by CBCTs to compute the 3D mesh. This process is tedious and time-consuming, and also requires extensive training, and has to be usually carried out by the dentist.

CBCTs are also commonly performed simultaneously with ICs, and their output has to then be registered to ensure an optimal mapping of the region of interest. Manual segmentation of CBCTs adds human error, and also shows high inter-subject variability.

Ultimately, this might lead to a non-viable alignment between the two image sources. The presence of metal implants introduces artifacts in the scan, increasing the segmentation's difficulty. Therefore, there is great interest in an automated segmentation tool that improves this workflow.

UOC's Applied Data Sciences lab, together with the digital dentistry start-up *Movumtech*, have joined forces to develop a deep learning based tool that will allow dentists to automatically select their anatomical ROI (i.e.: nerve, teeth, mandible, maxilla) and compute an STL mesh from the raw DICOM data.

For these reasons, a comprehensive database of CBCT scans will be assembled from healthy and pathological patients, with manually segmented labels with the oversight from *Movumtech*. Furthermore, IC scans will be used thus greatly improving segmentation accuracy thanks to its high resolution and imperviousness to metal artifacts.

When enough labels have been segmented, an atlas (expert validated reference consisting of the average of a set of templates) will be constructed to compare new input images to, extracting valuable information such as location, number of teeth present, and presence of artifacts.

With regards to the nerve-segmentation tool, we will be using the dataset from the *Toothfairy 2023 MICCAI* challenge, which contains CBCTs and dense labels for the IANs. Our approach follows the so-called triplanar or 2.5D approach, where the original 3D image is sliced along the three orthogonal axes [2] and each slice is predicted using the corresponding 2D model trained using the nnU-Net method [3].

A total of 130 patients were used in training. Data augmentation (gaussian blur, *flipping*) was used to improve the generalization ability of the models. Intensity was rescaled to a range of [0, 2048] to ensure images from different sources can be segmented.

Training was performed with the default nnU-Net hyperparameters except the max epochs, which was reduced to 360. Inference is performed on each slice and then merged obtaining three different 3D masks. These are thresholded and merged obtaining a binary mask. Several methods of postprocessing – majority voting, simple averaging or unanimous voting – will be studied. Preliminary results using probability averaging on 15 test set subjects show a DICE coefficient of 75%. This pipeline will be adapted and applied for teeth-segmentation after a sufficient number of images have been segmented.

Acknowledgements

FPC is partially funded by the National Institute for Health Research Biomedical Research Centre at UCL and UCLH.

References

[1] Tay ABG, Zuniga JR. Clinical characteristics of trigeminal nerve injury referrals to a university centre. International Journal of Oral and Maxillofacial Surgery. 2007;36(10):922-7. Available from: https://www.sciencedirect.com/science/article/pii/S0901502707001269.

[2] Dalbis T, Fritz T, Grilo J, Hitziger S, Ling WX. Triplanar U-net with orientation aggregation for new lesions segmentation. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. 2021:57.

[3] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021;18(2):203-11.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230675

Zero-Shot Prediction for Emotion Recognition Using Deep Spectrum Features

Elena ORTEGA BELTRÁN [a,1], Ismael BENITO-ALTAMIRANO [a,b] and Carles VENTURA [a]
[a] *Universitat Oberta de Catalunya*
[b] *Universitat de Barcelona, Departament d'Enginyeria Electrònica i Biomèdica*

Abstract. In this work we evaluate the zero-shot performance for emotion classification in audio recording datasets using a series of deep spectrum feature extractors which have been pre-trained on snore-analysus datasets.We evaluate here if those features are suitable for emotion-analysis on a zero-shot premise (no re-training of the CNNs). Once the feature extraction is complete, we evaluate training conventional classification models, i.e. an SVC, to classify a series of Spanish datasets which consist of male and female recordings expressing a set of emotions.

Keywords. Paralinguistics, Emotion recognition, Deep spectrum features

1. Introduction

One of the fields of Artificial Intelligence is the recognition of emotions from various sources. Among its applications we have, for example, social robots that are specifically designed to interact with humans in fields such as medicine[1,2]. In this work, we focus on the specific field of paralinguistics, the discipline that studies paralanguage, which is the set of phenomena that are not strictly verbal, of a vocal or gestural nature, that accompany the use of language and complete its meaning [3]. More specifically, this work focuses on the audio modality, where we propose to use a feature-based extractor, followed by a supported vector machine classifier (SVC), with pre-trained weights on another sound-related problem (snore analysis) [4], and we evaluate if zero-shot prediction is possible to solve our paralinguistic problem of emotion classification.

Results indicate that the proposed pipeline is suitable for emotion classification, achieving good results in zero-shot prediction. Moreover, we show how splitting the training of the SVC over gender –male/female– could improve our proposed solution.

2. Materials and methods

Our proposal is to use the so-called DeepSpectrum toolkit, a Python-based toolkit which extracts features from audio data using a pipeline that consists of converting audio data into a spectrum representation of this data, and then using pre-trained CNNs to extract

[1]Corresponding Author: Elena Ortega Beltrán, e-mail: eortegabeltran@uoc.edu.

features from those images [4]. Later, we propose using an SVM classifier, and train it to classify the feature representation of the audio data against the classes of the ELRA-S0329 dataset [5], which contains more than 6,000 recordings of Spanish recordings from one male and one female professional speakers. These recordings contain annotations of read text material for 11 classes: in neutral style with five different speech styles –loud, normal, slow, fast and soft– plus six expressions –anger, disgust, fear, joy, sadness and surprise–. The dataset is divided into train (80%) and test (20%) partitions.

We considered two experiments. On the one hand, we compared the results of training the SVC with features extracted with two pre-trained models: VGG16 and Xception. On the other hand, we divided the datasets by gender, and we trained the SVC model –using the VGG16 features– and we tested both trained models (male/female) against both subsets of the test dataset (male/female) as well as the merged one.

3. Results

The accuracy of the first test case (VGG16 architecture) is 82,4%, whereas the accuracy obtained when Xception is used is significantly lower (75%). For both implementations, results indicate that our proposed model (DeepSpectrum + SVC) presented a good response distinguishing the six emotion classes of the dataset (Anger, Disgust, Fear, Joy, Sadness and Surprise); however, it performs poorly dividing the neutral emotion tones (Fast, Slow, Loud, Soft and Normal). Futures works might include a separate training for the subset of neutral classes.

In the second experiment, we train the system both in the separated gender datasets and the complete one. We compare the different performance of both classifiers in the female and male datasets. We observe that the classifier trained in a specific gender subset achieves a better accuracy (83.88% and 83.19% for female and male respectively) than the one trained in the complete dataset (78.45% and 77.37% for female and male respectively).

Acknowledgements

This work is part of the project PLEC2021-007868, funded by MCIN and by the EU, and the project PID2022-138721NB-I00.

References

[1] Makiuchi MR, Uto K, Shinoda K. Multimodal emotion recognition with high-level speech and text features. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE; 2021. p. 350-7.

[2] Papakostas GA, Sidiropoulos GK, Papadopoulou CI, Vrochidou E, Kaburlasos VG, Papadopoulou MT, et al. Social robots in special education: a systematic review. Electronics. 2021;10(12):1398.

[3] Poyatos F. The multichannel reality of discourse: language-paralanguage-kinesics and the totality of communicative systems. Language Sciences. 1984;6(2):307-37.

[4] Amiriparian S, Gerczuk M, Ottl S, Cummins N, Freitag M, Pugachevskiy S, et al. Snore sound classification using image-based deep spectrum features. 2017.

[5] ELRA, Emotional speech synthesis database,:. http://catalog.elra.info/en-us/repository/browse/ELRA-S0329/.

This page intentionally left blank

Part 3

AI for Health: Theory and Applications

This page intentionally left blank

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230677

Bayesian Optimization with Additive Kernels for the Calibration of Simulation Models to Perform Cost-Effectiveness Analysis

David GÓMEZ-GUILLÉN [a,b,1], Mireia DÍAZ [b,c] Josep Lluís ARCOS [d] and Jesus CERQUIDES [d]

[a] *Universitat Autònoma de Barcelona (UAB)*
[b] *Institut Català d'Oncologia (ICO) - Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)*
[c] *Consortium for Biomedical Research in Epidemiology and Public Health - CIBERESP. Carlos III Institute of Health*
[d] *Institut d'Investigació en Intel·ligència Artificial - Consell Superior d'Investigacions Científiques (IIIA-CSIC)*

ORCiD ID: David Gómez-Guillén https://orcid.org/0000-0003-1787-6482, Mireia Díaz https://orcid.org/0000-0001-9360-4548, Josep Lluís Arcos https://orcid.org/0000-0001-7751-1210, Jesus Cerquides https://orcid.org/0000-0002-3752-644X

Abstract. The use of mathematical simulation models of diseases in economic evaluation is an essential and common tool in medicine aimed at guiding decision-making in health. Cost-effectiveness analyses are a type of economic evaluation that assess the balance between health benefits and the economic sustainability of different health interventions. One critical aspect of these models is the accurate representation of the disease's natural history, which requires a set of parameters such as probabilities and disease burden rates. While these parameters can be obtained from scientific literature, they often need calibration to fit the model's expected outcomes. However, the calibration process can be computationally expensive and traditional optimization methods can be time-consuming due to relatively simple heuristics that may not even guarantee feasible solutions. In this work, we investigate the use of Bayesian optimization to enhance the calibration process by leveraging domain-specific knowledge and exploiting inherent structural properties in the solution space. Specifically, we examine the effect of additive kernel decomposition and constraint handling for efficient search. Our preliminary results show that this improved Bayesian optimization procedure asymptotically improves the calibration process, leading to faster convergence and better solutions for larger simulation models.

Keywords. bayesian optimization, gaussian processes, additive kernels, constrained optimization, simulation models, cost-effectiveness models, cancer research

[1] Corresponding Author: David Gómez-Guillén, dgomez_ext@iconcologia.net

1. Introduction

Healthcare interventions are increasingly evaluated based on their cost-effectiveness due to usual budgetary constraints, ensuring equitable and efficient distribution of healthcare services. These constraints mean that not all available and effective interventions can be included in health plans. In many countries, it has become standard policy to assess the costs of new healthcare interventions in relation to their expected benefits before implementing them. Cost-effectiveness analysis (CEA) using mathematical simulation models is a crucial tool in this context, enabling us to assess the value of healthcare interventions and determine which ones offer the best value for money in the long term[1]. By comparing the costs and benefits of alternative interventions, policymakers and healthcare providers can prioritize strategies and allocate resources to achieve the maximum health benefits for the population. Ultimately, the goal of healthcare is to improve health outcomes, and CEA plays a vital role in achieving this objective[2].

CEA usually relies on simulation models that mimic disease processes to project the effects of different medical strategies on health outcomes over time[3]. There are different types of models but some of the most common simulate the traversal of a group of individuals through different health states (figure 1). These models can generate various outcomes, but they always produce two critical measures: the average cost and the average life expectancy, usually measured in Quality-Adjusted Life Years (QALYs)[4].

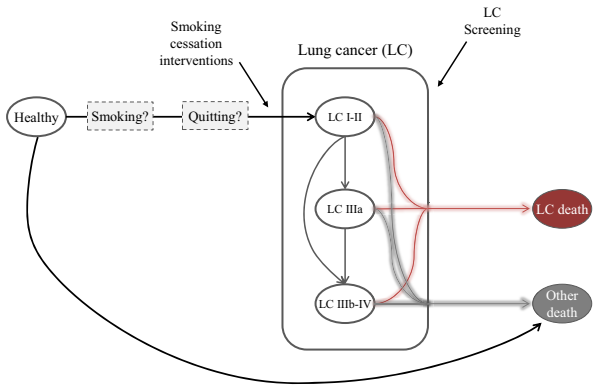


Figure 1. Lung cancer markov model state diagram

In order to execute the simulations, input parameters are required to describe the disease process, such as probabilities, hazard ratios, or disease burden rates extracted from scientific literature. Due to the inherent uncertainty of these values, it is often necessary to calibrate the model before proceeding with the analysis. Calibration consists in adjusting the input parameters until the resulting output approximates a target value identified in the scientific literature, such as disease incidence, prevalence, or mortality. This optimization process can be especially taxing for complex models, and may necessitate the use of advanced techniques to efficiently explore the solution space.

Moreover, calibrations can be highly dimensional optimization problems with many arbitrary constraints between parameters, dictated by the specific medical domain. In this work, we explore the challenges associated with calibrating simulation models and propose methods to overcome them. Our research provides valuable insight into novel ways

to calibrate these simulation models more efficiently using Bayesian Optimization (BO). We also investigate the circumstances under which it outperforms common methods used in the field.

2. Background

2.1. Bayesian Optimization

BO is a powerful technique for optimizing expensive, black-box functions that are difficult or time-consuming to evaluate[5]. The key advantage of this method is that it can find the global optimum of the function with relatively few evaluations, even in high-dimensional spaces. This is because the method actively seeks out the most promising areas of the input space to evaluate next, rather than simply evaluating points at random or using simple heuristics. The downside is that the method can be computationally expensive, especially for functions with a large number of input variables or when the surrogate model is complex.

BO has been successfully applied to a wide range of optimization problems in machine learning, including hyperparameter tuning, experimental design, and automatic algorithm configuration. It can help to quickly identify good values of the hyperparameters or experimental conditions, without having to exhaustively search the entire parameter space.

2.2. Gaussian Processes

Gaussian Processes (GPs) are non-parametric regression models that represent each observation as a random variable drawn from a normal distribution $f(x) \sim \mathcal{N}(\mu(x), k(x,x))$ [6]. The covariance function or kernel is the mechanism to give a GP its expressive power, and its choice will heavily depend on the kind of function we aim to model[7]. The squared exponential (SE) kernel $k(x,x') = \sigma^2 e^{-\frac{||x-x'||^2}{2l^2}}$ is a popular choice, despite significant drawbacks such as its locality and sensitivity to the curse of dimensionality[8]. In general, modeling complex high dimensional functions using a single kernel can be computationally expensive using local kernels, making this a first class research problem (e.g. [12][13]). Additive kernel decomposition[9] addresses this problem by breaking down the kernel into a sum of simpler kernels, each of which captures a different aspect of the relationship between the input variables. This approach can capture both local and global interactions between the input variables. Additionally, additive kernel decomposition can improve the interpretability of the model, as each kernel term can be associated with a specific interaction term. This can help users understand which orders of interaction are important for the current optimization problem.

Additive kernels suffer from the non-identifiability problem: the kernel hyperparameters are not uniquely identifiable from the observed data, which can lead to challenges in model selection and interpretation. To ensure a unique decomposition, Lu et al[11] proposed an extension of additive kernels by including an extra constant kernel $\tilde{k}_{add_0}(x,x')$ with an additional variance hyperparameter σ_0^2 and an orthogonality constraint to generate Orthogonal Additive Kernels (OAK)[11]. Assuming a normal input distribution $x_i \sim \mathcal{N}(\mu_i, \delta_i^2)$, the following constrained base kernel is derived:

$$
\begin{aligned}
k_{add_{OAK}}(x,x') &= \sum_{i=0}^{D} \sigma_i^2 \tilde{k}_{add_i}(x_i,x_i') \\
\tilde{k}_{add_j}(x,x') &= \sum_{1\le i_1<i_2<\ldots<i_j\le D} \left[\prod_{d=1}^{j} \tilde{k}_{i_d}(x_{i_d},x_{i_d}')\right] \\
\tilde{k}_i(x,x') &= e^{\frac{(x_i-x_i')^2}{2l_i^2}} - \frac{l_i\sqrt{l_i^2+2\delta_i^2}}{l_i^2+\delta_i^2} e^{-\frac{(x_i-\mu_i)^2+(x_i'-\mu_i)^2}{2(l_i^2+\delta_i^2)}}
\end{aligned}
\tag{1}
$$

One important advantage of these additive kernels is that we can interpret the σ_i^2 as the contribution of each individual order to the total kernel. Since many problems often rely on a few low-order interactions, we can truncate the higher orders and limit the computational cost while retaining most of the information present in the full decomposition. To achieve this, OAK kernels can be useful in accurately identifying each contribution and providing an accurate representation on the actual composition on the function.

3. Methodology

In this section we will describe the simulation model and the optimization methods we will use to calibrate it. These include BO and the rest of techniques that will be compared.

3.1. Description of the Simulation Model

We use a lung cancer model presented in a published cost-effectiveness analysis[14] as a fast benchmark for BO on simulation models. This Markov-based microsimulation model simulates a cohort's progression through six different health states from 35 to 79 years of age, in monthly intervals. The transition probabilities used in the model were age-specific, with distinct values for each 5-year age group (i.e. 35-39, 40-44, ..., 75-79). The state diagram for this model is pictured in figure 1.

Certain inherent constraints, such as ensuring that the sum of the probabilities in each row equals one or that certain probabilities are zero, were imposed on the matrices. This allowed the number of parameters to be optimized per age group to be reduced from 36 to 11. From the nine age groups, each one with a set of 11 parameters, only the first few of these were calibrated in this study. As a result, the problem was simplified to the calibration of 11 parameters, rather than the original $11 \cdot 9 = 99$ parameters associated with the full simulation. Furthermore, this model was designed to be computationally inexpensive, taking less than 10ms to simulate. By introducing arbitrary delays in the model we can observe the relationship of optimization times and model simulation times for different optimization methods.

The calibration target for the model was defined as the weighted sum of the euclidean distances between the observed and expected outputs of interest, namely lung cancer incidence (45%), lung cancer mortality (45%) and mortality from other causes (10%), computed for each age group.

3.2. Optimization Methods

In cost-effectiveness analysis, it is common to have an initial estimate of a good solution based on approximate values found in the scientific literature. For all optimization experiments conducted, a solution space of plus or minus $\pm 50\%$ was considered for each input variable, centered around this initial value.

First, we used different optimization methods to illustrate the performance differences between regular BO and classical methods. For this purpose we used python implementations of commonly used methods: a hill-climbing technique (Nelder-Mead[2][15]), metaheuristics (Simulated Annealing (SA)[3][16] and Particle Swarm Optimization[4][17]), and BO with GPs[5]. The default hyperparameter values were used for these methods, except for Particle Swarm Optimization, where the number of particles was set to 1,000 times the number of age groups calibrated.

For a second, distinct set of experiments we developed a new BO implementation with GPs from scratch using the R programming language. This implementation was used as a rapid prototyping environment to evaluate different enhancements to the optimization process for our specific domain, without being concerned by execution time at this stage. This implementation uses the Expected Improvement acquisition function, with Particle Swarm Optimization to search for the maximum. Finally, both the SE and the OAK kernels were implemented. Other runtime optimizations such as GPU use are beyond the scope of this work.

Note that the methods implemented in python were used exclusively for a fair execution time comparison between optimization methods in a common python runtime environment, while the methods implemented in R were used only for a fair error comparison among BO alternatives, without considering execution time.

3.3. Hyperparameter tuning

Before starting the BO procedure we learn the lengthscales $l_1, ..., l_D$ and the variances $\sigma_0^1, \sigma_1^2, ..., \sigma_n^2$ of the OAK kernel in a two stage process. This approach allows us to break down a potentially complex hyperparameter tuning task for high-dimensional problems into low-dimensional, manageable problems.

In the first stage, lengthscales are found by maximizing the marginal likelihood for each dimension separately. This approach is followed due to the implicit assumption that our simulation models have a strong additive component of order 1 and that a linear combination of one-dimensional kernels can be a reasonable approximation. These optimization subtasks are D simple univariate convex problems: for small and large lengthscales the kernel overfits and underfits, respectively, producing low-likelihood models. Each optimum is a unique value between these two extremes, quick to find using simple binary search.

In the second stage, the marginal likelihood is maximized for the whole set of variances, given the previously found lengthscales. This is a $(D+1)$-dimensional optimization subtask, solved using Nelder-Mead[15].

[2] https://docs.scipy.org/doc/scipy/reference/optimize.minimize-neldermead.html

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.dual_annealing.html

[4] https://pyswarms.readthedocs.io/en/latest/

[5] https://secondmind-labs.github.io/trieste/1.1.2/index.html

4. Results

Our interest in these experiments are about the comparative performance of BO versus the rest of methods used in the calibration of simulation models. We can see the relationship between simulation time and calibration time for each method in figure 2. For very fast models the inference overhead of BO dominates and other methods are able to calibrate faster by simulating the model many times. However, as the simulation time increases, the Bayesian method efficient approach in number of function evaluations results in faster calibration times. Specifically, for a model with 11 parameters and simulation times of less than 250ms, we observed the Bayesian approach outperform the alternative methods.

In contrast, as the dimensionality of our problem grows, the Bayesian method overhead increased significantly, as shown in the y-intercept of figure 2. The calibration times for the other methods also increased but, overall, the simulation time threshold where the Bayesian approach outperforms the other methods increases exponentially with the number of parameters, from 0.2 seconds to 0.35, 0.95 and 3.25 seconds. The bottom left plot in figure 2 projects that Bayesian calibration of all 99 parameters would be the best technique when each simulation takes approximately 5 minutes of computation.

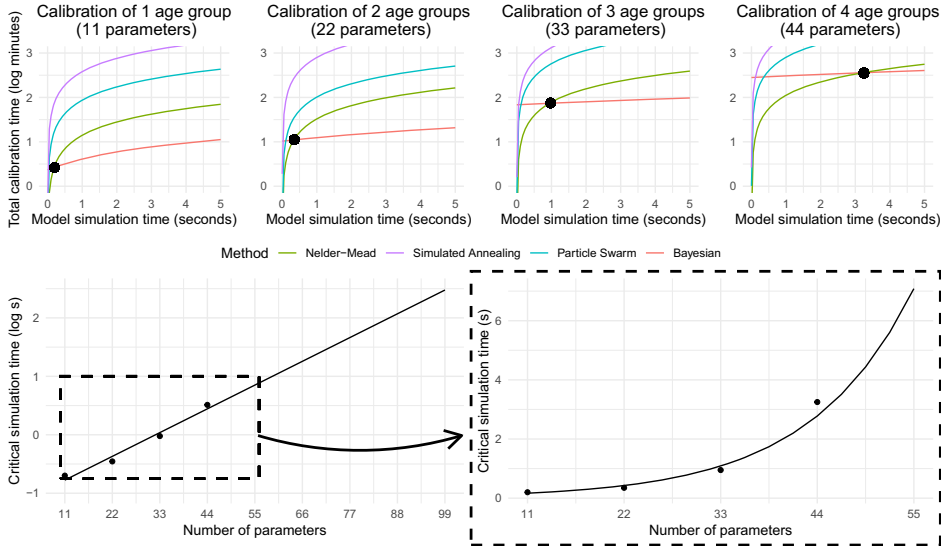


Figure 2. Total calibration time in log scale against model simulation time required to attain similar levels of error. The bottom left figure shows the exponential trend (in log scale) in the necessary simulation time before the bayesian method becomes the fastest method, as a function of the number of parameters. The bottom right figure is a zoomed-in plot of the same figure removing the log scale. The methods used in this comparison were all implemented in python.

In any case, the focus of our research in this work is the number of evaluations, where we see a sharp drop in error when using BO to achieve a similar level of accuracy compared to other methods, as shown in figure 3. Although each iteration requires a significant amount of time due to the Bayesian inference step, this overhead will become less relevant as the size of the model increases.

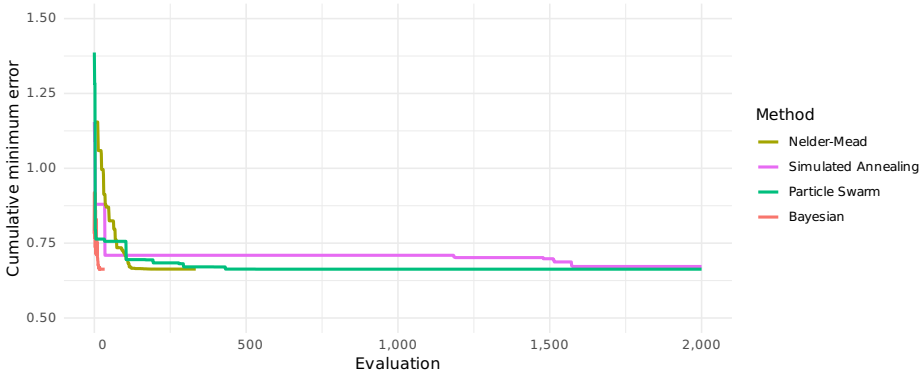


Figure 3. Time series of the lung cancer model calibration error (using 1 age group). The error is plotted against the number of evaluations by method.

While exploring the results of BO with OAK kernels we found that one of the variables had very significant explanatory power by itself, which could produce misleading results in the comparison. To address this issue, we introduced a third univariate SE kernel that considers only this variable. Figure 4 shows the average progression of the error during the optimization process for the three kernels and their interquartile range for a sample of 30 random executions. The univariate SE kernel shows a lower average error and lower spread than the full SE kernel, due to the reduction in dimensionality of the problem that allows for an easier exploration of the solution space, with barely any information loss. However, the OAK kernel under a normality assumption for the inputs is able to efficiently search the full 11-dimensional space to reach even better average results than the univariate SE kernel, while reducing the dispersion as the optimization progresses.

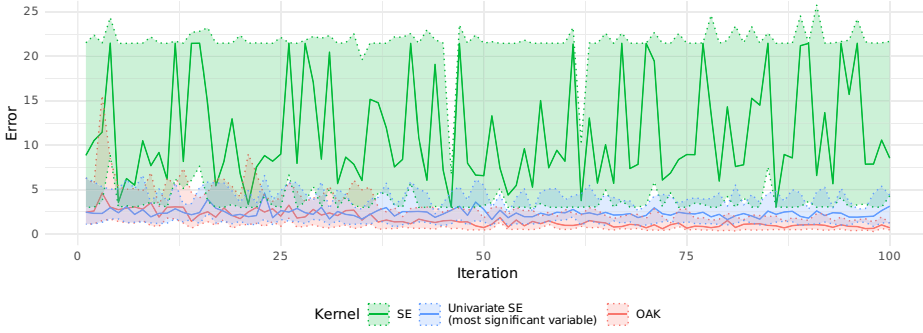


Figure 4. Time series of the median BO error with its interquartile range as the shaded area. We used three different kernels: the SE kernel (blue), the univariate SE kernel using only the most significant variable (green) and the OAK kernel under a normality assumption for the inputs (red). The methods used in this comparison were all implemented in R.

5. Discussion

BO is currently considered a state-of-the-art optimization method in various domains that involve costly function evaluations. Even though this result is already well known in the literature, an important research question would be determining the threshold at which the cost of the functions justifies the use of BO. Our findings show that simulation models with execution times of only a few seconds can be expensive enough to warrant the use of BO. In the cases when the function evaluations are not as costly, there are two other critical points in each BO iteration that must be taken into account: the surrogate model regression and the acquisition function optimization.

Regarding the former, SE kernels suffer greatly from the curse of dimensionality. In high-dimensional problems, the number of observations to explore the solution space quickly increases. This renders the regression unfeasible when trying to invert large Gram matrices to calculate the posterior predictive distribution. To address this issue, OAK kernels can reduce the number of necessary observations, mitigate the effects of an expanding Gram matrix and enhance the efficiency of the search. The observed scaling issues in figure 2, resulting from increasing dimensionality, justify the need for high-dimensional improvements such as OAK kernels. Nevertheless, our experiments showed that the asymptotic behaviour of BO persisted, making it the optimal choice for sufficiently large models.

A remarkable insight about additive kernels can be found in the comparison made in figure 4. As explained in the results, we considered a simulation model with eleven parameters, where one of the parameters was found to have a significant impact in the overall error. From a domain view, this parameter corresponds to the probability of death from other causes, which in this simulation has a greater impact on a greater amount of people than the rest. If we don't use additive kernels the optimization process has to explore all eleven dimensions and it is incapable of reducing the error over 100 iterations (green line). If we focus on this significant parameter by itself (blue line) we can see that the exploration finds much better solutions, but for more complex problems it might be difficult to manually isolate the important variables. The additive kernel (red line) is able to automatically detect this fact and perform what could be viewed as some kind of variable selection, while at the same time managing to refine better solutions with the rest of parameters.

The optimization of the acquisition function is the initial bottleneck, where the number of observations is still small enough so that the surrogate model regression is not yet a problem. As the search space remains constant, this acquisition optimization doesn't become much more expensive as more data is observed. We used Particle Swarm Optimization as an easy way to take advantage of parallelism in this area, but other approaches mentioned in the next section are being considered as well[10].

Lu et al[11] mention that an interesting direction of work would be to extend OAK kernels to BO leveraging the inferred low-order representation. In our tests we show that, even with a straightforward application of OAK kernels on this simple example, a slight improvement over the SE kernel is noticeable. This improvement is expected to be more meaningful for complex models, where more structure can be leveraged. It is interesting to note that these results hold even though some assumptions of the model were not met. Specifically, hyperparameter tuning was performed with a dataset sampled from a uniform input distribution, while the constrained kernels were calculated assuming

normality in the input. Even if these distributions were consistent, we still would have the problem of determining the input distribution for the actual optimization process, which would be neither normal nor uniform.

6. Future Work and Conclusions

One particular aspect that we did not incorporate into this article is constraint handling. Simulation models can be highly constrained problems, and these constraints are another expression of the structure of the solution space. We have been able to manage arbitrary constraints successfully using additional surrogate models and a new Constrained Expected Improvement acquisition function, as introduced by Gardnet et al[18].

We also mentioned in the discussion the convenience of exploiting the parallelization potential of the different areas of the optimization process. For that purpose we use the Particle Swarm method for the optimization of the acquisition function, but other more sophisticated venues for parallelization include batched optimization[19], parallel acquisition functions[20] or GPU approaches[21] among others.

Our research group recognizes the importance of efficiently calibrating increasingly complex models, injecting relevant domain knowledge in the process. In this work we have shown that using OAK kernels in a BO setting allows faster calibration times under very common circumstances in the health economics field. This is the beginning of several enhancements that are being implemented to have the tools to work with more challenging models in the future.

7. Acknowledgement

This work was supported by a grant from the Instituto de Salud Carlos III-ISCIII (Spanish Government) co-funded by European Regional Development Fund, a way to build Europe (CIBERESP CB06/02/0073, PI19/01118), also with the support of the Secretariat for Universities and Research of the Department of Business and Knowledge of the Government of Catalonia. Grants to support the activities of research groups (SGR 2017–2021). Grant number 2021SGR1029. We thank the CERCA Programme and Generalitat de Catalunya for institutional support.

References

[1] Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW, Drummond MF, et al. Methods for the Economic Evaluation of Health Care Programmes. Fourth Edition, Fourth Edition. Oxford, New York: Oxford University Press; 2015. 464 p. doi: 10.1002/(SICI)1099-176X(199903)2:1¡43::AID-MHP36¿3.0.CO;2-7

[2] Levin HM. Cost-Effectiveness Analysis: Methods and Applications. 2nd edition. Thousand Oaks, Calif: SAGE Publications, Inc; 2000.

[3] Gray, Alastair M., Philip M. Clarke, Jane L. Wolstenholme, Sarah Wordsworth, Alastair M. Gray, Philip M. Clarke, Jane L. Wolstenholme, Sarah Wordsworth. Applied Methods of Cost-effectiveness Analysis in Healthcare. Handbooks in Health Economic Evaluation. Oxford, New York: Oxford University Press, 2010. doi: 10.1093/pubmed/fds009

[4] Weinstein MC, Torrance G, McGuire A. QALYs: the basics. Value Health. 2009 Mar;12 Suppl 1:S5-9.

[5] Garnett R. Bayesian Optimization. Cambridge: Cambridge University Press; 2023. doi: doi:10.1017/9781108348973

[6] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge, Mass: MIT Press; 2006. 248 p. (Adaptive computation and machine learning). doi: 10.7551/mitpress/3206.001.0001

[7] Duvenaud D, Lloyd J, Grosse R, Tenenbaum J, Zoubin G. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In: Proceedings of the 30th International Conference on Machine Learning. PMLR; 2013. p. 1166–74.

[8] Bengio Y. On the challenge of learning complex functions. Prog Brain Res. 2007;165:521–34. doi: 10.1016/S0079-6123(06)65033-4

[9] Duvenaud DK, Nickisch H, Rasmussen C. Additive Gaussian Processes. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2011.

[10] Wilson JT, Hutter F, Deisenroth MP. Maximizing acquisition functions for Bayesian optimization. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2018. p. 9906–17. (NIPS'18).

[11] Lu X, Boukouvalas A, Hensman J. Additive Gaussian Processes Revisited. In: Proceedings of the 39th International Conference on Machine Learning [Internet]. PMLR; 2022 [cited 2022 Oct 5]. p. 14358–83.

[12] Durrande N, Ginsbourger D, Roustant O. Additive covariance kernels for high-dimensional Gaussian process modeling. Annales de la Faculté de Sciences de Toulouse. 2012;Tome 21(numéro 3):481.

[13] Binois M, Wycoff N. A Survey on High-dimensional Gaussian Process Modeling with Application to Bayesian Optimization. ACM Trans Evol Learn Optim. 2022 Aug 16;2(2):8:1-8:26. doi: 10.1145/3545611

[14] Diaz M, Garcia M, Vidal C, Santiago A, Gnutti G, Gómez D, Trapero-Bertran M, Fu M; Lung Cancer Prevention LUCAPREV research group. Health and economic impact at a population level of both primary and secondary preventive lung cancer interventions: A model-based cost-effectiveness analysis. Lung Cancer. 2021 Sep;159:153-161. doi: 10.1016/j.lungcan.2021.06.027

[15] Nelder JA, Mead R. A Simplex Method for Function Minimization. The Computer Journal. 1965 Jan 1;7(4):308–13. doi: 10.1093/comjnl/7.4.308

[16] Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by Simulated Annealing. Science. 1983 May 13;220(4598):671–80. doi: 10.1126/science.220.4598.671

[17] Bonyadi MR, Michalewicz Z. Particle Swarm Optimization for Single Objective Continuous Space Problems: A Review. Evolutionary Computation. 2017 Mar;25(1):1–54. doi: 10.1162/EVCO_r_00180

[18] Gardner JR, Kusner MJ, Xu Z, Weinberger KQ, Cunningham JP. Bayesian optimization with inequality constraints. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. Beijing, China: JMLR.org; 2014. p. II-937-II–945. (ICML'14).

[19] González J, Dai Z, Hennig P, Lawrence N. Batch Bayesian Optimization via Local Penalization. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) [Internet]. 2016. p. 648–57. (JMLR Workshop and Conference Proceedings; vol. 51).

[20] Wang J, Clark SC, Liu E, Frazier PI. Parallel Bayesian Global Optimization of Expensive Functions. Operations Research. 2020;68(6):1850–65. doi: 10.1287/opre.2019.1966

[21] Wang K, Pleiss G, Gardner J, Tyree S, Weinberger KQ, Wilson AG. Exact Gaussian Processes on a Million Data Points. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2019.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230678

Exploring the Role of Explainability in AI-Assisted Embryo Selection

Lucia URCELAY [*a], Daniel HINJOS[*a], Pablo MARTIN-TORRES[*a], Marta GONZÁLEZ[*a], Marta MÉNDEZ[c], Salva CÍVICO[c], Sergio ÁLVAREZ-NAPAGAO[a,b], Dario GARCIA-GASULLA[a,b]

[a] *Barcelona Supercomputing Center*
[b] *Universitat Politécnica de Catalunya - Barcelona TECH*
[c] *Hospital Clínic de Barcelona*

Abstract. In Vitro Fertilization is among the most widespread treatments for infertility. One of its main challenges is the evaluation and selection of embryo for implantation, a process with large inter- and intra-clinician variability. Deep learning based methods are gaining attention, but its opaque nature compromises their acceptance in the clinical context, where transparency in the decision making is key. In this paper we analyze the current work in the explainability of AI-assisted embryo analysis models, identifying the limitations. We also discuss how these models could be integrated in the clinical context as decision support systems, considering the needs of clinicians and patients. Finally, we propose guidelines for the sake of increasing interpretability and trustworthiness, pushing this technology forward towards established clinical practice.

Keywords. embryo selection, IVF, artificial intelligence, explainability, ethics

1. Introduction

Infertility is a common reproductive health problem that affects millions of people worldwide, causing social, psychological, physical and economic distress to the ones seeking to conceive [7]. In the coming years infertility rates are projected to grow due to environmental and lifestyle factors [18, 37]. In vitro fertilization (IVF) technology is used to overcome infertility, it involves the fertilization of an egg with sperm in the laboratory, followed by the transfer of the resulting embryos into the patient's uterus. The main challenge of IVF is the selection of the embryos that will be either selected for implantation, frozen (for later implantation) or discarded (if they exhibit undesirable features). This selection is to be performed during the early hours after embryo insemination, typically between three and five days after. During this time, embryos are monitored in time-lapse imaging incubators (TLI), facilitating uninterrupted embryo growth within stable culture conditions. This technology offers a dynamic perspective on in vitro embryonic development, augmenting the clinical effectiveness of IVF [29].

[*] Equal contributors

To assess quality, embryologists evaluate different morphological characteristics depending on the embryo development phase. Early development (days one to three) focus on cell number, symmetry and fragmentation rate, while embryos reaching bastocyst stage (day five) are further assessed by their expansion grade and the appearance of the inner cell mass (ICM) and the Zona Pellucida (ZP), as well as to the trophectoderm cells (TE) [30]. These morphological features represent the foundation of current development assessment guidelines such as Gardner's [15] or ASEBIR in Spain. These approaches are limited by the subjective assessment of embryologists, which causes inter and intra-observer variability. The success rates of IVF are restricted due to these limitations. According to [1], the global life birth rate via IVF in 2015 stood at 19.2% per oocyte retrieval with fresh embryo transfer and 24.8% with frozen embryo transfer.

Artificial Intelligence (AI), and specially Deep Learning (DL), due to its capacity for dealing with images, have recently been considered to assist in the embryo assessment and selection process. AI has the potential to facilitate and improve the process of embryo selection, increasing the implantation success rates, and reducing the chances of multiple pregnancies. AI can also mitigate inter and intra-observer variability, making results more reproducible and comparable [35]. Finally, AI can help reduce the financial, physical and emotional burden on patients, by optimising the treatment plan and minimising the need for repeated cycles of IVF.

Whenever AI is considered for an application which directly affects human life, trustworthiness must be assured. In this paper we review the current work on AI-assisted embryo selection with special focus on Explainable Artificial Intelligence (XAI), and how this is deployed. We analyze the limitations of current work and propose approaches that would yield AI systems that can provide better explanations for their recommendations. Moreover, we discuss key aspects in the integration of these systems in a clinical context where both clinicians and patients are involved in the decision-making process.

2. Related Work

This work discusses the integration of explainability components into AI solutions for embryo analysis in IVF, as well as the particularities of their integration into a human-centric decision making process. Previous solutions proposed in the bibliography are reviewed within the main discussion of this work (see §3.1), as it contextualizes our contribution. In contrast, the related work presented in this section comprehends previous analysis and surveys conducted in the field, which have motivated this paper.

Among recent surveys on AI-assisted embryo selection in IVF, most restrict themselves to the technicalities of the machine learning problem itself. That is, what is the problem to solve, which data is used and how, and which are the top-of-the-line performance metrics achieved by the models built. A few discuss the biological aspects of the problem with particular depth [22], while others focus on the variety of tasks one can tackle (*e.g.*, embryo segmentation, quality grading, etc.) [20]. The most ambitious tasks, those of predicting clinical pregnancy or fetal heartbeat from five days old fertilized embryo images, are reviewed in [23], together with the ploidy status (number of chromosomes) prediction task. [26] includes the annotation of data (*e.g.*, development phase, cell counting) in its review, while the particularities of the time-lapse imaging are

the focus of [28]. On top of embryos, [11] also reviews AI applications for spermatozoa and oocyte analysis. [13] goes beyond DL methods, and also consider Bayesian Networks and SVMs in their study. Remarkably, very few works [2, 40, 34] study the ethical, socio-economic, legal and cultural (ELSEC) aspects in AI-assisted embryo selection.

None of the previously discussed works includes a review on the XAI methods used for embryo analysis, or discussion on practical and responsible deployment. This work aims at filling that gap, considering that both these issues need to be properly assessed before clinicians can integrate AI in their day-to-day practice.

3. The Role of Explainable Artificial Intelligence in Embryo Analysis

Embryo analysis post-fertilization lasts for a maximum of 7 days. During that time, clinicians assess the status and evolution of embryos, as these evolve under controlled conditions. Through continuous visual monitoring and tracking of their morphological features, experts eventually discard, freeze or select embryos for transfer. Current embryo incubators include periodic microscopy, with images being recorded at frequency in the order of minutes (*e.g.*, ten) at several focal planes.

AI models trained on time-lapse embryo images have shown promise in embryo analysis. Within the domain of images, explainable AI methods are often feature scorers, representing their output as saliency maps on the input image (see Figure 1). By nature, XAI methods for DL are approximations to the model's real behavior. It is worth noting that XAI is critical to ensure that the decision-making process of an algorithm is as transparent and understandable to clinicians and patients as possible. The lack of explainability can lead to a lack of trust in the technology, which ultimately hinders its adoption.

In this section we review the use of XAI on embryo images. Among the methods found in the literature, we can differentiate between those applied after the model is trained (post-hoc), and those where the explainability is part of the model architecture and design (intrinsic). Post-hoc methods include model-agnostic ones such as LIME [33] and KernelSHAP [27], that obtain the explanations by perturbing the inputs and observing the changes caused on the outputs; also, model-specific methods such as Grad-CAM [38] and Deep SHAP [27], that use the parameters of neural network models to obtain the saliency maps. Among intrinsic methods, we can find attention-based CNNs, where the explainability is obtained from attention layers. A different approach is the use of Speeded Up Robust Features (SURF) [6] to extract local visual features, in conjunction with Gaussian Mixture Models (GMM) to obtain a Fisher vector per image [21].

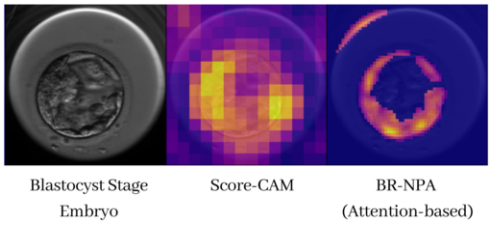


Figure 1. Saliency maps for Score-CAM and BR-NPA XAI methods from [17].

Ref.	Task	Architecture	XAI Method(s)	ClinicalEval
[10]	Embryo Quality Grading	DenseNet, ResNet	Grad-CAM++	No
[43]	Embryo Quality Grading	LWMA-Net	Grad-CAM	No
[21]	Embryo Quality Grading	SURF and GMM, ResNet-101, EfficientNet-B1	SURF, Grad-CAM	No
[4]	Embryo Segmentation	MASS-Net	Grad-CAM	No
[39]	Development Stage Identification	ResNet34, VGG16	Grad-CAM, SHAP, LIME	Yes
[32]	Embryo Quality Grading	ResNet50, VGG-16	Grad-CAM	No
[17]	Development Stage Identification	ResNet50, ABN	B-CNN, ABN, InterByParts, Grad-CAM++, RISE, Score-CAM, Ablation-CAM, AM	No
[12]	Fetal Heart Pregnancy, Live Birth	ResNet18, RandomForest	Grad-CAM, SHAP	No
[36]	Live Birth	ABN	ABN	Yes
[42]	Embryo Quality Grading	VGG-16	Grad-CAM	No
[41]	Embryo Quality Grading	Xception, Inception-ResNET-v2	CAM	No

Table 1. Summary table of studies that use XAI on embryo selection. The Clinical Evaluation column indicates whether the resulting saliency maps have been evaluated by experts or not.

3.1. Review on Current Work

Depending on the specific task, XAI is used and interpreted differently:

Embryo Quality Grading: the classification of embryos (*e.g.*, day 5 blastocysts), into classes for selecting the best for implantation. This challenge is the most addressed regarding XAI. Some studies apply Grad-CAM for visualization [32, 42] and observe that key regions that the model relies on seem to be consistent with clinical interpretation. [10] uses Grad-CAM++ for visualization, while statistically correlating their model outputs with the Gardner score. [41] directly use CAM saliency maps, and observe their model focus on known features like cellular fragmentation, blastomeres or vacuoles. A different approach compares Grad-CAM and SURF (computing a single Fisher Vector per image), finding the former more prone to erroneously focus on irrelevant areas [21].

Embryo Development Stage Identification: consists on the automated labelling of the embryo stage. Challenges include deformed cell shapes, poor visual features or similarities between embryos at different stages. [39] compares Grad-CAM and LIME classification activations, concluding that LIME explanations seem less consistent with biologically important regions. Authors also suggest SHAP could be useful to identify reasons for misclassification between adjacent cleavage stages.

Blastocyst Segmentation: four crucial morphological regions to segment: TE, ZP, ICM and BL (see §1). A multiscaling architecture that outputs segmentation masks for the four regions is introduced in [4], where authors apply Grad-CAM to different layers to show the evolution of activations. A U-Net has also been used to separate the background

from the blastocyst [21], using an ellipse on top of the segmentation mask to separate inner cell mass from trophectoderm.

Fetal Heart Pregnancy: is defined as the presence of fetal heartbeats in the uterus. Of course, the embryo implantation process does not guarantee a successful pregnancy, as it depends on many other factors (*e.g.*, age, progesterone levels, *etc.*). [12] uses Grad-CAM to detect morphological indicators, and SHAP to account for relevant metadata that influences the model's decision, concluding that embryo images are the best predictor for fetal heart beat, followed by age and pregnancy history.

Live Birth: prediction from time-lapse imaging, [36] uses an Attention Branch Network (ABN) [14]. Authors visualyze embryo features of relevance for *Live Birth* by exploring the attention mechanism. Results indicated that no common visual features could be associated to the predicted outcome of live or non-live birth.

To the best of our knowledge, other very important challenges in this field, such as *Blastocyst Prediction* or *Ploidy Detection*, are not addressed in the context of explainability. This highlights the need for further research in these tasks.

3.2. Limitations and Suggestions

A limited amount of research comprehends the use of AI-assisted IVF in combination with XAI methods. A representative list can be found in Table 1. In many cases XAI is only considered through the illustration and minor discussion of a few saliency maps [4, 21, 41, 43]. Others take one more step and analyze the map activations in order to correlate them to morphological features [39, 42] or to the objective [12]. Few share the saliency maps with expert embryologists for evaluation [39, 36]. Meanwhile, several studies conclude that saliency maps should not be used as the sole source of explainability in high risk medical domains [5, 16].

Considering the features that clinicians base their decisions on, and the difficulty of interpreting saliency maps, we suggest the use of expert models. Different models targeting different morphokinetical features, such as fragmentation score, ICM and TE grades, ZP and Pronucleus characteristics, and phase transition times. The ensemble of all these experts should be a white box model (*e.g.*, GLM, Decision Tree) to gain an intermediate level of explainability on how each feature contributes to the objective.

In the papers reviewed, the XAI method used for interpretability is rarely baselined or justified. Two studies compare two explainability methods, [39] (LIME vs Grad-CAM) and [21] (Bag-of-visual-words-based approach vs Grad-CAM). Only one thorough comparative study was found [17], in which nine different XAI methods, including five post-hoc ones, were evaluated using seven faithfulness metrics. This is, as well, the only study in which explainability methods are assessed through quality metrics such as the Increase in Confidence or the Average Drop [9].

Another limitation in the field is privacy. None of the previous works makes the code (except for [4]) or the data (except for [17]) publicly available, making reproducibility impossible and replicability impractical. These flaws, in a task with a high degree of subjectivity (inter and intra-observer variability) induce a high risk of confirmation bias. This issue could be mitigated through a quantitative benchmark measuring the model's attention to certain embryo regions of interest across a relatively big amount of samples.

That would require an additional segmentation model such as [4], which isolates the TE, ZP, ICM, and BL regions. A good explanation from a good model should overlap with the key segmented areas, and could be assessed with metrics like IoU. This approach is limited to blastocyst images, and subject to the accuracy of the segmentation model.

4. Integrating Explainable Artificial Intelligence in Decision Support Systems

The performance of AI methods on top-of-the-line metrics (*e.g.*, precision, recall, *etc.*) is often the endpoint of most contributions in the embryo assessment field. However, transforming a quantitatively performing model into a successful practical deployment is no easy task, esp. when considering the high-stakes involved in the context of IVF. The question on how to effectively and safely introduce this technology into a clinical setting where the different actors, including clinicians and patients, are involved, remains unsolved. Next we aim at advancing towards a solution, filling a gap in current literature.

4.1. Opaque Models Create Responsibility Gaps

The use of opaque AI models raises ethical and legal accountability concerns [2], particularly when clinicians can not explain the decision-making process. This creates a "responsibility gap", and without established accountability mechanisms it is challenging to determine who should be held responsible for any potential harm. For example, in cases of sub-optimal embryo selection or injury due to model recommendations, the decision-making process must be explainable to patients seeking to understand what happened or, in more extreme cases, seeking compensation for the damage caused. As of today, distrust in AI applications in medicine also comes from doctors' fear of legal repercussions if something goes wrong due to unclear liability regimes [31]. If clinicians base their decision on these opaque AI models, the evaluation of the decision-making process and, consequently, the determination of who is responsible will be greatly hampered.

It is then essential that the explanations provided by the model are as clear and understandable as possible by the clinician. For embryo image analysis this motivates the use of saliency maps obtained through feature attribution methods, which provide reliable (*i.e.*, true to the underlying model behavior) but also accessible evidence. To further ground the produced explanations in the context of clinical embryo selection, we suggest the mapping of this evidence-based support to the established annotation and evaluation guidelines and practices of the field, *e.g.*, through the association of feature attributions with the already established morphokinetic markers (such as a part of an image showing inappropriate cellular fragmentation), so that evidence provided is given in a language appropriate for the clinician and the problem. Still, due to the limitations of these XAI methods, it is recommended to adhere to ethical guidelines in AI, such as the one proposed by the EU [3]. By following this framework, other key requirements such as data privacy and governance or human agency and oversight can be accomplished.

4.2. Building Trust through Explanations

The successful integration of AI-assisted embryo selection systems in the medical context heavily depends on the acceptance by the clinicians and, consequently, their trust in them.

According to [31], there are several levels of trust which fall along a spectrum, ranging from complete distrust to over-reliance on AI systems. Studies have shown that over-reliance on the suggestions can cause clinicians to take less initiative [25] and also to be more likely to accept incorrect diagnosis [19]; this is known as *automation bias* [24]. On the other side of the spectrum lie clinicians which do not trust an algorithm that they do not understand [8], a phenomenon known as *algorithmic aversion*. As a result, not only training and education programs are necessary to ensure that clinicians understand the capabilities and limitations of AI-assisted embryo selection systems, but also making sure that the explanation or the model are not having negative impacts on the clinician's decision making process.

It is recommended to avoid offering the discretized outcome of models (*e.g.*, this embryo is good) in favour of more detailed information, such as outcome probabilities (*e.g.*, this embryo is good with a probability of 55%). For further bias avoidance, evidence may be presented only in the form of explanations, with the outcome hidden (*e.g.*, this area shows relevant visual features for determining if the embryo is good or not), so that the clinician interprets and integrates the visual features unbiasedly *w.r.t.* to the model's final prediction. Unfortunately, this approach is not without flaws either. It can induce *confirmation bias* (*i.e.*, a clinician may only review the evidence that supports its own hypothesis), while preventing the methodology to exploit visual patterns not perceivable by humans, which may be a limiting factor in embryo analysis. The proposed solution is to provide all information available, outcome probabilities and saliency maps, and to extend that information to all relevant cases (*e.g.*, an area shows relevant visual features in favour of the embryo being good, an area shows features in favour of it being bad).

4.3. Patient, Clinician, AI

IVF is a highly personal procedure, with a large impact in the physical and mental health of individuals. Patients undergoing IVF have the right to know how decisions are being made about their own healthcare, and to maintain their decision-making power. To guarantee that, AI explanations should be also accessible to them; they cannot be left behind by the complexity of technology. Integrating explainability into AI-assisted embryo selection systems can not only target clinicians, it must also empower patients. By providing them with clear and easy to understand explanations of certain decisions, patients can be more involved and better equipped for the decision-making process regarding their own care. XAI results should be presented when relevant, and under the supervision of the clinician for its adequate interpretation and contextualization. Finally, it has to be mentioned that, since the use of these systems directly involves the processing of their data, the patient should be educated about how it will be processed and the use that will be made of it, in addition to asking for their informed consent.

5. Conclusions

We present an analysis of the presence of XAI for embryo selection on the literature. While there is a relative abundance of models for embryo selection and many have shown promising results when trained on retrospective data, the performance of these models in actual cases is yet unknown and their integration in the medical decision process remains

unclear. Due to the opaque nature of these models, explainability is crucial to ensure their applicability. However, although some studies tackle the issue of explainability, the interpretations offered are often insufficient or do not rely on medical experts for evaluation. We enumerate a set of guidelines and suggestions that could help increase the interpretability and trustworthiness of these systems, with the goal of advancing towards the successful and safe integration of this technology in clinical practice.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no.957331 - knowlEdge.

References

[1] G. David Adamson et al. "International Committee for Monitoring Assisted Reproductive Technology: world report on assisted reproductive technology, 2011". In: *Fertility and Sterility* 110.6 (2018), pp. 1067–1080.

[2] Michael Anis Mihdi Afnan et al. "Ethical Implementation of Artificial Intelligence to Select Embryos in In Vitro Fertilization". In: AIES '21. New York, NY, USA, 2021, pp. 316–326.

[3] Pekka Ala-Pietilä et al. *Ethics Guidelines for Trustworthy AI*. en. Text. Dec. 2018.

[4] Muhammad Arsalan et al. "Human Blastocyst Components Detection Using Multiscale Aggregation Semantic Segmentation Network for Embryonic Analysis". In: *Biomedicines* 10.7 (2022), p. 1717.

[5] Nishanth Arun et al. "Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging". In: *Radiology: Artificial Intelligence* 3.6 (2021), e200267.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. "SURF: Speeded Up Robust Features". In: *Computer Vision – ECCV 2006*. Ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz. Springer Berlin Heidelberg, 2006, pp. 404–417.

[7] Jacky Boivin et al. "International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care". In: *Human reproduction* 22.6 (2007), pp. 1506–1512.

[8] Carrie J Cai et al. "Hello AI: uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making". In: *Proceedings of the ACM on Human-computer Interaction* 3.CSCW (2019), pp. 1–24.

[9] Aditya Chattopadhay et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 839–847.

[10] Sonya M. Diakiw et al. "An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos". In: *Reproductive BioMedicine Online* 45.6 (2022), pp. 1105–1117.

[11] Irene Dimitriadis et al. "Artificial intelligence in the embryology laboratory: a review". In: *Reproductive BioMedicine Online* 44.3 (2022), pp. 435–448.

[12] Noritoshi Enatsu et al. "A novel system based on artificial intelligence for predicting blastocyst viability and visualizing the explanation". In: *Reproductive Medicine and Biology* 21.1 (2022), e12443.

[13] Eleonora Inácio Fernandez et al. "Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data". In: *Journal of Assisted Reproduction and Genetics* 37.10 (2020), pp. 2359–2376.

[14] Hiroshi Fukui et al. "Attention Branch Network: Learning of Attention Mechanism for Visual Explanation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[15] D. K. Gardner and D. Sakkas. "Assessment of embryo viability: the ability to select a single embryo for transfer–a review". In: *Placenta* 24 Suppl B (2003), S5–12.

[16] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health* 3.11 (2021), e745–e750.

[17] Tristan Gomez, Thomas Fréour, and Harold Mouchère. *Comparison of attention models and post-hoc explanation methods for embryo stage identification: a case study*. 2022.

[18] A. C. Gore et al. "EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals". In: *Endocrine Reviews* 36.6 (2015), E1–E150.

[19] Yukinori Harada et al. "Effects of a differential diagnosis list of artificial intelligence on differential diagnoses by physicians: an exploratory analysis of data from a randomized controlled study". In: *International Journal of Environmental Research and Public Health* 18.11 (2021), p. 5562.

[20] Iza Sazanita Isa, Umi Kalsom Yusof, and Murizah Mohd Zain. "Image Processing Approach for Grading IVF Blastocyst: A State-of-the-Art Review and Future Perspective of Deep Learning-Based Models". In: *Applied Sciences* 13.2 (2023).

[21] Athanasios Kallipolitis et al. "Explainable computer vision analysis for embryo selection on blastocyst images". In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2022, pp. 1–4.

[22] Jihyun Kim, Jaewang Lee, and Jin Hyun Jun. "Non-invasive evaluation of embryo quality for the selection of transferable embryos in human in vitro fertilization-embryo transfer". In: *Clinical and Experimental Reproductive Medicine* 49.4 (2022), pp. 225–238.

[23] Evangelos Maziotis Konstantinos Sfakianoudis. "Reporting on the Value of Artificial Intelligence in Predicting the Optimal Embryo for Transfer: A Systematic Review including Data Synthesis". In: *Biomedicines* 10.3 (2022), p. 697.

[24] John D Lee and Katrina A See. "Trust in automation: Designing for appropriate reliance". In: *Human factors* 46.1 (2004), pp. 50–80.

[25] Ariel Levy et al. "Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–13.

[26] Claudio Michael Louis et al. "Review of computer vision application in in vitro fertilization: the application of deep learning-based computer vision technology in the world of IVF". In: *Journal of Assisted Reproduction and Genetics* 38.7 (2021), pp. 1627–1639.

[27] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: NIPS'17 (2017), pp. 4768–4777.

[28] Kersti Lundin and Hannah Park. "Time-lapse technology for embryo culture and selection". In: *Upsala Journal of Medical Sciences* 125.2 (2020), pp. 77–84.

[29] Yasuyuki Mio and Kazuo Maeda. "Time-lapse cinematography of dynamic changes occurring during in vitro development of human embryos". In: *American journal of obstetrics and gynecology* 199.6 (2008), 660–e1.

[30] Nahid Nasiri and Poopak Eftekhari-Yazdi. "An overview of the available methods for morphological scoring of pre-implantation embryos in in vitro fertilization". In: *Cell Journal (Yakhteh)* 16.4 (2015), p. 392.

[31] Cecilia Panigutti et al. "Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–9.

[32] Elena Payá et al. "Automatic characterization of human embryos at day 4 post-insemination from time-lapse imaging using supervised contrastive learning and inductive transfer learning techniques". In: *Computer Methods and Programs in Biomedicine* 221 (2022).

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144.

[34] Vasilija Rolfes et al. "Artificial Intelligence in Reproductive Medicine – An Ethical Perspective". In: *Geburtshilfe und Frauenheilkunde* 83.1 (2023), pp. 106–115.

[35] Zev Rosenwaks. "Artificial intelligence in reproductive medicine: a fleeting concept or the wave of the future?" In: *Fertility and Sterility* 114.5 (2020), p. 905.

[36] Yuki Sawada et al. "Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth". In: *Reproductive Biomedicine Online* 43.5 (2021), pp. 843–852.

[37] Thalia R. Segal and Linda C. Giudice. "Before the beginning: environmental exposures and reproductive and obstetrical outcomes". In: *Fertility and Sterility* 112.4 (2019), pp. 613–621.

[38] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.

[39] Akriti Sharma et al. "Explainable Artificial Intelligence for Human Embryo Cell Cleavage Stages Analysis". In: *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*. 2022, pp. 1–8.

[40] Sivan Tamir. "Artificial intelligence in human reproduction: charting the ethical debate over AI in IVF". In: *AI and Ethics* (2022).

[41] Prudhvi Thirumalaraju et al. "Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality". In: *Heliyon* 7.2 (2021).

[42] Shanshan Wang et al. "A Deep Learning Framework Design for Automatic Blastocyst Evaluation With Multifocal Images". In: *IEEE Access* 9 (2021).

[43] Chongwei Wu et al. "LWMA-Net: Light-weighted morphology attention learning for human embryo grading". In: *Computers in Biology and Medicine* 151 (2022).

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230679

Longitudinal Segmentation of Multiple Sclerosis Lesions Using nnU-Net Architecture

Eloy MARTÍNEZ-HERAS[a,1,*], Adrian VICENTE-GOMEZ[b,*], Francesc VIVÓ[a], Marcos DIAZ-HURTADO[b], Baris KANBER[c], Jordi CASAS-ROMA[b], Sara LLUFRIU[a] and Ferran PRADOS[b,c,d,e,1]

[a] *Neuroimmunology and Multiple Sclerosis Unit, Laboratory of Advanced Imaging in Neuroimmunological Diseases, Hospital Clinic Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) and Universitat de Barcelona, Barcelona, Spain*

[b] *e-Health Center, Universitat Oberta de Catalunya, Barcelona, Spain*

[c] *Centre for Medical Image Computing, University College London, London, United Kingdom*

[d] *Queen Square MS Centre, Department of Neuroinflammation, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, London, United Kingdom*

[e] *National Institute for Health Research Biomedical Research Centre at UCL and UCLH, London, United Kingdom*

Eloy Martínez-Heras - https://orcid.org/0000-0001-9937-3162
Adrián Vicente-Gomez - https://orcid.org/0009-0001-0421-559X
Francesc Vivó - https://orcid.org/0000-0002-6409-1197
Marcos Diaz-Hurtado - https://orcid.org/0000-0003-1528-5873
Baris Kanber: https://orcid.org/0000-0003-2443-8800
Jordi Casas-Roma - https://orcid.org/0000-0002-0617-3303
Sara Llufriu - https://orcid.org/0000-0003-4273-9121
Ferran Prados - https://orcid.org/0000-0002-7872-0142

Abstract. Multiple sclerosis is a chronic autoimmune disease that affects the central nervous systems. The detection of new lesions through conventional magnetic resonance imaging is particularly important in the management of people with multiple sclerosis. The advancements in machine learning technology in recent years have significantly transformed the analysis of medical images for multiple sclerosis, particularly in identifying and segmenting lesions., improving the accuracy and efficiency of this process. In this context, the objective of this work is to develop a system for the detection of new lesions in people with multiple sclerosis using consecutive magnetic resonance scans. The proposed system uses only pre-processed FLAIR images to train a nnU-Net, a type of deep learning architecture that has been proven very successful for image segmentation tasks. The resulting model is able to generate masks that highlight changes between baseline and follow-up images, identifying new lesions that may have appeared in the meantime. The model achieved an average Dice score coefficient

[1] Corresponding Authors: Eloy Martínez-Heras, EMARTIND@recerca.clinic.cat and Ferran Prados, fpradosc@uoc.edu.
* joint first authorship

of 0.58 on the evaluation set. Overall, this work demonstrates the potential of machine learning tools for improving the detection and monitoring of multiple sclerosis lesions in clinical practice, particularly in the context of longitudinal studies.

Keywords. Longitudinal, segmentation, lesions, multiple sclerosis, MRI

1. Introduction

Multiple sclerosis (MS) is a complex neurological disease that affects millions of people worldwide. Assessing the presence of new MS lesions and evidence of new disease activity is crucial for evaluating the efficiency of disease-modifying therapies [1,2]. Conventional magnetic resonance imaging (MRI) serves as an essential hallmark for MS progression, and an indispensable tool for diagnosing, monitoring disease activity, and gauging treatment response [3,4]. Despite its importance, MRI techniques have several limitations, including low specificity and sensitivity for assessing focal tissue damage, noise, and the presence of artifacts [5]. Despite this, longitudinal lesion segmentation is a research field with increasing interest. Recent advances in machine learning and the organization of several challenges in the field over the last decade, have eased the appearance of deep learning based solutions for detecting new lesions [6].

Although the task of detecting new lesions may seem straightforward, assessing the results can involve different considerations. Unlike other segmentation tasks like tissue segmentation or parcellation, the areas to detect new lesions can vary in size, with the majority being small or very small, and can appear in different locations of the brain tissue, predominantly in the white matter but also in the gray matter. Detecting the absence of new lesions is crucial from both a methodological and clinical perspective. While the appearance of new lesions is the primary sign of disease progression in MS, a stable MS patient may not have any new lesions. Therefore, if an algorithm correctly identifies the absence of new lesions in people with MS who have not developed new lesions, it should be considered a good segmentation, even if most segmentation metrics evaluate it as 0. As a result, the effectiveness of a new lesion segmentation technique should be evaluated based on its ability to detect new lesions and its ability to prevent over-segmentation in the absence of new lesions.

Manual delineation, while considered the gold standard, provides higher reliability in detecting enlarging lesions and new lesions in regions of accumulated damage, such as periventricular areas. However, it is time-consuming and subject to inter-observer variability, which can significantly impact the accuracy of the lesion assessment [7,8]. In response to these challenges, automatic lesion segmentation has emerged as a promising alternative, It offers more consistent segmentations, reproducible results and improved processing speed. The MS research community would greatly benefit from a common dataset for accurate estimation and comparison of proposed methods. Publicly available datasets, such as those released by ISBI 2015 [9] and MSSEG-2 challenges, provide a limited but valuable resource [10]. The recent MSSEG-2 challenge, which focuses solely on new lesions, has been another vital step for benchmarking the current methods on the field towards the potential adoption of these methods in clinical practice.

In recent years, techniques for longitudinal MS lesion segmentation have evolved significantly [6,9,10]. From image subtraction and thresholding to Bayesian

generative models and deep neural network-based methods, such as convolutional neural networks (CNN), substantial progress has been made in this field. The nnU-Net is a powerful and flexible 3D CNN that has demonstrated remarkable performance in various medical imaging tasks [11]. Its benefits include robustness against data class imbalance, which is a common issue in MS segmentation, and the ability to adapt to different datasets with minimal modifications. We hypothesize that the utilization of nnU-Net will significantly enhance the detection of new MS lesions in longitudinal MRI studies.

2. Method

In this section we introduce the used dataset, the data preprocessing strategy and the implemented method for segmenting longitudinal MS lesions.

2.1. Dataset

The dataset used for the training model has been obtained by combining two different open datasets: MSSEG-2 (https://portal.fli-iam.irisa.fr/msseg-2/data/) and Open MS Data (https://github.com/muschellij2/open_ms_data). Respectively, each of the datasets contributes with 40 and 20 distinct patients, each with their baseline and follow-up FLAIR images (ranging from 80 to 1000 days apart between scans), as well as the mask for new lesions. Therefore, there will be a total of 60 patients. It is important to note that 11 of them belonging to the MSSEG-2 dataset did not have any new lesions, so their mask for new lesions is blank. This aspect should be considered when elaborating the model of training with this dataset.

2.2. Preprocessing

MRIs are taken under different acquisition protocols, and although efforts are made to replicate the same parameters of the initial acquisition (baseline) when obtaining follow-up images, it is necessary to standardize certain characteristics among the different images of the dataset involved. The following preprocessing steps ensure that the model being trained is based on more homogeneous MRI characteristics across timepoints, resulting in a more reliable final model: orientation to MNI, skull stripping and intensity inhomogeneity correction (see Figure 1).

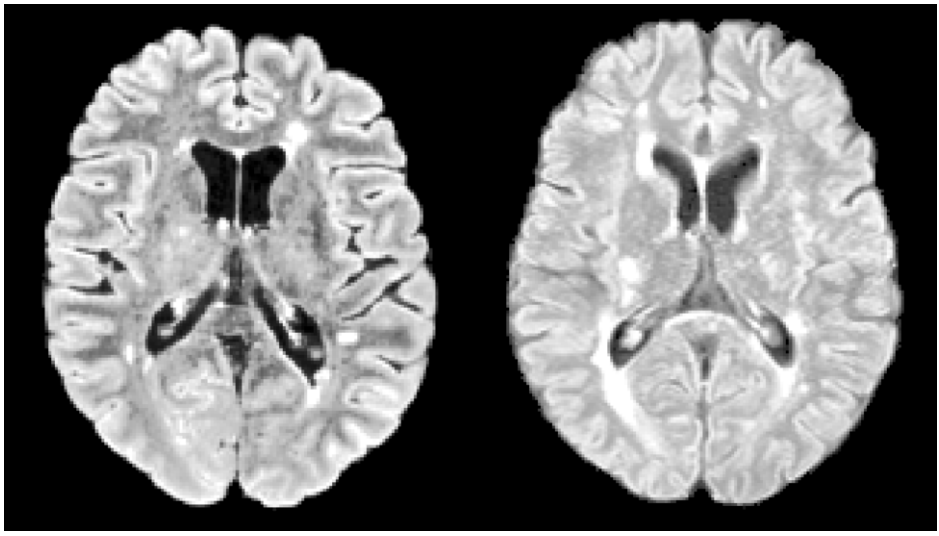


Figure 1. Example of the MSSEG-2 (left) and the Open MS Data (right) images.

A consistent orientation has been established for all images in the final dataset. Specifically, the Montreal Neurological Institute (MNI) coordinate system ("Neurological") is used. This registration and reorientation phase entails aligning the FLAIR images with the MNI template. This procedure uses a 6-degree-of-freedom (6 DOF) rigid registration transformation to ensure consistency across the dataset and improve the learning process. After orienting the image, we proceeded with the skull removal from the FLAIR images. This was carried out using the HD-BET algorithm [12]. The result is crucial for accurate subsequent analyses and model training avoiding the appearance of spurious results outside the white matter and gray matter brain tissues. Finally, we correct the image intensities for inhomogeneities due to coil uniformities, field strength or other biases [13]. For bias inhomogeneity correction, we utilized the N4 algorithm [14] to achieve a more uniform intensity of the whole FLAIR images of the dataset. Before starting the training process, images were cropped to include only the brain area. A z-score intensity normalization with mean 0 and standard deviation 1 was applied to each subject to regularize the intensity values.

2.3. nnU-Net architecture

The nnU-Net (no-new-Net) is a framework proposed by Insesee et al. [11] that proposes an architecture based on the U-Net (see Figure 2), which is able to adapt to the proposed dataset since up to now each case of use forced to specialize the architecture of the network. Baseline and a follow-up FLAIR scan were used as model inputs. The model output is a binary 3D prediction mask of new lesions in the follow-up scan. The nnU-Net architecture includes a contracting and an expanding path. Each resolution level in the contracting path contains two convolutional layers with a $3 \times 3 \times 3$ convolution kernel, followed by instance normalization and a leaky ReLU (rectified linear unit) activation, similar to ReLU activations used in the original architecture but with negative slope with value 0.01 which allows a more stable network training and improved model performance.

At each level of the contracting and expanding paths a combination of convolutional layer, stride and activations were used (see Figure 2). By incorporating skip connections, features extracted from the contracting path are concatenated with features obtained from the expanding path at each respective resolution level.

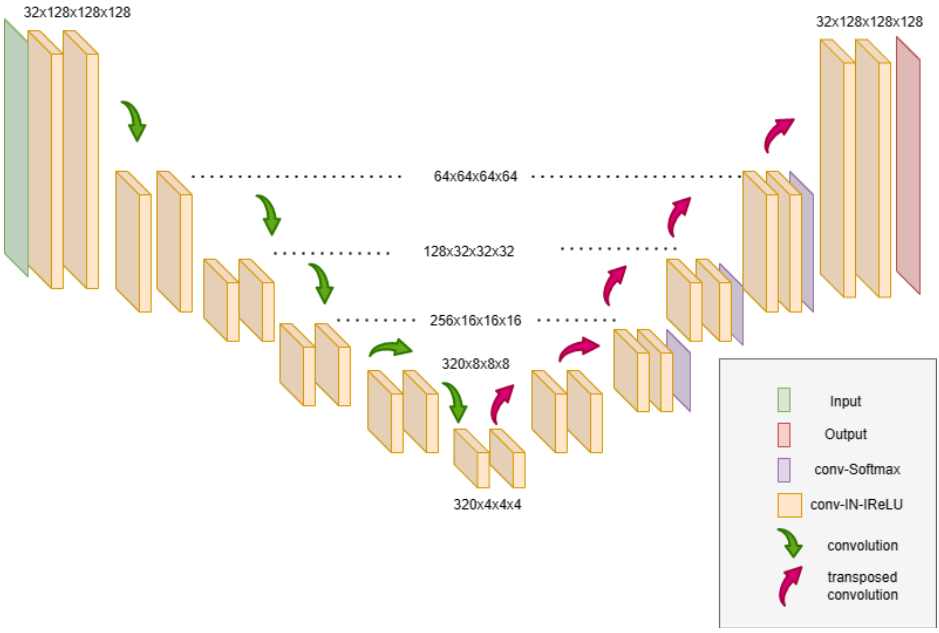


Figure 2. Schema of our nnU-Net architecture.

2.4. Hyperparameters

We used the Adam optimizer with an initial learning rate of 0.01, defining an epoch as iteration over 250 training batches, if the value of training and validation losses is not reduced by at least 0.005 on passing 30 epochs, the value of the learning rate is reduced by a factor of 5. On the other hand, no early stopping criterion was used, but a maximum threshold of 300 epochs was established, carrying out a checkpoint to be able to resume training in case of equipment failure every 25 epochs.

2.5. Training

The Dice score was selected as the evaluation metric for training progress. Training was carried out using a 5-fold cross-validation technique, which offers a dependable assessment of the model's capabilities while maintaining manageable computational requirements. In this process, for each fold, a random selection of 80% (48) was allocated for model training, and the remaining 20% (12) were designated for validation. Furthermore, an inherent feature of nnU-Net's default architecture is its use of data augmentation to enrich our training data and enhance the model's ability to generalize.

2.6. Evaluation metrics

Two classical evaluation metrics have been used for assessing the goodness of the proposed solution: Dice score coefficient and recall. Mean and standard deviation have been computed for each metric and dataset. If no lesions were present in the ground truth, and no lesions are detected by the method, it has been assumed the best result, for example 1 in Dice score coefficient instead of 0.

3. Results

The nnU-Net model achieved a good performance in terms of Dice score coefficient in the training process, with the measure ranging between 0.3 and 0.6. The average value for the final model of each k-fold is presented in Table 1.

Table 1. Average Dice score coefficients obtained for the final model of each k-fold.

k-fold	Dice score
0	0.48
1	0.37
2	0.52
3	0.50
4	0.49

The testing set consisted of 12 MR FLAIR images, out of which 3 did not have new lesions. The evaluation metrics obtained for the test set are shown in Table 2.

Table 2. Individual results, mean and standard deviation of the evaluation metrics for the test set.

Image	Dice score	Recall
MS_00001	0.37	0.75
MS_00006	1.00	1.00
MS_00007	0.00	0.00
MS_00010	0.81	0.73
MS_00014	0.00	0.00
MS_00017	1.00	1.00
MS_00019	0.50	0.39
MS_00021	0.71	0.63
MS_00028	0.57	0.46
MS_00042	0.65	0.74
MS_00047	0.38	0.28
MS_00056	1.00	1.00
All	**0.58±0.35**	**0.62±0.36**

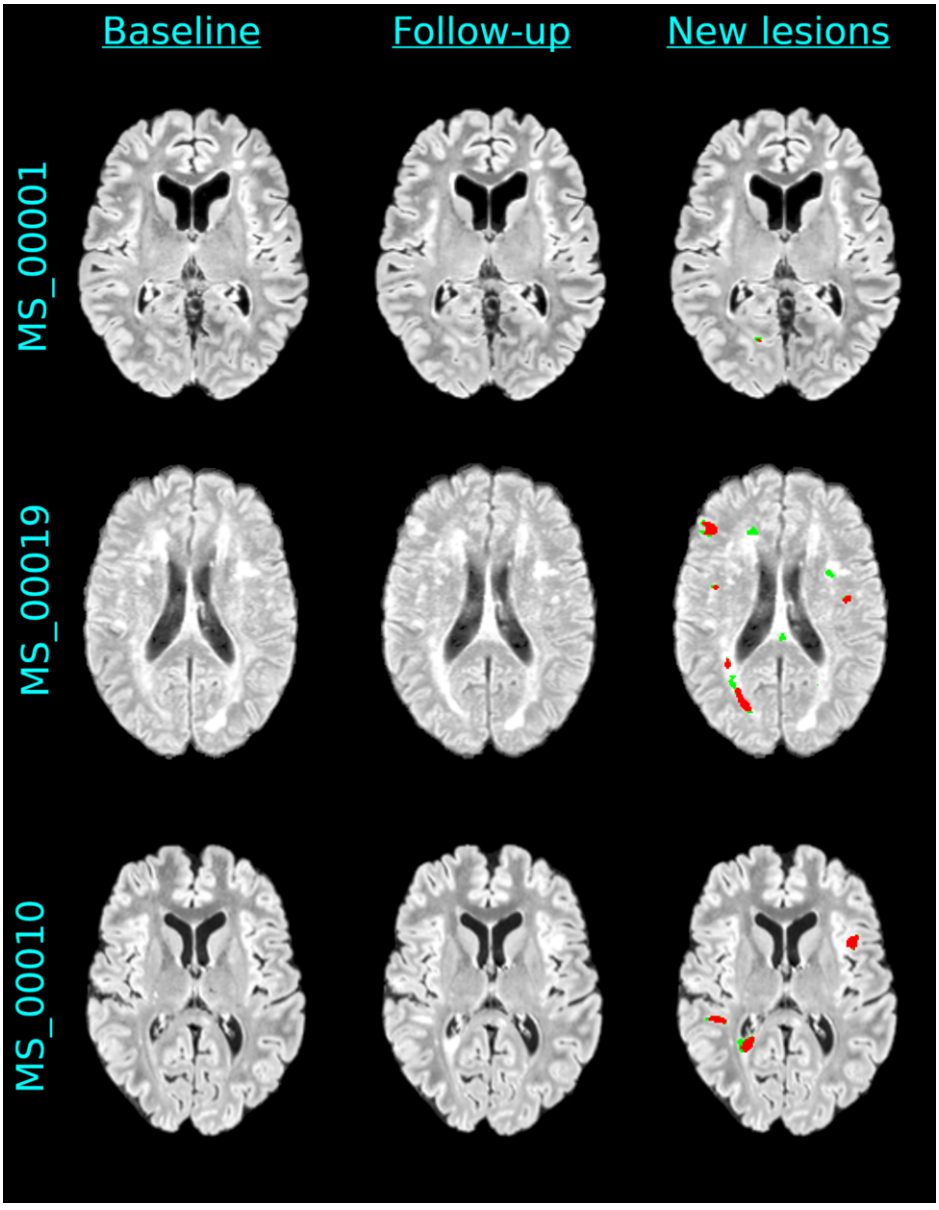


Figure 3. The mask produced by the model is shown in green, while the expert-segmented area is represented in red. The figure displays three distinct examples of people withMS, with Dice score coefficients of 0.37, 0.50, and 0.81, respectively.

4. Discussion

The primary aim of this work was to develop a model that can accurately detect new MS lesions in MRI scans taken at different timepoints. By comparing our

results with those obtained in MSSEG-2 Challenge, we observed that the model generated achieved an average Dice score coefficient of 0.58, which reflected a good performance compared to the majority of models proposed in MSSEG-2. Our work presents a self-customisation of a publicly available nnU-Net architecture [11]. The presented method uses the same architecture as Basaran et al. [15] but there are important implementation differences that need to be highlighted. Unlike them, we changed the preprocessing steps order by first registering to the MNI space, followed by skull stripping, and finally applying bias field correction [15]. The reason for first performing registration followed by skull stripping, is because we wanted to take advantage of the skull as an immutable part of the head for getting an accurate alignment between the two timepoints. We also normalize the intensity of the images using a z-score to minimize differences between images coming from different scanners. Moreover, during the training process, we used Dice score coefficient only for assessing the goodness of the results and the hyperparameter optimization follows two different strategies, Adam vs Nesterov. We also used different model hyperparameters to adapt the training to our more limited hardware.

While the model exhibited proficiency in identifying new lesions, it faced difficulties in accurately identifying the smallest lesions as exemplified by the MS_00001 case in Figure 3. This case features small new lesions and a suboptimal Dice score coefficient of 0.37 in comparison to other instances. The training process emphasized optimizing the Dice score coefficient value, which led to better detection of larger lesions at the expense of smaller ones. This is because larger lesions have a more substantial impact on this similarity coefficient. Despite this drawback, the selected model exhibited a relatively high success rate in identifying larger lesions, as demonstrated in Figure 3. In the accuracy values of Table 2 reveal solid performance, with a 62% correct identification rate for voxels with lesions. While the model's performance may seem moderate, it's crucial to remember that it acts as a diagnostic support tool, requiring expert validation before making final decisions. By effectively directing physicians to potential new lesions, the model substantially reduces the time needed to analyze the entire brain volume.

This work has certain limitations that must be acknowledged. Firstly, the method was developed using a reduced dataset and will require further testing with real clinical data and larger datasets to assess its overall effectiveness more thoroughly. We used the Dice score coefficient alone for assessing the training progress, it would be good to explore other metrics.

5. Conclusions

This work presents a nnU-Net model capable of detecting new lesions between MRIs performed on the same subject followed up for MS, where both images were taken months or years apart. The presented solution achieves good performance using publicly available data. Future work should explore the performance of the proposed method using a larger dataset from a clinical setting where scans can be obtained from different scanner manufacturers and with variable acquisition parameters.

Code availability

The code derived from this publication is publicly available at: https://github.com/ADaS-Lab/nnUnet-LongMS/. In order to ease the accessibility of nnU-Net, we have developed a Docker container that simplifies the execution of the models on new data. This container includes all the required dependencies, libraries, and scripts for running the pre-trained models. By using this container, users can easily deploy the models using their own data without having to worry about installing additional software or configuring complex environments.

Data availability

The data described in this manuscript are in the public domain at MSSEG2 (https://portal.fli-iam.irisa.fr/msseg-2/data/) and Open MS Data (https://github.com/muschellij2/open_ms_data) websites.

Acknowledgments

This work was partially funded by the Instituto Carlos III (ISCIII), co-funded by the European Union through the Plan Estatal de Investigación Científica y Técnica y de Innovación 2015-2024 (PI15/00587, PI18/01030 and PI21/01189 to SL) and National Institute for Health Research Biomedical Research Centre at UCL and UCLH.

References

[1] S.E. Hughes, and G. Macaron, Fast Facts: Multiple Sclerosis: 5th edition. A new era of disease modification and treatment, Karger Medical and Scientific Publishers, 2021.

[2] H.L. Weiner, and J.M. Stankiewicz, Multiple Sclerosis: Diagnosis and Therapy, John Wiley & Sons, 2012.

[3] M.P. Wattjes, O. Ciccarelli, D.S. Reich, B. Banwell, N. de Stefano, C. Enzinger, F. Fazekas, M. Filippi, J. Frederiksen, C. Gasperini, Y. Hacohen, L. Kappos, D.K.B. Li, K. Mankad, X. Montalban, S.D. Newsome, J. Oh, J. Palace, M.A. Rocca, J. Sastre-Garriga, M. Tintoré, A. Traboulsee, H. Vrenken, T. Yousry, F. Barkhof, À. Rovira, Magnetic Resonance Imaging in Multiple Sclerosis study group, Consortium of Multiple Sclerosis Centres, and North American Imaging in Multiple Sclerosis Cooperative MRI guidelines working group, 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis, *Lancet Neurol.* **20** (2021) 653–670.

[4] A.J. Thompson, B.L. Banwell, F. Barkhof, W.M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M.S. Freedman, K. Fujihara, S.L. Galetta, H.P. Hartung, L. Kappos, F.D. Lublin, R.A. Marrie, A.E. Miller, D.H. Miller, X. Montalban, E.M. Mowry, P.S. Sorensen, M. Tintoré, A.L. Traboulsee, M. Trojano, B.M.J. Uitdehaag, S. Vukusic, E. Waubant, B.G. Weinshenker, S.C. Reingold, and J.A. Cohen, Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria, *Lancet Neurol.* **17** (2018) 162–173.

[5] R. Bakshi, A.J. Thompson, M.A. Rocca, D. Pelletier, V. Dousset, F. Barkhof, M. Inglese, C.R.G. Guttmann, M.A. Horsfield, and M. Filippi, MRI in multiple sclerosis: current status and future prospects, *Lancet Neurol.* **7** (2008) 615–625.

[6] M. Diaz-Hurtado, E. Martínez-Heras, E. Solana, J. Casas-Roma, S. Llufriu, B. Kanber, and F. Prados, Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review, *Neuroradiology*. **64** (2022) 2103–2117.

[7] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S.C. Pop, P. Girard, R. Améli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight,

A. Khademi, A. Mahbod, C. Wang, R. McKinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Lladó, M.M. Santos, W.P. Santos, A.G. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F.J. Vera-Olmos, N. Malpica, C. Guttmann, S. Vukusic, G. Edan, M. Dojat, M. Styner, S.K. Warfield, F. Cotton, and C. Barillot, Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure, *Sci. Rep.* **8** (2018) 13650.

[8] O. Commowick, M. Kain, R. Casey, R. Ameli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, S. Camarasu-Pop, T. Glatard, S. Vukusic, G. Edan, C. Barillot, M. Dojat, and F. Cotton, Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset, *Neuroimage*. **244** (2021) 118589.

[9] A. Carass, S. Roy, A. Jog, J.L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C.H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C.A.M. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S.K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L.O. Iheme, D. Unay, S. Jain, D.M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P.A. Calabresi, C.M. Crainiceanu, L.M. Ellingsen, D.S. Reich, J.L. Prince, and D.L. Pham, Longitudinal multiple sclerosis lesion segmentation: Resource and challenge, *Neuroimage*. **148** (2017) 77–102.

[10] O. Commowick, B. Combès, F. Cervenansky, and M. Dojat, Automatic methods for multiple sclerosis new lesions detection and segmentation, Frontiers Media SA, 2023.

[11] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, and K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, *Nat. Methods*. **18** (2021) 203–211.

[12] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K.H. Maier-Hein, and P. Kickingereder, Automated brain extraction of multisequence MRI using artificial neural networks, *Hum. Brain Mapp.* **40** (2019) 4952–4964.

[13] M. Ganzetti, N. Wenderoth, and D. Mantini, Intensity Inhomogeneity Correction of Structural MR Images: A Data-Driven Approach to Define Input Algorithm Parameters, *Front. Neuroinform.* **10** (2016) 10.

[14] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, and J.C. Gee, N4ITK: improved N3 bias correction, *IEEE Trans. Med. Imaging*. **29** (2010) 1310–1320.

[15] B.D. Basaran, P.M. Matthews, and W. Bai, New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation, *Front. Neurosci.* **16** (2022) 1007453.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230680

Resource Allocation in Home Care Services Using Reinforcement Learning

Bernat SALBANYA [a,1] Irene UNCETA [a] and Jordi NIN [a]
[a] *Universitat Ramon Llull, ESADE*

Abstract.
This paper proposes an innovative method based on reinforcement learning techniques to optimally allocates home care resources in controlled urban environments. We demonstrate the value of this approach when compared with traditional allocation heuristics by simulating a real-life scenario in the city of Barcelona. Our results show that the proposed tool can help home care services accommodate the increase in demand by optimizing human and time resources through more effective planning. In the long term, this will improve the overall quality of this service and positively affect the well-being of the elderly.

Keywords. Healthcare, Allocation, Reinforcement Learning, Urban Services

1. Introduction

A critical societal challenge in large urban areas is improving personal care, attention, and quality of life for elderly people. Today, 20% of the European population is aged 65 years or more. By 2100 older people will account for 31.3% of the EU's population. As the population ages and the amount of people requiring assistance grows, so will the demand for public services that assist them during the final years of their life. This increase in demand for assistance comes together with a declared preference to remain at home [1]. Hence, public efforts are being channeled toward improving home care services that offer assistance with personal and instrumental activities of daily living such as meal preparation, housekeeping, medication reminders, or laundry. This will inevitably put tremendous pressure on the home care system, which has historically struggled with the need to provide high-quality assistance with very limited resources.

Caregivers have a long-standing record of precarious employment [2]. They work for long hours and often for a small salary. While their time is precious, it is often wasted due to the system's inefficacy and the need to respond to emergency requests. They attend to many people on an everyday basis, a situation which results in long commuting times and which prevents the establishment of trustful relationships between caregiver and care recipient. The quality of home care services is directly related to the job conditions of those who work in this sector. Hence, any proposal aimed at improving a city's home care services should, therefore, not forget the well-being of caregivers.

[1] Corresponding Author:
Bernat Salbanyà <bernat.salbanya@esade.edu>, Av. Torre Blanca, 59, 08172, Sant Cugat del Vallès, Spain

In light of this, many European cities, including Barcelona [3], have started to promote a new model for delivering home care services based on the notion of a "distributed or virtual nursing home" [4]. Here, the home of a dependent elderly receives all the services of a room in a nursing center. At the same time, the neighborhood provides all the typical services that a residential care home would otherwise offer. Tasks are distributed among so-called "social urban sectors" [5], each managed by a team of caregivers who establish a close relationship with the people they attend. This model has many advantages for the elderly, who can stay home, maintain their social contacts and enjoy a more peaceful way of living [6]. It also has several advantages for caregivers, who are offered a more stable working environment and the possibility to self-organize daily tasks in small agile teams who attend a controlled group of elderly people [7]. However, this system still displays allocation problems for those specialized workers, such as psychologists or physical therapists, who collaborate with several teams simultaneously. As a result, it can inadvertently increase stress for individuals forced to commute to different places to perform their tasks. Moreover, if not designed to be sufficiently flexible, this system can also introduce new inefficiencies by focusing exclusively on the reality of each separate urban sector and overlooking the bigger picture. This may create differences in the quality of service and the work conditions of caregivers across different areas. Matching home care resources to needs, therefore, remains a significant challenge [8].

Recent work has demonstrated the potential of reinforcement learning techniques for supporting intelligent healthcare applications, for example, in patient-to-physician communication [9]. Unlike classical optimization techniques, RL algorithms can adapt and learn from experience to effectively navigate uncertain and ever-changing scenarios. This learning paradigm allows the algorithm to explore different strategies and improve its performance through trial and error, a capability not readily available in traditional optimization approaches. In the home health care field, RL has been shown to improve the results of previous planning schemes by providing an efficient framework to handle constraints [10]. In this article, we further exploit this approach by proposing a framework for daily optimal resource allocation in-home care services. To our knowledge, no such framework exists today. We experiment with reinforcement learning techniques in an environment modeled as a Markov decision process. In collaboration with a local home care services company, we have designed the concrete specifications of the environment and the algorithms to model the current home care system in Barcelona. Our experiments yield promising results for optimal resource allocation systems in-home care.

2. Related work

In reinforcement learning, the *agent* interacts with its *environment* through *actions* to explore the *state space*. The goal is defined by the *reward function*, which maps the state-action pair to a reward indicating the desirability of that *state*. The agent's behavior is governed by a *policy*, which maps available states to allowed actions. The agent takes action, updates its state, and observes outcomes based on the policy. Through trial and error, the agent discovers actions that yield the most reward under the given context, and learning is mediated by experience. The information gathered at each state informs the time-delayed reward function, and the agent's objective is to maximize this function by the end of the learning process [11].

2.1. Q-learning methods

The most straightforward approach to learning in this context is Q-learning. Essentially, Q-learning attempts to learn the value of each action in each state independently of one another by assigning a so-called q-value to each state-action pair. During the learning process, the q-values are iteratively updated to approximate the optimal values $q_*(s,a)$. The optimal q_*-value for any state-action pair (s,a) can be defined to be the expected value of the immediate reward gained from taking action a in a state s plus the highest discounted (with the discount rate γ) optimal q_*-value that is possible in the next state s'. Optimal q-values are unknown. We estimate them using the temporal difference (TD) learning method [12]. At each time step, we update the current q-estimate as follows:

$$q^{new}(s,a) = q(s,a) + \alpha\delta_t \tag{1}$$

where α indicates the learning rate and δ_t the temporal dependent error. This error can be interpreted as a measure of how close the q-estimate of a state-action pair (s,a) is before and after incorporating the environment feedback from performing this action. It is usually computed by the expected stochastic error between two consecutive steps.

The performance of Q-learning inherently relies on updating the q-value estimated numerous times. The number of q-values equals $|S| \times |A|$, for S, the state space that contains all the possible states of the environment, and A, the action space that contains all the possible transformations. It is easy to see that the problem quickly becomes unfeasible for large spaces. This is often called 'the curse of dimensionality [13]. Deep Q-learning (DQN) techniques compute q-estimates using artificial neural networks to mitigate this issue. DQN estimates the hidden parameters θ_i that map a one-hot encoded [2] state vector $(s_1, s_2, \ldots, s_m)$ to the corresponding q-estimates for its actions defined as $(q_*(s,a_1),\ q_*(s,a_2), \ldots, q_*(s,a_n))$. In this configuration, we only need to estimate the hidden parameters $\theta_{h,i}$ of the neural network.

2.2. Policy gradient algorithms

Q-learning methods optimize a value function for state-action pairs and indirectly derive a policy based on those value estimates. Instead, policy gradient algorithms directly learn a policy π by parameterizing it as a neural network $\pi(\theta)$. This network is trained to output the probability of taking action a in a state s using the following equation [14]:

$$\pi_\theta(s,a) = \mathbb{P}[a|s;\theta] \tag{2}$$

The main advantage of policy optimization methods is that they directly search for an optimal action-selection policy. They are also more stable/reliable and better suited for various situations. Even so, Q-learning methods often outperform multiple-agent policy optimization methods [15]. One needs to adjust the policy optimization network weights to produce action probabilities.

[2] A one-hot encoded state vector of state s_i refers to a sparse vector of m features where each element j equals 0 if $j \neq i$ and 1 if $j = i$.

There are more complex reinforcement learning methods. However, as we will see when working with small urban sectors, it is not required to use complex techniques to obtain satisfactory results. Having described the main ideas and methods behind reinforcement learning, in what follows, we discuss how the home care resource allocation problem can be modeled under this framework. For this purpose, we formalize the constituent elements of our Markov decision process as a reinforcement learning problem.

3. Methodology

We propose a unique system of multiple agents trained centrally. To this aim, we model this system as a Markov decision process, which is computationally simple and closely related to reinforcement learning [16]. To pose our reinforcement learning problem, we identify the caregivers, states, and allowed actions as follows:

- **Caregivers:** Caregivers (g) have a working schedule and can be assigned several services during their available time slots.
- **Services:** Services (r) represent an elderly's need for a specific home care service at a given time.
- **Caregiver-Service Pairs:** A given caregiver-service pair is said to be 'allowed' if a specific condition is met: the caregiver needs to have the appropriate qualifications for the service type and needs to be available for the entire duration of the service.
- **The State Space:** A state s_i is a 'complete allocation' of caregivers to services so that every caregiver-service pairing is allowed under the abovementioned constraints. In turn, the state space S is the set of all possible states s_i. The state is represented as a matrix where rows represent caregivers and columns represent services. The elements of this matrix indicate whether there is service allocation and the corresponding working hours.
- **State Transitions:** A state s_t provides a description of the system at a given time t. A state transition is moving from one state s_t to the next state s_{t+1}.
- **The Action Space:** An action a is a 'free' choice to move from a state s_t to a new state s_{t+1}. Therefore, the action space a_{t+1} is the set of all transformations from state s_t to another legal state s_{t+1}. We can therefore define an action a_{t+1} as:

$$a_{t+1} : A_t \to A_{t+1}; A_t \mapsto a_{t+1}(s_t) \tag{3}$$

$$s_{t+1} = a_{t+1}(s_t), \text{ where } \begin{cases} dom(a_{t+1}) = A_t \\ ran(a_{t+1}) = A_{t+1} \end{cases} \tag{4}$$

  for dom the domain of the function and ran the range of the function.
- **Reward Function:** The reward function represents the feedback from the environment every time a caregiver acts. In our case, the reward function can be represented in terms of the ingredients introduced above as:

$$r : S \times S \to R;\ S \times S \mapsto r(s_t,\ s_{t-1}) \tag{5}$$

Essentially, our reward function consists of two types of elements:

* The rewards happen every time step. The proposed reward schema includes two scores: (a) the *similarity score* quantifying the similarity among caregivers and elderly profiles in the given pairs; (b) the *consistency score* accounting for whether caregiver-service pairings are maintained from state to state.
* A delayed penalty for the net overtime or idle time happening at the end of each month. This delayed (or sparse) reward poses a significant problem for the reinforcement learning algorithm since it is unclear how to attribute such a delayed reward over the sequence of actions that lead to it [17].

- **Policy:** The policy π in a RL setting refers to the strategy based on which the agent decides which action to perform given the caregivers' state.

4. Barcelona use case

Barcelona aspires to become a benchmark for person-centered care through the Barcelona 2016-2020 Health Plan [3]. This plan reformulates how home care services are provided and includes a novel care model tailored to the city's complex needs [18]. The "Programa d'Atenció Domiciliària" (PADES) is a home health care and personal assistance program that allocates resources to the different organizational areas or social urban sectors, aligned with the 2016-2020 Health Plan. In each sector, a reduced team of caregivers assists dependent people with their everyday activities.

Determining how resources are allocated at the different levels is a complex task of meeting the care demands of service users, with the local conditions and the availability and circumstances of caregivers, altogether ensuring a cost-effective system [19]. Given the impact of optimal allocation on the quality of the service, substantial literature has focused on understanding the decision-making process in different countries. Most papers have studied this problem from a macro perspective. They have examined, for example, how central authorities distribute work among local providing units, how funds are allocated to these units, and how allocation patterns influence the outcome [20].

A matter that has received less attention is understanding the mechanisms for allocating resources in smaller, controlled environments, such as that of social urban sectors [21]. Social urban sector managers work under a fixed budget, manage a regular team of caregivers and provide service to a stable population of elderly people. Here, we study this problem from a reinforcement learning perspective. The main challenge is accounting for the inherent dynamics of the system inside the rewarding function without overlooking service quality and customer satisfaction.

To this aim, we have consulted a private home care company that offers service in different organizational areas throughout the city. Individual interviews with caregivers working at the mentioned company reveal that the favorability of resource allocation in caregiver-elderly systems in this context depends essentially on three main factors: similarity, consistency, and cost. *Similarity* refers to how similar the profiles of the caregiver and the elderly are. *Consistency* is a crucial factor. If an elderly consistently gets assigned the same caregiver, they get the opportunity to build a personal relationship. Finally, in any business activity, *cost* plays an important role. Considering caregivers' patient preferences and dislikes is essential when designing optimal home care resource allocation

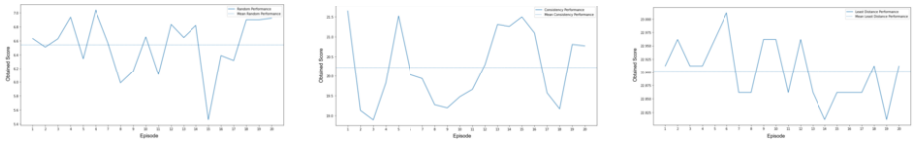
systems. According to the above, favorable allocations are defined as those that (1) contain pairs with highly similar people, (2) contain pairs that consistently reappear, and (3) are cost-effective. Besides the favorability aspect, additional constraints must be considered when pairing caregivers and elderlies, such as the caregiver's availability during specific time slots to prepare lunch. An optimal resource allocation framework should also consider the system's dynamics. Caregivers' and elderly people's needs and situations can change daily due to unexpected events. Due to this, we use a Markov decision process to model the constituent elements of the described scenario.

5. Experiments

This section presents the experiments to validate the proposed methodology and analyze its performance. The setting was designed with a home care services company in the Barcelona metropolitan area. This setting included 15 caregivers and 6 services per caregiver. To create the reinforcement learning environment, discussions with the company were held to understand the real-life situation. The configuration for the environment includes a time step representing one day, with each episode lasting two months (60 steps). Active caregivers become inactive, and inactive caregivers become active at the end of every time step with a probability of 0.01. The final reward is calculated at the end of each time step as a weighted sum with the following weights: 20% similarity score, 30% consistency score, 30% incremental overtime/idle score, and 20% end-of-month overtime/idle score, grounded on qualitative business-wise judgment. The Q-learning algorithms will be trained with a ε-greedy policy and tested with a Q-greedy policy, and their performance will be tracked throughout 20 episodes. The rewards generated by the environment and accumulated by the algorithms are obtained by computing the weighted average of unit-less scalars.

5.1. Benchmark heuristics

Three benchmarks assuming different knowledge levels were considered to evaluate the proposed methodology's performance improvement. The first benchmark assumed zero knowledge, where the system took random actions at every step from the set of all potential legal actions. The second benchmark defined a consistent state heuristic where the system could remember and replicate the previous action to maximize the state consistency score. The third benchmark was the minimum distance heuristic, where the system knew the distances among services and used this knowledge to minimize caregivers' commuting distance. Figure 1 shows the obtained performance scores for the benchmarks. The average score for the Random action benchmark was around 6.54, much lower than the other two. The average score for the consistent state heuristic was 20.22, and the average score for the minimum distance heuristic was 22.90. The minimum distance heuristic outperformed the other two benchmarks and was taken as the best approach for comparison with the reinforcement learning algorithms' performance. The first objective was to obtain improved performance over this score when deploying the reinforcement learning algorithms.

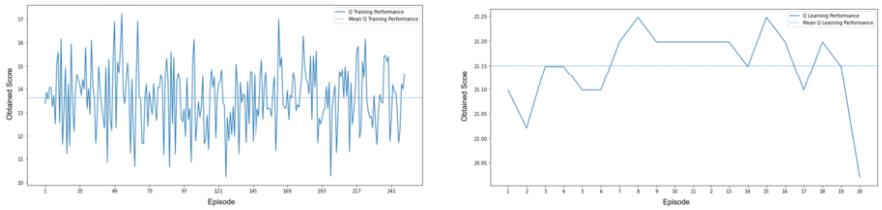


(a) (b) (c)

Figure 1. Obtained performance scores by episode for (a) Random actions, (b) Consistent state heuristic, and (c) Minimum distance heuristic.

5.2. *Q-Learning*

Once having established an appropriate benchmark, we can analyze the performance of our proposed Q-learning approach. Figure 2 shows the performance scores for this model's training and testing steps.



(a) (b)

Figure 2. Q-learning (a) train and (b) test performance.

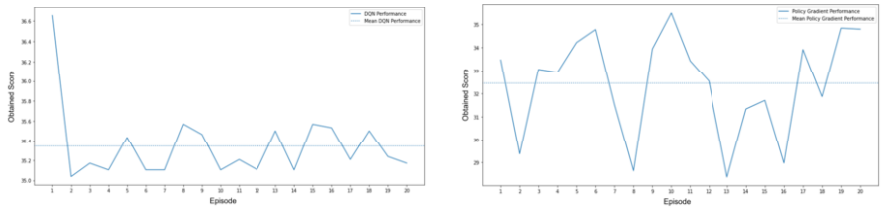
Comparing both plots shows that test performance is better than train performance. This happens because we used the ε-greedy policy during training, whereas we used a greedy policy during testing. In addition to that, we notice no upward/downward trends in the training performance. The training score looks like a noise trend centered at 13.7 without any apparent signs of improvement throughout the process. However, this is not the case when using the learned Q-table in the test phase with a greedy policy. In this setting, the average performance rises to 21.15, and the noise almost disappears. It is also important to highlight that test performance does not outperform the benchmarks. This result is because the q-table we attempted to estimate is of shape $|S| \times |S|$. In our case, $|S| = 3639$, resulting in 13.242.321 distinct state-action pairings for which we need to estimate a q-value. Even after 250 two-month simulations, around 99.96% of the values in the q-table were still equal to their initial values. Therefore, it is reasonable that the training process did not appear to evolve over time and that the subsequent test performance did not outperform the benchmarks. We conclude that the Q-learning algorithm severely suffers from the curse of dimensionality in the setting that we are considering. Therefore, it does not lead to any improvement in the benchmarks.

5.3. *Deep Q-Learning*

Deep Q-Learning (DQN) promises to overcome the dimensionality issues in Q-learning. We evaluate this assumption by building a Q-network having as an input the one hot encoded state vector and comprising two fully-connected hidden layers with 32 and

16 nodes, respectively. All activation functions are *ReLu functions*, and optimization is achieved via the *Adam algorithm*. We complete the training process under a ε-greedy policy for $\varepsilon = 0.1$. We follow a fully Q-greedy policy for the same parameter value during testing. Figure 3 (a) shows the results for this setting.

We notice that DQN outperforms Q-learning by achieving an average episode reward of around 35.34. In addition, it also improves all benchmark performances by reaching a 54.32% more reward than the best-performing heuristic. On top of this, the performance is also remarkably stable, considering the dynamics in the system. Combined with the strong performance, this fact indicates that the DQN algorithm learns a suitable function that maps an input state to its corresponding action q-values.



(a) (b)

Figure 3. (a) Deep Q-learning performance and (b) Policy gradient performance.

5.4. Policy Gradient methods

Finally, we also analyze the performance of the policy gradient *REINFORCE algorithm*. This algorithm has more reliable and stable performance and is better suited to a broader range of situations than Q-learning methods. We use a network that consists of one 64-node hidden layer with a ReLu activation and choose the Adam algorithm as an optimizer. Figure 3 (b) illustrates the test score values obtained for this setting. We notice that the performance is comparable to the DQN algorithm, if slightly lower. This fact is consistent with our prior beliefs. Despite the increased stability and broader application domain, DQN outperforms the policy gradient method by a small margin.

6. Discussion

The aging European population has created a growing demand for home care services to assist the elderly with personal and instrumental activities of daily living. In Barcelona, a case manager matches a reduced team of caregivers to individuals in a specific geographical area. A process that leads to inefficiencies, and significant idle and overwork times. Caregivers can be viewed as agents in a reinforcement learning framework, and the case manager is a central controller that assigns tasks and allocates resources. The reward scheme depends on the desirability of the caregiver/service pairings and the resulting idle/overtime compensations. Staff changes, and emergencies, introduce additional complexity to the system. Approaching the problem of optimizing home care services allocation from a reinforcement learning perspective offers several advantages. Firstly, it allows for flexible modeling of the system's requirements and constraints, with the re-

ward function incorporating metrics such as consistency or similarity to ensure desirable pairings. Secondly, the delayed reward schemes introduced by the framework enable different time scales to be considered, which is particularly useful for systems that require long-term planning. Thirdly, this approach demonstrates the potential of reinforcement learning in aiding complex decision-making processes for real-life issues.

We compare the proposed RL techniques with three benchmark heuristics to further validate them. These include pure random allocations, consistency-based allocations, and distance-minimizing allocations, which serve as baselines to evaluate the proposed techniques. While Q-learning methods suffer from the curse of dimensionality due to many possible allocations, Deep Q-learning methods perform better by focusing on estimating network parameters rather than actual q values. Although policy optimization methods offer a different approach, they cannot outperform Deep Q-learning techniques.

Method	Avg Episode Reward	Method	Avg Episode Reward
Random Actions	6.54	Consistent State Heuristic	20.22
Min Distance Heuristic	22.90	Q Learning	21.15
Deep Q Learning	35.34	Policy Gradient REINFORCE	32.46

Table 1. Score performance comparison

Table 1 shows the performance comparison of different methods. Random actions result in the worst performance, while heuristic methods, achieve significantly better performances and closely approximate the decision-making in a real-life scenario. The Q-learning method does not outperform heuristic results, mainly due to the curse of dimensionality. Most entries in the final q-table remain equal to their initialization values. The greedy system keeps selecting the same state because it always chooses the same action as a tie-breaker when the q-value of all actions in a state is equal to zero. Therefore, arriving at a performance level comparable to the consistent state heuristic makes sense. However, Deep Q-learning and policy gradient suffer much less from the high dimensionality of the problem, and they significantly outperform the best heuristic benchmark result. DQN handles complex state-action relations, leveraging neural networks' flexibility to model a state's mapping to action value estimates. DQN outperforms policy gradient methods in the use case and seems best suited for the scenario. The policy definition in our DQN results is essential. During training, we use an 'ε-greedy policy' to balance exploration and exploitation. During testing, we switch to a 'greedy policy' to choose actions with the highest reward. However, more policies could be defined to better capture the system's dynamics. The system has a limitation in that it does not allow for long dependencies between states due to the Markov property. Still, this limitation is mitigated by accommodating the long-time dependencies inside the reward function.

7. Conclusions

In this paper, we have proposed and validated a reinforcement learning setting for optimal resource allocation in-home care services. We have formalized the allocation problem as a Markov decision process. Our experimental results show that this approach is a viable alternative to classical methods for resource allocation, even in complex environments with time-delayed penalties and interdependent state rewards. Given the key

role that home care services will play in the upcoming years and the importance of optimal resource allocation systems in large urban areas, the proposed research can have a significant impact on enhancing the quality of life for elderly people.

References

[1] E. Commission, i2010: independent living for the ageing society (2006).

[2] FeSP-UGT, Estudio sobre las condiciones laborales de las trabajadoras de asistencia a domicilio, https://generoprl.org/wp-content/uploads/2018/10/ESTUDIO.pdf (2018).

[3] J. Rodriguez-Pereira, J. de Armas, L. Garbujo, H. Ramalhinho, Health care needs and services for elder and disabled population: Findings from a barcelona study, Journal of Environmental Research and Public Health 17 (8071) (2020).

[4] B. Q. C. Fei, Virtual nursing-home, to solve the pension problem in rural communities to explore the beneficial, Social Security Studies 3 (2011).

[5] K. V. Krinichansky, Smart Solutions for Smart Cities, Emerald Publishing Limited, 2019. doi:10.1108/978-1-78973-881-020191010.

[6] P. M. Duque, M. Castro, K. Sörensen, P. Goos, Home care service planning. the case of landelijke thuiszorg, European Journal of Operational Research 243 (1) (2015) 292–301.

[7] O. Bossert, A. Kretzberg, J. Laartz, Unleashing the power of small, independent teams, McKinsey Quarterly 18 (3) (2018) 67–75.

[8] E. Pavolini, C. Ranci, Restructuring the welfare state: Reforms in long-term care in western european countries, Journal of European Social Policy 18 (2) (2008) 246–259. doi:https://doi.org/10.1177/0958928708091058.

[9] A. Alelaiwi, Resource allocation management in patient-to-physician communications based on deep reinforcement learning in smart healthcare services, in: 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), 2020, pp. 1–5. doi:10.1109/ICMEW46912.2020.9105982.

[10] S. Lofti, A. Khalifa, A. Bellahsen, Planning in home health care structures using reinforcement learning, ICLR AI for social good workshop (2019).

[11] S. Racanière, T. Weber, D. P. Reichert, L. Buesing, A. Guez, D. J. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li, R. Pascanu, P. W. Battaglia, D. Hassabis, D. Silver, D. Wierstra, Imagination-augmented agents for deep reinforcement learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5690–5701.

[12] J. Tsitsiklis, B. Van Roy, An analysis of temporal-difference learning with function approximation, IEEE Transactions on Automatic Control 42 (5) (1997) 674–690. doi:10.1109/9.580874.

[13] F. Y. Kuo, I. H. Sloan, Lifting the curse of dimensionality, Notices of the AMS 52 (11) (2005) 1320–1329.

[14] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, in: Advances in Neural Information Processing Systems (NeurIPS), 1999, pp. 1057–1063.

[15] T. T. Nguyen, N. D. Nguyen, S. Nahavandi, Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications, IEEE Transactions on Cybernetics 50 (9) (2020) 3826–3839. doi:10.1109/TCYB.2020.2977374.

[16] R. A. Howard, Dynamic Programming and Markov Processes, The Technology Press of The Massachusetts Institute of Technology, 1960.

[17] B. Han, Z. Ren, Z. Wu, Y. Zhou, J. Peng, Off-policy reinforcement learning with delayed rewards (2021). arXiv:2106.11854.

[18] S. C. of Barcelona, Pla de salut, http://www.csb.cat/plansinformacio-sanitaria/pla-de-salut/ (2016).

[19] K. Fraser, G. B. Lisa, D. Laing, J. Lai, N. S. Punjani, Case manager resource allocation decision-making for adult home care clients: With comparisons to a high needs pediatric home care clients, Home Health Care Management & Practice 30 (4) (2018) 164–174. doi:10.1177/1084822318779371.

[20] R. A. Anderson, P.-C. Hsieh, H.-F. Su, Resource allocation and resident outcomes in nursing homes: Comparisons between the best and the worst, Research in Nursing & Health 21 (4) (1998) 297–313.

[21] K. D. Fraser, C. Estabrooks, M. Allena, VickiStranga, Factors that influence case managers' resource allocation decisions in pediatric home care: An ethnographic study, International Journal of Nursing Studies 46 (3) (2009) 337–349.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230681

Assessing VTE Risk in Cancer Patients Using Deep Learning Synthetic Data Generation and Domain Adaptation Techniques

Sergi BECH [a], Bárbara LOBATO [b], Oriol PUJOL [a], José Manuel SORIA [b], Andrés MUÑOZ [c]

[a] *Dept. Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain*
[b] *Unitat de Genòmica de Malalties Complexes, Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, Barcelona, Spain*
[c] *Medical Oncology, Hospital General Universitario Gregorio Marañón; and Cancer and Thrombosis Working Section, Spanish Society of Medical Oncology (SEOM), Madrid, Spain*

Abstract. This article focuses on the use of deep learning synthetic data generation methods to assess the risk of future treatments and medication for preventing venous thromboembolism (VTE) in cancer patients, based on a small dataset of genetic and clinical variables. The study employs CopulaGANs to generate synthetic tabular data, which is then used to train a Deep Learning-based classifier using domain adaptation techniques. The trained model is fine-tuned using real data and performs better than current state-of-the-art medical scores in assessing VTE risk. Additionally, the resulting Precision-Recall curve offers flexibility in selecting different and better operational points for VTE risk assessment.

Keywords. Synthetic Data Generation, Generative Adversarial Networks, Domain Adaptation, Application to Biomedical Data

1. Introduction

Cancer is one of the most common and fatal diseases nowadays. In 2020, over 19 million people were newly diagnosed and 10 million died worldwide [1]. Thanks to recent advances in oncologic treatments, many types of cancers are now regarded as chronic illnesses and, even in some cases, curable [2]. Cancer patients are in greater risk of suffering venous thromboembolism (VTE) when compared to the general population [3]. Although VTE can have dire effects, it can be easily prevented with anticoagulants. In fact, more effective anticoagulants are now accessible [4]. However, cancer patients also are in greater risk of bleeding [5]. Thus, the clinical management of such patients is difficult due to the delicate balance between the benefit of receiving anticoagulants to prevent VTE and the risk of severe bleeding. Besides, the longer the life expectancy in chronic cancer patients, the more likely it is that they will suffer VTE at some point [6].

The first attempt to create a tool able to classify cancer patients by VTE risk prior to anti-cancer treatment is the Khorana Risk Score (KRS) [7], which utilises five clinical features and assigns a score depending on their observation. Patients are then classified into low, intermediate or high risk. Despite its modest performance, the KRS has become the reference score and many of the current scores in clinical practice are modified versions of this.

With the increase in accessibility to genetic analysis and the evidences that support the role of genetics in VTE [8], [9], new scores based on combining genetic variables and phenotypic information have been developed. The first of such scores was TiC-Onco score [10]. This score has been recently replaced by ONCOTHROMB score [11], which establishes the current state-of-the-art and outperforms KRS when access to genetic information is available.

Due to the costs and potential impact of these studies, the creation and validation of these scores usually entails small datasets. For example, the validation of the ONCOTHROMB score only uses a few hundreds of patients. Although using the correct methodologies can lead to models with a certain degree of generalization, the small amount of samples combined with the large dimensionality (imposed by the combination of phenotypic and genetic features) certainly hinders predictive models from achieving the highest performance.

The machine learning community has recognized the problems of small data combined with the curse of dimensionality. Dimensionality reduction and feature selection techniques have been widely used to solve this issue. With the advent of deep learning techniques and the increased need of data for training complex models, there has been a resurgence of methodologies focused in the creation of new samples. Overcoming some of the problems of noise-based models and interpolation methods such as SMOTE, synthetic data generation constitutes many of the current efforts of the community. In this trend, one can find different approaches depending on the specific application, ranging from variational autoencoders [12], generative adversarial networks [13] or diffusion models [14], among others.

In this work we explore the creation of tabular synthetic data compatible with features in ONCOTHROMB score with the goal of inducing a learning bias in such model. To that end, a variant of Conditional Tabular GAN — a CopulaGAN — is used to generate a large dataset of compatible data with the real ONCOTHROMB cohort. Then, using transfer learning techniques [15], a pretrained model with synthetic data is fine-tuned to the real data. The results show that the obtained model improves not only the generalization of the score but also the range of operational points in the Precision-Recall curve. This is of particular interest, as it allows to improve the evaluation of the two risks involved, namely complications due to VTE and those due to anticoagulant treatment.

The article is organized as follows: section 2 introduces the relevant concepts and techniques that will be referred to throughout this article. Section 3 briefly describes the proposed approach. Section 4 describes and report the obtained results and discusses them. Finally, conclusions are drawn and future lines are depicted.

2. Background and related works

2.1. Assessing VTE risk using ONCOTHROMB score

The ONCOTHROMB score was created using both genetic and clinical features with the aim to predict whether a patient diagnosed with cancer will develop VTE within the next six months since the diagnosis. The cohort of patients used to train this classifier is described in [10]. The predictive features used to train a logistic regression are: VTE risk according to tumor type as defined in the KRS, stage, whether the body mass index (BMI) is greater than 25 Kg/m2, and a Genetic Risk Score (GRS) that condenses the number of risk alleles of 9 Single Nucleotide Polymorphisms (SNPs) know to be associated with VTE. The genotyping and data collection procedures were performed using the protocols established in the Spanish ONCOTHROMB 12-01 study [11].

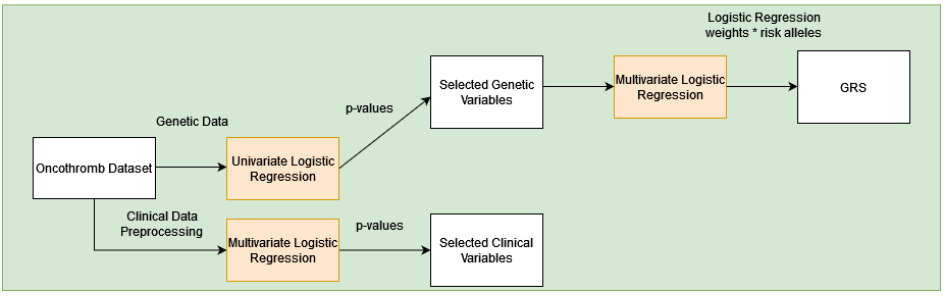


Figure 1. ONCOTHROMB score preprocessing pipeline

The steps in which the score was built are described in [11] and showed in Figure 1. The selection of the genetic variables is done using an univariate logistic regression, and variables with a p-value < 0.25 were chosen. Next, a multivariate logistic regression was performed using the number of risk alleles for the selected genetic variant. The weight of each genetic variant was multiplied by the number of risk alleles, thus generating a GRS for each patient. On the other hand, a multivariate logistic regression was also performed with the aim of selecting the clinical variables, and those with p-value ≤ 0.25 were kept. If any missing values were found for the selected variables for a patient, that patient was excluded from the analysis.

2.2. Deep Generative Models for tabular data

Deep generative models use deep learning techniques to generate synthetic data as similar as possible to real-world data. One of the most common types of deep generative models used for tabular data is the Generative Adversarial Network (GAN) [16]. These models were first developed to generate synthetic images, but several types of GANs were specifically developed for tabular data — some examples include the TabGAN, Conditional Tabular GAN (CTGAN) [17], Tabular Variational Autoencoder-GAN (TVAE-GAN), and CopulaGAN [13]. Firstly, TabGAN models generate synthetic tabular data by learning the underlying distribution of the original data using an adversarial loss. Secondly, CTGAN generates synthetic tabular data based on a given set of conditional vari-

ables. Thirdly, TVAE-GAN is a hybrid model that combines a variational autoencoder and a GAN; the VAE is used to encode the input data into a latent representation, and the GAN is used to generate new samples from the learned latent space. Finally, CopulaGAN is an extension of the CTGANs that utilizes copulas (a function that represents the correlation or dependence between variables with independence of their marginal distributions) to model the dependence structure of real-world data, and then generates synthetic data that preserves this same dependence structure using a GAN.

More specifically, the fitting process of a CopulaGAN begins by standardizing non-categorical variables using Gaussian normalization. In order to generate synthetic data, the CopulaGAN samples from the learned joint distribution using CTGAN, and the resulting data is transformed back to its original distribution by applying the inverse cumulative distribution function, ensuring that the generated synthetic data has the same marginal distributions as the original input data. The code for CopulaGAN and GaussianNormalizer can be found in the sdv library [18].

3. Transferring knowledge from synthetic data to the ONCOTHROMB cohort

In the ONCOTHROMB cohort, each patient contains clinical and genetic information that can be distilled into a small set of aggregated variables. The main hypothesis of this work is that, by means of deep learning generative methods, a large set of compatible data can be synthesized and used to induce a learning bias that can be transferred to the real setting, thus improving the generalization and operational points of the original ONCOTHROMB score.

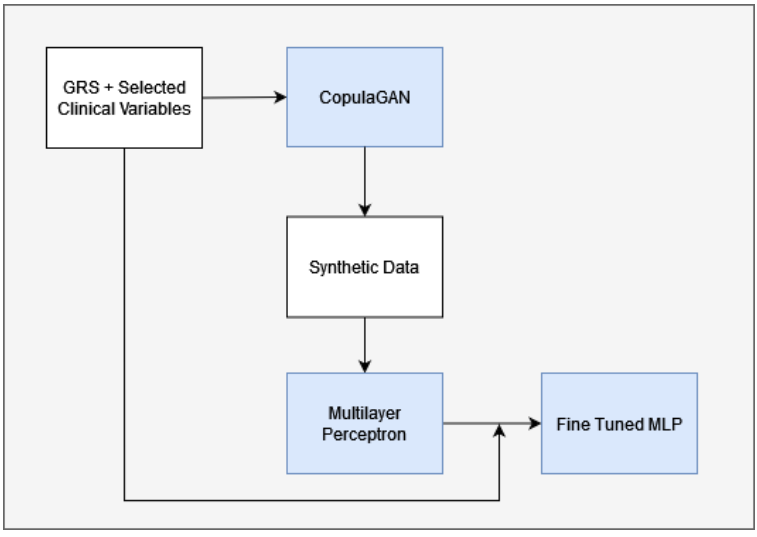


Figure 2. Methodology proposal

For that purpose, we first generate synthetic tabular data utilizing generative models. It is crucial to ensure that the synthetic data closely resembles the distribution of real data and, as a result, performs well in a classification task. Therefore, we propose to

train a multilayer perceptron (MLP) only with synthetic data and assess its performance. Lastly, a fine-tuning process of the MLP is carried out using real-world data. Figure 2 summarizes the proposed methodology.

Synthetic Data Generation: The first step of our proposal is to generate synthetic data samples using the CopulaGAN. As shown in Figure 2, we generate synthetic samples containing the GRS and the clinical variables that are selected on [11]. Then, the CopulaGAN is trained and 150,000 observations of synthetic data are sampled. The quality of the synthetic data is evaluated both visually and by training classifiers for the prediction of VTE.

Fine-tuning with Real Data: The synthetic data generated by the CopulaGAN is used to train a MLP. Then, we proceed to fine-tune the model with real data in order to shift the biases of the classifier. We experiment with various fine-tuning strategies, such as adding more layers to the pretrained model and freezing all layers except for these new ones, as well as retraining all layers for a small number of epochs and several learning rates. Both strategies produced indistinguishable results. For the sake of simplicity, in this article we report the results using fine-tuning with reduced learning rates.

4. Experiments and Results

In this section we explain the experimental setting, describing the details of the data, methods and models used, parameter set-up, and performance measures as well as the results obtained. The results are divided in three different subsections: first, the plausibility of the generated synthetic data is assessed; then, the results of the model trained on synthetic data are reported and discussed; and, finally, the transferred model is evaluated.

Data description: As already mentioned, the dataset used in this study comes from the ONCOTHROMB12-01 cohort, and it comprises clinical and genetic information on 390 patients with 19 clinical features plus 54 genetic features representing the number of risk alleles present in a set of genetic variants known to be linked to VTE. There are some missing values in the dataset, primarily in the genetic data. No imputation is used to deal with the missing values; instead, patients with missing values in key variables are dropped. Thus, 29 instances are dropped and the information of 361 patients is used in subsequent experiments. The same four features used in the original ONCOTHROMB (mentioned in Subsection 2.1) are the ones used in further analyses.

Methods and models used: The first step in this study is to generate synthetic tabular data from the preprocessed dataset. A CopulaGAN model is used to generate 150,000 samples of synthetic data. The CopulaGAN is trained with 5 discriminator updates for each generator update as in [19]; we also set the number of epochs to 1,500 and the batch size to 60, and the rest of parameters are the default ones from [18].

A MLP is trained with synthetic data only with the aim of learning the inductive biases of our problem. The MLP has 3 hidden layers with 128, 64, and 16 neurons, using the Rectified Linear Unit (ReLU) activation function with dropout rates of 0.2 and 0.1 after the first and second layers, respectively. The model was trained using binary cross-entropy as the loss function with Adam optimizer, a learning rate of 1×10^{-3}, and a sigmoid activation function on the output layer. Early stopping was applied by monitoring the validation precision-recall. Then, domain adaption techniques are applied to this pretrained deep learning model. Specifically, the model is fine-tuned using real

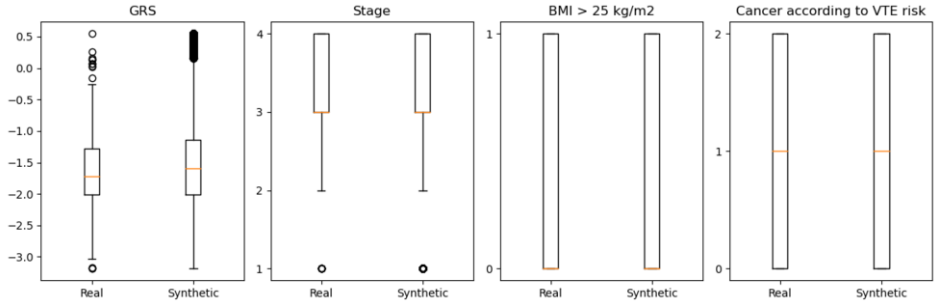


Figure 3. Comparison of real and synthetic data feature's distribution

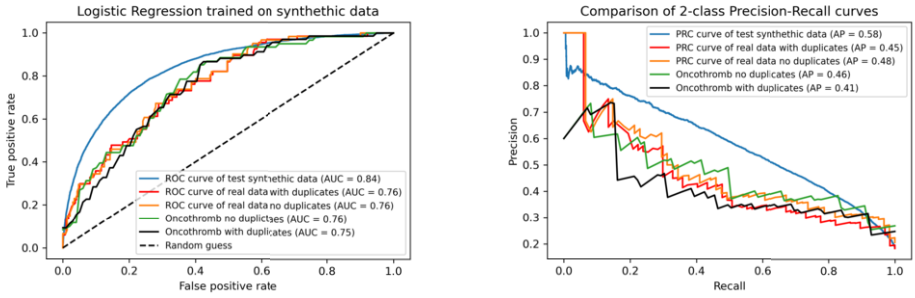
data to adapt and learn its specific patterns. In this step, the architecture of the model remains exactly the same and all layers are trained again, but this time with real data and for a smaller number of epochs and smaller learning rate (1×10^{-4}). To balance the class distribution of the real data, we computed class weights by dividing the total number of examples by twice the number of negative (or positive) examples. These class weights help to ensure that the model does not become biased towards the majority class during fine-tuning step. The early stopping criteria used during the previous training step is still used for this training as well.

Performance metrics: The evaluation of our method is done by comparing the ROC (Receiver Operating Characteristic) curve and PRC (Precision-Recall Curve) with the ONCOTHROMB score's curves. Additionally, we report the area under the curves, AUC-ROC and AUC-PRC, respectively.

We observed that, due to the variable selection process conducted in the ONCOTHROMB score, the original data set contains duplicated observations. This may undesirably bias the results. In order to provide a fair comparison, we evaluate all metrics in the dataset with duplicates (as done in the original paper) and also without duplicates.

Experimental set-up: The experiments are divided in three blocks:

- **Experiment 1: Synthetic data quality assessment.** The CopulaGAN is trained over all the preprocessed dataset. To assess the quality of the synthetic data, a graphical comparison is performed between the distributions of both synthetic and real data.
- **Experiment 2: Assessment of synthetic data trained classifiers.** Subsequently, the synthetic data was divided into training and test sets, and a logistic regression classifier was trained on this data. This logistic regression is evaluated using the synthetic fake data and also the real data in order to compare it to the baseline. The purpose of this evaluation was to verify that the logistic regression model trained with synthetic data produced similar results to that trained with real data. Afterwards, a MLP consisting of three layers was trained using the synthetic training data, and precision-recall was utilized as the stopping criteria. This model is evaluated with the test data and also with the real data to be compared with the baseline.
- **Experiment 3: Assessment of the transfer learning approach.** As a last step, the real data is divided into 5 stratified folds and is used to adapt the model to our domain by fine-tunning the MLP.



(a) (b)

Figure 4. Synthetic and real data evaluation on classification task. The blue curves represents the logistic regression model tested on test synthetic data. The green curves are the ONCOTHROMB model's and the yellow/red curves are from the logistic regression tested on the real data.

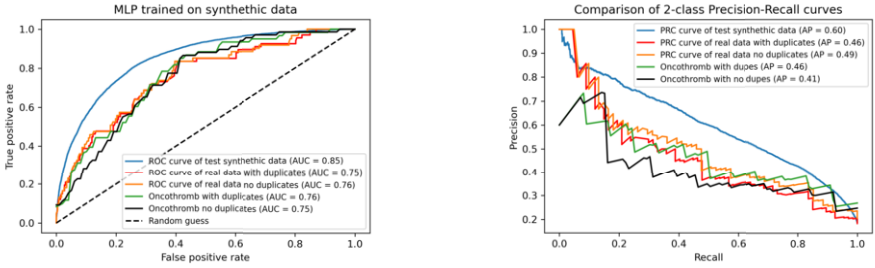
Experiment 1: Synthetic Data Qualitative Assessment

Synthethic data generated with CopulaGAN is first evaluated by comparing the cumulative sum distribution per feature (not shown) as well as the feature's data distribution represented in boxplots.

Figure 3 suggests that the synthetic data preserves the distribution and patterns of the original data despite the fact that synthetic data is generated in a non-constrained parameter space that does not necessarily resemble that of the real data. Moreover, we proceed to evaluate how good is this synthetic data compared to the real data when classifying patients by VTE risk.

Experiment 2: Assessment of synthetic data trained classifiers

The information captured in synthetic data is demonstrated to be useful for classifying real data through the performance of logistic regression trained on synthetic data (Figure 4). We observe that the differences between the model trained exclusively using synthetic data is able to achieve comparable results to the model trained using the original data. The ROC curve in Figure 4 (a) indicates that our synthetic data still contains valuable information. Similar conclusions can be drawn from Figure 4 (b), where syn-



(a) ROC curve (b) PRC curve

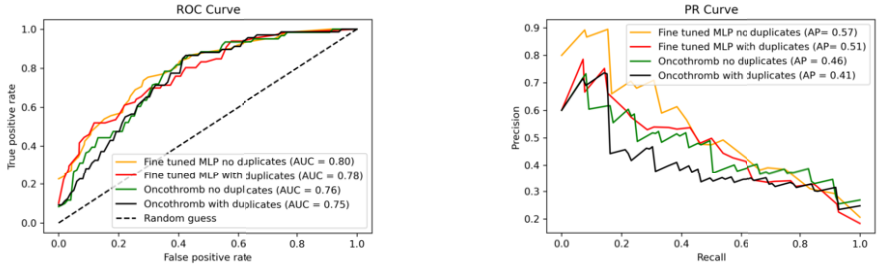
Figure 5. Synthetic data and real data evaluated on classification task. The blue curves represents the MLP model tested on test synthetic data. The green curves are the ones from ONCOTHROMB replications and the yellow/red curve are from the MLP tested on the real data.

thetic models are on pair with the models trained with real data in terms of PR curve. These results support our hypothesis that the generated synthetic data contains valuable information for making predictions.

Figure 5 shows the same kind of evaluation when training a MLP. Again, we observe in both the ROC curve, Figure 5(a), and PR curve in Figure 5(b), that the MLP classifier replicates the findings shown in the former experiment.

Experiment 3: Assessment of the transfer learning approach

As described in the experimental set-up, we report the results after using the transfer learning approach. Figure 6 shows the curves obtained after the fine tuning.



(a) ROC curve (b) PRC curve

Figure 6. Real data evaluated on classification task. The blue curves are the one from our ONCOTHROMB replications and the yellow/red curves are from the fine tunned MLP tested on the real data.

When looking at the ROC curves in Figure 6 (a) we observe a consistent improvement of the AUC-ROC values compared to the baseline in both settings; i.e. considering duplicates and non-duplicates. This is shown by the clear improvement in the TPR for the first half of the FPR in the ROC curves. However, for large values of FPR all methods seems to converge to the same behavior. A more drastic improvement is found in the Precision-Recall curves. Figure 6 (b) shows the obtained curves. A detailed examination of these segments according to the two different set-ups, namely 'duplicates' and 'non-duplicates', provides evidence that the proposed approach improves the AUC-PRC by 10%. All the relevant figures and its performance metrics are summarized in Table 1.

	ONCOTHROMB score	Trained with synthetic data	Fine-tuned with real-world data
Non-duplicates	0.46/0.76	0.49/0.76	**0.57/0.80**
Duplicates	0.41/0.75	0.46/0.75	**0.51/0.78**

Table 1. Results of AUC-PRC/AUC-ROC for the baseline model, the model trained exclusively with synthetic data, and the transfer learning approach.

The large improvement in AUC-PRC might have important consequences in the clinical practice. If we examine the plots in Figure 6 (b) we observe that the fine-tuned approaches clearly improve the operational point range of the score. In the case of considering the original dataset (curves in black and red), we observe that for recalls between 0.2 and 0.8, the proposed methodology achieves important gains. If we focus on the modified dataset (curves green and yellow), we observe the same effect in the range of recalls between 0.0 and 0.65. Again, the proposed method clearly outperforms the

baseline. As said earlier, this has might have a huge impact in clinical practice as it allows the physician to leverage the different risks involved in the process.

For the sake of discussion, the recall value stands for the rate of VTE detected from all the VTE population. Thus, the larger the recall, the best one can prevent VTE complications by administering anticoagulant treatment. However, as expected, this value trades-off with the precision. In this case, precision is related to the risk created by the administration of the anticoagulant treatment. The operational point of the original ONCOTHROMB article is $(0.8, 0.3)$. This is justified by the flat area in the black precision curve for the range of values between 0.4 and 0.8 recall. Precision of 0.3 means that the original score would recommend to overmedicate 70% of detected patients. The new score proposed in this article allows for the exploration and selection of different operational points, with a clear reduction of the risk of overmedication while keeping a large sensitivity value.

5. Conclusions

In this paper, we explored the use of synthetic data generation and transfer learning methodologies for improving the operational points in the PRC and ROC curves of a classifier that predicts the risk of VTE in cancer patients. Results show clear improvements in the AUC-ROC and AUC-PRC. These findings have impact in the clinical practice as it may allow to reduce the risk of complications due to the treatment of VTE.

The next steps in this line of work follow two main avenues. First, the obtained results are very promising but further validation is recommended. Thus, new data is being collected that may further confirm these improvements. Second, we plan on focusing on the influence of the genetic variants in VTE. In this line of thought, we plan on exploring the use of causal machine learning algorithms to better understand the underlying mechanisms of cancer in VTE.

6. Acknowledgements

This work is supported by the Sociedad Española de Trombosis y Hemostasia (SETH), Sociedad Española de Oncología Médica (SEOM), Asociación ActivaTT por la Salud and PI/20/00325 FIS project, and partially supported by MCIN/AEI/10.13039/501100011033 under project PID2019-105093GB-I00 and PID2022-136436NB-I00.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71(3):209-49. doi:10.3322/caac.21660.

[2] Pituskin E. Cancer as a new chronic disease: Oncology nursing in the 21st Century. Canadian OnCOlOgy nursing JOurnal. 2022;32(1):87.

[3] Blom JW, Doggen CJM, Osanto S, Rosendaal FR. Malignancies, Prothrombotic Mutations, and the Risk of Venous Thrombosis. JAMA. 2005 02;293(6):715-22. doi:10.1001/jama.293.6.715.

[4] Desai R, Koipallil GK, Thomas N, Mhaskar R, Visweshwar N, Laber D, et al. Efficacy and safety of direct oral anticoagulants for secondary prevention of cancer associated thrombosis: a meta-analysis of randomized controlled trials. Scientific Reports. 2020;10(1):18945. doi:10.1038/s41598-020-75863-3.

[5] Al-Samkari H, Connors JM. Managing the competing risks of thrombosis, bleeding, and anticoagulation in patients with malignancy. Hematology 2014, the American Society of Hematology Education Program Book. 2019;2019(1):71-9. doi:10.1182/bloodadvances.2019000369.
[6] Mahajan A, Brunson A, Adesina O, Keegan TH, Wun T. The incidence of cancer-associated thrombosis is increasing over time. Blood advances. 2022;6(1):307-20. doi:10.1182/bloodadvances.2021005590.
[7] Khorana AA, Kuderer NM, Culakova E, Lyman GH, Francis CW. Development and validation of a predictive model for chemotherapy-associated thrombosis. Blood, The Journal of the American Society of Hematology. 2008;111(10):4902-7. doi:10.1182/blood-2007-10-116327.
[8] Zöller B, Li X, Ohlsson H, Ji J, Sundquist J, Sundquist K. Family history of venous thromboembolism as a risk factor and genetic research tool. Thrombosis and haemostasis. 2015;114(11):890-900. doi:10.1160/TH15-04-0306.
[9] Zöller B. Genetics of venous thromboembolism revised. Blood. 2019 11;134(19):1568-70. doi:10.1182/blood.2019002597.
[10] Muñoz Martin AJ, Ortega I, Font C, Pachón V, Castellón V, Martínez-Marín V, et al. Multivariable clinical-genetic risk model for predicting venous thromboembolic events in patients with cancer. British journal of cancer. 2018;118(8):1056-61. doi:10.1038/s41416-018-0027-8.
[11] Muñoz A, Ay C, Grilz E, López S, Font C, Pachón V, et al. A Clinical-Genetic Risk Score for Predicting Cancer-Associated Venous Thromboembolism: A Development and Validation Study Involving Two Independent Prospective Cohorts. J Clin Oncol. 2023. doi:10.1200/JCO.22.00255.
[12] Wan Z, Zhang Y, He H. Variational autoencoder based synthetic data generation for imbalanced learning. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI); 2017. p. 1-7. doi:10.1109/SSCI.2017.8285168.
[13] Abedi M, Hempel L, Sadeghi S, Kirsten T. GAN-Based Approaches for Generating Structured Data in the Medical Domain. Applied Sciences. 2022;12(14):7075. doi:10.3390/app12147075.
[14] Azizi S, Kornblith S, Saharia C, Norouzi M, Fleet DJ. Synthetic Data from Diffusion Models Improves ImageNet Classification; 2023. doi:10.48550/arXiv.2304.08466.
[15] Iman M, Arabnia HR, Rasheed K. A Review of Deep Transfer Learning and Recent Advancements. Technologies. 2023 mar;11(2):40. doi:10.3390/technologies11020040.
[16] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27:2672-80. doi:10.1145/3422622.
[17] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN; 2019. doi:10.48550/arXiv.1907.00503.
[18] Team SD. SDV: Synthetic Data Vault. GitHub; 2022. Version 0.11.0. https://github.com/sdv-dev/SDV.
[19] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN; 2017.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230682

Hipertension Demand Forecasting Using Cross-Correlation and Lagged Multiple Linear Regression Models for Anticipating Health Resources Needs

Guillem HERNÁNDEZ GUILLAMET [a,b,c,1], Beatriz LÓPEZ [c] and Oriol ESTRADA [a,b] and Francesc LÓPEZ SEGUÍ [a,b]

[a] *Research Group on Innovation, Health Economics and Digital Transformation (Institut de Recerca Germans Trias i Pujol)*
[b] *Hospital Universitari Germans Trias i Pujol (Institut Català de la Salut)*
[c] *eXiT research group, University of Girona*
ORCiD ID: Guillem Hernández Guillamet https://orcid.org/0000-0002-1053-3084, Beatriz López https://orcid.org/0000-0001-9210-0073, Oriol Estrada https://orcid.org/0000-0003-1414-5000, Francesc López Seguí https://orcid.org/0000-0003-0977-0215

Abstract. This article presents an algorithm that uses a combination of cross-correlation analysis and lagged multiple linear regression models to predict the time-series of future demand for clinical visits associated with a certain diagnosis, specifically hypertension, in the Catalan health-care system. The algorithm aims to provide a robust and explainable feature selection set of predictors. The study demonstrates that it is possible to predict demand associated with a diagnosis through the demand for previous clinical visits, and identifies important predictors for example case hypertension-related visits. The data used is from the primary care services of the Catalan Institute of Health, and the methodology can be applied to optimize resource allocation in the healthcare system.

Keywords. multiple linear regression, forecasting, healthcare demand

1. Introduction

Many studies and algorithms have explored the possibility of predicting future demand to anticipate resource consumption and optimize their distribution. The case study is particularly important in the healthcare sector, which is heavily strained by the shortage of professionals and the aging population in many developed countries. In the case of the Catalan healthcare system, this problem is exacerbated due to systemic underfunding, overload and talent drain of professionals, entrenched waiting lists, the commercializa-

[1] Corresponding Author: Guillem Hernández Guillamet (ghernandezgu.germanstrias@gencat.cat ; +34 648977302)

tion of healthcare, and a gradual aging of the population that is living longer (with all its clinical consequences). It is also important to point out the effect that the COVID-19 pandemic has had on the healthcare system. Most of these issues have been aggravated, and the normal functioning of the system has been distorted. With the lockdown, the treatment of multiple conditions and patients has been postponed, and it is necessary to recover the lost activity [1][2]. In this context, the ability to predict the future demand of the healthcare system is essential for the efficient planning of resources and modeling the post-pandemic demand. It is particularly interesting not only to predict the demand but also the reason for consultation, in this case, the diagnoses associated with the medical visit, to distribute healthcare professionals according to their expertise and develop a more flexible system.

In this article, we present an algorithm that uses a combination of cross-correlation analysis and a lagged multiple linear regression (lagged MLR) framework to predict the time series of future demand for clinical visits associated with a certain diagnosis, in this case, hypertension. The algorithm seeks to achieve the best prediction while penalizing the progressive entry of new predictors, trying to achieve the best forecasting with the minimum number of variables. Therefore, the algorithm not only achieves good predictive results but also indicates which diagnoses may be important for predicting future demand for visits associated with hypertension.

This paper is organized as follows. In Section 2 we provide a description of related proposals. Section 3 is devoted to outlining the methodology and the algorithm. In Section 4 the main results of its application are presented and discussed. Finally, section 5 presents the main conclusions of the study.

2. Related work

The hypothesis we aim to demonstrate is that the demand associated with a visit related to a specific diagnosis can be predicted through the demand for previous clinical visits. There are very few studies that focus on developing demand prediction models in the medical field. Most research revolves around predicting epidemiological diagnoses and identifying relevant predictor variables. Some studies attempt to predict epidemiological visits, such as flu outbreaks [3]. Other studies explore the possibility of predicting the spread of COVID-19 using climate variables [4][5][6][7][8]. Additionally, there are studies specifically dedicated to predicting visit demands using Deep Learning [9] or predicting unplanned visits for diabetic patients [10]. Our approach differs from the previous ones since it aims to predict demand, not only cases. Moreover, the use of regression models provides a familiar framework for health professionals and allows for a feature selection procedure that can be later used for prediction with AI methods.

On the other hand, one interesting research related work is the prediction of emergency attendances, since they require from an accurate resource planning as we are trying to pursue. In [11] a combination of MLR and Artifical Neural Networks (ANN) is used to capture the complexity of emergency attendances in a touristic area, with a huge variability between different periods of the year (seasonality). It is evident that seasonality plays a key role, as well as other variables in prediction, but we have not included them due to this ”evidence.” Our work aims to understand which clinical variables act as a proxy for the target, independently of the season.

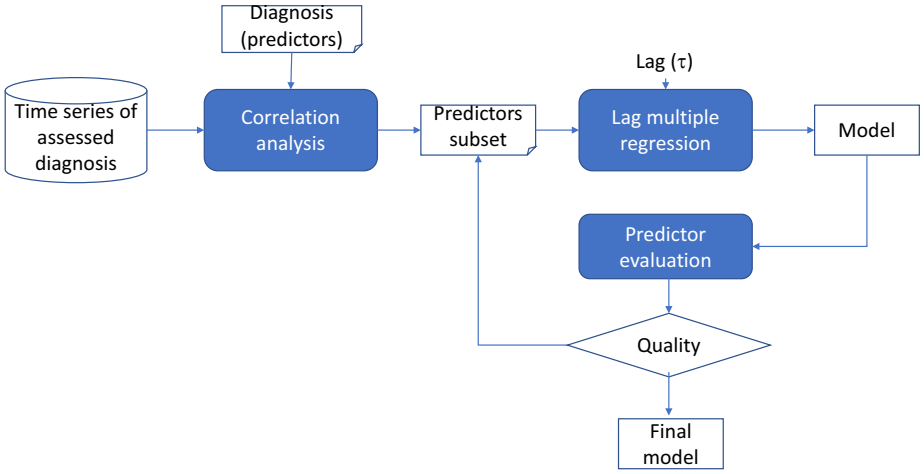


Figure 1. Flow chart of the methodology.

There are some other works that link emmergency departments to some secfic resources, as for example [12]. Linear regression analysis is also used in [12], as we do in our work, showing how simple tools can provide a big improvement in healthcare systems for resource management.

3. Methodology

The framework used can be divided into three major steps (Figure 1). The starting point is time series of diagnoses. Diagnoses play both the role of predictors and target. Our aim is to predict the number of visits associated with a certain diagnosis based on the information of other diagnoses. This prediction is an estimation of the demand for services. Firstly, a cross-correlation analysis is developed between the different time series of demand for diagnoses, and the number of predictors is reduced to a non-correlated subset. Subsequently, a MLR model is trained to predict the target diagnosis. MLR models offer a high degree of interpretability and are easily understandable for healthcare professionals who use them regularly. These models allow for the identification and quantification of the relationship between predictor variables and the target variable, providing a clear understanding of the factors that influence the outcomes. Additionally, their simplicity and familiarity in the healthcare field facilitate their utilization and acceptance among professionals. For these reasons, we have selected these models over other more complex or less interpretable options. This method is applied iteratively; so the MLR model starts with a single predictor and analyzes the progressive entry of new ones. Moreover, the regression model is parameterized by a lag value to build predictive models according to a given time horizon.

With this proposal, we aim to develop a robust model with greater explainability than alternative models such as deep learning (DL). Moreover, we find the minimum number of predictors for each forecasting horizon.

3.1. Dataset

The data used in this study comes from the Primary Care Services of the Catalan Institute of Health (ICS), the main public health provider in Catalonia, Spain. The database is retrospective and contains all primary care visits from the period 2010-2019.

Each visit is associated with a diagnosis d_i that belongs to a set of diagnoses $D = d_1, \ldots, d_N$. In total, the database contains information on 6,301,095 patients, and the diagnoses are coded according to the medical ontology CIM-10 [13]. Only the first three letters of the diagnostic codes have been used, with the codes aggregated in the ontology until the disease group level. This reduces the amount of variables in the dataset from over 9000 to 1846. Visit information is gathered in a time period of L days, $T = t_1, \ldots, t_L$.

The initial database is transformed into a matrix $M = DxT$, where each cell M_{ij} corresponds to the number of occurrences of diagnosis d_i on day t_j. The final dimension of matrix M is 6,743,438 (N = 1846, L = 3653). Observe that the matrix is sparse since not all possible diagnoses are used in a given day. Each row of the matrix represents a time series for a given diagnosis in the period of time L.

Each time series is transformed with a rolling mean of window size = 14 to eliminate the spikes generated by the weekend effect, where there are no visits to primary care. In addition, the data is scaled to have minimum and maximum values of 0 and 1, respectively to obtain easy-to-interpret outcome metrics.

3.2. Cross-correlation analysis and collinearity effects

Given an objective diagnosis to predict d^g, the first part of the model aims to eliminate any possible collinearities present in the predictor set $D - d^g$. Initially, the model computes the Pearson correlation coefficients r_{ij} between all pairs of predictors d_i and d_j. Subsequently, correlation t-tests are performed to evaluate the probability that both predictors have a significant linear relationship [14]. The t-statistic TS_{ij} associated with each combination of predictors is calculated as follows:

$$TS_{ij} = \frac{r_{ij}\sqrt{L-2}}{\sqrt{1-r_{ij}^2}} \tag{1}$$

where L is the size of the sample, that is, the constant length diagnoses time-series. The statistic follows a t-distribution with L-2 degrees of freedom. Assuming a p-value of 0.1, pairs of variables with a TS_{ij} greater than 0.9 have a significant correlation and could therefore cause multicollinearity. To evaluate the degree of collinearity among significantly correlated variables, the algorithm uses the variance inflation factor (VIF), defined as:

$$VIF_j = \frac{1}{1-c_j^2} \tag{2}$$

Parameter c_j indicates the coefficient of determination of variable d_i regressed on the remaining predictors. The algorithm iteratively eliminates variables with the highest degree of collinearity until it is left with a subset of predictors with maximum collinearity of $max(VIF) = 20.0$ (after experimentation), considering collinearity to be non-significant [15]. Therefore, we get a set of $\Delta \subseteq D - d^g$ remaining predictors that the model can use; they do not have a significant correlation or have passed the VIF test.

3.3. Lagged MLR

Our target variable is d^g, which we aim to forecast at time t according to the minimal set of predictors $P \subseteq \Delta \subseteq D - \{d^g\}$. To that end, the following MLR model is defined:

$$model(d^g, P, t, \tau) = \alpha_0 + \sum_{i \in P} \sum_{j \in [t-\tau, t-1]} \alpha_{ij} * M_{ij} \tag{3}$$

where $t \in T$, α_0 is the constant coefficient of the model, α_j is the regression coefficient of predictor $d_i \in P$ at time j, and $\tau < L$ is the lag value applied. Note that lag value can take value 0 and therefore predict the target variable using as predictors variables in the same day.

3.4. Predictors selection

The methodology that the algorithm follows to select the predictors that form the MLR model is the forward stepwise estimation method (see Algorithm 1). Initially, the variable in P with the highest correlation coefficient with the diagnosis d^g is chosen. This predictor and the target variable d^g are fitted to a linear regression according to Equation 3. The adjusted coefficient of determination R^2 is used to evaluate the goodness of fit. It indicates the percentage of the variation explained by the added prediction with respect to the baseline model. Next, the following predictor with the highest correlation with respect to diagnoses d^g is added to the model and R^2 recomputed. To decide whether the addition of the new predictor to the model produces a significant improvement, we use the F-test statistic, defined as:

$$F1 = \frac{(SSR_2 - SSR_1)/(p_2 - p_1)}{\frac{SSE_2}{L - p_2 - p_1}} \tag{4}$$

where subindexes 1 and 2 correspond to the models with the new predictor removed or added, respectively. The variables *SSR* and *SSE* refer to the sum of squares error due to regression (variability explained by the regression) and the sum of squares error due to error (variability not explained by the regression) of the model.

The significance value marked by the F-test statistic is (p-value ≥ 0.1), indicating that the model significantly improves with the addition of the new predictor.

Sometimes, due to the forward stepwise estimation strategy, it may be found that the addition of one predictor does not show significance, but adding the next one does. For this reason, the algorithm allows for up to 3 iterations to exceed the p-value of 0.1 (assuming this 10% of possible error of significance).

Finally, among all of the models computed, the model with the best MAPE (Mean Absolute Percentage Error) is chosen. In case of a tie, the model with the fewer number of predictors is preferred (simpler model). Predictors of the best model are the ones selected.

3.5. Final model assessment

To assess the correctness of the final model multiple tests are performed to ensure it meets all required assumptions. First, the joint F-test is performed to compare the best

Algorithm 1 Predictors selection

Require: $d^g, \Delta = \{d_1, \ldots, d_{|\Delta|}\}, \tau, t$
Ensure: $P \subset \Delta$
$list_of_models \leftarrow null$
$counter \leftarrow 0$
$sort(\Delta, d^g, correlation)$
$P \leftarrow first(\Delta)$
$SSR_1 \leftarrow SSR(model(d^g, P, t, \tau), M)$ ▷ See Equation 4
for $\delta_i \in \Delta - first(\Delta)$ **do** ▷ Iterated according to the sort outcome
  if $counter = 3$ **then**
    break
  end if
  $model_i = model(d^g, P \cup \{\delta_i\}, t, \tau)$ ▷ See Equation 3
  $SSR_2 \leftarrow SSR(model_i, M)$
  $SSE_2 \leftarrow SSE(model_i, M)$
  $F1 \leftarrow f_{test}(SSR_1, SSR_2, SSE_2, |P|, |P \cup \{\delta_i\}|)$ ▷ See Equation 4
  append(P, $\{\delta_i\}$)
  append(list_of_models, $model_i$)
  if p-value(F1) > 0.1 **then**
    $counter = counter + 1$
  end if
  $SSR_1 \leftarrow SSR_2$
end for
$best \leftarrow minimumMAPE(list_of_models)$
$P \leftarrow predictors(best)$

model against the simpler linear model of the target variable (see Equation 6). This is what we call the baseline model. If the p-value was higher than a 0.05 significance level, we concluded that there is enough statistical evidence that the final model fits the observations better than the intercept-only. p represents the number of predictors used in the final model. The join F-test is computed as follows:

$$F2 = \frac{SSR}{SSE} \frac{(L-p-1)}{p} \tag{5}$$

$$baseline(d^g, t, \tau) = \beta_0 + \sum_{j \in [t-\tau, t-1]} M_{ij}, d_i = d^g \tag{6}$$

Next, Kolmogorov-Smirnov (K-S) test was conducted at a 95% confidence interval to test for normal distribution in residuals [16], and the White test for heteroscedasticity at a 95% confidence interval was applied to assess the non-constant variance of regression errors and test the assumptions of the regression model [17].

4. Results

We applied the method to compute the demand for clinical visits associated with the diagnosis of hypertension, labeled "H10" according to the CIM-10 ontology; thus $d^g =$ "$H10$". Our method is applied to make a future prediction of the target variable changing progressively the lag parameter until reaching a one-month prediction horizon,

$\tau \in [0,30]$. The system trains a model for each lagging time that can vary in predictors. This means that the model best forecasting $H10$ at horizon 0 ($\tau = 0$) may have different predictors than the one forecasting the future visits associated with $H10$ $\tau \in [1,30]$.

After applying the lagging in the matrix M (this is, shifting the target variable to the future with respect to all predictors in P accordingly to τ), the train-test splitting strategy is performed to the time series with a partition of 80% for training and 20% for testing to evaluate the forecasting. Cross-validation isn't used to preserve the structure of the time series. In this case, if the study period of the database is 10 years, the first 8 years are used for training, and the remaining 2 years are used for testing. It is worth noting that as the lag $\tau \in [1,30]$ increases in the experimentation, the training set remains the same, but the test set shortens accordingly. Since the testing period is still relatively long (730 days), the results remain robust.

The models are evaluated through the analysis of the number of predictors, the F-statistic (4) and joint F-statistic (5); and the RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) of the training and test predictions. Both RMSE and MAPE are representations of the prediction error. RMSE presents the error as units of the variable being predicted. Therefore, since we have scaled the database within the range [0-1], RMSE addresses the error on this scale. For better interpretation, MAPE represents the error as a percentage on the scale [0-100]. Additionally, MAPE allows for comparing results between experiments, even if the test set does not have the same length.

Table 1 summarizes the best models for each lag. The algorithm achieves accurate predictions in both the on-time model ($\tau(lag) = 0$) and the future prediction models ($\tau(lag) > 0$). The number of variables that form the models ranges from 4 when predicting 9 to 12 days ahead to 25 when predicting 6 or 7 days ahead (one week). All models present a constant subset of predictor diagnoses mainly related with chronicity and elderly populations:

- M15: Primary generalized (osteo)arthrosis
- E78: Pure hypercholesterolaemia
- E03: Congenital hypothyroidism with diffuse goitre
- H90: Conductive hearing loss, bilateral
- F17: Mental and behavioural disorders due to use of tobacco

Details of the results are provided in Figure 2 to 4. Figure 2 shows the on-time model (lag value τ = 0) and Figure 3 plots the model outcome when making a prediction one month ahead (lag value τ = 30). In the case of Figure 2, the prediction is significantly better (see MAPE and RMSE results of both models in table 1), although the model also captures the general trend in the one-month prediction. The x-axis in Figures 2 and 3 represents the day, while the y-axis represents a scaled range of [0,1] for the number of visits associated with the diagnosis.

Figure 4 demonstrates the behavior of the algorithm when increasing the lagging time. The x-axis represents the prediction lag τ, while the y-axis represents the percentage of prediction error achieved by each model compared to the actual time series. The error of prediction increases from 1.2% to 2.8% when the lag is increased $\tau \in [0,30]$. The stepped effect that we see in the Figure is due to the behavior of the algorithm with the application of the forward stepwise estimation method. The algorithm does not find it worthwhile to add or remove new predictors if there is no significant increase in the lag. For small and progressive increases, the algorithm tends to maintain the best model

number variables	F1	p-value (F1)	F2	p-value (F2)	MAPE train	MAPE test	RMSE train	RMSE test	predictors (diagnoses)	LAG
9	1.67	0.20	0.87	3,51E+05	1.39	1.35	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83	0
9	0.26	0.61	-0.06	1,00E+06	1.37	1.34	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83	1
9	1.19	0.27	0.31	5,80E+05	1.48	1.43	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83	2
14	0.70	0.40	-14.32	1,00E+06	1.35	1.20	0.02	0.01	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	3
14	2.06	0.15	-22.75	1,00E+06	1.40	1.26	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	4
20	1.31	0.25	-9.88	1,00E+06	1.42	1.20	0.02	0.01	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	5
26	0.55	0.45	-15.87	1,00E+06	1.37	1.14	0.02	0.01	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	6
26	0.25	0.62	-10.77	1,00E+06	1.35	1.15	0.02	0.01	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	7
12	0.81	0.37	39.09	4,65E-04	1.67	1.53	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F32	8
5	2.27	0.13	-0.24	1,00E+06	2.01	1.67	0.03	0.02	M15,E78,E03,H90,F17	9
5	0.06	0.81	-0.29	1,00E+06	2.08	1.72	0.03	0.02	M15,E78,E03,H90,F17	10
5	0.64	0.43	1.76	1,85E+05	2.13	1.76	0.03	0.02	M15,E78,E03,H90,F17	11
5	2.70	0.10	4.73	2,98E+04	2.18	1.78	0.03	0.02	M15,E78,E03,H90,F17	12
12	1.77	0.18	-45.02	1,00E+06	1.96	1.70	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F32	13
12	2.09	0.15	-48.26	1,00E+06	1.96	1.70	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F32	14
16	0.86	0.35	-19.88	1,00E+06	1.88	1.48	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	15
16	1.98	0.16	-28.14	1,00E+06	1.97	1.50	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	16
16	2.17	0.14	-28.76	1,00E+06	2.00	1.51	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	17
8	2.54	0.11	20.38	6,62E+00	2.35	1.92	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90	18
8	1.26	0.26	13.72	2,16E+02	2.37	1.94	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	19
8	0.58	0.45	9.12	2,55E+03	2.34	1.91	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	20
8	0.16	0.69	4.67	3,09E+04	2.33	1.90	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	21
8	0.10	0.75	3.60	5,78E+04	2.42	1.95	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	22
8	0.08	0.77	3.13	7,71E+04	2.51	1.99	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	23
8	0.08	0.78	3.03	8,20E+04	2.56	2.01	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	24
8	0.02	0.88	1.59	2,08E+05	2.60	2.02	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	25
8	0.01	0.92	-1.02	1,00E+06	2.62	2.04	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	26
8	0.05	0.82	-2.35	1,00E+06	2.58	2.03	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	27
8	0.06	0.80	-2.57	1,00E+06	2.57	2.03	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	28
8	0.00	0.97	-0.41	1,00E+06	2.65	2.08	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	29
8	0.04	0.84	2.10	1,47E+05	2.74	2.14	0.04	0.03	M15,E78,E03,H90,F17,H40,I25,F90	30

Table 1. Best models found for each lag value. Columns: number of predictors in best model for each lag; F1 value for best model; significance F1 best model; F2 value for best model; significance F2 for best model; MAPE train error (%); MAPE test error (%); RMSE train error; RMSE test error; predictors for best model; LAG is the time horizon in days (τ)

of the previous lag, assuming a small increase in prediction error. It is not until the increase in error is significantly high that the model changes the predictors to better adapt to the variable to be predicted (the moment when the error decreases or remains stable, as observed in the figure). This behavior can be observed in Table 1, which shows the predictors selected for the best model at each lag.

5. Discussion

Results should be further validated in clinical practice. Most of these diagnoses are highly related to an aging population, as well as hypertension, which explains the possible underlying relation detected by the model. We can assert that the five variable identified as best predictors (M15, E78, E03, H90, F17) are highly correlated with the demand for hypertension diagnoses, and monitoring these variables could help predict trends in medical visits related to hypertension six to seven days in advance with a small forecasting error.

Having models with the ability to predict future demand would be key to improving various aspects of the healthcare system. On a macro level, the system could operate much more flexibly, anticipating the need for professionals in specific areas and times,

Figure 2. On-time hypertension test prediction model ($\tau = 0$). Y-axis: Normalized amount of hypertension diagnoses in 14 days.

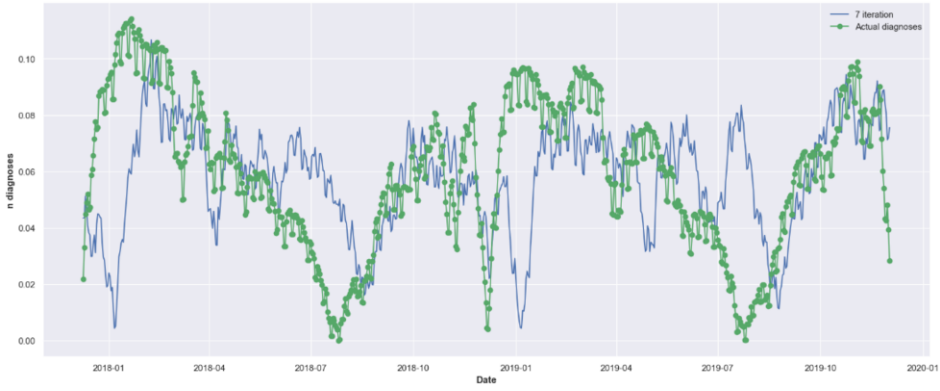


Figure 3. 30 days lagged hipertension test prediction model ($\tau = 30$). Y-axis: Normalized amount of hypertension diagnoses in 14 days.

maintaining health control over the population and also manage resources in a much more efficient way. These models would also be useful in defining disease indicators. On a micro level, healthcare professionals could have better control over their schedules by knowing what profiles they may encounter in the coming days. In a context of high demand like the one we are currently experiencing and that is expected in the future, tools like these can be key to achieving greater efficiency and flexibility in the system.

The model has strengths such as simplicity and interpretability. MLR models are widely used in the medical field and offer a high level of interpretation and acceptance by professionals. The methodology we propose can be used to select those variables most related to a target diagnosis through the feature selection method. The selection of such predictors could be a first step towards considering adding other variables related to visit diagnoses, such as medical prescriptions or pharmacological use, to improve prediction. To that end, other models like recurrent neural networks (RNN) or convolutional neural networks (CNN) could be explored.

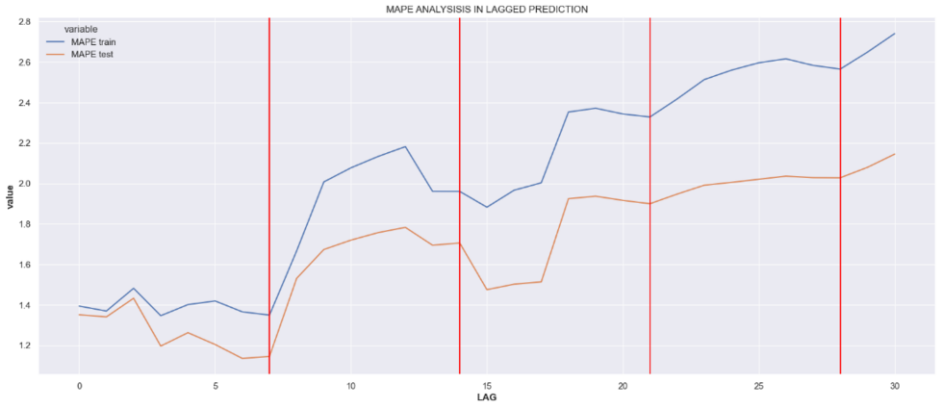


Figure 4. MAPE error lagging model.

In our work, we focus in in predicting hypertension that is one of the most common diagnostics. Despite predicting hipertension involves a quick and inexpensive test, our aim is to obtain the predictors and to understand how they behave in relation to the target variable, what their relationship is, and the time correlation that occurs between them. Moreover, other complex diagnoses as metabolic sindrome could be targered in a future.

6. Conclusions

We present in this paper a method to perform feature selection and forecast the demands of a given diagnosis (hypertension in this case) based on other diagnoses. The method is capable of indicating which predictors (other diagnostics) should be considered to predict future demand for the target, thereby increasing interpretability. In this manner, we obtain relevant information and greater explainability compared to other models, such as deep learning techniques.

The method has been tested with data of primary care services of Catalonia (ICS) in a cohort of 6,301,095 patients during years 2010-2019. The developed algorithm is capable of predicting the trend in demand for hypertension-related visits with an accuracy ranging from 98.6% to 97.86% when predicting one month in advance. Furthermore, the proposed algorithm is capable of indicating which variables are significant predictors for demand. Some of these variables have already been demostrated to have a causal relationship with hipertension in other related works [18].

Future work involves to use other variables, such as medical prescriptions, together with the predictors found; in that case, other machine learning forecasting tools, that could find some non-linearity on data, and eventually improve the forecasting results achieved by the MLR methods sould be explored.

Acknowledgements

This study was conducted with the support of the Generalitat de Catalunya, via the Industrial Doctorate Plan and via the Consolidated Research group 2021 SGR 01125.

References

[1] Lopez Segui F, Hernandez Guillamet G, Pifarré Arolas H, Marin-Gomez FX, Ruiz Comellas A, Ramirez Morros AM, et al. Characterization and Identification of Variations in Types of Primary Care Visits Before and During the COVID-19 Pandemic in Catalonia: Big Data Analysis Study. J Med Internet Res. 2021 Sep;23(9):e29622. Available from: https://www.jmir.org/2021/9/e29622.

[2] Garcia-Olive I, Lopez Seguí F, Hernandez Guillamet G, Vidal-Alaball J, Abad J, Rosell A. Impact of the COVID-19 pandemic on diagnosis of respiratory diseases in the Northern Metropolitan Area in Barcelona (Spain). Medicina Clínica. 2023;160(9):392-6. Available from: https://www.sciencedirect.com/science/article/pii/S0025775323000131.

[3] Upadhyay RK, Kumari N, Rao VSH. Modeling the spread of bird flu and predicting outbreak diversity. Nonlinear Analysis: Real World Applications. 2008;9(4):1638-48. Available from: https://www.sciencedirect.com/science/article/pii/S1468121807000879.

[4] Sil A, Kumar VN. Does weather affect the growth rate of COVID-19, a study to comprehend transmission dynamics on human health. Journal of Safety Science and Resilience. 2020;1(1):3-11. Available from: https://www.sciencedirect.com/science/article/pii/S2666449620300049.

[5] Sarkodie SA, Owusu PA. Impact of meteorological factors on COVID-19 pandemic: Evidence from top 20 countries with confirmed cases. Environmental Research. 2020;191:110101. Available from: https://www.sciencedirect.com/science/article/pii/S0013935120309981.

[6] Ahlawat A, Wiedensohler A, Mishra SK. An Overview on the Role of Relative Humidity in Airborne Transmission of SARS-CoV-2 in Indoor Environments. Aerosol and Air Quality Research. 2020;20(9):1856-61. Available from: https://doi.org/10.4209%2Faaqr.2020.06.0302.

[7] Towfiqul IAR, M H, Azad AAK, Roquia S, Zannat TF, Islam KS, et al. Effect of meteorological factors on COVID-19 cases in Bangladesh. Environment, Development and Sustainability. 2021;23(6):9139-62. Available from: https://doi.org/10.4209%2Faaqr.2020.06.0302.

[8] Morató JP, Pelegrí JL, Rey MM, Abelló AO, Vallès X, Roca J, et al. Environmental predictors of SARS-CoV-2 infection incidence in Catalonia (northwestern Mediterranean). Research Square. 2022. Available from: https://doi.org/10.21203/rs.3.rs-2206639/v1.

[9] Wang WW, Li H, Cui L, Hong X, Yan Z. Predicting Clinical Visits Using Recurrent Neural Networks and Demographic Information. 2018:353-8.

[10] Arielle S, Drake A, Emily G, L WT, Benson H, Cheryl W. Predicting unplanned medical visits among patients with diabetes: translation from machine learning to clinical implementation. BMC Med Inform Decis Mak. 2021;31;21(1):111.

[11] López B, Torrent-Fontbona F, Roman J, Inoriza JM. Forecasting of emergency department attendances in a tourist region with an operational time horizon. 2021. Available from: https://dugi-doc.udg.edu/handle/10256/19433.

[12] Spencer RJ, Amer S, George EJS. A retrospective analysis of emergency referrals and admissions to a regional neurosurgical centre 2016–2018. British Journal of Neurosurgery. 2020. Available from: https://pubmed.ncbi.nlm.nih.gov/33292027/.

[13] World Health Organization. ICD-10 : international statistical classification of diseases and related health problems: tenth revision, 2nd ed. World Health Organization https://appswhoint/iris/handle/10665/42980. 2004.

[14] Freedman D, Pisani R, Purves R. Statistics (international student edition), 4th edn. WW Norton & Company, New York. 2007.

[15] Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis - a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Quality and Quantity. 2018 Jul;52(4):1957–1976.

[16] Jr FJM. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association. 1951;46(253):68-78. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769.

[17] White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica. 1980;48(4):817-38. Available from: http://www.jstor.org/stable/1912934.

[18] Guillamet GH, Seguí FL, Vidal-Alaball J, López B. CauRuler: Causal irredundant association rule miner for complex patient trajectory modelling. Computers in Biology and Medicine. 2023;155:106636. Available from: https://www.sciencedirect.com/science/article/pii/S0010482523001014.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230683

Challenges in the Exploitation of Historical Clinical Data for the Classification of Diabetic Retinopathy Patients

Jordi PASCUAL-FONTANILLES [a,1] Aida VALLS [a,c] Antonio MORENO [a,c]
Pedro ROMERO-AROCA [b,c]

[a] *ITAKA, Dept. Enginyeria Informàtica i Matemàtiques*
Universitat Rovira i Virgili, Tarragona, Catalonia, Spain
[b] *Servei d'Oftalmologia, Hospital Universitari Sant Joan de Reus, Catalonia, Spain*
[c] *Institut d'Investigació Sanitària Pere Virgili, Tarragona, Catalonia, Spain*
ORCiD ID: Jordi Pascual-Fontanilles https://orcid.org/0000-0002-7528-5819, Aida Valls https://orcid.org/0000-0003-3616-7809, Antonio Moreno https://orcid.org/0000-0003-3945-2314, Pedro Romero-Aroca https://orcid.org/0000-0002-7061-8987

Abstract.
Analysing medical data stored in Electronic Health Records is of great interest to build clinical decision support systems. There is a lot of hidden knowledge in these databases, obtained from the continuous work of medical practitioners when attending and diagnosing patients. However, it is not easy for Machine Learning methods to exploit these data. Their use always requires a careful and complex pre-processing stage. In this paper, we study the case of diagnosis of diabetic retinopathy, which is made by ophthalmologists. The characteristics of a 12-year dataset about Type-2 diabetic people are analysed. Several numerical and categorical variables were selected by experts as relevant risk factors for this disease. We explain the challenges that are being faced in order to generate a dataset composed by time series with the same length and intervals. The final aim of the research is to build a clinical decision support system that can make a personalised prediction of the evolution of the disease.

Keywords. Time series classification, Clinical decision support systems, Classification, Diabetic retinopathy

1. Introduction

This short paper analyses the characteristics of a real dataset about Type-2 diabetic people. The dataset contains different clinical and analytical variables of different types (numerical and categorical), which have been extracted from the Electronic Health Records (EHR) of diabetic patients. It contains data collected between 2010 and 2021. The final goal of our research project is to construct a classifier that distinguishes between several

[1] Corresponding Author. Email: jordi.pascual@urv.cat.

levels of the diabetic given the historical data collected in his/her successive visits. The purpose of this short paper is to present the complexities of the available temporal data, which will guide the pre-processing stage of data preparation, which is usually a crucial step to build proper classifiers.

2. Diabetic retinopathy classification

Diabetic retinopathy (DR) is an ocular complication produced by the increase of blood sugar levels due to diabetes, which damage the back of the eye (retina). The lesions it can produce cause vision loss and even blindness if not detected and treated at an early stage [1]. According to the medical ETDRS standard classification [2], the following categories are considered: healthy ($DR = 0$), mild ($DR = 1$), moderate ($DR = 2$) and severe ($DR = 3$). In 2021, the number of diabetic patients worldwide was estimated at 537 million people. In Spain it is expected to affect about 11.1% people by 2030, reaching 3.8 million inhabitants [3]. The prevalence of DR among type-2 diabetic people in Spain in 2022 was 15.28%, including 1.92% for the severe level ($DR = 3$) [1].

Computer methods to assist in DR diagnosis are being developed. Many works detect DR with the analysis of eye fundus images. Because they are costly to obtain, other clinical decision support systems (CDSS) based solely on EHR data are nowadays on the focus of research. Machine learning methods are trained with EHR to build DR classifiers. Several studies on the literature analyse how different kinds of classifiers perform on EHR-based DR datasets [4,5,6,7,8,9]. They use techniques such as random forests, XGBoost, logistic regression, support vector machines or k-nearest neighbours. There is not a consensus in the selection of the best classifier, as there are varied results on each study. In [10][11], Retiprogram was presented as a CDSS to assess Type-2 Diabetes Mellitus patients' risk of developing DR. It takes into account the current patient's conditions, including the analytical data from the last blood analysis. It is an ensemble method consisting on a Fuzzy Random Forest classifier, with an accuracy of 81%, sensitivity of 80%, and specificity over 84%. First, it was developed as a binary CDSS [11]. Later, it was extended to deal with the ordinal multi-class case, being able to detect the levels of DR severity [12].

3. Challenges of the DR Data Available in Electronic Health Records

The goal of our research is to improve DR severity detection by means of a time series classifier. Therefore, we need to work with multi-variate series data representing the changes of several clinical and analytical risk factors, stored in the EHR.

For this research, we have obtained a dataset with 231,064 diabetic patients from Catalonia (Spain), with their medical data from 2010 to 2021. Figure 1 depicts the frequency of patient's visits by ophthalmologists. Only patients with at least 2 visits are considered. It can be observed that patients are usually visited with a frequency over 18 months; therefore, the diagnosis must be made with relatively short sequences of data.

Usual machine learning techniques for time series classification assumes a dataset with a homogeneous time series (i.e. with equal length and intervals). However, the data extracted from EHR is not homogeneous. Considering the 12-year dataset on diabetic patients, we present here a list of challenges that need to be tackled:

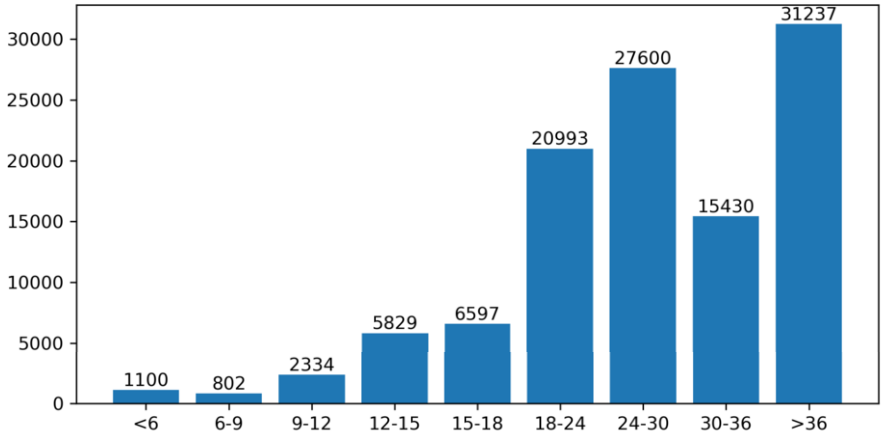


Figure 1. Diabetic patients visits mean frequency in months

- **Short sequences**: around half of the patients have a single visit on their EHR, and must be discarded. Moreover, just 1% of them have at least 6 visits in 12 years, which we consider the minimum amount of information needed to construct a reliable temporal sequence for a patient. Thus, from the whole dataset, only 2108 patients could be considered for training and testing. Moreover, the length of the sequences is still unusually short for time series classifiers, which usually manage long samples.
- **Irregular visits frequency**: most time series algorithms require data spaced at regular time intervals. This is not the case here, where we can find two situations when establishing a period of 1 year. First, we can find a few cases of patients having multiple records for the same period. Second, in some sequences there is a lack of a visit in a certain year, whose data should be interpolated to fill in the gap.
- **Different data alignment**: all the time sequences should be aligned on certain time points with the same number of occurrences. However, the time span for different patients varies, as it depends on the date of diagnosis of diabetes. For the trends analysis, we need a sequence of the same number of consecutive years, even if they do not start and end at the same date.
- **Missing data**: the values of the risk factors must be extracted from the EHR for each time point. These values may be stored in different sources (blood test reports, doctor's notes, etc.) and may have been recorded at different dates. Thus, each patients' visit cannot be found in the EHR as a single element, and has to be reconstructed from the multiple sources that are considered on a visit. If any of those values is missing, the visit data cannot be included in the time series, unless we apply some imputation method.
- **Labelling mistakes**: some mistakes may be found on the diagnosis data at EHR, due to human errors. For instance, we found that for some patients the level of DR decreases over time, which is not a common situation. Similarly, in the case of treatment type, we found some cases in which no medication was reported for a patient treated with insulin before, which is very unusual. Therefore, data values must be carefully curated taking into account medical expertise before being used.

- **Data imbalance**: the DR dataset is highly imbalanced towards the negative class. This is expected due to the low incidence of this disease. The patients distribution is: 1730 patients in $DR = 0$ (82.1%), 209 in $DR = 1$ (9.9%), 133 in $DR = 2$ (6.3%) and 36 in $DR = 3$ (1.7%). Consequently, the availability of examples with a high degree of DR is scarce. Because of this over-representation, the classification models have more difficulty to correctly identify and distinguish the positive classes. Some oversampling technique could be used to minimise this problem.

The presented challenges require tailored solutions that take into account the medical knowledge related to the DR disease. Time series classification methods could only be used if satisfactory pre-processing techniques are applied on the raw ERH data. These preparation methods are our next line of work. Moreover, these challenges should also be analysed for other diseases diagnoses that have a temporal evolution.

Acknowledgements

This study is funded by Instituto de Salud Carlos III and the European Union with project PI21/00064, and by URV projects 2022PFR-URV-41 and 2021PFR-B2-103. The first author has a pre-doctoral grant (2022 FI_B1 00036) from Generalitat de Catalunya and Fons Social Europeu.

References

[1] Romero-Aroca P, López-Galvez M, Martinez-Brocca MA, Pareja-Ríos A, Artola S, Franch-Nadal J, et al. Changes in the Epidemiology of Diabetic Retinopathy in Spain: A Systematic Review and Meta-Analysis. Healthcare (Switzerland). 2022 7;10:1318.

[2] Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110:1677-82.

[3] Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes research and clinical practice. 2010;87(1):4-14.

[4] Zhao Y, Li X, Li S, Dong M, Yu H, Zhang M, et al. Using Machine Learning Techniques to Develop Risk Prediction Models for the Risk of Incident Diabetic Retinopathy Among Patients With Type 2 Diabetes Mellitus: A Cohort Study. Frontiers in Endocrinology. 2022 5;13:885.

[5] Tsao HY, Chan PY, Su ECY. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. BMC Bioinformatics. 2018 8;19:111-21.

[6] Ogunyemi OI, Gandhi M, Tayek C. Predictive Models for Diabetic Retinopathy from Non-Image Teleretinal Screening Data. AMIA Summits on Translational Science Proceedings. 2019;2019:472.

[7] Sun Y, Zhang D. Diagnosis and Analysis of Diabetic Retinopathy Based on Electronic Health Records. IEEE Access. 2019;7:86115-20.

[8] Ogunyemi O, Kermah D. Machine Learning Approaches for Detecting Diabetic Retinopathy from Clinical and Public Health Records. AMIA Annual Symposium Proceedings. 2015;2015:983.

[9] Reddy GT, Bhattacharya S, Ramakrishnan SS, Chowdhary CL, Hakak S, Kaluri R, et al. An Ensemble based Machine Learning model for Diabetic Retinopathy Classification. International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020. 2020 2.

[10] Romero-Aroca P, Valls A, Moreno A, Sagarra-Alamo R, Basora-Gallisa J, Saleh E, et al. A Clinical Decision Support System for Diabetic Retinopathy Screening: Creating a Clinical Support Application. Telemedicine and e-Health. 2019;25:31-40.

[11] Saleh E, Błaszczyński J, Moreno A, Valls A, Romero-Aroca P, de la Riva-Fernández S, et al. Learning ensemble classifiers for diabetic retinopathy assessment. AI in Medicine. 2018;85:50-63.

[12] Pascual-Fontanilles J, Lhotska L, Moreno A, Valls A. Adapting a Fuzzy Random Forest for Ordinal Multi-Class Classification. Frontiers in Artificial Intelligence and Applications. 2022 10;356:181-90.

This page intentionally left blank

Part 4

AI for Good, Ethics, Society, Serious Games

This page intentionally left blank

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230685

A General Model for the Metainformation of Complex Questionnaires for Automatic Preprocessing and Reporting Under INSESS Methodology

Xavier ANGERRI [1] and Karina GIBERT
Intelligent Data Science and Artificial Intelligence Research Center
Universitat Politècnica de Catalunya—BarcelonaTech

ORCiD ID: Karina Gibert https://orcid.org/0000-0002-8542-3509

Abstract. In the field of analytics and intelligent data analytics one of the open challenges regards to the metainformation model that allows an intelligent reporting system to preprocess and process correctly data coming from any kind of questionnaire, independently of number of variables, types of variables involved, including multiple response variables or other complex structures like compositional variables or temporal variables among others. INSESS is an intelligent system designed and used for the first time in INSESS-COVID19 project, specialized in automatic intelligent reporting for a parameterized questionnaire that includes different types of variables. From one INSESS-consultation to another, the questionnaire can change completely and the consultation may be oriented as well to different types of target populations going from citizens, to associates to an NGO or to special nonpublic populations and the pre and processing still works. The flexibility required to do this, relies on a well-founded and sufficiently general conceptual model for the metainformation associated to the questionnaire, so that the analytical module can properly reason on top of the metainformation and make intelligent decisions about the kind of preprocessing required or the kind of descriptive analysis to be applied. The metainformation model has been developed from a knowledge representation perspective.The paper presents the context of the research, the conceptual model defined and 4 real use cases where the INSESS methodology was used for different kinds of private/public consultations proving the flexibility of the tool.

Keywords. Metainformation, Preprocessing, questionnaire, INSESS, conceptual model

1. Introduction

In the field of analytics and intelligent data analytics one of the open challenges regards to the metainformation model that allows an intelligent reporting system to preprocess

[1] Corresponding Author: Xavier Angerri, xavier.angerri@upc.edu.

and process correctly data coming from any kind of questionnaire, independently of number of variables, types of variables involved, including multiple response variables or other complex structures like compositional variables or temporal variables...
INSESS is an intelligent system designed and used for the first time in INSESS-COVID19 project (see results in [1]), specialized in automatic intelligent reporting for a parameterized questionnaire that includes different types. From one INSESS-consultation to another, the questionnaire can completely change and the consultation may be oriented to different types of target populations going from citizens, to associates or NGO or to special non-public populations, and the preprocessing and processing of the data should still work. In this paper 4 real applications are assessed: a) INSESS-COVID19 project on social vulnerability of lockdown, b) Popular culture association (Colla castellera-Human towers) c) Maturity of Digital Transformation in European Third Sector Entities and d) Energy consumption in BCN households. The flexibility required to do this, relies on a well-founded and sufficiently general conceptual model for the metainformation associated to the questionnaire, so that the analytical module can properly reason on top of the metainformation and make intelligent decisions about the kind of preprocessing required or a certain kind of variable and the kind of descriptive analysis to be applied. The metainformation model has been developed from a knowledge representation perspective. The main novelty of this paper is the formalization of an abstract metainformation model, and the associated rule based reasoner module in charge of making the automatic reporting for any kind of questionnaire. The existence of this abstract metainformation model makes possible to process new questionnaires without changing a single code line of the INSESS system.
The paper presents in section 2 the context of the research and the state of the art relatd with automatic reporting and the metadata file. In section 3 is defined the conceptual model with all the steps presented. In section 4 real use cases where the INSESS methodology was used for different kinds of private/public consultations proving the flexibility of the tool. Section 5 is dedicated to show how the results appear in the developed automatic report. Conclusions are in section 6.

2. State of the art

2.1. Metainformation model

The research on the metainformation structure is an open problem. The amount of data is increasing every day, and managing it is complex. Thus, the research in the metadata field is based on creating software and technologies to store all the metainformation of a big database, like in [2] or [3], where a methodology is described for implementing an extensible tool for editing the metainformation by the domain expert.

Moreover, authors are not aware on other research on the relationship between the metainformation model and automatic reporting.

2.2. Automatic reporting

In general, papers were reporting methodology is presented like in [4] are not showing any automatic reporting tool because the field of automatic reporting is more linked to real practice than to academic and scientific publications. Nevertheless [5] and [6]

published their research on the added value of automatic reporting in the fields of medical image processing and [7] [8] in very specific and concrete applications.

3. Methodological Proposal

The main contribution of this paper consists in a general metadata model that can be used with any associated dataset, which is machine readable and allows automatic intelligent reporting. To develop this technology, the information of the variables is described and formalized to be read by an R script, which also imports the original database. With both imported files, several algorithms are developed to create an automatic Microsoft Office Word Data Report using RMarkdown technology.

Nevertheless, the main contribution of this paper is that this structure is exportable to any other software that can be inputed to the analytical reasoner. The constructs formalized in S3.3 have to be instanciated for each real questionnaire and it will work for all consultations using the same questionnaire.

3.1. The framework project INSESS-COVID19

In March 2020 the COVID19 pandemic was declared and the majority of the states decided to order several months of lockdown to all population. Many people were dying and the hospitals were collapsed. At that moment most of the research done in the field of COVID19 was focused on the prediction of the infection rates, propagation of the disease, survival rates; indeed, most of the research in COVID19 topics is done under a medical approach. However, the project INSESS-COVID19 was focused on Social Services, largely forgotten in the management of the pandemic, although being a field with a strong need of including data as an asset for management and improvement of the Social Services system itself as well as for improvement of services to citizens.

To obtain Social Services user's status, was designed a questionnaire to collect data directly from them, and to get them available for analysis in real-time. The INSESS-COVID19 questionnaire contains 195 questions distributed in 20 different blocks related with basic social aspects in life. In [9] the selected methodology and the main results of the survey are published and in [10] some of the analysis techniques are published.

The project started in April 2020, the first Social Services users answered the questionnaire in July 2020 and the last ones by December 6, 2020. The results of the project were described in a general report. It was presented to the Catalan Government by December 15, 2020 in a public event only 9 days after data collection closed, thanks to the powerful Artificial Intelligence technology developed which uses the metainformation methodology explained in this paper.

3.2. Variable types

To create a good metadata file it is very important to properly model and formalize the several variable types present in data. In [9] some variable types were originally defined. Our proposal is designed to support datasets with numerical, Likert, ordinal, nominal variables, open questions, which are quite common, plus some special types:

- **Multivalued variables:** A multivalued variable is a qualitative variable X with S possible modalities D = {m_1, m_2,…,m_S} so that an individual can simultaneously take several values from D.
- **Grid Variable:** Grid Variable (X,Q) is a qualitative variable X with S modalitites in D. For each modality $m_s \in D$, a value (from Q) indicating the qualification of m_s is given (Q is a Likert or ordinal set of values)
- **Temporal basic variables (TBV):** Temporal variable (X,T) defines X as a qualitative variable (nominal, ordinal, binary or Likert) with S modalities in D that is measured n_T times along time, providing n_T columns in the dataset as temporal replicas of X. We can denote these replicas as X_{t1}, X_{t2}, …, X_T.
- **Multivalued Grid Variable:** (X,Q) is a qualitative multivalued variable X with S modalities in D, replicated at every value of Q (Q might also be several timestamps). For each modality $m_s \in D$, and value of Q a binary column m_{sq} indicates the presence/absence of m_s in group q.
- **TQQ variable:** Temporal Qualified Qualitative (X, T, Q) is a qualitative variable X with S modalities in D, replicated n_T times. For each modality m_s, a value (from Q) indicates the qualification of m_s (Q being Likert or ordinal set).

3.3. Metainformation model

To make possible the automatic analysis of data, the conceptual metainformation model has been designed with all the necessary knowledge to generate the automatic report and having the original database in a separate structure. The Metadata Model (namely MdM) is a datamatrix containing as many rows as necessary, with a minimum 1 row per question for numerical and open questions. For qualitative variables with S modalities, the number of rows are 1+S. For TQQ and Grid Variables the number of rows is: 1+max(Q,S). This are the 16 columns in the table. Some of the rows in MdM describe variables. Other describe the modalities of qualitative variables. The model combines several constructs:
(k, $\tilde{k}$, B, ℓ_B, ℓ^*_B, O=Bloc); (k, ℓ^*_B, , it_Q, $\mathbb{Q}$, τ, $O = question$); ($\mathbb{Q}$, A, A*, O=modality); ($\mathbb{Q}$, τ, J, J*, O=modality, k_o, k_f, n_c) , τ *in* {TQQ, TBV, Grid}
It can be implemented in R or any other support like DataBases management system:

- k (Col): Number of column where the variable (X_k) is in the original dataset ($\mathcal{I}$ is a set of individuals $\mathcal{I} = \{i_1 \dots i_n\}$ described by K variables {X_1,X_2,…,X_k,…X_K}). This column is not fulfilled in rows where the modality for a categorical question is described. For TQQ variables k is the position of the columns in the dataset containing the modalities.
- $\tilde{k}$ (Eticol): This parameter is only filled in the MdM rows referring modalities. For these rows $\tilde{k} = k$ It indicating the position of qualitative variable X_k.
- B (Block): In case the questionnaire is designed by thematic blocks the block number is specified. Each block with several variables of the same topic.
- ℓ_B: Specifies the name of the block
- ℓ^*_B : Short label for the block to be used in the report
- it_Q : index of the Question

- $\mathbb{Q}$ (Question): Complete text of the question as it is appearing in the original questionnaire. We call r_k the row of the MdM where is described.
- A (Answers): If X_k is qualitative, rows $r \in \{r_k + 1, \ldots, r_k + S\}$ are used to describe $s \in D_k$ the modalities of X_k
- J: The default value is "inexistent". It is only used for Grid or TQQ variables. The values of corresponding Q are described in rows $r \in \{r_k + 1, \ldots, r_k + S\}$.
- A* (Rephrasing): A short expression for both questions and possible answers to be used in the statistical tables and graphs, as the long texts will overlap and make reading difficult
- J* (Colcurt): Short label for Q values in TQQ and Grid variables. In case the question is not a TQQ neither grid is not fulfilled.
- O (Object): Indicates the kind of information represented in the row: "question", "modality", "name of Block" or separator (used between variables)
- τ (Type of variable): Gives the type of variable represented in $\mathbb{Q}$ (only in rows with O≠missing): "Numerical"; "Simple response" for qualitative or Likerts; "Multivalued variable" for Multivalued (Nominal, Ordinal or Likert); "grid" for TQQ, Temporal basic and Grid Variables;;; "Open" for textual answers;.
- τ_D (Descriptive type): Identifier of the descriptive procedure associated (some used for more than one type of variable). Each descriptive procedure is a rule-based function generating specific combination of numerical and graphical descriptive tools, some specifically designed in this work. τ_D is specified only when τ is defined.

Table 1 Number of variables per project

Type	Descriptive kit	Type	Descriptive kit
Numerical	Numerical	TBV	NomxT, LikertxT
Simple response	Nominal, Ordinal, Likert	Multivalued Grid	NomMVxNom + TQQ
Multivalued variable	NomMV OrdMV, LikertMV	TQQ	NomxOrdxT
Grid	NomxLikert, OrdxLikert + TBV	Open	NomxOrdxT

- k_0 (PackIni): For the row r where $\tau_r \in$ {"Multiple response", "Grid", "TBV", "Multivalued Grid", "TQQ"}, $k_0 = k_{r-1_f} + 1$ indicates the column in the database where the information about this variable $\tilde{k}_r$ starts (basically one column after the previous variable ending position).
- k_f (PackFi): For the row r where $\tau_r \in$ {"Multiple response", "Grid", "TBV", "Multivalued Grid", "TQQ"}, k_f indicates the last column containing information about variable $\tilde{k}_r$. See Table 3 for the values of k_f.
- n_C(nCols): Is the number of possible modalities in the cells of the data structure associated with the variable, sometimes correspond to number of modalities of Q or T or X depending on the type of variables. Only available for variables with $\tau_r \in$ {"Multiple response", "Grid", "TBV", "Multivalued Grid", "TQQ"}.

Table 2 Value of k_f and n_c per variable type

Type	k_f	n_C
Multivalued variable	$k_f = k_0 + S - 1$	2
Grid	$k_f = k_0 + S - 1$	n_Q
TBV	$k_f = k_0 + n_T - 1$	S
Multivalued Grid	$k_f = k_0 + S \cdot n_q - 1$	2
TQQ	$k_f = k_0 + S \cdot n_T - 1$	n_Q

3.4. Preprocessing step

Once the questionnaire has been created together with the corresponding metadata file, the preprocessing step can be addressed. In this work, the preprocessing step is implemented by R, where techniques specified in [11] are applied like missing imputation or recodification. The preprocessing step has five main steps:

- Error detection and correction: where errors in the database are detected and treated using the appropriate technique, as shown in [11].
- Missing Data Treatment: Those questions that are answered by all participants, contain the modality "No Answer" in qualitative questions and it remains as a additional modality. For filter questions the cells are initialized to "Filter"; thus no missings are generated for the subgroups that skip the question.
- Duplicities: Google forms generates an automatic data-time variable with timestamp of the answers. In case 2 repeated rows in same timestamps, redundance is eliminated. In case of variables with correlation 1, or redundant, one of those is also eliminated.
- Rephrasing: The original Variable and modalities names might be too long in many real applications to allow correct labelling of the plot's titles, axis, variable or individual labels. Using the information from MdM given in $\tilde{k}$, $\mathbb{Q}$, A* and J* questions, variable names and modalities are recoded into short forms understandable and compatible with the graphical representation. For those variables represented in more than one column of the dataset, like TQQ questions or Grid Variables, a specific process is used considering the n_D columns involved.. Type of variable column in metadata file determines the appropriate algorithm to rephrase each question.
- PreprocessingValidation: To validate that the rephrasing has been done correctly contingency tables are built between original and preprocessed variables. The result should be like a diagonal matrix, validating that all modalities from the original variable correspond to its new short name.

3.5. Reporting step

The preprocessed database contains short variable and modality names, no missing data, and no duplicates. This is the input for the automatic reporting process, that produces a large document containing the descriptive statistics and some inferential indicators for each variable). An R script is able to read the preprocessed database and apply a different treatment to each variable according to its type and descriptive kit. For each variable the corresponding code is created in real time by processing the metainformation given in MdM. For example, for TQQ variables. A trajectory map is done visualizing the evolution of qualification of each ms modality along time (the corresponding tables of effectives and proportions is also shown), a bivariate multiple barplot displays distribution of Q vs T and the corresponding cross table T x Q (with conditional frequencies and effectives is associated), a grid of pies (TxQ), and transition tables (cross tables of Q in t_{-1} vs t) in effectives and proportions to see how the qualification of ms modified along time, step by step and globally along the entire sequence t_1-t_n.

In $\tilde{k}$, τ, k_0, k_f, *Eticols,* $\mathbb{Q}$, A, A*,J, J* is appearing the information to build the report using the developed software.

4. Applications

The methodology has been applied in four real cases:

- INSESS-COVID19 (InCov19):The first INSESS-COVID19 consultation got data about social vulnerabilities raised as a consequence of the first lockdown due to COVID19. Data came from 971 vulnerable citizens and results were delivered to Catalan Social Services Department to improve decision making in COVID19 management. The results were presented to the Catalan government in [1] and the main context is explained in section 2.1 of this paper.
- Colla Castellera Jove de Barcelona annual surveys (CCJB21, CCJB22): *Colla Castellera Jove de Barcelona*, is a non-lucrative entity belonging to the Catalan tradition called *Castellers*. (A group who build human tower raised without any type of mechanical aid [12]). The association conducts a yearly opinion survey among their associates and the results are presented in the General Assembly.
- Kids IT inspirational activity (kids): A specific inspirational activity is designed to raise interest in Data Mining to small kids. Initially a female computer engineer visits a class from primary school and as part of the workshop she proposes an activity. The kids has 7 days to collect some data and deliver in a collaborative data file. A few days later, the same scientist come back to the school and share with the kids the results of analysing collected data. In 2022 the analysis was done with INSESS technology and kids could learn about the most electricity consuming tool in their homes.[13]
- DIMCARE (DCE): It is an European project for understanding the digitalization degree of the people in care sector [14]. A questionnaire designed in the project and distributed among all Third Sector entities of three pilot cities (Mataró, Prato, Varaždin) were received and analyzed using INSESS tools.

Table 3 is summarizes the main characteristics of each dataset:

Table 3. Number of variables per project

	InCov19	**CCJB21**	**CCJB22**	**kids**	**DCE**
Columns	195	30	47	32	40
Individuals	971	61	71	175	48
Numerical variables	5	0	0	12	0
Categorical variables	38	6	8	15	23
Multivalued variables	3	0	0	0	7
Grid variables	6	5	8	0	2
TBV	13	2	0	0	0
Multivariate Grid	0	0	0	0	
TQQ variables	4	0	0	0	0
Pages of the resulting report	320	26	41	46	61
Time to obtain the result	30 mins.	5 min	6 min	5 min.	5 min
Data colletion start	01-07-20	15-12-21	15-12-22	23-3-22	external
Data colletion finish	06-12-20	31-12-21	31-12-22	27-3-22	11-01-2023
Report delivered	15-12-20	15-01-22	15-02-23	23-5-22	18-01-2023

5. Results

After applying this methodology, automatic reporting generates a Word document, with a complete layout and all numerical and graphical information associated to simple and complex data types described (fig1).

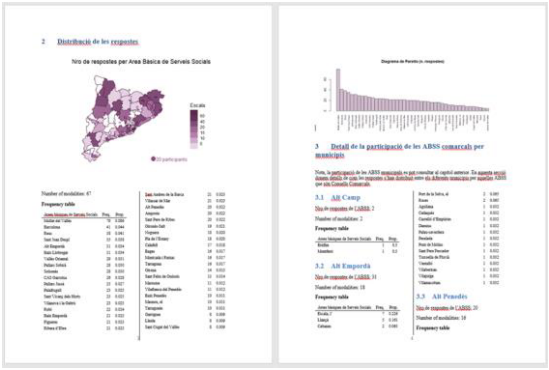


Figure 1. Automatic Report Generated

As said before, the analyses is adapted to the type of variable and the R procedure finds relevant information in dataset according to all information represented in MdM and some rules to proceed in each case.

5.1. Numerical Variables

For numerical variables, a classical descriptive analysis is done, with: Boxplot, histogram, five-number summary and specific evaluation of sample error analysis

5.2. Categorical variables

For categorical variables, a traditional analysis is done with: barplot(Fig 2), pie chart, frequency table. Specific procedures are used to compute the positions of the bars inside a page with the number of modalities, automatic labelling of data..... etc

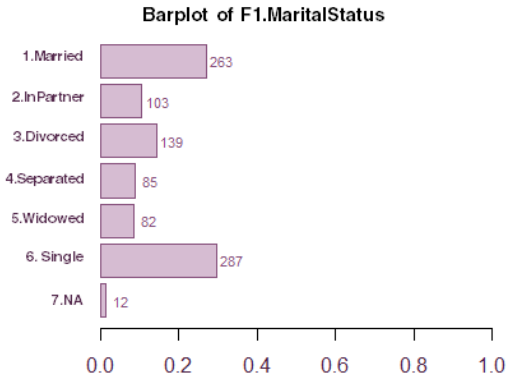


Figure 2. Barplot for categorical variables description

5.3. Multivalued variables

Multivalued variables use : marginal barplot, marginal piechart and frequency table and multivalued frequency table.

5.4. Grid Variables

A multiple barplot is built to describe this 2 dimension variables and a grid of pies. An absolute frequency cross table, a relative frequency cross table

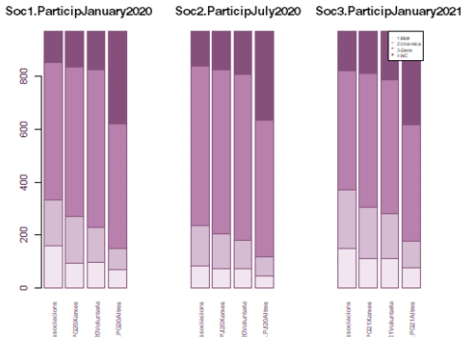


Figure 3 .Multiple barplot

5.5. Temporal basic variables

For this type of variables has been developed a trajectory map (Figure 3) where each individual represents one line and its trajectory along the time is shown in this plot. It is presented in [7]. For this type of variables also is shown a Trajectory Frequency Table a multiple barplot, a grid of pies and transition Tables.

5.6. TQQ variables

A multiple stacked plot (Figure 4) is built to describe this 3 dimension variables. Several multiple barplots, several grid of pies, several cross tables with counts and proportions are appearing.

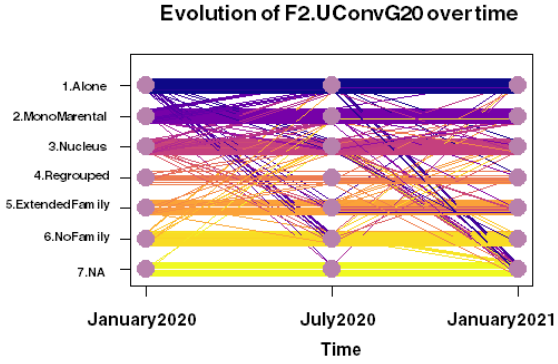


Figure 4. Trajectory map

6. Conclusions

Thanks to INSESS methodology the structure of the metainformation was crucial for the developed intelligent reporting system. This paper contributes to the field with an open

metainformation structure that could be applied with any software or programing language capable to read, manage and describe databases.

This new structure allows to make a fast preprocessing step and an automatic report is built, where several tools are used depending on the type of variable, which is also indicated in the metainformation. Moreover, as the MdM is instanciated a priori, together with the questionnaire, and the proposed INSESS technology is flexible to any pair dataset+MdM, getting final results from data becomes an accelerated process as the first draft report is obtained in very few minutes after closing the data collection process. Thus, the time between data collection completion (or the closing of the questionnaire) and the publication of the results extremely short (see table 3), making INSESS a very powerfull tool to support complex decision making based on data, with especial added value for those decisions requiring information that is not available in the systemic databases and require direct consultation to target population (vulnerable population, associates of an NGO, third sector entities, households from a certain district…. among others).

Currently, we are working in more complex structures for variables like those recording the number of members in a family (numerical) with a certain degree of dependency (ordinal or Likert) and distributed per age groups (ordinal) among others. A deep comparison between INSESS methodology and alternative approaches is being in progress.

References

[1] GIBERT, Karina; CODINA, Toni; ANGERRI TORREDEFLOT, Xavier. Informe INSESS-COVID19: identificació de necessitats socials emergents com a conseqüència de la Covid19 i efecte sobre els serveis socials del territori. 2020.

[2] DAI, Hao, et al. The State of the Art of Metadata Managements in Large-Scale Distributed File Systems—Scalability, Performance and Availability. IEEE Transactions on Parallel and Distributed Systems, 2022, 33.12: 3850-3869. doi: 10.1109/TPDS.2022.3170574

[3] GRIBOVA, V. V., et al. Implementation of a model of a metainformation-controlled editor of information units with a complex structure. Automatic Documentation and Mathematical Linguistics, 2016, 50: 14-25. doi: 10.3103/S0005105516010052

[4] BOYNTON, Petra M.; GREENHALGH, Trisha. Selecting, designing, and developing your questionnaire. Bmj, 2004, 328.7451: 1312-1315.

[5] MESSINA, Pablo, et al. A survey on deep learning and explainability for automatic report generation from medical images. ACM Computing Surveys (CSUR), 2022, 54.10s: 1-40. doi: 10.1145/3522747

[6] KAUR, Navdeep; MITTAL, Ajay; SINGH, Gurprem. Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey. Multimedia Tools and Applications, 2022, 81.10: 13409-13439. Doi: doi.org/10.1007/s11042-021-11272-6

[7] MATHEW, Timothy H. Chronic kidney disease and automatic reporting of estimated glomerular filtration rate: a position statement. The Medical Journal of Australia, 2005, 183.3: 138-141. doi: 10.5694/j.1326-5377.2005.tb06958.x

[8] LEI, Yujiao, et al. Research of Automatic Generation for Engineering Geological Survey Reports Based on a Four-Dimensional Dynamic Template. ISPRS International Journal of Geo-Information, 2020, 9.9: 496.doi: 10.3390/ijgi9090496

[9] GIBERT, Karina; ANGERRI, Xavier. The INSESS-COVID19 Project. Evaluating the impact of the COVID19 in social vulnerability while preserving privacy of participants from minority subpopulations. Applied Sciences, 2021, 11.7: 3110. doi: 10.3390/app11073110

[10] ANGERRI, X.; GIBERT, K. Preprocessing and Artificial Intelligence for Increasing Explainability in Mental Health. International Journal on Artificial Intelligence Tools, 2023., 32.2 doi: 10.1142/S0218213023400110

[11] GIBERT, Karina; SÀNCHEZ–MARRÈ, Miquel; IZQUIERDO, Joaquín. A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. AI Communications, 2016, 29.6: 627-663. doi: 10.3233/AIC-160710

[12] CATALUNYA.COM. *Castells* [online]. [Accessed 26 April 2023]. Available from: https://www.catalunya.com/castells-1-3-12567?language=en

[13] Web d'fcri. *100tifiques.cat* [online]. [Accessed 26 April 2023]. Available from: https://100tifiques.cat/

[14] DIMCARE - digital missions for Care Social Economy's resilience. *Yunus Social Business Centre University of Florence* [online]. 25 January 2023. [Accessed 26 April 2023]. Available from: http://sbflorence.org/en/dimcare-social-economy/

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230686

Respect for Autonomy in the Machine Learning Pipeline

Paula SUBÍAS-BELTRÁN [a,b,1], Oriol PUJOL [c], Itziar DE LECUONA [b,d], and Vicent RIBAS RIPOLL [a]

[a] *Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain*
[b] *Bioethics and Law Observatory - UNESCO Chair in Bioethics, Universitat de Barcelona, Barcelona, Spain*
[c] *Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain*
[d] *Dept. of Medicine, Universitat de Barcelona, Barcelona, Spain*
ORCiD ID: Paula Subías-Beltrán https://orcid.org/0000-0003-1167-1259, Oriol Pujol https://orcid.org/0000-0001-7573-009X, Itziar de Lecuona https://orcid.org/0000-0002-5081-5756, Vicent Ribas Ripoll https://orcid.org/0000-0002-7266-6106

Abstract. The machine learning (ML) community recognizes the potential impact of ML systems on human rights, especially regarding privacy and discrimination. To address these concerns, the community has conducted various studies on fairness, accountability, and transparency in developing and deploying ML systems. Despite these efforts, the importance of autonomy, a fundamental principle underlying many human rights, has often been overlooked. This oversight is concerning as it could jeopardize individuals' decision-making and exercise effective control over their lives, resulting in a violation of their rights. This article examines the principle of autonomy and its significance in the ML pipeline from a transdisciplinary perspective. The authors argue that autonomy remains a theoretical concept and does not translate well into the real use of ML, leading to contradictory outcomes. The absence of an effective approach to integrating autonomy leads to the persistence of disparities.

Keywords. human autonomy, machine learning, ethics of AI, bioethics and human rights

1. Introduction

We face several demands to constantly consume and produce as a result of the current economic and social structure. As people and organizations are concerned about failing to keep up with fast-moving events [2], these pressures create a climate in which notions such as "if we do not do it, someone else will" grow more prominent. Thus, these demands for haste frequently prevent us from effectively considering how our activities affect society and the environment.

[1] Corresponding Author: Paula Subías-Beltrán, paula.subias@eurecat.org.

By streamlining systematic operations, automation plays a critical role in advancing development by freeing up time and resources to be applied to more complicated tasks. By being able to process enormous volumes of data and find insights that might not be obvious through manual analysis alone, Machine Learning (ML) enables us to address more complicated problems. In light of this, ML-based technologies are spreading ubiquitously. These technologies are intrinsically vehicles for the companies that develop them to achieve their goals, such as expanding their reach, boosting user engagement, and so on. Because of their nature, the main objective of these technologies is not the respect of societal values, but it is a requirement that we must demand.

Autonomy is the capacity of individuals to make free and self-determined judgments. The premise is that people require autonomy in order to conduct their life according to their own views, values, and needs, as long as they do not damage others. Autonomy is a core value in the development and promotion of the principles and human rights we embrace as European society. Making decisions regarding privacy and intimacy requires autonomy, and maintaining privacy is crucial to sustaining intimate relationships. These are interconnected and interrelated principles that form the cornerstone of the society in which we live.

The importance of autonomy in algorithmic decision-making is often underappreciated in the Artificial Intelligence (AI)[2] community. Instead, emphasis is being placed on other crucial values, such as fairness and privacy, but autonomy-related issues are not being sufficiently addressed. The topic of autonomy in AI calls for a more philosophical approach, requiring more critical and in-depth views on how to include the value of autonomy in technology. There is still work to be done to fully achieve effective respect for autonomy, despite the fact that advancements have been achieved in this area [5,8,14,22].

To address these concerns, the AI community needs an effective and systematic strategy, one that is accessible to individuals with technical backgrounds and may be less conversant with philosophical frameworks. The innovation of this work is that it addresses the risks posed in each phase of the ML pipeline in a methodical manner, allowing it to be easily transferred to real-world development carried out by AI practitioners. This paper begins by describing the concept of autonomy and stating its importance as a core principle in European society in §2. The full examination of the hazards to autonomy at the various stages of the ML pipeline will then follow in §3. Finally, §4 concludes by summarising the research's findings.

2. Autonomy as a foundational value

Autonomy is a very complex concept that has been debated by different disciplines. The notion that unites them all, though, is self-government and self-determination, which elevates autonomy to the status of a distinguishing quality of free moral actors. People are therefore at their most autonomous when they behave in accordance with their own interests, beliefs, and desires [9]. The disputes around autonomy are frequently brought on by the term autonomy being confused with one specific conception of autonomy rather than a more comprehensive perspective.

[2]In folk language usage, AI and ML are often interchanged, but in this context, we use "AI" to describe the general approach discussed in the public discourse, reserving "ML" for addressing specific real-world implementations.

In bioethics, one of the four basic principles that govern the field is autonomy. This principle is addressed within the decision-making process and is connected to individual responsibility. Beauchamp and Childress, two of the main authors of the reference work that examines the principles of Bioethics [3], claimed that to demonstrate respect for autonomous agents implies "to acknowledge their right to hold views, to make choices, and to take actions based on their personal values and beliefs".

From a liberal philosophical approach, autonomy is seen as an individual right necessary to make one's own decisions or as a psychological ideal of independent thinking and rational self-control [12]. According to this view, individuals should be free to make their own choices, as long as they do not harm others. On a different note, there is the concept of "relational autonomy", which is an umbrella term indicating a variety of linked ideas rather than referring to a single understanding of autonomy. Relational autonomy is grounded in the belief that people are socially constituted and that social relationships and a variety of intersecting social variables, including race, class, and gender, impact people's identities. Thus the focus of this approach is to analyse the effects of the intersubjective and social aspects of selfhood and identity on concepts of individual autonomy as well as moral and political agency [15]. This view is rooted in feminist philosophy and challenges traditional views of autonomy. Autonomy is caricatured in feminist critiques of autonomy as a self-sufficient individualist male who adheres to libertarian doctrine. According to this view, autonomy is created not just by individual decisions but also by cultural and social structures that favor certain groups over others. In line with feminist views, full autonomy can be attained only when society is more egalitarian and everyone has equal access to important resources and opportunities.

These many philosophical viewpoints on autonomy have significant ramifications for how autonomy is interpreted and valued in different scenarios, including AI technologies. It is crucial for AI practitioners to be aware of the importance of autonomy and to engage in critical reflection on how autonomy is addressed during each phase of the ML pipeline. This was backed up by the European Commission's High-Level Expert Group on AI in its well-known Ethics Guidelines for Trustworthy AI [1]. They listed respect for human autonomy as one ethical principle to take into account in the context of AI, despite the fact that they did not even describe autonomy and did not offer any practical advice for making this respect effective.

In this work, we understand autonomy as a principle that enables people to behave independently of others or in connection to their society, in accordance with their own values, beliefs, and desires. In democratic environments, autonomy is considered a vital element of human freedom, and self-determination since it enables people to take part in the political process as educated and engaged citizens [20]. And any widely used technology, such as ML, can have an impact on it. Excessive reliance on ML systems, for example, may reduce the overall level of human skill required, impairing our ability to make educated and independent decisions and, as a result, reducing our autonomy in the medium or long term. ML-based solutions have the potential to improve social justice or exacerbate existing disparities. Therefore, it is the duty of ML practitioners to think about the wider ramifications of their work in order to comprehend the potential effects and lessen any harm. This calls for a dedication to moral standards and a readiness to consider the implications of their job critically. By doing this, ML professionals may strive toward a society in which their solutions promote everyone's rights and principles, such as respect for autonomy. Our study targets ML applications where decisions are

made based on the output produced, which includes both intended designs and situations where there is an excessive reliance on the result. The focus of our research is thus restricted to ML models that are utilized for decision-making. Next, we take a thorough look at how autonomy is affected by each of the major stages of the ML pipeline.

3. Autonomy in the ML pipeline

Our strategy to promote autonomy has been to examine the impact of autonomy within the ML pipeline, with the goal of engaging more technical profiles in this dialogue. We seek to raise awareness of the ethical aspects of AI and encourage AI practitioners to prioritize respect for autonomy in their work by identifying and considering the potential dangers to autonomy at each phase of the ML pipeline. In order to structure and facilitate the discussion around the impact of autonomy within the ML pipeline, we have simplified the pipeline into the following phases, which are connected in the pipeline shown in Figure 1:

- **Design** involves setting expectations for the performance of the model, as well as considering the ethical and social implications of the model's use in practice.
- **Data** entails its collection or creation, and its transformation so that it will be used to train and validate the model. It includes data collection, data cleaning, imputation, preprocessing, data transformation, and supervised annotation, among others.
- **ML model** is primarily concerned with the internal workings of the model, rather than the input and output of the system. It includes the definition of the functional hypothesis space and model architecture, loss function, optimization algorithm, regularization definition, and other kinds of inductive biases as well as dependencies of third-party components and libraries.
- **Utilization** involves not only the technical aspects of deploying the model but also the practical considerations of how the system will be used and monitored over time.

Next, we delve into the identified phases and expand on how ML may threaten the exercise of effective autonomy.

3.1. The world as it is

"The world as it is" serves as the foundational factual starting point for the design and development of ML-based systems. This is because the world as it is provides a source of data that represents current dynamics. However, it is up to us to choose whether to step in to alter present dynamics or to remain silent and let everything play out as it has in the past. We as a species do not have common values and even the parts we do share change over time. The world as it ought to be provides an opportunity to adopt a vision of the world based on our ideals rather than how things are now. It inspires us to think of a time when technology is employed to advance human flourishing rather than to uphold injustice or cause damage. It serves as a reminder that we have the ability to mold our environment and technological advancements to match our goals and ideals. An important design choice depends on two different perspectives, one where we presume

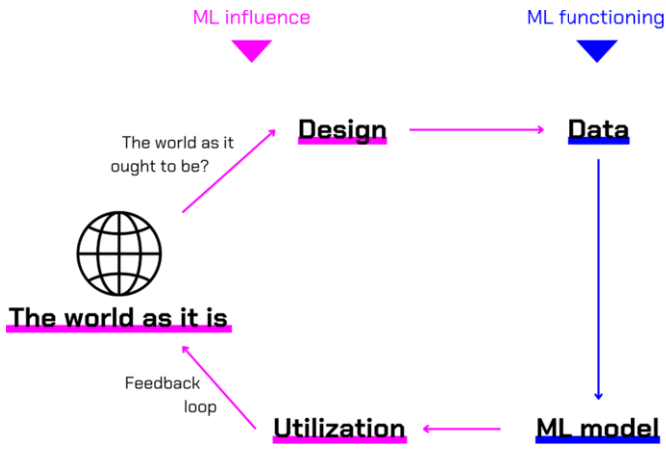


Figure 1. Phases that form the ML pipeline and their impact on the world. Pink highlights correspond to the phases that are external to the technicalities of ML in itself, while blue identifies the phases that explicitly affect the functioning of ML.

that the data is accurate and that the observed disparities are due to differences that have a causal motivation, and the other where we assume that structural artifacts and historical factors are to blame for the observed differences. This is assuming that all relevant factors are considered in the problem. If we were to do a fairness analysis, the conclusion from the latter viewpoint might be that we are all equal and that any found inequities should be corrected while considering independence criteria.

The discrepancy between the world as it is and as it ought to be is one form of misalignment. But the alignment problem [7] can also manifest when one uses proxies, non-representative data, or mix up ground truth with the information captured by data. Misalignment impedes the correct assessment of ML, rendering the ability to exercise autonomy more challenging. Our attempts to be precise may be hampered by the limitations of our own understanding and the fact that the map we construct based on that understanding does not exactly match the territory we are modeling.

These practices, which rely on the excessive use of ML systems, may jeopardize the value of autonomy by eroding individual autonomy and decision-making. The three actors of interest are the ML-based solution, the organization that owns it, and the end user. The solution unites the two key stakeholders, and its execution reveals the importance organizations place on end-user autonomy. In social media, we see personalized feeds that automatically limit the information we receive; in automated processes, we find a lack of adaptive automation based on risk; in diagnostic tools, we find absolute judgments decontextualized from their uncertainty; etc. These examples show how organizations advocate full automation or information limitation in the name of efficiency, objectivity, and personalization while minimizing end-user autonomy. Respect for autonomy needs to happen at all levels, between individuals and from institutions. Devaluing autonomy can have major ethical consequences and weaken confidence in the system's legitimacy.

3.2. Design

As stated in the AI Act, ML systems cannot be used for manipulating or deceiving people [10]. These unethical behaviors may have an impact on how the decision space is

formed and the ability to weigh the consequences of actions. Therefore, it is essential to carefully evaluate the effects of the design decisions in order to create the necessary safeguards to stop misuse or abuse.

One tool that may foster respect for people's autonomy is meaningful human control over the system. This technique may increase the system's accountability and transparency, enabling people to contest or challenge the decisions made [17]. At the same time, it will avoid over-reliance on ML and ensure that humans remain in charge of the decision-making process. In this line of gaining system control, the adequate allocation of responsibilities is key so that users can evaluate the risks that the given solution entails. This helps prevent agency laundering [19], in which responsibility is transferred away from people who built or deployed the system, and responsibility gaps [16], in which no one is held accountable for the actions or choices of the system.

Another technique for defending the respect of autonomy is related to the automation capabilities of ML-based solutions. Automation entails the cession of autonomy by definition. In the decision-making process, the cession of autonomy means delegating decision-making to the model. A successful system should be able to give up control of the appropriate items at the appropriate times to the end-user. This concept is known as adaptive automation and it is founded on the premise that systems should be modular and always supported by an explainability layer that allows useful information to be exchanged with their users [21]. This strategy will give users control at the right times and adapt to the level of risk they are willing to accept.

3.3. Data

Data is the cornerstone of any ML-based system. Starting with data collection, ML systems that rely on large amounts of interconnected data may violate individuals' privacy rights because of their ability to expand users' profiles. In addition, this may endanger users' autonomy by allowing sensitive information to be utilized without their explicit consent and by allowing decisions affecting their lives to be made without their involvement or comprehension. This might result in judgments being made based on unreliable assumptions and correlations, which would restrict people's autonomy by preventing them from taking charge of their own life and making their own decisions. This might reinforce structural injustices and go against the idea of respecting autonomy, which also entails respecting consent. Sensitive data should not be utilized, even tangentially, if individuals have not given their authorization to do so since it may lead to unjust and discriminatory effects. Furthermore, because some user types are a minority, pre-processing procedures like data cleaning and normalization might change the distribution of the data and ignore them, which reduces their representation in the entire sample. As a result, the representation may be insufficiently representative, contributing to the persistence of systemic disparities and, consequently, the persistent deterioration of their autonomy.

Data can be collected, but it can also be manufactured. This is the case with taxonomies. We will concentrate on taxonomies that characterize people's non-visual qualities without their agreement in order to focus on examples that worsen autonomy. This was the situation with ImageNet previous to 2020, when people were classified based on their presumed occupation, sexual orientation, and gender [24]. In this instance, a group of people categorized others without their consent, infringing on their capacity for self-determination. It is always necessary to include those who are affected by the taxonomy in the decision-making process to ensure that their autonomy is respected.

In any case, data must be accurate in order to solve the problem at hand. But we find sometimes that what the model actually predicts is not always what we think it is predicting. Selecting a "ground truth" that really is the ground truth is not always straightforward. There are cases where complexity is unavoidable, and the only quantifiable data may come from human abstractions, such as the "number of prior offenses", which does not correspond with the desired measure, such as the probability of reoffense. This misalignment between what the data gather and the interpretation we assign to the data is a risk factor for the system's correct behavior.

Data is also utilized to predefine the decision space of the solution. The decision space selected may influence how many alternatives users perceive themselves to have, prompting them to exclude particular possibilities. A limited decision space risks excluding alternative possibilities and assuming that the indicated decision space is complete. However, the data may not contain all possibilities and hence may not truly represent the variety of the whole set. In this scenario thinking beyond the box can be tough, and it usually does not happen at all.

3.4. ML model

Models are the outcome of a series of decisions, such as which metrics to optimize, which models to train, which hyper-parameters to tweak, etc. But also their level of understandability, the level of control they provide to end users, etc. Control over these decisions is lost when ML is outsourced. This is true for foundation models as well as other ready-to-use solutions. The loss of control may be avoided if ML-based systems were sufficiently understandable so that users could get all of the information required to create their own judgment about ML's behavior and limitations. The general lack of transparency in these systems makes it harder to scrutinize them, limiting users' agency which might ultimately lead to a misconception of the assurance on the system.

Models are intended to do a certain task while operating within established boundaries. When models are compelled to answer questions outside of their intended scope, they may deliver erroneous responses. The risk of models making mistakes because they are forced by design to give an answer may jeopardize autonomy insofar as the user may not know the underlying uncertainty the model faced and, thus, erroneously rely on the system. Rejection algorithms [6] are a recognized but less used solution for these situations since they provide users the choice to withhold their response when the uncertainty is greater than the specified threshold.

When ML is used to make judgments or interact with the environment, such as when using reinforcement learning or prediction systems for decision-making, the balance between the algorithm's exploratory and exploitative capabilities must be evaluated. While the former broadens the variety of options, the latter maximizes user adaption. It is common to devote greater weight to exploitation than to exploration in response to commercial strategies. However, this comes at the price of people's autonomy. Giving exploratory capacity less credit involves restricting users' access to information, which already compromises their autonomy, but can be exacerbated by making them more likely to overlook opportunities by threatening the creation of their decision space.

The opt-out mechanism is one technique that allows for the exercise of autonomy with respect to voluntary exclusion from a system. As stated in the Article 21 of the GDPR [11], consent can be withdrawn by the user at any time, and ML-based systems

must be prepared to respond accordingly. Furthermore, systems must also account for individuals exercising their right to rectification (Article 16 of the GDPR) so that no judgments are made based on wrong data, which might further erode people's autonomy.

Last but not least, we consider it crucial to examine how the context of the users is integrated into the model's decision-making. This is addressed by decision theory, which is concerned with how decisions are formed depending on the information available and the context in which it is perceived. The most prevalent technique in ML is to maximize expected utility [23]. This translates to maximizing the probability of the correct outcome or, alternatively, minimizing the empirical risk. However, it is not always the best strategy to choose for combining decision and prediction theories. For example, prospect theory [13] examines how people view their loss and gain perspectives asymmetrically. The degree to which the model is applicable to the users' environment depends significantly on the decision theory used, which has a negative impact on adaptation and, consequently, on the users' degree of effective autonomy.

3.5. Utilization

Both human flourishing and the greater entrenchment of systemic disparities are possible in the post-ML world. To avoid creating a dystopian society, it is crucial to evaluate how ML solutions affect various subgroups of the population in order to spot any possible problems early on. The world is changing, and so is the data. Data is the backbone of ML-based solutions. And for that reason, it is critical that it remains as accurate as possible. If not, the solutions would become outdated since they would be lacking relevant knowledge. This may occur as a result of improvements in the quality that technology can deliver, individuals' change in behavior, populations becoming more diverse, etc. In any case, the actual utilization of ML-based solutions should directly affect the first phases of the ML pipeline to ensure that the adequacy of the solution is not affected.

It is essential to have been transparent about the system's intended usage for users to successfully utilize it. Therefore, it is crucial that ML practitioners convey all decisions made throughout the system's development so that the user can benefit from this knowledge. Lack of knowledge of this information may cause assurance to be interpreted incorrectly, overconfidence in the system, and opportunities to be missed, among others. But not every communication is effective. It is crucial to consider the potential misunderstanding between model creators and end users. Sharing assumptions and making the method as self-explanatory as possible can help close this gap. By doing this, we can encourage a better knowledge of the model and its implications, enabling users to act independently and make more educated decisions, fostering thus the exercise of their autonomy.

Another challenge is deskilling, which in this context occurs when the excessive usage of ML systems leads to a reduction in the required human expertise. Although ML automation might be advantageous to increase efficiency, it can also have negative consequences, such as lowering the overall level of human ability and knowledge required in a specific profession. As a result, decision-making may become more reliant on ML models, which are not always precise and can contribute to wider social and economic inequalities. As a result, our autonomy has been put at risk.

One of the greatest risks of ML models is that they become reality. "The world is its own best model" [4], hence, rather than modeling the world, appropriate perception-to-

action systems can be employed to engage with it directly. Performative predictions [18] are forecasts that have an effect on the target. Thus, performativity might take the shape of a distribution shift as the decision-maker applies a prediction model. The ambiguity of performativity's effects extends beyond ML practitioners to end-users, who might be concerned that more interaction with the system could lead to future deterioration of their interaction with the system.

4. Conclusions

As the number of accessible ML systems expands significantly, it becomes more worrying that ethical issues are not sufficiently integrated throughout their development. The rights of individuals, notably their autonomy, are directly impacted by this absence. Instead of fostering autonomy, ML systems actually act against it by weakening people's ability to make free and informed decisions. The difficulty of incorporating ethical concerns from the start is being made more difficult by the speed at which ML systems are being developed and used in daily life. This helps to explain why there have not been as many efforts to operationalize autonomy effectively, which would have allowed for the translation of abstract ideas into logical languages that allow for its practical application. In order to ensure that autonomy is properly understood, operationalized, and deployed, we propose an ethical analysis from the inception of ML systems.

In this study, we analyze the potential hazards that could emerge in each of the indicated ML pipeline phases and pose a threat to the successful exercise of our autonomy. We emphasize the need of recognizing the constraints of modeling the world as it as well as examine all aspects of the alignment problem from the start. We highlight the need of anticipating mechanisms to ensure human control in order to avoid excessive reliance and responsibility gaps in the design phase, the impact of the data chosen on the perpetuation of current practices, and the effect of data selection on the formation of the perceived decision space. We also emphasize the need of putting the end-users at the heart of the ML model in order to avoid neglecting their context. Finally, we analyze the risks associated with the utilization of ML systems, which may result in deskilling or even changing the world as it is through feedback loops.

We propose a transdisciplinary approach in a society that takes for granted that the decisions we are going to make with the support of ML systems prevail over our own. We question that this basis does not allow us to make autonomous decisions and we conduct a critical diagnosis to identify violations of the respect for autonomy at the different phases of the ML pipeline. Our work aims to provide useful advice that guides AI practitioners in identifying potential risks when developing ML solutions that endanger the respect for autonomy and, ideally, advise effective actions to take to reduce the hazards that are identified.

5. Acknowledgements

This work is partially supported by MCIN/AEI/10.13039/501100011033 under project PID2019-105093GB-I00 and PID2022-136436NB-I00.

References

[1] AI HLEG. Ethics Guidelines for Trustworthy AI. Technical report, High-Level Expert Group on Artificial Intelligence, Brussels, 4 2019.

[2] Zygmunt Bauman. *Liquid life*. Polity, 2005.

[3] Tom L Beauchamp, James F Childress, and Others. *Principles of Biomedical Ethics*. Oxford University Press, USA, 2001.

[4] Rodney A Brooks. Intelligence without reason. In *The artificial life route to artificial intelligence*, pages 25–81. Routledge, 2018.

[5] Rafael A. Calvo, Dorian Peters, Karina Vold, and Richard M. Ryan. Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry. *Philosophical Studies Series*, 140:31–54, 2020.

[6] C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

[7] Brian Christian. *The Alignment Problem: machine learning and human values*. W.W. Norton & Company, New York, 1 edition, 2020.

[8] Itziar de Lecuona, María Jesús Bertrán, and et al. Guidelines for reviewing health research and innovation projects that use emergent technologies and personal data. *Review of biomedical research and innovation that uses emergent technologies and personal data: the challenges facing research ethics committees in a data-driven society*, 2020.

[9] Edward L. Deci and Richard M. Ryan. Intrinsic Motivation and Self-Determination in Human Behavior. *Intrinsic Motivation and Self-Determination in Human Behavior*, 1985.

[10] European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Technical report, European Commission, Brussels, 4 2021.

[11] General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. Technical report, Official Journal of the European Union, 4 2016.

[12] Thomas E Hill. Kantian autonomy and contemporary ideas of autonomy. *Kant on moral autonomy*, pages 15–31, 2013.

[13] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

[14] Arto Laitinen and Otto Sahlgren. AI Systems and Respect for Human Autonomy. *Frontiers in Artificial Intelligence*, 4:151, 10 2021.

[15] Catriona Mackenzie and Natalie Stoljar, editors. *Relational autonomy: feminist perspectives on autonomy, agency, and the social self*. Oxford University Press, New York, 2000.

[16] Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology 2004 6:3*, 6(3):175–183, 2004.

[17] Frank Pasquale. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

[18] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative Prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 11 2020.

[19] Alan Rubel, Clinton Castro, and Adam Pham. Agency Laundering and Information Technologies. *Ethical Theory and Moral Practice 2019 22:4*, 22(4):1017–1041, 10 2019.

[20] Alan Rubel, Clinton Castro, and Adam Pham. *Algorithms and Autonomy: The Ethics of Automated Decision Systems*. Cambridge University Press, 2021.

[21] Mark W Scerbo. Theoretical perspectives on adaptive automation. In *Automation and human performance: Theory and applications*, pages 37–63. CRC Press, 2018.

[22] Paula Subías-Beltrán, Oriol Pujol, and Itziar De Lecuona. The forgotten human autonomy in Machine Learning. In Desara Dushi, Francesca Naretto, Cecilia Panigutti, and Francesca Pratesi, editors, *Proceedings of the Workshop on Imagining the AI Landscape after the AI Act (IAIL 2022)*, Amsterdam, 9 2022.

[23] John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 3 2007.

[24] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230687

Sequence Pattern Mining for Citizens Behaviour Learning in Fair-Purpose Social Games. Some Preliminary Results

Beatriz LÓPEZ [a,1], Roger CAMPDEPADROS [b] and Albert SABATER [c]
[a] *eXiT research group, University of Girona*
[b] *Grup de Recerca Interdisciplinar en Gènere i Desigualtats Socials (IGESI), University of Girona*
[c] *Observatori d'Ètica i Intel·ligència Artificial (OEIAC), University of Girona*
ORCiD ID: Beatriz López https://orcid.org/ 0000-0001-9210-0073, Roger Campdepadros https://orcid.org/ 0000-0003-3744-162X, Albert Sabater https://orcid.org/ 0000-0003-3532-6572

Abstract. Artificial Intelligence can be used to understand how citizens interact in a society with the aim to work together towards a better world. The application of AI in this domain has shown immense potential, particularly in leveraging games as powerful tools to simulate individual roles within a society. This research endeavors to explore how sequence pattern mining are applied to data coming from a game, in which humans are posed with the dilemma of progressing through their jobs, while enjoying life, trying to catch an equilibrium between happiness and healthiness. Preliminary results on a game played with 13 citizens shows the feasibility of sequence learning for this purpose. Our findings shed light on the intricate interplay of individual aspirations, communal synergies, and the pursuit of equilibrium, laying the foundation for future advancements in AI-driven methodologies for sociological investigations.

Keywords. pattern mining, sequence learning, social games

1. Introduction

Studying how Artificial Intelligence (AI) will change public service provision and its underlying societal systems has become key in different contexts worldwide, as governments and organizations increasingly recognize its potential to revolutionize the way services are delivered, streamline administrative processes, and enhance overall efficiency. For instance, one of the most significant impacts of AI on public service provision is the potential to improve decision-making through data analysis and predictive algorithms. This data-driven approach enables policymakers and public service providers to make more informed decisions, allocate resources more effectively, and address societal needs proactively.

[1] Corresponding Author: beatriz.lopez@udg.edu

However, which policies, behavioural changes and institutional developments are necessary and appropriate to prevent or support certain scenarios has focused so far on how AI-based social assessment technologies for public service provision categorise present and future human behaviour on scales such as legal recipients/fraudulent recipients, deserving/non-deserving or acceptable/not-acceptable ones. Since delegating decisions on such value judgements to AI raises ethical and social considerations related to responsibility, accountability, transparency and the quality of social decision making about the distribution of resources, further research needs to be conducted from a multidisciplinary perspective of social sciences and AI. This is the case of the AI FORA project [2] in which we are involved. In this project, heterogeneous societal organizations interact in their specific socio-cultural context to negotiate and co-produce AI social assessment technology. This approach is particularly useful in social service provision as often public opinion and discourse in these areas can be emotional and empathetic, not least because societal core values are affected and at stake, and decisions on providing or refusing public social services have far-reaching consequences for the concerned individuals.

Because games have a unique ability to engage and empower participants and they can play a significant role in facilitating the co-production of AI social assessment technology and achieving more responsible and empathetic social service provision, we employ gamification elements, such as challenges, rewards, and competition in order to motivate stakeholders to contribute their knowledge, insights, and concerns more willingly.

Games have been previously used as a key tool to understand self-identity and values and by designing collaborative games that require input from different stakeholders, societal organizations can work together in an interactive and inclusive environment that recognise the citizens major implications in society [1]. Games allow participants to explore different scenarios and consequences and data gathered from games can be further used by agent-based social simulations to understand the impact of behaviours at the long run. In the context of AI social assessment technology, stakeholders can use simulation games to understand the potential impacts of AI-driven decisions on social service provision. This promotes thoughtful reflection on the technology's implications and helps identify and address potential biases and unintended outcomes [2,3,4]. Despite their potential complexity, games not only offer the opportunity to simulate what-if scenarios about possible automated decisions regarding in social provision, but also provide the potential to engage the wider public in discussions surrounding AI and social service provision [5,6]. Increased public understanding and input contribute to a more democratic and transparent approach to AI development and the use of concepts as ethnicity [7] and transparency [8] in decision-making models in virtual environments before implementing them in real-world scenarios allows for identifying potential flaws and biases, minimizing the risk of harmful consequences in actual social service provision.

In addition, games generate empirical data, and machine learning methods could be used to analyze them, as we demonstrate in this work. In particular, we use the data from the unemployment game[3] that has been set up to understand the concept of socially fair regarding provision of jobs. To that end, the game has been designed with the aim that players have the possibility to agree in the algorithm that assigns jobs to citizens. Players interact in several rounds, selecting some actions to choose in each round. Therefore, a

[2] https://www.ai-fora.de/

[3] https://github.com/micrology/AIFORA

sequence of actions per player are gathered at the end of the game, and sequence pattern mining is applied as a suitable method to discover common patterns.

By applying machine learning techniques, such as sequence pattern mining, to the data generated by the unemployment game, we can gain valuable insights into how players approach the concept of socially fair job provision. The gathered sequences of actions represent the decision-making processes and preferences of the players throughout the game, offering a rich dataset to analyze. Thus, sequence pattern mining allows us to identify recurring patterns in the players' actions, which can shed light on their strategies and priorities through patterns. For example, these patterns can offer valuable clues about potential biases and areas for improvement in the job assignment algorithm. Moreover, machine learning methods can help quantify the impact of different player actions on the overall fairness and effectiveness of the job provision process. The combination of serious games and sequence pattern mining provides a powerful approach for understanding complex social dynamics and decision-making processes, unveiling underlying patterns and trends that may not be apparent through traditional analytical methods.

This paper is organized as follows. In Section 2 we provide a literature review of the work related to us. Section 3 provides a basic explanation of the game we use. Section 4 details the methodology designed, and Section 5 describes and discuss the results achieved. Finally, the paper ends with some conclusions and future work.

2. Related work

Sequence pattern mining has been applied to games according to the purpose the game has been set up. For example, in [9], the authors use the Prefixspan algorithm [10] to learn patterns in a game with tutoring purposes. Tutoring systems offer the possibility to conduct users towards a useful achievement; on the other hand, in our work we aim to analyse self-identify values in a game.

Learning environments are also explored in [11] but with the goal to understand players' emotions. To that end, Taub et al. use the SPAM algorithm[12]. Taub et al. work is similar to us since they use sequence pattern mining to analyse different students skills regarding the games progresses. In educational games, sequence pattern mining helps track learners' progress and identify areas where they may be struggling. This information allows educators to tailor their instructional approach and provide personalized feedback, enhancing the learning experience. Another similarity of the Taub and colleagues approach is that we also use information on different person's skills.

[13] focuses on finding the given context of a frequent sequence. Context is a mattern of future research of our work. Although our game allow changes according to different external events (i.e. decrease of available jobs), our preliminary approach does not consider such dynamics of the game.

In this work, we use VEPRECO[14], a more powerful algorithm than its ancestors, Prefixsan and SPAM, and we apply sequence learning with the aim to understand self-identity and values, according to [1].

3. The game in a nutshell

In this research, the social game serves as an interactive platform where a group of users can participate and experience various aspects of social service provision and decision-making. Each player is assigned an identity card that outlines their citizen profile, which includes information such as their marital status (single or parents), level of education, current state (happiness, wealth, training level, communication skills), and job status (unemployed, poor job, good job). The game aims to simulate real-life scenarios, and players are expected to act according to their assigned identity and state.

During each round of the game, players have the opportunity to visit six different places or services, each offering unique interactions and choices:

- The Labour Office: Here, players can apply for a new job based on a predefined algorithm that determines job assignment. The outcome of this interaction can impact the player's job status and overall well-being.
- The Working Place: If the player has a job, they can attend their workplace, reflecting the daily work routine and the experience of having a job.
- Staying at Home: Players can choose to stay at home, which may signify taking a break or focusing on personal well-being.
- The Travel Agency: This allows players to simulate going on holidays, offering an opportunity for leisure and relaxation.
- The School: Players can visit the school to acquire new knowledge and skills, reflecting the importance of education in their lives.
- The City Hall: The city hall acts as a central meeting place where players can interact with each other, propose changes to the algorithm governing job assignment, participate in voting processes, and make donations.

Throughout the game, players make choices at each place, and these decisions impact their happiness, wealth, skills, training level, and job position, if applicable. For example, getting a new job might increase a player's happiness and wealth, while attending the school can enhance their skills and training level. It is also worth noting that the game offers certain exceptions to the rule of visiting a single place in a round. For instance, players can go to work after obtaining a new job from the labor office, showcasing how work and job transitions are interconnected. Additionally, players can always visit the city hall to make donations, emphasizing the importance of social engagement and contributing to the community.

Overall, depending on the chosen action, the agents change their happiness, wealth, skills and training level, as well as the job position, if it is the case. In summary, the game used in this research provides a dynamic and immersive environment for players to experience social service provision, decision-making, and societal interactions. By assigning unique citizen profiles and offering multiple places to visit with different outcomes, the game allows players to explore various aspects of life and gain insights into the complexities of socially fair job provision and the factors that influence individual well-being and opportunities. Further details of the game can be found elsewhere[4].

[4]https://github.com/micrology/AIFORA

4. Methodology

Figure 1 shows an overview of the methodology followed. The starting point is the data collected from the social game. Data has been collected in the basis of each entity providing services in the game; therefore, we need to track users behavior from all the data to build individual sequences per each user. Once a database of sequences is obtained, two possible branches are explored to obtain frequent patterns. First frequent patterns from the full dataset are generated. And second, the sequence database is split according to a given dimension of the profile of users (see Table 1). This second step is repeated for each dimension. Next, sequence learning mining is applied in each split and the corresponding sequence patterns are found. The patterns that characterize a given dimension are finally obtained by selecting the ones that are present for a given value of the dimension, and do not in the opposite one.

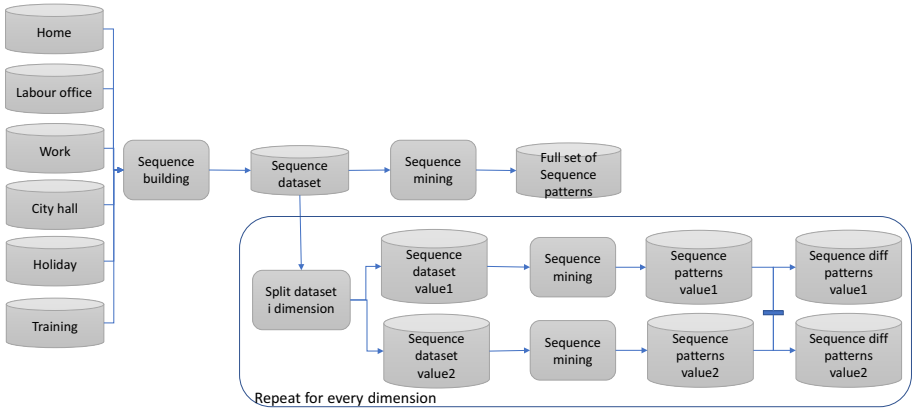


Figure 1. Flow chart of the methodology

Dimension	value	
Gender	men	women
Family	single	parent
Ethnicity	white	minority
Skills	high	low
Vulnerability	high	low
Residence area	rich	poor

Table 1. Dimensions that configure the players profiles.

4.1. Dataset

We get the data from a game played by 13 players ($N = 13$). Each action available to choose by the agents is labelled according to the map provided in Table 2. There are a total of $M = 6$ places to visit. A total of $R = 10$ rounds have been played.

From each round quantitative data about the players is gathered, as well as qualitative data (i.e. what changes in the algorithm has been proposed; which externalities could

be arisen, as firing a person from a job, ...). We focus in this work on quantitative data, as a preliminary exploratory analysis, and we left for future work the use of qualitative data.

Place, activity (if many)	label
Home	H
Labour office	L
Work	W
City hall, proposal to change the algorithm	C1
City hall, voting	C2
City hall, donation	D
Travel agency	Hol
School	T

Table 2. Mapping of the places a player could attend, and the activities she could perform (if more than one), and the labelled used for the method presented in this work.

We have a total of $R = 10$ sets of data, one per round, sorted according to the visits of players ($P = \{p_1, \ldots, p_N\}$) to the different places. We note by $\langle D_1, \ldots, D_R \rangle$ the sequence of data gathered in each round, where each $D_j = \{S_1, \ldots, S_M\}$. Each S_i contains the particular activities available in the service i that has been chosen by the players, $S_i = \{a^i_1, \ldots, a^i_{|S_i|}\}$, each $a^i_j = \{p^{i,j}_1, \ldots, p^{i,j}_{|a^i_j|}\}$ with $p^{i,j}_k \in P$. Note that a^i_j could be empty if no player has chosen this action in a given round.

4.2. Preprocessing

From the sequence $\langle D_1, \ldots, D_R \rangle$, we design Algorithm 1 to obtain a sequence of activities per user in each round, $\{Seq_1, Seq_2, \ldots, Seq_N\}$, each $Seq_j = \langle A_1, \ldots A_R \rangle$, and $A_i \subset \mathscr{A}$, being $\mathscr{A}$ the full set of available activities (see Table 2). Observe that A_i is a set of actions (itemset), since a user can eventually perform more than one action in a given round.

The algorithm bassically iterates over the different rounds (step 2), and on the services (nested iteration of step 3), to collect the different activities performed by each players (iteration of step 3). An example of the pre-processing output is the following:

$$\langle [T], [L, W], [W], [H], [C1], [W, D], [W], [W], [Hol], [W] \rangle$$

4.3. Sequence pattern mining

Given a data base of sequences of itemsets (set of events or items that occur at once), sequential pattern mining algorithms learn sequential patterns in the form of $\langle [a, b, ...], [a', b', ...], ..., [a'', b''] \rangle$ that are frequent in a the database, meaning that the itemsets in $[a, b, ...]$ happens befere itemsets in $[a', b', ...], ..., [a'', b'']$, etc. This kind of algorithms requires a support threshold $\gamma \in [0, 1]$ that determines the frequency required by the pattern in the database to be considered a pattern. A low *gamma* value means that a lot of patterns could be gathered, which could be eventually difficult to interpret; a high *gamma* value could led to find no patterns at all. After some preliminary trials, we finally set up $gamma = 0.8$, mining that a pattern will have the 80% of sequences in the

Algorithm 1 Sequence building

Require: $R, P = \{p_1, \ldots, p_N\}, \langle D_1, \ldots, D_R \rangle$
each $D_j = \{S_1, \ldots, S_M\}$
each $S_i = \{a_1^i, \ldots, a_{|S_i|}^i\}$
each $a_j^i = \{p_1^{i,j}, \ldots, p_{|a_j^i|}^{i,j}\}$ with $p_k^{i,j} \in P$
Ensure: $Seq = \{Seq_1, Seq_2, \ldots, Seq_N\}$
1: $sequence_j \leftarrow null$ for all $p_j \in P$
2: **for** $r \in [1, R]$ **do**
3: $itemset_j \leftarrow null$ for all $p_j \in P$
4: **for** $i \in [1, M]$ **do**
5: **for** every $a_j^i \in S_i$ **do**
6: **for** $k \in [1, N]$ **do**
7: **if** $p_k \in a_j^i$ **then**
8: append(a_j^i, $itemset_j$)
9: **end if**
10: **end for**
11: **end for**
12: **end for**
13: append($sequence_j$, $itemset_j$) for all $p_j \in P$
14: **end for**

database that match the pattern. This *gamma* value allows to get a manageable amount of patterns; with a lower *gamma* value, the number of patterns would be huge, and difficult to interpret; while with a higher *gamma* value, very few patterns would be found.

An example of the patterns found in our setting is $[[T], [L], [W, D], [W]]$, which can be read as follows: players first go to the school to get some training (T); next (not necessarily the immediate next action) they go to the labour office (L); afterwards they go to the work office, and make some donation (W,D); and finally they go again to work (W).

Sequential pattern mining have a high computational cost, and there has been a lot of research to make the mining process efficient. In this wok we use the VEPRECO algorithm [14] that has been shown to be a competitive algorithm.

5. Results

In this section, we provide the results in terms of the number of the patterns found, for the full data set (first branch of the methodology, see Figure 1), and for each dimension (second branch on the Figure). Regarding the full set of patterns, we found 46 patterns, with the length shown in Table 3 . The maximum length of a pattern was 5. The identification of distinct patterns for each dimension allows us to understand how different attributes impacted decision-making and interactions within the game environment.

In Table 4 the results of the analysis for each individual dimension of the players' profiles are presented. In doing so, we provide a detailed breakdown of the patterns found within each dimension, along with additional statistics that offer valuable insights into player behavior and decision-making. Table 4 begins by showing the distribution of the different profile dimensions among the players. Here it is worth noting that there is a bias regarding gender representation in the dataset. There are more players from one gender

Description	#
1-length patterns	6
2-length patterns	14
3-length patterns	14
4-length patterns	9
5-length patterns	3

Table 3. Length of patterns found for the full set of sequences.

compared to the other and this imbalance should be addressed in future gamification experiments to ensure a more equitable and representative dataset. However, the remaining dimensions (other than gender) show a balanced representation, with an equal number of players (6 versus 7) falling under each given condition. For each dimension, the table presents the total number of patterns found (# patterns) specific to that dimension. These patterns represent recurring sequences of actions and interactions associated with a particular attribute of the players' profiles. The table also provides the number of patterns that make a given dimension-value distinct from the other dimensions (# diff patterns), along with the percentage of distinct patterns from the total patterns. This highlights the uniqueness and specificity of certain patterns associated with each profile dimension. For example, only 45.71% of the patterns found for men are truly distinct from the patterns found for women, suggesting some overlapping behavior between the genders.

Table 4 also includes the mean and standard deviation of the pattern lengths for each dimension. From the analysis, it is observed that, on average, patterns had a length of around 4 actions in the game. The standard deviation indicates slight variations from this average length, suggesting some players followed more extended sequences of actions compared to others. Finally, at the bottom of table 4, three actions are highlighted as possible key distinguishable features between dimensions. For instance, going to the work office (W) is found to be much more common in women and single individuals compared to other profiles. Additionally, players living in poor areas were less likely to visit the work office frequently. This highlights how certain attributes influence players' choices and job-seeking behavior in the game. Donations (D) were more often performed by men, white individuals, and high-skilled and high-vulnerable players, suggesting differences in charitable behavior across various demographics. Holidays (Hol) were found to be more often selected by white and high-skilled individuals, indicating distinct preferences for leisure activities among different groups.

5.1. Discussion

The results shown in this work are a preliminary analysis of sequence learning from information coming from a social game. Patterns are important to understand the behaviour of people in the game according to their skills and the selection of the different actions they have available. In the case of social games, patterns provide insights about self-identify regarding, for example, fairness in job assignment. Some patterns are consistent with expectations, such as women and single individuals going more often to work, or individuals of certain ethnic backgrounds frequently selecting holidays. However, the length of the patterns indicates a diverse range of behaviors exhibited by players. While the findings are insightful, the interpretation of these patterns requires further in-depth

	Genre		Family		Ethnics		Skills		Vulnerability		Residence area	
	men	women	single	parent	white	minority	highS	lowS	highV	lowV	rich	poor
Distribution	9	4	6	7	6	7	6	7	6	7	6	7
# patterns	35	37	55	213	363	49	190	47	49	235	151	66
# diff patterns	16	18	20	11	43	7	42	8	7	49	11	20
% diff patterns	45.71%	48.65%	36.36%	5.16%	11.85%	14.29%	22.11%	17.02%	14.29%	20.85%	7.28%	30.30%
Pattern length (mean)	2.3	3.1	3.0	3.5	4.1	2.9	3.4	2.9	2.9	3.7	3.4	3.0
Pattern length (std)	0.8	1.2	1.2	0.9	1.2	1.1	1.0	1.1	1.1	1.2	1.0	1.2
Diff patterns length (mean)	2.5	4.1	4.0	3.7	4.2	3.2	3.6	3.0	3.0	4.0	3.6	3.7
Diff patterns length (std)	1.2	0.7	1.4	0.8	1.3	0.9	0.9	0.9	0.8	1.1	0.8	1.1
Work (W)												
City Hall- Donation (D)												
Holidays (Hol)												

Table 4. Description of the data and the found patterns. Black boxes at the bottom represent a high correlation between players profiles (columns) and the actions at the three bottom rows. Grey boxes means a low correlation; and the cell in white means that no correlation has been found.

analysis. In addition, several considerations have been identified as potential areas of improvement and expansion for future research.

First, it is important to consider a dynamic perspective and context awareness. Currently, the analysis focuses on static patterns without considering changes in the game environment, such as alterations in the job assignment algorithm. Incorporating sequence learning from a dynamic perspective, along with context-aware information, could enhance the understanding of how players' decisions are influenced by changing conditions. Second, the game could be further improved by using an integration of player condition information. We acknowledge the omission of critical information at the onset of the game, such as players' happiness or wealth, which can significantly impact their decision-making. Indeed, integrating this additional player condition data poses a challenging problem for sequence learning. Third, while quantitative analysis provides valuable insights, complementary approaches are needed to gain a deeper understanding of the patterns observed. Within this context, qualitative data, such as player annotations and observations, can contribute to a more nuanced view and rich interpretation of the findings, including further information on the context dimension to the quantitative patterns. Thus, to provide truthful and meaningful insights, it is essential for the machine learning community to consider and incorporate qualitative data and ethical perspectives in their analyses.

In conclusion, the preliminary analysis of sequence learning from the social game data has revealed interesting patterns in player behavior. However, further research is needed to integrate additional information, consider dynamic perspectives, and incorporate qualitative data to gain a more comprehensive understanding of the intricacies of player interactions within the game. By addressing these challenges and limitations, we believe that it is possible to provide deeper insights into the complexities of social dynamics, leading to more informed and responsible AI-driven decision-making in social service provision.

6. Conclusions

Social games are a powerful mechanism to self-identify on fairness, and machine learning provides a tool to understand the human behaviors in the games. This works presents

a sequence learning approach over data gathered in a social game. Results show that some interpretations can be gathered from the patterns found, though some results could be biased due to gender issues in the players distribution. Therefore, future work should repeat the analysis in a non-biased game. Moreover much more future work is required to provide truthful patterns that include the complexity of the dynamics of the game, in order to infer behaviors that could be used by social science. This dynamics include the consideration of qualitative information (i.e. externalities).

Acknowledgements

This work received support from the Generalitat de Catalunya via the Consolidated Research groups 2021 SGR 01125 and 2021 SGR 00838 as well as the AI FORA research project funded by the VolkswagenStiftung in the programme: Artificial Intelligence and the Society of the Future.

References

[1] Oliver TH, Doherty B, Dornelles A, Gilbert N, Greenwell MP, Harrison LJ, et al. A safe and just operating space for human identity: a systems perspective. The Lancet Planetary Health. 2022 11;6:e919-27.
[2] Ahrweiler P. Agent-based simulation for science, technology, and innovation policy. Scientometrics. 2017 1;110:391-415.
[3] Epstein JM. Generative social science : studies in agent-based computational modeling. Princeton University Press; 2006.
[4] Epstein JM. Agent-based computational models and generative social science. Complexity. 1999;4:41-60.
[5] Herget F, Kleppmann B, Ahrweiler P, Gruca J, Neumann M. How Perceived Complexity Impacts on Comfort Zones in Social Decision Contexts—Combining Gamification and Simulation for Assessment. Springer Proceedings in Complexity. 2022:203-15.
[6] Szczepanska T, Antosz P, et al . GAM on! Six ways to explore social complexity by combining games and agent-based models. International Journal of Social Research Methodology. 2022;25:541-55.
[7] Anzola D, Barbrook-Johnson P, Gilbert N. The Ethics of Agent-Based Social Simulation. Journal of Artificial Societies and Social Simulation. 2022 10;25.
[8] Özge Dilaver, Gilbert N. Unpacking a Black Box: A Conceptual Anatomy Framework for Agent-Based Social Simulation Models. Journal of Artificial Societies and Social Simulation. 2023 1;26.
[9] Bermudez RS, Sison AM, Medina RP. Extended PrefixSpan for Efficient Sequential Pattern Mining in a Game-based Learning Environment. APIT '20: Proceedings of the 2nd Asia Pacific Information Technology ConferenceJanu. ACM; 2020. p. 118-22.
[10] Pei J, Han J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, et al. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. Proceedings - International Conference on Data Engineering. 2001:215-24.
[11] Taub M, Azevedo R. Using Sequence Mining to Analyze Metacognitive Monitoring and Scientific Inquiry based on Levels of Efficiency and Emotions during Game-Based Learning. Journal of Educational Data Mining. 2018;10(3):1-26.
[12] Ayres J, Flannick J, Gehrke J, Yiu T. Sequential pattern mining using a bitmap representation. Proceedings of the eighth ACM SIGKDD International Conference on Knowledge discovery and Data Mining - KDD. ACM Press; 2002. p. 429-35.
[13] Dumartinet J, Foppolo G, Forthoffer L, Marais P, Croitoru M, Rabatel J. Contextual Sequential Pattern Mining in Games: Rock, Paper, Scissors, Lizard, Spock. Research and Development in Intelligent Systems XXXII. 2015:375-80.
[14] Mordvanyuk N, Bifet A, López B. VEPRECO: Vertical databases with pre-pruning strategies and common candidate selection policies to fasten sequential pattern mining. Expert Systems with Applications. 2022 10;204:117517.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230688

Reversal Error in Mathematics: Training Against It Using $\overset{\curvearrowleft}{\varepsilon}$ *coach*

Vicent SANTAMARTA [a] Pablo GARCIA-SEGARRA [a]
Noelia VENTURA-CAMPOS [b] Aida MORENO-RUS [b] and Zoe FALOMIR [a,1]
[a] *Universitat Jaume I, Departament d'Enginyeria i Ciència dels Computadors*
[b] *Universitat Jaume I, Departament d'Educació i Didàctiques Específiques*

Abstract. This paper presents $\overset{\curvearrowleft}{\varepsilon}$ *coach*, a serious game prototype which aims at training players to detect and avoid reversal error when representing word problems of the type *"N times as many S as P"* algebraically as an equation.

The $\overset{\curvearrowleft}{\varepsilon}$ *coach* visualizes the relative size of the entities in the relation following a qualitative model based on containers. This serious game also provides a quantitative visualization of the entities related using graphics, it helps players to resolve the formulated equation using discrete values and it shows players' progress.

Keywords. Education, reversal error, SP-type problems, qualitative representations, relative size, reasoning, serious games, serious games, AI for good.

1. Introduction

In mathematics education, the reversal error or the "students-professors problem" (*SP-type* problem), identified by Clement *et al.* [1], refers to reverse the relationship between two variables in an algebraic word problem, e.g. «Write an equation using the variables S and P to represent the following statement: *There are six times as many students as professors at this university*. Use S for the number of students and P for the number of professors». In these statements, Clement *et al.* [1] observed that most of the wrong answers were in the form of $6 \cdot S = P$, which they called reversal error because students reversed the order of the letters, compared to the correct answer $6 \cdot P = S$. And as Clement *et al.*[1] identified: *appears to be rather resilient and requires considerable attention and discussion before students can learn to overcome it* [2,1]. Factors influencing the reversal error were identified by Cohen *et al.*[3] and recently extended by Jankvist and Niss [4].

The literature reports few applications developed to study the reversal error [5]. To our knowledge, the only interactive tool used to address the reversal error is POETIC (Purdue Optimization modeling Education Tool—Interactive equation Component) [6], whose results showed that interactive visualization of equations can reduce the occurrence of this error. However, there was no substantial improvement in error elimination in their study in which the presence of syntactic obstructions in problem statements or the type of magnitude used were not taken into account. Soneira *et al.* [7] showed that there are significant differences in the production of reversal errors depending on the syntactic

[1]Corresponding Author: Z. Falomir, Universitat Jaume I, Avda. Sos Baynat s/n, E-12017 Castelló.

configuration as well as the type of magnitude involved in the statement. A recent study [8] investigated human brain differences during the process of problem-solving with reversal error showing significant differences between participants who committed reversal error and those who did not. The group that committed reversal error needed more resources and a greater cognitive demand, resulting in greater brain activation of the bilateral fronto-parietal network. The group that did not commit reversal error showed brain activation only in the left fronto-parietal network indicating that participants who are competent problem-solvers developed solid algebraic knowledge, in terms of the meaning of the variable and the logic of the process of constructing an equation. Ventura-Campos *et al.* [8] suggested that the reversal error may be a manifestation of a brain function resulting from inadequate learning. As studies in the literature [8,6] showed that traditional abstract teaching methods used so far do not help eliminate this type of error, interventions based on more concrete and visual methods are needed, where participants understand the relative size relation between the variables in a SP-type equation.

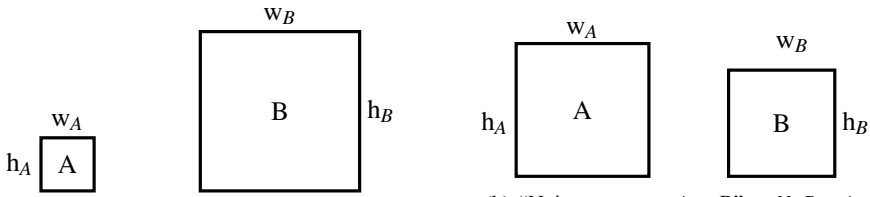
To the best of our knowledge, applications in the literature related to reversal error were designed and implemented as tests for researchers to gather data in order to conduct their research. Inspired by the interactive videogames developed in the literature to train spatial reasoning skills [9,10], we developed the $\overset{\curvearrowleft}{\varepsilon}$ *coach* serious game to train users to avoid committing reversal error by showing them the relation among the compared entities visually, in two ways: (i) qualitatively showing the containers of both entities and (ii) quantitatively by helping users resolving their written equations interactively.

The rest of the paper is organized as follows. Section 2 presents a qualitative model for comparing the size of the entities in *SP-type* problems. Section 3 presents the $\overset{\curvearrowleft}{\varepsilon}$ *coach* serious game prototype. Finally a discussion is provided.

2. Modelling Qualitatively the Container Relative Size in the *SP-type* problems

In the literature, Qualitative Spatial Representations and Reasoning (QSR) [11,12] model and reason about properties of *space* (i.e. topology, location, direction, proximity, geometry, size, etc.) and their evolution between continuous neighbouring situations. In cognitive science, qualitative models [13] have also been successful to solve oddity tasks [14], Raven's Progressive Matrices [15], 3D perspective description matching [16] and paper folding-and-punching reasoning [17]. As far as we are concerned, QSR has been never applied to model the container relative size in *SP-type* problems.

In order to qualitatively visualize the relative size relation *"N times as many A as B"*, a container-box is drawn without elements whose size must represent this relation.



(a) "N times as many B as A" $\rightarrow N \cdot A = B$. (b) "N times as many A as B" $\rightarrow N \cdot B = A$.

Figure 1. Visualizing containers: in (a) A is *smaller-than* B; whereas in (b) A is *larger-than* B.

In order to determine logically which entity, A or B, is larger, a qualitative model can be used:

$$S = \langle A_S, Q_{RS}, OP_{RS} \rangle$$

where A_S refers to the arity or the number of entities implied in each relation, that is 2 in *SP-type* problems (i.e., A and B); Q_{RS} refers to the reference system defined as the set of relations between the objects which depends on the level of granularity considered; and OP_{RS} refers to the operations that define all possible relations between the entities.

Relative size is a continuous magnitude that can be discretised using the following reference system $Q_{RS} = \{S_{Rel}, S_{LAB}, S_{INT}\}$, where S_{Rel} refers to the relation for comparing the entities; S_{LAB} refers to the set of labels for relatize size; and S_{INT} refers to the values of S_{Rel} related to each label. In general, $S_{LAB} = \{S_1, S_2, ..., S_K\}$ and $S_{INT} = \{[0, s_1], (s_1, s_2], ..., (s_{K-1}, s_K]\}$ where K is the number of concepts used for defining relative sizes. As an example,

$S_{Rel_1} = A/B$

$S_{LAB1} = \{$*smaller-than* ($<<$), *same-as* ($==$), *larger-than* ($>>$)$\}$

$L_{INT1} = \{(0, 1), 1, (1, \infty)\}$

For *SP-type* problems like *"N times as many A as B"*, we must take into account all the possible relations between entities, which take the form of equations:

$$S_{Rel_2} = Eq_{A,B,\square} = \{\ A = N \square B, \xrightarrow[\text{relation}]{\text{reversed}} B = N \square A, \text{ and its derivatives } \}$$

where N is the factor and takes the values: $N \in \mathbb{N}$; and $\square$ is the operator symbol and takes the values: $\square \in \{+, \cdot, -, /\}$.

$Eq_{A,B,\square}$		symmetric by $=$	symmetric by $\square$	symmetric by $=$
	$A = N \square B$	$N \square B = A$	$A = B \square N$	$B \square N = A$
reversed	$B = N \square A$	$N \square A = B$	$B = A \square N$	$A \square N = B$

Equations symmetrical by $\square$ are equivalent only if the operators are commutative:

$A = N \square B \ \equiv \ A = B \square N \iff \square = \{+, \cdot\}$

$B = N \square A \ \equiv \ B = A \square N \iff \square = \{+, \cdot\}$

In order to qualitatively visualize each entity A or B as larger or smaller in the serious game, relations between the operator symbol $\square$, the factor N, and the order of them in the equation must be taken into account. Table 1 and Table 2 shows how to infer if A is *smaller-than* B or A is *larger-than* B, or if A is the *same-as* B by analysing the size relations represented as equations in *SP-type* problems.

Note also that:

A *smaller-than* B $\xleftrightarrow{\text{involves}}$ B *larger-than* A

B *smaller-than* A $\xleftrightarrow{\text{involves}}$ A *larger-than* B

A *same-as* B $\xleftrightarrow{\text{involves}}$ B *same-as* A.

Table 1. Qualitative relative size between entities(A,B) for *SP-type* problems: factor N located after the operator $\square$.

				A *larger-than* B	A *smaller-than* B	A *same-as* B
			$\square$	$A >> B$	$A << B$	$A == B$
	$A = B \square N$	$B \square N = A$	$\cdot$	$N > 1$	–	$N = 1$
r	$B = A \square N$	$A \square N = B$		–	$N > 1$	$N = 1$
	$A = B \square N$	$B \square N = A$	$+$	$N > 0$	–	$N = 0$
r	$B = A \square N$	$A \square N = B$		–	$N > 0$	$N = 0$
	$A = B \square N$	$B \square N = A$	$-$	–	$N > 0$	$N = 0$
r	$B = A \square N$	$A \square N = B$		$N > 0$	–	$N = 0$
	$A = B \square N$	$B \square N = A$	$/$	–	$N > 1$	$N = 1; N = A/B$
r	$B = A \square N$	$A \square N = B$		$N > 1$	–	$N = 1; N = A/B$

Table 2. Qualitative relative size between entities(A,B) for *SP-type* problems: factor N located before the operator $\square$.

				A *larger-than* B	A *smaller-than* B	A *same-as* B
			$\square$	$A >> B$	$A << B$	$A == B$
	$A = N \square B$	$N \square B = A$	$\cdot$	$N > 1$	–	$N = 1$
r	$B = N \square A$	$N \square A = B$		–	$N > 1$	$N = 1$
	$A = N \square B$	$N \square B = A$	$+$	$N > 0$	–	$N = 0$
r	$B = N \square A$	$N \square A = B$		–	$N > 0$	$N = 0$
	$A = N \square B$	$N \square B = A$	$-$	$N > A + B$	$N < A + B$	$N = A + B$
r	$B = N \square A$	$N \square A = B$		$N < A + B$	$N > A + B$	$N = A + B$
	$A = N \square B$	$N \square B = A$	$/$	$N > A \cdot B$	$N < A \cdot B$	$N = A \cdot B$
r	$B = N \square A$	$N \square A = B$		$N < A \cdot B$	$N > A \cdot B$	$N = A \cdot B$

From Table 1 and Table 2 we can infer which entity A or B is larger and represent their containers visually. In order to know the exact quantitative relation between A and B, the equation relating A and B must be resolved by assigning values to A, to B, and to the operator ($\square$) and the factor (N). The minimum container area of an entity (C_0) is defined by $C_0 = w_0 \cdot h_0$ where h_0 is the height and w_0 is the width of the image representing the icon. So, the area for the container representing A is $C_A = A \cdot C_0$ and the area for the container representing B is $C_B = B \cdot C_0$, and their wide sides and high sides (as shown in Figure 1) are: $w_A = h_A = \sqrt{C_A}$ and $w_B = h_B = \sqrt{C_B}$.

3. The $\overset{\curvearrowleft}{\varepsilon}$ *coach* serious game prototype

The $\overset{\curvearrowleft}{\varepsilon}$ *coach* prototype[2] has been designed and implemented to train users to avoid reversal error. It has been developed using Unity3D 2021.3.12f1 and it is targeted to Android and iOS devices. The $\overset{\curvearrowleft}{\varepsilon}$ *coach* serious game runs on any tablet/mobile with an operating system Android 4.1 Jelly Bean (API 16) or upper version.

The $\overset{\curvearrowleft}{\varepsilon}$ *coach* targets teenagers at their last stage at the high school as users, so that they resolve their reversal error problem before starting their jobs or their studies at university.

[2] *Spatial Reasoning Games* channel: https://www.youtube.com/@spatialreasoninggames

Section 3.1 explains the gameplay at $\overset{\curvearrowleft}{\varepsilon}$ *coach* and Section 3.2 describes its design and development.

3.1. The gameplay at $\overset{\curvearrowleft}{\varepsilon}$ coach prototype

The $\overset{\curvearrowleft}{\varepsilon}$ *coach* prototype includes 8 levels, composed of a set of 10 exercises, which are selected and downloaded from a dataset stored in Firebase Firestore. As Figure 2 shows, the gameplay for each exercise involves the following interactive steps: **(1) Problem analysis:** users are presented with the written problem; **(2) Equation writing:** users write a equation using the calculator displayed on screen which depends on the read problem; **(3) Equation reasoning:** users must write with their own words the sentence that better represents their equation; **(4) Equation resolution:** if users do not commit reversal error, their equation is broken down so that it is resolved by providing numerical values to the variables in order to validate the equation; otherwise, if users commit reversal error, the game shows them a visual representation of both entities sizes, this representation can be qualitative (i.e. showing the relative size of both containers) or quantitative (i.e. showing the concrete images of entities inside a container).

The game starts on the path menu where users can visualize which levels are finished and which are available to complete. By touching a round button on the path, players can start solving the exercise corresponding to that level. Once players solve an exercise, the progress bar level is displayed. After users complete at least 8 exercises, they can go back to the main screen to continue the path in the map.

3.2. Development of $\overset{\curvearrowleft}{\varepsilon}$ coach as a serious game prototype

Figure 3 outlines the main components of the $\overset{\curvearrowleft}{\varepsilon}$ *coach* serious game prototype:

(1) Loading exercises: written exercises are imported from a file in the cloud stored in Firebase's Firestore, a flexible and scalable NoSQL data model. The data are stored in documents with fields mapping values and these documents are stored in collections, which are used to build queries. For each exercise, the data stored is: a collection with the statement, the labels corresponding to the related entities (i.e. A, B), the type of operation ($\square$), the factor N, the possible equations as well as the reversed equations which relate both entities.

(2) Relative Visual Representation: in order to help players visualize the relation between the two entities in the *SP-type* problem, the $\overset{\curvearrowleft}{\varepsilon}$ *coach* displays two containers of related size. Users who committed reversal error are shown two displays of relative sizes. Each display represents one of the entities in the *SP-type* problem and their size depend on the values given by the written statement in the exercise. An example of a *SP-type* problem statement can be: *"There are 3 times as many trees as birds"*, so the display corresponding to the entity *trees* (T) must be 3 times larger than the display representing the entity *birds* (B). This relation can be shown qualitatively by the size of the containers as explained in Section 2 or it can be shown quantitatively since these displays can also include as many graphics as mentioned in the statement (see Figure 3).

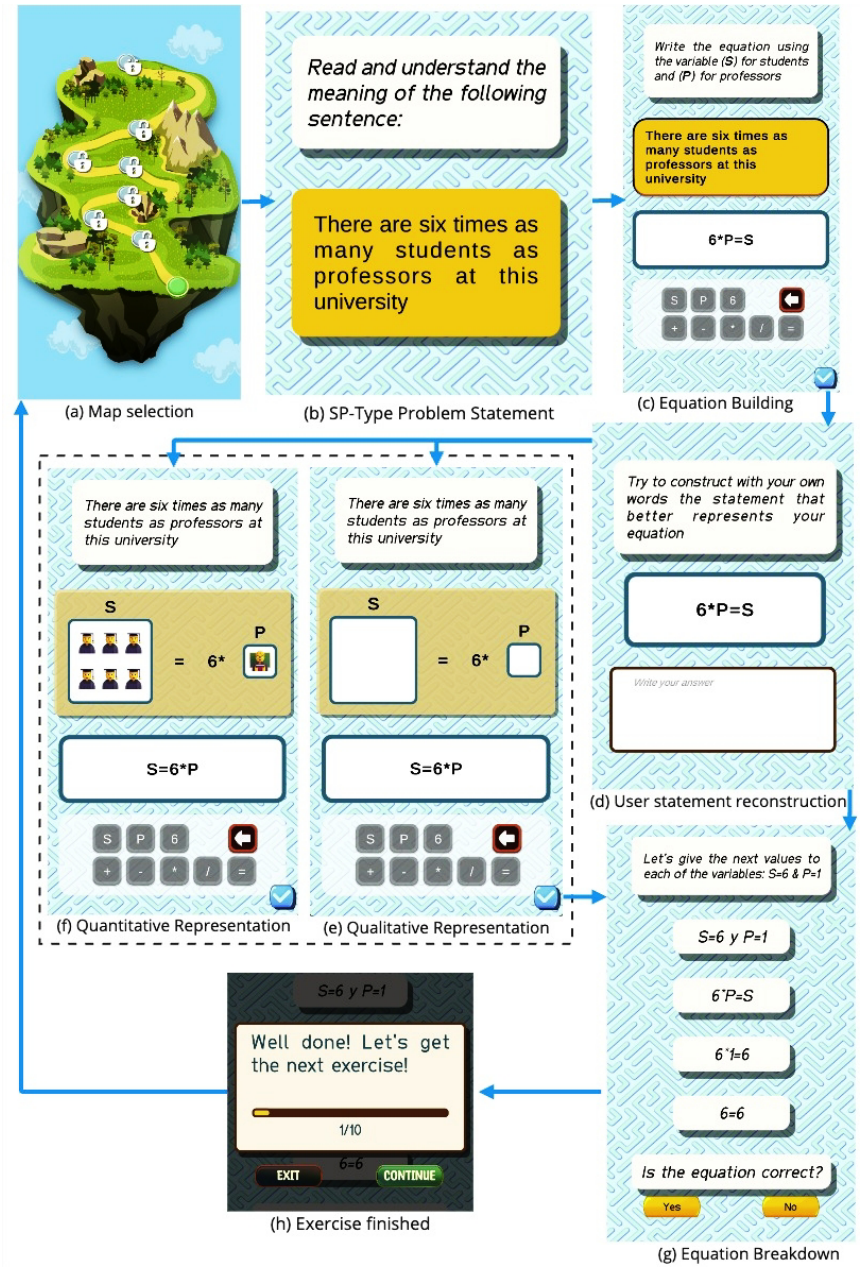


Figure 2. Gameplay at $\hat{\varepsilon}$ *coach*: overview of steps to complete for each exercise.

(3) User data collection: if allowed by users, the $\hat{\varepsilon}$ *coach* stores data about their performance at each level for generating statistics. In order to follow the rules of the European General Data Protection Regulation[3], approved and adopted by the EU Parliament

[3]General Data Protection Regulation: https://eugdpr.org/

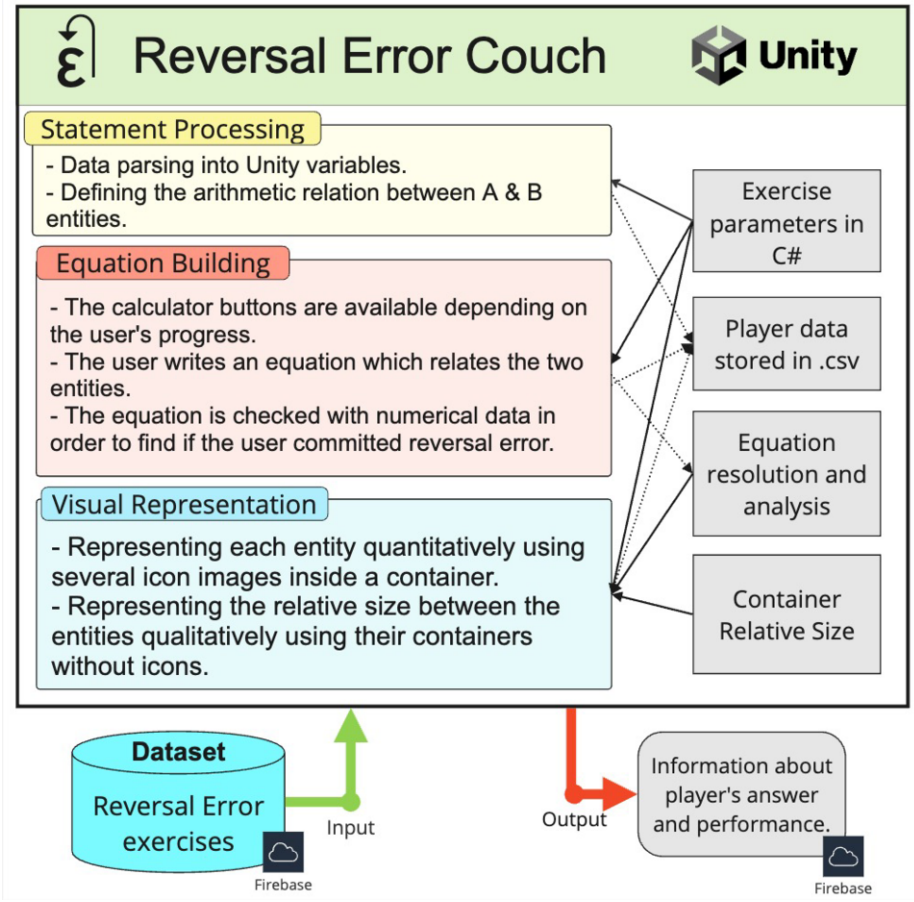


Figure 3. Diagram showing the parts composing $\overset{\curvearrowleft}{\varepsilon}$ *coach*.

in April 2016, this serious game shows initially a pop-up message that notifies the user about data storage purpose. Users can disable this option by unchecking the corresponding toggle in the settings menu.

4. Discussion

The $\overset{\curvearrowleft}{\varepsilon}$ *coach* serious game intends to train students to detect and avoid reversal error when formulating *SP-type* problems. It is based on a map and presents an intuitive and interactive gamified experience. Players are presented with algebraic word problems formulated using typical narratives and then they are assisted on finding the corresponding relation between the entities presented. For that, two approaches are proposed: (i) a qualitative visual representation of containers whose size represent the relative relation of size between the entities and (ii) reasoning behind the formulated equation by breaking it down until it is resolved.

A dataset of *SP-type* problems in Spanish has been built which is imported by the $\overset{\curvearrowleft}{\varepsilon}$ *coach* prototype. This dataset is intended to be used in a study where participants who

usually commit reversal error will be trained to avoid it using the $\overset{\curvearrowleft}{\varepsilon}$ *coach*. If the results of this study show improvement in participants' performance, $\overset{\curvearrowleft}{\varepsilon}$ *coach* is intended to be transferred to society as an educational tool.

As future work, we intend to carry out controlled tests on our prototype in our lab in order to finalize the product and also to gather gameplay data to find out which approach do players prefer in order to clarify the entities relative size relation: if (i) visualization of containers or (ii) equation breakdown resolution.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Science under grants PID2021-123152OB-C22 and PID2020-118763GA-100 both funded by MCIN/AEI/10.13039/501100011033 and by the European Union and FEDER/ERDF (European Regional Development Funds). The project UJI-A2021-04 financed by Universitat Jaume I and the funding by the Ministry of Labor and Social Economy under INVESTIGO Program grant INVEST/2022/308 are also acknowledged.

Z. Falomir acknowledges the Ramon y Cajal fellowship (RYC2019-027177-I/AEI/10.13039/501100011033) awarded by the Spanish Ministry of Science, Innovation and Universities.

References

[1] Clement J. Algebra word problem solutions: Thought processes underlying a common misconception. Journal for research in mathematics education. 1982;13(1):16-30.

[2] Clement J, Lochhead J, Monk GS. Translation difficulties in learning mathematics. The American Mathematical Monthly. 1981;88(4):286-90.

[3] Cohen E, Kanim SE. Factors influencing the algebra "reversal error". American Journal of Physics. 2005 11;73(11):1072-8.

[4] Jankvist UT, Niss M. The Students–Professors Problem—The Reversal Error and Beyond. Implementation and Replication Studies in Mathematics Education. 2021;1(2):190 226.

[5] González-Calero JA, Arnau D, Laserna-Belenguer B. Influence of additive and multiplicative structure and direction of comparison on the reversal error. Educational Studies in Mathematics. 2015 05;89:1-15.

[6] Kim SH, Phang D, An T, Yi JS, Kenney R, Uhan NA. POETIC: Interactive solutions to alleviate the reversal error in student–professor type problems. International Journal of Human-Computer Studies. 2014;72(1):12-22.

[7] Soneira C, Gonzalez-Calero JA, Arnau D. An assessment of the sources of the reversal error through classic and new variables. Educational Studies in Mathematics. 2018 09;99.

[8] Ventura-Campos N, Ferrando-Esteve N, Epifanio I. The underlying neural bases of the reversal error while solving algebraic word problems. Scientific Reports. 2022;12:1-14.

[9] Falomir Z, Oliver E. Q3D-Game: A Tool for Training Users' 3D Spatial Skills. In: Novais P, Konomi S, editors. Intelligent Environments 2016 - Workshop Proceedings of the 12th International Conference on Intelligent Environments, IE 2016, London, United Kingdom, September 14-16, 2016.. vol. 21 of Ambient Intelligence and Smart Environments. IOS Press; 2016. p. 187-96.

[10] Falomir Z, Tarin R, Puerta A, Garcia-Segarra P. An interactive game for training reasoning about paper folding. Multim Tools Appl. 2021;80(5):6535-66. Available from: https://doi.org/10.1007/s11042-020-09830-5.

[11] Cohn AG, Renz J. Qualitative Spatial Reasoning, Handbook of Knowledge Representation. F Harmelen VL, Porter B, editors. Wiley-ISTE, London: Elsevier; 2007.

[12] Ligozat G. Qualitative Spatial and Temporal Reasoning. Wiley-ISTE, London: MIT Press; 2011.

[13] Forbus KD. Qualitative modeling. Wiley Interdisciplinary Reviews: Cognitive Science. 2011;2(4):374-91. Available from: http://dx.doi.org/10.1002/wcs.115.

[14] Lovett A, Forbus K. Cultural commonalities and differences in spatial problem-solving: A computational analysis. Cognition. 2011;121(2):281 287. Available from: http://dx.doi.org/10.1016/j.cognition.2011.06.012.

[15] Lovett A, Forbus K. Modeling visual problem solving as analogical reasoning. Psychological Review. 2017;124(1):60 90.

[16] Falomir Z. A Qualitative Model for Reasoning about 3D Objects using Depth and Different Perspectives. In: Lechowski T, Wałęga P, Zawidzki M, editors. LQMR 2015 Workshop. vol. 7 of Annals of Computer Science and Information Systems. PTI; 2015. p. 3-11. Available from: http://dx.doi.org/10.15439/2015F370.

[17] Falomir Z. Towards A Qualitative Descriptor for Paper Folding Reasoning. In: Proc. of the 29th International Workshop on Qualitative Reasoning; 2016. Co-located with IJCAI'2016 in New York, USA.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230689

Recommenders for Improved Lesson Planning in Formal Education

Pere GARCIA-CALVÉS [a] and Jordi SABATER-MIR [a,1] and Pablo ARAMENDIA [a] and Jordi COROMINAS [b] and Bryan Miguel PEREZ RUCHAT [b]
[a] *IIIA, CSIC, Campus UAB, E-08193. Bellaterra, Catalonia (Spain)*
[b] *Didactic Labs, C/Emili Grahit, 91. 17003, Girona, Catalonia (Spain)*

Abstract. This paper presents a system that combines a multi-recommender mechanism and an iterative algorithm to assist teachers in formal education in planning learning activities for students. It aligns with the latest educational trends, emphasising the central role of competencies. The system allows for initial planning based on specified competencies and knowledge, among other preferences. After that, it helps with alternative suggestions during the course when deficiencies in the learning of competencies are detected.

Keywords. Recommender systems, education.

1. Introduction

One of the primary responsibilities of a teacher in formal education is to meticulously plan the sequence of learning activities that students will engage in throughout the academic year. This task, when executed effectively, can be quite intricate and time-consuming. It involves taking into account numerous factors such as the competencies and knowledge students are expected to acquire as per educational regulations, the teacher's personal preferences regarding learning topics, the duration and complexity of the learning units, and more.

Furthermore, after the sequence is established and the course starts, things may not always proceed as intended. The level of competencies and knowledge attained by students upon completing a learning activity might not meet expectations. As a result, the subsequent scheduled learning unit may not be the most suitable and a readjustment is necessary.

The tool presented in this paper has been specifically created to assist teachers in formal education in this undertaking. It not only aids in creating an initial plan of learning activities but also provides alternative suggestions when any issues arise during the academic course. The tool's core is a multi-recommender system that, following an iterative process, assists the teacher in creating the sequence of learning units.

The academic course is organised in what we call *Learning Blocks* or LB from now on. A LB refers to a series of learning activities usually centred around a specific topic

[1]Corresponding Author: Jordi Sabater-Mir, IIIA, CSIC, Campus UAB, E-08193. Bellaterra, Catalonia (Spain); E-mail: jsabater@iiia.csic.es.

or theme and aimed at enhancing a set of competencies and knowledge. It has a flexible duration, typically ranging from one to two weeks up to a trimester.

While there is a long tradition of applying recommender systems in the field of education [1], the current proposal stands out due to several distinctive features:

- It is a tool oriented to teachers instead of students. The tool is intended to help with course preparation and to suggest possible adaptations when certain conditions are detected once the course has started.
- The proposed learning activities need to be suitable for an entire class. Our focus is not on personalizing the learning path of individual students, but rather on accommodating a collective group of students who are expected to engage in the same learning activities, typically within a face-to-face learning environment.
- It is adapted to suit a competency-based learning (as established by the Spanish education law, the LOMLOE).

2. The stages of the system

Figure 1 illustrates the general idea behind the multi-recommender system. From a set (which can be quite large) of learning activities, a preliminary filtering is performed, and the results are presented one by one to a set of specialised recommenders that analyse the learning activity from different perspectives. Each recommender can be seen as an expert in a specific area who, considering contextual variables as well, analyses the problem (in this case, the learning activity) from a particular viewpoint and provides its opinion. As a result of this analysis, the user receives a set of opinions to aid in selecting one of these learning activities. The ordered list of learning activities is the outcome of aggregating the specific opinions, weighted by the user's importance assigned to each analysis area.

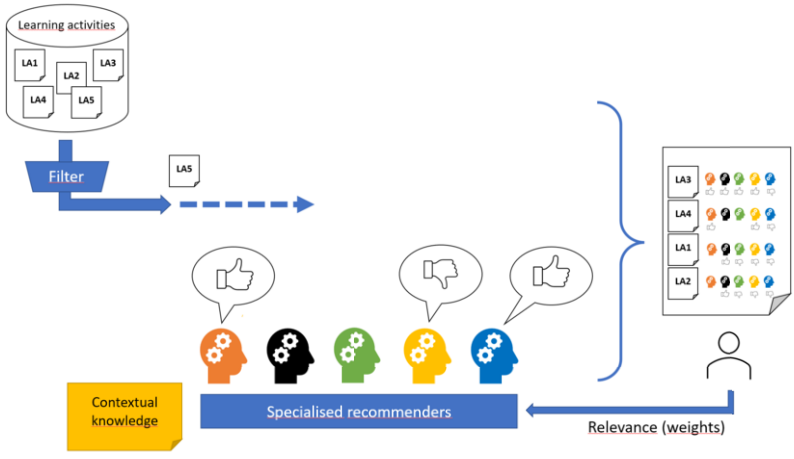


Figure 1. General outline of the multi-recommender system.

The use of the tool by the teacher involves three distinct stages, two performed before the course begins and the third during the academic year.

The **Preparation phase** entails the professor outlining the course structure and their preferences. This involves organising the course into a series of LBs. For each LB, the professor specifies the competencies and knowledge he/she would like to work on as well as the number of hours for each competency/knowledge. This is taken by the system as a soft constraint to, later on, propose learning activities for that LB. In this phase, other course characteristics, such as the number of desired instructional units for each LB, are also determined.

After the Preparation stage, the next phase is the **Programming phase**. During this phase, the professor utilises a multi-recommender system to aid in identifying the most appropriate learning activities that fulfil the requirements established in the Preparation phase. To ensure the teacher's involvement in the final selection of each learning activity, the multi-recommender system has been designed as an iterative mechanism. In each iteration, the system presents a set of learning activities from which the professor can choose the most suitable option. As the programming progresses, the system takes into account the already assigned learniong activities when suggesting the next ones. By the end of the Programming Phase, the professor has a sequence of learning activities that closely align with the course structure designed during the Preparation Phase.

Once the course has been designed, the instructional period commences, and students proceed to execute the learning activities in the predetermined sequence from the Preparation phase. This new phase is referred to as the **Execution Phase**. By conducting evaluations on the learning activities, the system can track the students' progress in attaining competencies and knowledge. During the Execution phase, this monitoring data is used as a source for recommendations to assess the continued validity of the next selected learning activity from the Programming phase.

If a deviation in the acquisition of competencies and knowledge is detected, indicating the need to modify the next learning activity, the multi-recommender system supports the professor in finding alternative learning activities. These alternatives are designed to preserve the original structure while addressing the identified shortcomings that arise during the course's progression. Notice that the system does not provide personalised alternatives to each student because the system is intended to provide learning activities that are performed by the entire class. However, there is a recommender (see REC_6 in section 3 that suggests tailored learning activities that cater to the specific needs of diverse student groups at different stages of the learning journey.

For instance, consider a scenario where there are two distinct groups of students in the learning activities sequence: one group that is progressing well and another group facing challenges. REC_6 adeptly handles this situation by initially recommending activities that suit the needs of the first group. As the sequence unfolds, REC_6 dynamically adjusts its recommendations to better suit the second group, providing them with appropriate learning opportunities and, in turn, striving to achieve a sense of balance and equity among the students' learning experiences. This approach ensures that both high-performing students and those encountering difficulties receive the support and guidance they require to progress optimally.

The different stages are supported by a prototype GUI that leads the teacher in the process.

3. The multi-recommender

The multi-recommender system serves as the centerpiece of the overall system. Its primary function is to assist the teacher in choosing the most suitable learning activity at a certain point in the sequence of learning activities that compound an LB.

Beginning with a potentially extensive collection of learning activities stored in a database, an initial filtering process is applied to narrow down the options. The resulting subset is then presented to a group of specialised recommenders, each acting as an expert in a particular domain. Considering also contextual variables, these recommenders analyse the problem at hand, evaluating the usefulness of each learning activity from their unique perspective and providing their opinion. The user is subsequently provided with a collection of opinions to assist in selecting a learning activity. The final ordered list of learning activities is generated by aggregating these individual opinions, weighted based on the user's assigned importance to each recommender. This empowers the user to determine the most relevant "experts" and make an informed final decision.

To give an opinion on a teaching unit, the multi-recommender takes as an input different elements: the properties of the teaching unit, the general preferences of the teacher, the sequence of the already selected teaching units in that LB and the evaluations of the class till that moment (if we are already in the Execution phase). Not all recommenders use all the information.

We have implemented 7 recommenders that help the teacher to analyse the learning activities from different perspectives.

- **REC_1**: *Competency-based recommender*. Evaluates learning activities based on competency compatibility, teacher preferences, and already selected activities.
- **REC_2**: *Knowledge-based recommender*. It does the same as the competency-based recommender but considering knowledge instead of competencies.
- **REC_3**: *Keyword-based recommender*. Considers matching keywords between learning activities and teacher preferences.
- **REC_4**: *Rating-based recommender*. Evaluates learning activities based on quality ratings.
- **REC_5**: *Flagship-based recommender*. It determines the utility of each learning activity based on its relevance according to the editor's parameters.
- **REC_6**: *Recommender based on students' competency level*. This recommender takes into account the suitability of the competencies it works on, relative to the current competency level of the Class Group, while also ensuring equal opportunities for every student.
- **REC_7**: *Recommender based on students' knowledge level*. It does the same as the previous recommender but considers knowledge instead of competencies.

Recommenders 6 and 7 are used only in the Execution phase. The teacher has the flexibility to determine the weight assigned to each recommender and can choose to discard certain recommenders if desired.

4. The iterative algorithm

The Programming Phase of the multi-recommender system is based on an iterative process to find a sequence of learning units that make up the LBs and collectively meet the specifications determined in the Preparation phase.

Instead of using a system based on global optimization that would propose the entire sequence to the teacher, we have opted for an iterative process.

The choice of a recommender system based on an iterative algorithm instead of using a system based on global optimization is given by the nature of the problem itself. Deciding on a sequence of didactic units is a highly subjective task that depends heavily on the teacher. There is no ”correct solution,” and there are many elements that depend on the teacher and that the system is unaware of. Asking the teachers to provide all the information that the system would need to make an accurate proposal is costly and impractical. Therefore, it has been decided to use a user-in-the-loop approach and to involve the teacher in the sequence creation process, granting them decision-making capacity at all times, with the assistance of the multi-recommender system.

The iterative algorithm is structured as follows:

- **Adaptation**, to pass from the global objective of the LB to the objective of each iteration.
- **Evaluation.** Each learning unit is evaluated by the recommenders.
- **Concentration.** Obtain the best k learning units from the multi-recommender. The list of the “best” is built from the opinions of each recommender and the weights assigned by the user.
- **Selection.** After receiving the proposal from the multi-recommender, the user is given the freedom to choose a learning unit, which does not necessarily have to be from the previous list proposed by the multi-recommender. The user has the ability to review the opinions from the recommenders as well as other characteristics of the available learning units.
- **Termination.** Finish the iteration if the objectives of the LB have been accomplished.
- **Adjustment.** The objective for the next iteration is adjusted based on the selection made by the user in the current iteration.

The objective of the iterative algorithm is to assist the teacher in populating each of the defined LBs with appropriate learning activities during the Programming phase. The system assumes that the sequence of learning units within each LB is independent of the previous LBs. The system’s primary concern with regard to previous LBs is only to prevent the repetition of learning activities that have already been selected. However, through the GUI that has been built, the teacher has a comprehensive view of the overall coverage of competencies and knowledge achieved by the total number of learning units already selected.

5. Conclusions

We have developed a tool that combines a multi-recommender mechanism with an iterative algorithm to support teachers in designing the sequence of learning activities for

students throughout the academic year. There is already a first prototype that has been tested with generated learning units with different characteristics to test the Programming phase, and simulated evaluations to test the Execution phase. Figure 2 shows an example screenshot of the tool.

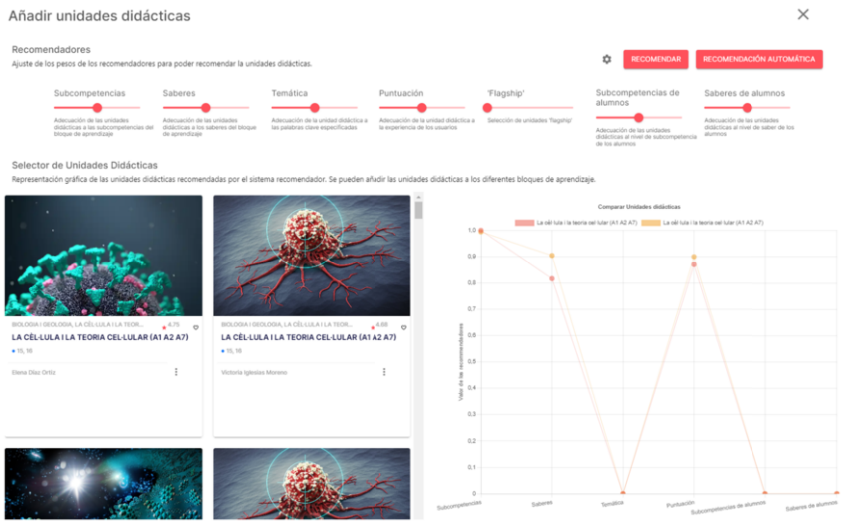


Figure 2. Example screenshot of the tool.

Acknowledgements

This research has been funded by the Generalitat de Catalunya through the project "ADDIA: Plataforma d'aprenentatge competencial personalitzat mitjançant l'ús d'intel·ligència artificial" (INNOTEC - ACE014/20/000039) and by the CSIC through the project "MARA: Desarrollo de mecanismos avanzados de interacción humano-máquina en el contexto de recomendadores argumentativos" (Intramural CSIC - 202050E132)

References

[1] Urdaneta-Ponte, María Cora, Amaia Mendez-Zorrilla, and Ibon Oleagordia-Ruiz. 2021. "Recommendation Systems for Education: Systematic Review" Electronics 10, no. 14: 1611. https://doi.org/10.3390/electronics10141611

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230690

Capturing Young People's Interest in Food Waste Reduction: An Approach Using Qualitative Reasoning

Konstantina ZACHARAKI [a,1], Jennifer NGUYEN [a], Queralt PRAT-I-PUBILL [a] and Núria AGELL [a]
[a] *ESADE Business School, Ramon Llull University*
ORCiD ID: Konstantina Zacharaki https://orcid.org/0000-0001-8264-8457,
Jennifer Nguyen https://orcid.org/0000-0002-7498-7536,
Queralt Prat-i-Pubill https://orcid.org/0000-0003-3505-4490,
Núria Agell https://orcid.org/0000-0001-9264-2147

Abstract. According to the United Nations Food and Agriculture Organization (FAO) more than one third of the food produced globally is lost or wasted. This tremendous amount of food waste highlights the urgent need to investigate this phenomenon promptly with powerful methods in order to reduce it. In the present work, we consider a qualitative reasoning approach in an attempt to understand people's interest in the food waste (FW) problem. To do so, we have run an in-person taste experiment and have acquired data from 188 participants. We apply a measure based on a cosine similarity to capture the degree of interest towards the environment as individuals respond to the New Environmental Paradigm scale (NEP). Previously, they had to decide on whether a piece of fruit allegedly coming from the supermarket is better than another one allegedly coming from an alternative source such as applications designed to reduce food waste. We are using the taste experiment and the questions as a base to gain a more in-depth understanding of the drivers behind FW.

Keywords. Qualitative reasoning, similarity measures, sustainable decision-making, food waste, experimental data.

1. Introduction

Households contribute the most to FW, accounting for 53% of total waste in 2020. Addressing this issue is crucial for a sustainable food supply chain [1]. Numerous research studies are now utilizing artificial intelligence and machine learning to research FW ([2]; [3]; [4]; [5]; [6]; [7]). In this paper we examine if a ranking method that considers a cosine similarity measure is more suitable to capture people's attitudes towards FW than a classic psychometric scale. To do so, we will use a cosine similarity measure to capture individuals' profiles towards FW [8]. Our objective is to understand individuals' behavior when they must make decisions in terms of food purchases or throwing away food. There is a plethora of sources where people can buy products that

[1] Corresponding Author: Konstantina Zacharaki, konstantina.zacharaki@esade.edu.

combat FW, selling the surplus of food or products that are suboptimal, such as the 'Too good to go' application. This platform is a network that includes a variety of different establishments like restaurants, supermarkets, and bakeries.

We ran an in-person experiment with 188 participants. Participants were led to believe that they were trying two different kinds of apple, one from the Supermarket and another one from the 'Too good to go' application. After the tasting, they had to answer a series of questions. Our goal in this paper is to investigate whether we can extract people's profiles based on environmental attitudes. We are using a qualitative reasoning approach based on a linguistic computational method to aggregate and compare the qualitative responses from the NEP scale [9]. We consider the cosine similarity to compute the distance between vectors, where each one of the components corresponds to the frequency with which the linguistic terms were used in the NEP scale. Our approach, by applying the cosine similarity measure to participants' qualitative answers to the NEP scale, allows us to differentiate between two different fruit preference groups, where classic statistics fail to do the same.

2. Preliminaries

2.1. New Environmental Paradigm Scale (NEP)

The NEP scale is the most popular unidimensional measure used to quantify environmental concern and pro-environmental tendencies. The NEP scale has been used in many different topics, ages and cultures ([10]; [11]; [12] ; for a meta-analysis see [13]). It was originally designed in 1978 by Dunlap and Van Liere based on the premise that within environmentalism, the relationship between humans and nature is implicitly entangled. In the 15-item scale (1=Strongly disagree, 5= Strongly agree), which is the one used in the present manuscript, the eight odd-numbered questions were framed in an ecocentric manner, while the even-numbered questions are framed in an anthropocentric manner. The goal of using the scale is to get a value per participant that will classify them as eco-friendly or not. Two different approaches have already been proposed, in previous studies, when analyzing the responses. The first method (method 1) entails reverse coding the even-numbered questions. As these questions are framed in a human-centric way, responses from the lower part of the scale in these questions indicate pro-environmental ideas. Then the evaluations of each participant are averaged together, therefore the values can range from 0 to 5. This technique was implemented for example in [11] in their application of the NEP scale in a Greek context. The second method (method 2) is described in [12] where the environmental worldview of employees in Kosovo is studied. The authors counted the frequency of the 'Agree' and 'Strongly Agree' items in the odd-numbered questions (ecocentric) and the frequency of 'Disagree' and 'Strongly Disagree' in the even-numbered questions (anthropocentric). They calculated the NEP score by aggregating the frequency of these responses. The values could range from 0 to 15.

2.2. Similarity measure based on Linguistic Terms

In this subsection the qualitative reasoning methodology, based in linguistic computation, that is going to be used in the experiment is presented. We introduce the concept of

linguistic terms, as well as some key definitions related to them that will be used to capture similarities in participants' answers.

Let S be a totally ordered set of basic linguistic terms (BLTs), $S = \{s_1, \ldots, s_n\}$, with $s_1 < \ldots < s_n$, corresponding to the values of a rating scale [14]. In order to determine the strength of people's attitudes towards environmentalism, we introduce a distance measure, between vectors expressed in S terms, based on a weighted cosine similarity. Given the vectors $P_i = (P_{i1}, \ldots, P_{ik})$ and $P_j = (P_{j1}, \ldots, P_{jk})$ corresponding to the linguistic responses of a set of k items of two different observations, we compute new vectors $(F_{i1}, \ldots, F_{in})$ and $(F_{j1}, \ldots, F_{jn})$ with frequencies corresponding to some linguistic terms. Then the similarity index is computed as:

$$Sim(P_i, P_j) = \frac{\sum_{l=1}^{n} w_l \cdot F_{il} \cdot F_{jl}}{\sqrt{\sum_{l=1}^{n} w_l \cdot F_{il}^2} \cdot \sqrt{\sum_{l=1}^{n} w_l \cdot F_{jl}^2}}$$

where F_{il} and F_{jl} are the absolute frequencies respectively and w_l is the normalized weight for each one of the linguistic terms. Note that the greater the similarity between two vectors, the value of the similarity measure is closer to one.

In the proposed method (method 3), similarly to method 1, we reversed coded the odd numbered questions. Then, in a similar way as in method 2, we computed F_1 and F_2 as the cumulative frequencies of $\{s_1, s_2, s_3\}$ and the cumulative frequencies of $\{s_4, s_5\}$, respectively. The rationale is to use the frequencies approach, as in method 2, but taking into account the cosine similarity to the optimal ecocentric vector (pure environmentalist).

3. Experiment

3.1. Participants

One hundred-eighty-one participants were tested ($Mean_{age}$ =19, SD_{age} = 1, female =97). An additional 7 participants were tested but eliminated from the final sample due to the following reasons: technical error (n=5), not enough data (n=2). Participants were undergraduate students from the ESADE Business School and were given extra-credit scores in exchange. The current experiment was approved by the local ethical committee of ESADE (009/2023), and all data were treated confidentially. Participants signed a written consent form before the experiment began. Data collection is still ongoing.

3.2. Materials

A local Catalan variety of apples (Golden Empordà) was chosen for the experiment. The fruit was always fresh and was cut no more than 10 minutes before testing. To ensure symmetrical pieces, we used an apple cutter after peeling. The exact same apple was served in two transparent separate bowls that were later identified as coming from the supermarket or the 'Too good to go' application. This type of procedure is used widely in marketing, cognitive science, and other disciplines when the research question

revolves around the external factors that can influence taste. For instance, participants were offered the exact same chocolate but critically it was shaped differently using a 3-D printer. Participants' taste differed depending on the shape (circular or angular), this is due to a mental association between shapes and other features like taste or sounds [15]. This type of experimental settings is well accepted and typically include deception as the same product is used but labelled differently to investigate whether labelling the product can change the perception of the consumers.

3.3. Paradigm

We adapted the paradigm used in [16] to our research question. Participants were tested in a sound attenuated room in the Decision Lab located at the ESADE Sant Cugat Campus. They were seated in a chair opposite to a table where two transparent bowls were placed with the slices of apple. In the middle a recyclable, cardboard paper cup was placed with water. Participants were first asked to sign the consent form and whether they had any food allergies. The experimenter asked if they had heard of the 'Too good to go app' before. Regardless of their response, the experimenter gave the same brief description to all participants to ensure that they all have the same understanding of the alleged origin of the fruit. Then they were asked to try one kind of apple. The experimenter informed them where the apple came from (Supermarket or 'Too good to go') and that they can have as many slices as they like to make sure they have an idea about the taste of the fruit. Once they were done with the first apple, they were instructed to have some water and then to try the other apple. The order and side of presentation of the two kinds of apple was counterbalanced across participants. After the tasting they were guided to an adjoining room, where a Microsoft Surface Tablet with a physical keyboard was placed. They were asked to answer a series of questions privately. Among them, they had to answer a closed-ended question on which apple they preferred (Supermarket, 'Too good to go', or Both). They were also asked to share their opinion on the taste and the quality of the fruit, and to fill in the New Environmental Paradigm scale [9].

3.4. Data analysis

The data presented in this manuscript belongs to a bigger project on FW. We computed the NEP scores (method 1, method 2) and the cosine similarity distance (method 3). Out of the 181 participants, 48 preferred the fruit coming from the supermarket (26%), 70 had no preference (39%) and 63 preferred the fruit coming from the application (35%). We calculated the proportion of participants that were above the mean, i.e., environmentalist in each method. The thresholds considered to classify participants in two groups according to the answers they gave on the NEP scale for the two psychometric methods are 3 for method 1 and 7 for method 2.

We ran three two-proportions z-test comparing the two groups separately for the two methods (Supermarket vs. 'Too good to go', 'Too good to go' vs. Both, Supermarket vs. Both). Our preliminary results show that method 1 and method 2 were unable to capture any statistical differences between the three profiles of the people showing different preferences (see table for values).

In method 3, the threshold selected was 0.707 which is the similarity of an indecisive person (the one that answers 7 times 1, 2 or 3 and 7 times 4 or 5) to a pure environmentalist. Where the similarity was greater than this opinion, a person was considered to have an environmentalist attitude. In this case, the proposed method was able to detect a significant difference between the 'Too good to go group' and the Supermarket group ($Z= -1.71$, $p= 0.04$). However, it found no differences between either group or Both.

Table 1: Comparison of the results obtained using the three methods.

Method 1	**Above Threshold**	**Below Threshold**	**Sum**
Supermarket	45 (0.94)	3 (0.06)	48 (1)
Both	63 (0.90)	7 (0.10)	70 (1)
Too good to go	57 (0.91)	6 (0.09)	63 (1)

Method 2	**Above Threshold**	**Below Threshold**	**Sum**
Supermarket	36 (0.75)	12 (0.25)	48 (1)
Both	57 (0.81)	13 (0.18)	70 (1)
Too good to go	56 (0.88)	7 (0.12)	63 (1)

Method 3	**Above Threshold**	**Below Threshold**	**Sum**
Supermarket	32 (0.67)	16 (0.33)	48 (1)
Both	51 (0.73)	19 (0.27)	70 (1)
Too good to go	51 (0.81)	12 (0.19)	63 (1)

4. Conclusions

In this paper we contribute a new real-life application to capture differences across profiles towards FW by means of a similarity measure. Traditional measures based on psychometric scales are used in most cases to analyze scale data. Our study based on the results from the NEP scale show that the majority of participants were more environmentalists. In addition, our results suggest that considering the NEP scale as a qualitative scale and using a similarity measure based on linguistic computation, we can distinguish the different profiles while the classic psychometric tools are unable to do so.

A limitation of the present experimental work is that our sample is comprised of only young adults. This is due to two reasons. Firstly, we know that younger people waste more food therefore it is worthwhile to study this particular population in an attempt to try to fight this issue to its root. Secondly, our recruitment is mainly done at the University's premises which makes it challenging to test various age groups in person so that they can try the fruit themselves.

In this paper we have combined experimental work with a qualitative reasoning approach, considering linguistic computation in the context of FW. Our findings hold the potential for replication in the future and we are optimistic that our work will inspire others to study this pressing phenomenon using innovative techniques.

Acknowledgement

This research has been partially supported by the PERCEPTIONS Research Project (PID2020-114247GB-I00), funded by the Spanish Ministry of Science and Information Technology.

References

[1] *Eurostat*. (2023). https://ec.europa.eu/eurostat/web/waste/data/database
[2] Huang, J., & Koroteev, D. D. (2021). Artificial intelligence for planning of energy and waste management. *Sustainable Energy Technologies and Assessments*, *47*(January), 101426. https://doi.org/10.1016/j.seta.2021.101426
[3] Kayikci, Y., Demir, S., Mangla, S. K., Subramanian, N., & Koc, B. (2022). Data-driven optimal dynamic pricing strategy for reducing perishable food waste at retailers. *Journal of Cleaner Production*, *344*(August 2021), 131068. https://doi.org/10.1016/j.jclepro.2022.131068
[4] Pocol, C. B., Pinoteau, M., Amuza, A., Burlea-Schiopoiu, A., & Glogoveţan, A. I. (2020). Food waste behavior among romanian consumers: A cluster analysis. *Sustainability (Switzerland)*, *12*(22), 1–17. https://doi.org/10.3390/su12229708
[5] Queiroz, A. C. M., Meirelles, F. S., & Nascimento, A. M. (2022). *AIS Electronic Library (AISeL) Applying Artificial Intelligence to Reduce Food Waste in Small Grocery Stores Applying Artificial Intelligence to Reduce Food Waste in Small Grocery Stores Completed Research Paper*.
[6] Sriman, B., Santhosh, R., Santhosh Guru, R., Saras Prasad Raju, V., Ruma Prasad, H., & Sailendiran, R. (2023). Tracking of Food Waste in Food Supply Chain Using Machine Learning. *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF) Https://Ieeexplore.Ieee.Org/Abstract/Document/10083755*, 1–4.
[7] Abouelsaad, I. A., Teiba, I. I., El-Bilawy, E. H., & El-Sharkawy, I. (2022). Artificial Intelligence and Reducing Food Waste during Harvest and Post-Harvest Processes. In *IoT-Based Smart Waste Management for Environmental Sustainability* (pp. 63–82). CRC Press.
[8] Nguyen, J., Armisen, A., Sánchez-Hernández, G., Casabayó, M., & Agell, N. (2020). An OWA-based hierarchical clustering approach to understanding users' lifestyles. *Knowledge-Based Systems*, *190*, 105308.
[9] Dunlap, R. E., Van Liere, K. D., Mertig, A. G., & Jones, R. E. (2000). Measuring endorsement of the new ecological paradigm: A revised NEP scale. Journal of Social Issues, 56(3), 425–442.
[10] Boeve-de Pauw, J., & Van Petegem, P. (2012). Cultural Differences in the Environmental Worldview of Children. *International Electronic Journal of Environmental Education*, *2*(1), 1–11.
[11] Ntanos, S., Kyriakopoulos, G., Skordoulis, M., Chalikias, M., & Arabatzis, G. (2019). An application of the new environmental paradigm (NEP) scale in a Greek context. *Energies*, *12*(2), 1–18.
[12] Zhushi-Etemi, F., Ceka, R., Çadraku, H., Bytyqi, P., Shala-Abazi, A., Fetoshi, O., Ymeri, P., & Ismaili, M. (2021). The use of NEP scale to evaluate the environmental worldview of the employees in the city of Kacanik in Republic of Kosovo. *Environment, Development and Sustainability*, *23*(6), 8195–8211.
[13] Hawcroft, L. J., & Milfont, T. L. (2010). The use (and abuse) of the new environmental paradigm scale over the last 30 years: A meta-analysis. *Journal of Environmental Psychology*, *30*(2), 143–158.
[14] Montserrat-Adell, J., Agell, N., Sánchez, M., Prats, F., & Ruiz, F. J. (2017). Modeling group assessments by means of hesitant fuzzy linguistic term sets. *Journal of Applied Logic*, *23*, 40–50.
[15] Ogata, K., Gakumi, R., Hashimoto, A., Ushiku, Y., & Yoshida, S. (2023). The influence of Bouba- and Kiki-like shape on perceived taste of chocolate pieces. Frontiers in Psychology, 14.
[16] Sörqvist, P., Hedblom, D., Holmgren, M., Haga, A., Langeborg, L., Nöstl, A., & Kågström, J. (2013). Who needs cream and sugar when there is eco-labeling? Taste and willingness to pay for "eco-friendly" coffee. PLoS ONE, 8(12).

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230691

Explainable Machine Learning Models for Predicting COVID-19 Cases in Catalonia Based on Wastewater Monitoring Data

Marc RIBALTA[a, b,1], Josep PUEYO-ROS[c], Lluis COROMINAS[c,d], Ramón BÉJAR[b], Carles MATEU[b], Edgar RUBIÓN[a]
[a] *Eurecat, Technology Centre of Catalonia, Unit of Applied Artificial Intelligence, Bilbao 72, Barcelona, 08005, Spain*
[b] *GREiA - UdL, Jaume II, 69, Lleida, 25001, Spain*
[c] *Catalan Institute for Water Research (ICRA-CERCA), Emili Grahit 101, Girona, 17002, Spain*
[d] *University of Girona, Plaça de Sant Domènec, 3, Girona, 17004, Spain*

Abstract. Surveillance of infectious diseases that can spread quickly among the population has become a critical goal nowadays due to the dramatic effect of diseases like SARS-CoV-2. One promising method to be able to monitor the spread of such diseases among the population of a city is the analysis of biological compounds in the sewage network of different cities. In this paper, we summarize the results of training a prediction model for SARS-CoV-2 cases based on historical biological data collected from different wastewater treatment plants (WWTP) located in different parts of Catalunya. We consider different approaches for the prediction problem and develop models based on extreme gradient boosting. Finally, we evaluate the quality of the results of our model and study the relevance of the different variables of the model using SHapley Additive exPlanations (SHAP) analysis.

Keywords. Explainable AI, Machine Learning, Wastewater-based epidemiology

1. Introduction

During the COVID-19 pandemic, the research community found different ways of monitoring the SARS-CoV-2 within the population to estimate the trend of infection cases. One of them is sewage surveillance to understand the prevalence of COVID-19 within the population. Different gene targets can be used to quantify the prevalence, such as the N1, N2, or E [1]. The prediction of SARS-CoV-2 cases using machine learning (ML) models is a difficult task due to its uncertainty in population behavior and unknown virus evolution, but gene target values can be processed and used to enhance them.

In this paper we consider the predictive problem in Catalunya, using historic data for COVID-19 cases plus biological indicators obtained from a total of 56 WWTPs across different points of Catalunya [2]. The general goal is to obtain predictive models for covid cases in the next period based on previous COVID-19 cases and the current presence of these biological markers in wastewater. We consider different attributes related to either the water samples or the population linked with the WWTP where the

[1] Corresponding Author: Marc Ribalta, E-mail: marc.ribalta@eurecat.org

water comes from. As ML algorithms, we use ensemble trees obtained with extreme gradient boosting. Given the challenging nature of this problem and that some attributes have a high extraction cost, we not only consider the performance of the models obtained, but we also study the relevance of the different attributes, using SHapley Additive exPlanations (SHAP) analysis. Our final aim is to provide relevant conclusions that improve the data acquisition procedure for a more reliable and sustainable monitoring network for future similar infectious diseases.

2. Materials and methods

2.1. Sampling points and collected data

The biological data from wastewater data was obtained from a total of 56 WWTPs selected based on two criteria: 80% coverage of the served population (about 6M inhabitants) and high territorial coverage (41 out of 42 regions covered). The sampling frequency was 1 week for some WWTP stations and biweekly for others, amounting to a total of 2000 samples. Each sample was obtained from a flow-proportional, 24h-composite set of INLET subsamples, where each subsample was obtained every 20 minutes[2]. Regarding the number of covid cases, we used the public data for municipal covid cases publicly reported[3].

The attributes used from each sample in the ML algorithm can be divided into four different groups: 1) parameters linked with the characteristics of the water flow including quality parameters such as TSS, COD and BOD, 2) a scaled value of the wastewater gene targets N1/N2, 3) WWTP population data such as inhabitants assisted and estimated citizens mobility, and 4) number of total covid cases in Catalunya in the last 14 days.

We used gradient boosting to learn ensembles of prediction trees as the ML model. We used different metrics to evaluate the error: the mean absolute error (MAE) and the coefficient of determination (R^2). The parameters used with xgboost are the number of trees: 500, maximum depth: 15 nodes, learning rate: 0.01, and subsample ratio: 0.5.

2.2. Predictive exercises

The problem of predicting future covid cases from biological samples in the setting considered here is very challenging, so we have considered four different specific variants of the problem, which predict the covid cases for each municipality:

1. In our first approach, we want to predict the number of cases in the next week. We use the wastewater data (N1/N2 gen presence) plus the number of covid cases in the two previous weeks, the number of inhabitants, and estimated mobility of citizens, just to know how many people leave or enter the municipality.
2. A second approach that we had is instead of predicting the total amount of cases, we use the number of cases for each hundred thousand inhabitants, given the same input data used in the first problem.

[2] The final data set can be found at Zenodo: https://zenodo.org/record/4955582#.ZGIwONJBxBW

[3] The COVID-19 cases data can be found at: https://analisi.transparenciacatalunya.cat/Salut/Incid-ncia-de-la-COVID-19-a-Catalunya/623z-r97q/data

3. Given that we have found that depending on the analysis lab, the accuracy of the measured viral load differs, we have also considered learning a model for each analysis lab (three in total) to predict the average cases for every 100k inhabitants. Therefore, one of the three models is used depending on the municipality.
4. Finally, we have considered the hardest problem: predicting covid cases using only wastewater data, without using the data from past covid cases count.

3. Results

3.1. Model predictions

Table 1 shows the results of evaluating the models using a test set of cases from November 2020 to April 2021. The R^2 explains the total variation of the model predictions, and the MAE is the average number of wrongly predicted cases. Combining them, we can understand the robustness of the models' errors. The *total cases prediction* model shows a high R^2 which indicates the reliability of the MAE value of 161. In contrast, the *cases per 100k inhabitants* model shows a low R^2, indicating that the 149 MAE is mixed by predictions with no error and predictions with a high error. The same happens for the *cases per laboratory* model, and this is because the models learn to predict only for some of the municipalities. Figure 1 shows a clear example of low and high-error municipalities. The *total cases prediction without past cases information* model is different since the R^2 is low and the MAE is high, meaning that the model error may be represented across the municipalities.

Predictive exercise	Coefficient of determination (R^2)	Mean Absolute Error (MAE)
Total cases prediction	0.90	161
Cases per 100k inhabitants	0.45	149
Cases per laboratory	0.28	226
Total cases prediction without past cases information	0.65	333

Table 1. Scores of the models created for each predictive exercise

Figure 1 shows the result of predicting using the second model for two municipalities. While both cases are standardized, their population density and local culture are different. For instance, Sant Feliu de Llobregat holds a population of 44.474 (2018) and an area of 11.8 km^2, and Vielha e Mijaran has a population of 5493 (2018) and an area of 211.7 km^2. Furthermore, they also differ in being an urban and a rural area. In this case, the model predicts better the cases for the higher-density municipality.

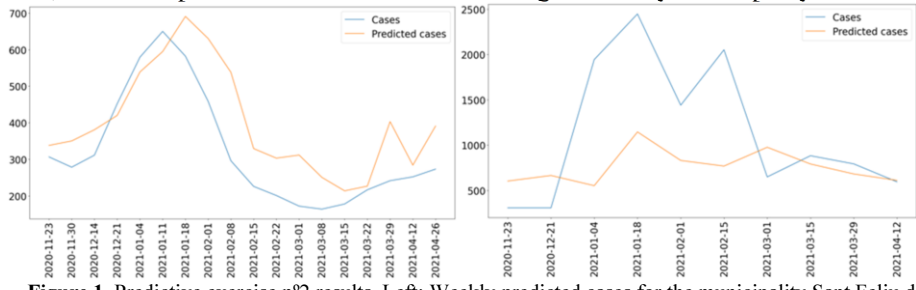


Figure 1. Predictive exercise nº2 results. Left: Weekly predicted cases for the municipality Sant Feliu de Llobregat. Right: Weekly predicted cases for the municipality Vielha e Mijaran.

3.2. Explainability

In this subsection, we identify the most relevant features of the prediction models we have developed. Our analysis is based on SHapley Additive exPlanations (SHAP). At a general level, and across all the ML models developed, the most important variables are the past number of cases and the wastewater indicator. Figure 2 shows the relevance of the past cases, being the most impactful feature in both cases and the importance of the wastewater indicator "N", which is higher when predicting the cases per 100k inhabitants. The other features vary in importance depending on the predictive exercise. To predict the total cases, the inhabitants assisted by the WWTP, the citizens' mobility, and the wastewater quality parameters are the most relevant, but to predict the cases per 100k inhabitants, only the wastewater quality ranks high.

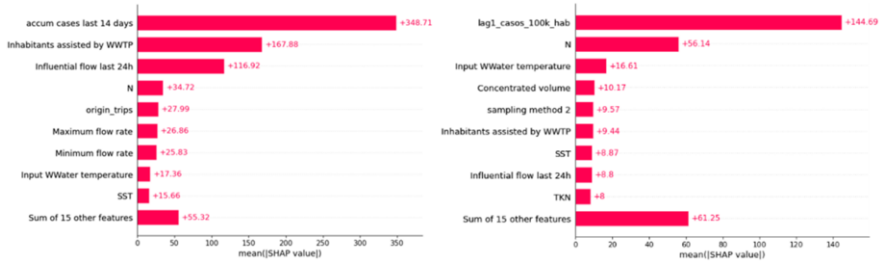


Figure 2. SHAP value for the most important variables. Left: Total cases prediction. Right: Cases per 100k inhabitants.

To improve future feature engineering, we study the impact of the individual feature values for the model. Figure 3 shows the impact of past cases and the wastewater indicator for the model prepared to predict the cases per 100k inhabitants. The SHAP for both features indicates that the bigger the value, the higher the importance to the model. When the value is low, the importance is near zero and in some cases, the feature value has a negative SHAP value, meaning it confuses the model. Those negative values demand future feature engineering to ensure stable behavior.

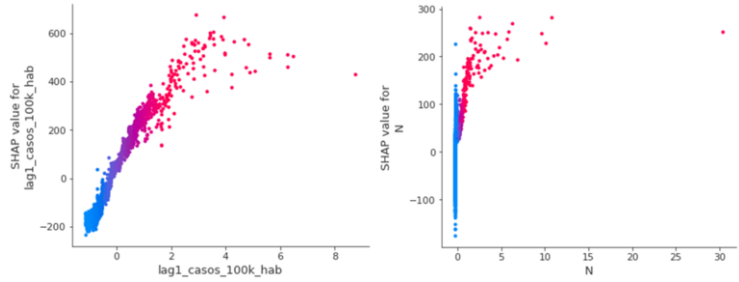


Figure 3. SHAP value for the standardized COVID-19 cases variable (left) and the N value (right). The X axis indicates the value of the feature, and the Y axis indicates the SHAP value for the feature.

4. Conclusions

The machine learning model is good at predicting how COVID-19 cases will change over time, but it struggles to predict the highest number of cases during a peak. The SHAP importance of the features varies depending on the objective variable to be predicted, but the number of people living in an area and the N indicator are always

relevant. To improve the model, samples should be collected more often. Right now, the frequency is one per week, but if it increases to three to seven samples, the model can better understand how wastewater samples relate to the number of COVID-19 cases daily.

Funding

This research has received funding from the Innovation and Technological Development Center (CDTI) and the European Regional Development Fund (ERDF) with the corresponding consolidation (COI-20201289). Marc Ribalta also acknowledges funding from AGAUR DI-2019-066. Ramon Bejar was partially funded by Spanish Project PID2019-111544GB-C22/AEI/10.13039/501100011033 (MINECO / FEDER). Ramon Bejar and Carles Mateu also ackowledge the funding by 2021 SGR 01615.

References

[1] Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A. Presence of SARS-coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in the Netherlands. Environ Sci Technol Lett. 2020;7:511–516 doi: 10.1021/acs.estlett.0c00357

[2] Guerrero-Latorre L, Collado N, Abasolo N, et al. The Catalan Surveillance Network of SARS-CoV-2 in Sewage: design, implementation, and performance. Sci Rep. 2022;12(16704), doi: 10.1038/s41598-022-20957-3

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230692

Exploring Learning Techniques for Developing Socially-Aware Service Robots: Best Practices for Social Comfort

Silvia Guillén Ruiz [a,1], Juan P. Bandera Rubio [a], Alejandro Hidalgo-Paniagua [a] and Antonio Bandera [a,2]

[a] *University of Málaga, Dpto. Tecnologia Electronica, ETSI Telecomunicacion, 29010–Málaga, Spain*

ORCiD ID: Silvia Guillén Ruiz https://orcid.org/0000-0003-0386-101X, Juan P. Bandera Rubio https://orcid.org/0000-0003-3814-0335, Alejandro Hidalgo-Paniagua https://orcid.org/0000-0002-1823-9592, Antonio Bandera https://orcid.org/0000-0003-3147-0307

Abstract. In the past few years, there has been an increase in commercial and research focus on service robots operating in daily surroundings. These machines are anticipated to function independently in busy settings, enhancing movement efficiency and safety parameters, as well as social acceptance. Expanding conventional path planning modules to include socially aware criteria, while sustaining speedy algorithms that can adapt to human behavior without causing distress, presents a significant challenge. To address this challenge, learning methods have gained significant relevance. Among the various techniques, deep reinforcement learning, end-to-end, and inverse reinforcement learning have been the most promising. However, it is difficult to determine which techniques are superior, and sometimes, developers may obtain poor results due to inadequate data or experimental procedures during the learning stage. Therefore, it is essential to evaluate and discuss the best practices and options for an effective training stage that can improve results. As we are specifically referring to social robots, the evaluation of results should take into consideration social comfort as a key factor.

Keywords. Socially aware robotics, deep learning, multi-behaviour navigation, social navigation, social comfort

1. Introduction

One of the challenges in social robotics is expanding conventional path planning modules to include socially aware criteria. Learning methods such as deep reinforcement learning (DRL) approaches like CADRL, SARL, and SOADRL have been proposed for learning

[1]Corresponding Author: Silvia Guillén, silguirui@uma.es

[2]This work has been partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825003 (DIH-HERO SUSTAIN and DIH-HERO GAITREHAB), and projects TED2021-131739B-C21 and PDC2022-133597-C42, funded by the Gobierno de España and FEDER funds.

navigation policies in socially aware criteria. End-to-end learning has shown promise in scenarios where pedestrian positioning is challenging. Inverse reinforcement learning (IRL) offers a potential solution to automate reward function computation, with approaches like maximum likelihood estimation and Bayesian modeling showing promising results. However, poor results can occur due to insufficient data or experimental procedures, emphasizing the importance of evaluating and discussing effective training practices. Social comfort is a crucial factor in evaluating the outcomes of learning methods for social robots.

2. Methodology

This section outlines the criteria that were employed to select the articles included in this review and the criteria to analyze the experiments of the papers.

2.1. Article selection criteria

We meticulously compiled a selection of papers focused on social navigation, with particular emphasis on the topic of learning. Our approach involved a comprehensive review of the literature, drawing from authoritative works such as [19], [4], [12], and [37], as well as recent citations. We identified the most relevant papers that delved into the areas of social navigation, comfort, and learning.

3. Experiments

3.1. Deep Reinforcement Learning and End-to-End approaches

On these collected of papers there are several techniques o combinations of them use on each paper.[1],[2] and [7] use value networks and deep RL for learning social norms in collision avoidance. [32] compares reward functions,[6] evaluates a controller training approach, and [14] employes optimized parameters (OP-DDPG) for collision avoidance training. [9] evaluates Probabilistic roadmap (PRM-RL) on different maps. [22] compares their model with existing ones in randomized environments,[3] trains a gaze-modulated RL network for robot navigation and compare it with SARL.[11] propose a socially normative navigation algorithm,[23] compare their policy with NH-ORCA, and [15] use deep RL to train a robot in a simulated environment with changing obstacles.

[33] proposes curiosity-driven exploration, while [21] presents a simulative learning environment for socially concomitant navigation, both demonstrating their effectiveness through experiments. [24] introduces a deep Q-network (DQN) agent surpassing previous algorithms by learning challenging tasks from pixels and game scores alone. Simulation is widely used for training motion planners in mobile robotics. [34] trains a motion planner in the V-REP simulator for nonholonomic robots, while Fan et al. (2018) efficiently explore the state space with multiple robots using the Stage mobile robot simulator. [29] employs an observable Markov decision process (POMDP) approach to train navigation policies in diverse environments, while [5] trains a navigation algorithm in the Gazebo simulator for obstacle avoidance and autonomous navigation. In contrast,

[10] and [28] focus on expert demonstrations and evaluating path planning algorithms through simulation.

Two studies, [20] and [26], evaluate their navigation methods using simulations.[20] assess the navigation capability of a DQN agent in an unknown indoor environment while avoiding obstacles. [26] trains a neural network-based motion planner to safely guide the robot through environments cluttered with obstacles. Although [26] also evaluates their method in real-world experiments, their simulation results demonstrate its transferability to previously unseen virtual and real-world environments.

In real-world experiments, various navigation methods were evaluated. SA-CADRL [1] and GA3C-CADRL [7] policies were implemented on robotic vehicles in pedestrian environments, while [14] performed real indoor experiments with the IVO robot.[9] used a PRM-based approach, [22] employed deep RL with 3D lidar and stereo camera, and [33] tested their algorithm using a laser ranging sensor.[21] evaluated the transferability of their policy by comparing it with humans in real-world scenarios. [20] evaluated the DQL algorithm for autonomous mobile robot navigation in both simulation and real-world experiments. [26] focused on generating training data for navigating a maze-like area, while [27] demonstrated the generalization and robustness capabilities of models trained purely in simulation. [34] evaluated the real-world performance of robot navigation to reach targets.[8] successfully tested a collision avoidance policy on various mobile platforms in different real-world environments.[10] trained a robot to navigate in a real-world scenario using data collected by human experts. [5] trained a model in a simulation environment and demonstrated its adaptability to real-world uncertainties by testing it on a real tracked robot without fine-tuning.

3.2. Inverse Reinforcement Learning

In [38], real-world routing preferences are modeled using GPS data, while [16] demonstrates successful navigation path learning from demonstrations using a crowd motion simulator. [13] generates feasible velocity trajectories resulting in smooth paths for robot navigation, while [35] introduces a software platform with various contributions including GPU-based learning algorithms and teleoperation. [18] evaluates their approach's generalization to new situations using datasets of interacting pedestrians, while [36] collects data for a robot's path planning algorithm using lidar and remote control. [31] uses Bayesian Inverse Reinforcement Learning to learn reward functions for adventure games, and [25] demonstrates the effectiveness of their learning approach in enabling robots to learn complex navigation behaviors with experiments on a real robot and a pedestrian simulator.

However for real experiments [30] and [18] demonstrate successful social behavior for mobile robots in crowded environments. [35] showcases the potential for robots to provide helpful services in public spaces. [17] and [25] focus on robot behavior similar to human drivers and respecting personal spaces, respectively. The IRL planner was faster but more conservative, while the robot in [25] maintained personal spaces with the help of a social relation tracker.

4. Social Comfort

4.1. Deep Reinforcement Learning and End-to-End approaches

These methods ensure human-aware navigation and achieve socially compliant in dynamic pedestrian environments. They include SA-CADRL [1],long short-term memory (LSTM) extension [7], the SOADRL algorithm [22], the neural network-based approach [21], the danger-zone prediction method proposed by [32], the approach proposed by [14] that combines machine learning techniques with the Social Force Model, the method proposed by [11] that models both human-robot cooperation and inter-human interactions, the attention mechanism-based system proposed by [28], and the approach proposed by [29] that adapts to pedestrians in real-time while ensuring trajectory planning.

4.2. Inverse Reinforcement Learning

Several studies aim to develop socially compliant path planning frameworks for robots in human environments. [16] replicates human pedestrian behavior, while [30] imitates human behavior for socially compliant motions. [35] replicates human interactions, and [18] focuses on learning pedestrian behavior to predict and navigate through them. Meanwhile, [36], [17], and [25] propose approaches that generate socially adaptive paths by learning from demonstrated paths and taking social variables into account.

5. Discussion and Conclusion

The aim of this paper is to analyze the collected papers' ideas to provide future navigation algorithms with a source of methods that can be combined to improve social navigation and achieve better results.

To ensure effective learning models, diverse datasets should be used instead of relying on limited recorded datasets. The selected datasets should encompass various locations and environments for different purposes, not just a specific service or objective. Examples of valuable datasets include ETH, EWAP, Motion Capture, ZARA, and UCY, as used by [21]. Other notable datasets include Students [3] and WALLS [29]. Providing these references is an important contribution of the paper.

Not all studies are evaluated through real experiments due to the difficulties of simulating and generating scenarios. These limitations may include a lack of available robots or test subjects. Nonetheless, studies like those by [30], which test algorithms in a tourist environment, [34], which test algorithms in an office, and [8], which test them in a crowded street and mall, provide crucial insights into different scenarios and random people for evaluating and achieving robust coverage of situations.

In terms of social comfort, various strategies can be incorporated into navigation. For instance, [3] and [21] add social norms, [32] presents the concept of danger zones, [22] learns motion patterns, [10] and [28] propose adaptive robots, and [35] replicate human-social interaction, and [18] discusses cooperative navigation.

References

[1] Yu Fan Chen, Michael Everett, Miao Liu, and Jonathan P. How. Socially aware motion planning with deep reinforcement learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1343–1350, 2017.

[2] Yu Fan Chen, Miao Liu, Michael Everett, and Jonathan P. How. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 285–292, 2017.

[3] Yuying Chen, Congcong Liu, Bertram E. Shi, and Ming Liu. Robot navigation in crowds by graph convolutional networks with attention learned from human gaze. *IEEE Robotics and Automation Letters*, 5(2):2754–2761, 2020.

[4] S.F. Chik, Yeong Fai, Eileen Su, T.Y. Lim, Yuvashini Subramaniam, and P.J.H. Chin. A review of social-aware navigation frameworks for service robot in dynamic human environments. *Journal of Telecommunication, Electronic and Computer Engineering*, 8:41–50, 01 2016.

[5] Jinyoung Choi, Christopher Dance, Jung-eun Kim, Kyung-sik Park, Jaehun Han, Joonho Seo, and Minsu Kim. Fast adaptation of deep reinforcement learning-based navigation skills to human preference. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3363–3370, 2020.

[6] Daniel Dugas, Juan Nieto, Roland Siegwart, and Jen Jen Chung. Navrep: Unsupervised representations for reinforcement learning of robot navigation in dynamic human environments. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7829–7835, 2021.

[7] Michael Everett, Yu Fan Chen, and Jonathan P. How. Motion planning among dynamic, decision-making agents with deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3052–3059, 2018.

[8] Tingxiang Fan, Xinjing Cheng, Jia Pan, Dinesh Manocha, and Ruigang Yang. Crowdmove: Autonomous mapless navigation in crowded scenarios. *CoRR*, abs/1807.07870, 2018.

[9] Anthony Francis, Aleksandra Faust, Hao-Tien Lewis Chiang, Jasmine Hsu, J. Chase Kew, Marek Fiser, and Tsang-Wei Edward Lee. Long-range indoor navigation with prm-rl. *IEEE Transactions on Robotics*, 36(4):1115–1134, 2020.

[10] Wei Gao, David Hsu, Wee Sun Lee, Shengmei Shen, and Karthikk Subramanian. Intention-net: Integrating planning and deep learning for goal-directed autonomous navigation. In *Conference on Robot Learning*, 2017.

[11] Xingyuan Gao, Shiying Sun, Xiaoguang Zhao, and Min Tan. Learning to navigate in human environments via deep reinforcement learning. In Tom Gedeon, Kok Wai Wong, and Minho Lee, editors, *Neural Information Processing*, pages 418–429, Cham, 2019. Springer International Publishing.

[12] Yuxiang Gao and Chien-Ming Huang. Evaluation of socially-aware robot navigation. *Frontiers in Robotics and AI*, 8, 2022.

[13] Brian Gerkey and Kurt Konolige. Planning and control in unstructured terrain. *Computer Science*, 01 2008.

[14] Oscar Gil, Anaís Garrell, and Alberto Sanfeliu. Social robot navigation tasks: Combining machine learning techniques and social force model. *Sensors*, 21(21), 2021.

[15] Martin Gromniak and Jonas Stenzel. Deep reinforcement learning for mobile robot navigation. In *2019 4th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pages 68–73, 2019.

[16] Peter Henry, Christian Vollmer, Brian Ferris, and Dieter Fox. Learning to navigate through crowded environments. In *2010 IEEE International Conference on Robotics and Automation (ICRA)*, pages 981–986, 2010.

[17] Beomjoon Kim and Joelle Pineau. Socially adaptive path planning in human environments using inverse reinforcement learning. *International Journal of Social Robotics*, 8:51–66, 01 2016.

[18] Henrik Kretzschmar, Markus Spies, Christoph Sprunk, and Wolfram Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11):1289–1307, 2016.

[19] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013.

[20] Min-Fan Ricky Lee and Sharfiden Hassen Yusuf. Mobile robot navigation using deep reinforcement learning. *Processes*, 10(12), 2022.

[21] Mingming Li, Rui Jiang, Shuzhi Ge, and Tong Lee. Role playing learning for socially concomitant mobile robot navigation. *CAAI Transactions on Intelligence Technology*, 3, 05 2018.

[22] Lucia Liu, Daniel Dugas, Gianluca Cesari, Roland Siegwart, and Renaud Dubé. Robot navigation in crowded environments using deep reinforcement learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5671–5677, 2020.

[23] Pinxin Long, Tingxiang Fan, Xinyi Liao, Wenxi Liu, Hao Zhang, and Jia Pan. Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6252–6259, 2018.

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[25] Billy Okal and Kai O. Arras. Learning socially normative robot navigation behaviors with bayesian inverse reinforcement learning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2889–2895, 2016.

[26] Mark Pfeiffer, Michael Schaeuble, Juan Nieto, Roland Siegwart, and Cesar Cadena. From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, page 1527–1533. IEEE, 2017.

[27] Mark Pfeiffer, Samarth Shukla, Matteo Turchetta, Cesar Cadena, Andreas Krause, Roland Siegwart, and Juan Nieto. Reinforced imitation: Sample efficient deep reinforcement learning for mapless navigation by leveraging prior demonstrations. *IEEE Robotics and Automation Letters*, 3(4):4423–4430, 2018.

[28] Ashwini Pokle, Roberto Martín-Martín, Patrick Goebel, Vincent Chow, Hans M. Ewald, Junwei Yang, Zhenkai Wang, Amir Sadeghian, Dorsa Sadigh, Silvio Savarese, and Marynel Vázquez. Deep local trajectory replanning and control for robot navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5815–5822, 2019.

[29] Claudia Pérez-D'Arpino, Can Liu, Patrick Goebel, Roberto Martín-Martín, and Silvio Savarese. Robot navigation in constrained pedestrian environments using reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1140–1146, 2021.

[30] Noé Pérez-Higueras, Rafael Ramón-Vigo, Fernando Caballero, and Luis Merino. Robot local navigation with learned social cost functions. In *2014 11th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, volume 02, pages 618–625, 2014.

[31] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2586–2591, Hyderabad, India, 01 2007.

[32] Sunil Srivatsav Samsani and Mannan Saeed Muhammad. Socially compliant robot navigation in crowded environment by human behavior resemblance using deep reinforcement learning. *IEEE Robotics and Automation Letters*, 6(3):5223–5230, 2021.

[33] Haobin Shi, Lin Shi, Meng Xu, and Kao-Shing Hwang. End-to-end navigation strategy with deep reinforcement learning for mobile robots. *IEEE Transactions on Industrial Informatics*, 16(4):2393–2402, 2020.

[34] Lei Tai, Giuseppe Paolo, and Ming Liu. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 31–36, 2017.

[35] Dizan Vasquez, Billy Okal, and Kai O. Arras. Inverse reinforcement learning algorithms and features for robot navigation in crowds: An experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1341–1346, 2014.

[36] Yao Wang, Yuqi Kong, Zhiyu Ding, Wenzheng Chi, and Lining Sun. Nrtirl based nn-rrt* path planner in human-robot interaction environment. In Filippo Cavallo, John-John Cabibihan, Laura Fiorini, Alessandra Sorrentino, Hongsheng He, Xiaorui Liu, Yoshio Matsumoto, and Shuzhi Sam Ge, editors, *Social Robotics*, pages 496–508, Cham, 2022. Springer Nature Switzerland.

[37] Kai Zhu and Tao Zhang. Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology*, 26(5):674–691, 2021.

[38] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230693

Adding Edge Local Differential Privacy to the Dynamic Stochastic Block Model

Sudipta PAUL [a,1], Julián SALAS [b] and Vicenç TORRA [a]
[a] *Department of Computing Science, Umeå Universitet, Umea, Sweden*
[b] *Internet Interdisciplinary Institute,Universitat Oberta de Catalunya, Barcelona, Spain*

Abstract. In today's networked systems a massive amount of data is produced every day. These data can be modelled using graphs, where the nodes typically correspond to users or devices and the edges to the connections between them. Almost all networks change over time, with new nodes and edges appearing or disappearing as the system matures. Therefore dynamic graph models are more adequate to analyse such networks than static graphs, and appropriate tools need to be implemented to protect them. In this paper we obtain an edge differentially private version of dynamic stochastic block model. We show experimentally that the trends in the dynamic stochastic block model obtained from the original data are well preserved with the additional privacy guarantees.

Keywords. Local Differential Privacy, Dynamic Stochastic Block Model, Dynamic Graph, Social Network, Edge privacy

1. Introduction

A massive amount of data is produced every day in today's networked systems, such as social networks, biological networks, internet peer-to-peer networks, and other technical networks [1,2]. These data can be modelled using graphs, where the nodes typically correspond to users or systems and the edges to the connections between them. However, data changes with time, as well as the corresponding network models. Therefore static graphs are inadequate to represent such complex network architectures.

It is well recognized that naive anonymization of a graph can result in disclosure. Attackers can utilize their side-knowledge to infer private information of the graph. For example, through de-anonymization [3], degree [4], 1-neighbourhood [5], or sub-graph [6] attacks. For static graphs, appropriate privacy models have been devised and implemented, which can be broadly divided into two categories: those that adhere to k-anonymity [7] and those that adhere to differential privacy [8]. While using k-anonymity the existing solutions explore the neighbourhood of the nodes where the anonymised graph has at least $k-1$ nodes with the same degree for each node, or, the same neighbourhood, or, $k-1$ number of automorphic subgraphs etc. As for, differential privacy the approaches are mainly centered on triangle counting, smoothing the sensitivity, subgraph counting, clique counting or generating synthetic graphs [9,10,11]. Also, edge privacy is more extensively used than node privacy in real-world applications as it provides effi-

[1] Corresponding Author: Sudipta Paul, spaul@cs.umu.se

cient privacy protection, while keeping good data utility. Nevertheless, for the dynamic graphs, there is not so much literature on privacy models or solutions.

In this work, we focus on edge differential privacy for dynamic graphs. We apply the parallel mechanism [12] to guarantee edge differential privacy in the dynamic stochastic block model [13]. We carry out an empirical evaluation on the Enron dataset [14] to demonstrate how such algorithm preserves the edges' privacy while maintaining the utility of the dynamic stochastic block model.

2. Noise-graph Mechanism

In this section, we define the noise-graph mechanism [15] that we use to randomize the edges in the snapshot graphs an thus protect their privacy.

We denote by $G(V,E)$ the graph with the set of nodes V and set of edges $E = E(G)$.

Definition 2.1. Let $G_1(V,E_1)$ and $G_2(V,E_2)$ be two graphs with the same set of nodes. Then the addition $G = G_1 \oplus G_2$ is the graph $G = (V,E)$ where $E = (E_1 \setminus E_2) \cup (E_2 \setminus E_1)$.

We denote by $G' \in \mathscr{G}(n,p)$, a random graph drawn from the *Gilbert model* (or the Erdös-Renyi model), in which there are n nodes and each edge in G' is chosen with probability p.

Definition 2.2 (Noise-graph mechanism [15])**.** For any graph G with n nodes, and two probabilities p_0 and p_1, we define the following noise-graph mechanism:

$$\mathscr{A}_{p_0,p_1}(G) = G \oplus G_0 \oplus G_1,$$

where $E(G_0) = E(G') \setminus E(G)$ for $G' \in \mathscr{G}(n, 1-p_0)$ and $E(G_1) = E(G'') \cap E(G)$ for $G'' \in \mathscr{G}(n, 1-p_1)$.

3. Differential Privacy for Dynamic Graphs

In this section we provide the definitions of dynamic graph and local differential privacy applied specifically to edges in a dynamic graph. Then, we present the definition of the parallel protection mechanism and a condition so that these mechanism is ε-edge locally differentially private [12].

We follow the definition of a dynamic graph [16] which is a network observed at an initial state G_0 which is a graph, and at a set of T further snapshots, evenly spaced at integer times $t = 1,\ldots,T$. These other graphs are denoted as $G_1, G_2, ..., G_T$. Note that it may be assumed that all the graphs have the same set of nodes, considering that $V(G_i) = \bigcup_{i=0,\ldots,T} V(G_i)$.

We denote by $\mathbb{1}_{uv(t)}$ the indicator function of edge uv in G_t, that is $\mathbb{1}_{uv(t)} = 1$ if $uv \in E_t$, and $\mathbb{1}_{uv(t)} = 0$ otherwise. Similarly, $\mathbb{1}_{\mathscr{A}(uv(t))}$ is the indicator function of edge uv in $\mathscr{A}(G_t)$, the randomized graph.

Definition 3.1 (Edge-local differential privacy for dynamic graphs)**.** An edge randomization algorithm $\mathscr{A} : \mathscr{G} \to \mathscr{G}$, satisfies ε-edge local differential privacy (is ε-eLDP) if for every pair of nodes $u, v \in V$, any timestamp $t \in \{1, \dots T\}$ and $x, x', y \in \{0, 1\}$:

$$P(\mathbb{1}_{\mathscr{A}(uv(t))} = y \mid \mathbb{1}_{uv(t)} = x) \leq e^{\varepsilon} P(\mathbb{1}_{\mathscr{A}(uv(t))} = y \mid \mathbb{1}_{uv(t)} = x')$$

Definition 3.2 (Parallel protection mechanism)**.** Let $G = G_0, G_1, \dots, G_T$ be a dynamic graph. Then, we define the parallel protection of the dynamic graph with parameters p_0 and p_1 as the protection process that provides $\tilde{G} = \tilde{G}_0, \tilde{G}_1, \dots, \tilde{G}_T$ with $\tilde{G}_i = \mathscr{A}_{p_0,p_1}(G_i)$ for $i = 0, \dots, T$. We denote the parallel protection of a dynamic graph G with parameters p_0 and p_1 as $\mathscr{A}^{||}_{p_0,p_1}(G)$.

Remark 3.1. The mechanism $\mathscr{A}^{||}_{p_0,p_1}$ is ε-eLDP if $e^{\varepsilon} \geq \max\left\{\frac{1-p_1}{p_0}, \frac{p_1}{1-p_0}, \frac{p_0}{1-p_1}, \frac{1-p_0}{p_1}\right\}$.

4. Experimental Results

We perform an experiment on a dynamic social network constructed from the Enron corpus [14], which consists of about 0,5 million email messages between 184 Enron employees from 1998 to 2002, with the same pre-processing as Xu and Hero [13]. In addition to the email data, the roles of most of the employees within the company (directors, CEOs, presidents, vice-presidents, managers, traders, and others) are available.

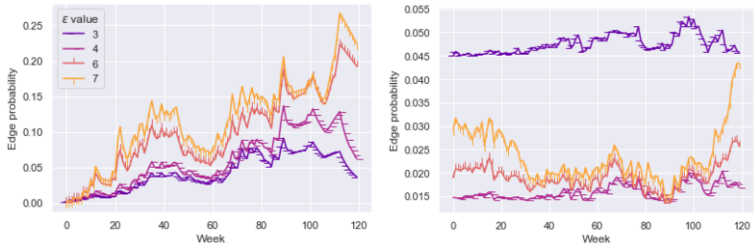
We fit the Dynamic-Stochastic-Block-Model [13] to the data protected with differential privacy for ε values from 1 to 10. We note that for $\varepsilon \geq 7$ the estimated probabilities are the same as without any protection, and for $\varepsilon < 3$ the probabilities are almost constant, for clarity we included only few values in Fig. 1.

Note that in Fig. 1 the trends of the original data are preserved for large enough ε and flattened for smaller values. Still, it can be observed in Fig. 1a the increase in edge probabilities from Enron CEOs to presidents as Enron's financial situation worsened, while in Fig. 1b other employees edge probabilities remain at their baseline levels until Enron fell under federal investigation. More specifically, Fig. 1a shows that the effects of CEO Skilling resignation in week 89 and the resignation of CEO Lay in week 111 are preserved in the dynamic stochastic block model with differential privacy, since the two most prominent peaks in the edge probabilities are present in the protected data.

5. Conclusions

In this work, we applied the parallel mechanism to guarantee edge differential privacy in the dynamic stochastic block model. We carried out an experimental evaluation on the Enron dataset, and observed that the trends in the edge probabilities obtained through the dynamic stochastic block model are well preserved for ε values larger than 3. This shows that it is possible to analyse dynamic network data with the added privacy guarantees.

It remains as future work to carry out a more detailed evaluation of the differentially private mechanisms for dynamic networks proposed, applying them to different models and inference procedures to study the dynamics of other time-evolving networks.



(a) CEOs to presidents (b) Others to others

Figure 1. Estimated edge probabilities obtained by fitting the Dynamic Stochastic Block Model to the protected data. Note that for $\varepsilon = 7$ (yellow line) the estimated probabilities are the same as those of the Dynamic Stochastic Block Model without any protection.

References

[1] Backstrom L, Huttenlocher D, Kleinberg J, Lan X. Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining; 2006. p. 44-54.

[2] Asharov G, Demmler D, Schapira M, Schneider T, Segev G, Shenker S, et al. Privacy-preserving interdomain routing at internet scale. Cryptology ePrint Archive. 2017.

[3] Takbiri N, Shao X, Gao L, Pishro-Nik H. Improving privacy in graphs through node addition. In: 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE; 2019. p. 487-94.

[4] Pedarsani P, Grossglauser M. On the privacy of anonymized networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining; 2011. p. 1235-43.

[5] Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks. In: 2008 IEEE 24th International Conference on Data Engineering. IEEE; 2008. p. 506-15.

[6] Hay M, Miklau G, Jensen D, Towsley D, Weis P. Resisting structural re-identification in anonymized social networks. Proceedings of the VLDB Endowment. 2008;1(1):102-14.

[7] Samarati P. Protecting respondents identities in microdata release. IEEE transactions on Knowledge and Data Engineering. 2001;13(6):1010-27.

[8] Dwork C. Differential privacy: A survey of results. In: Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5. Springer Berlin Heidelberg; 2008. p. 1-19.

[9] Jiang H, Pei J, Yu D, Yu J, Gong B, Cheng X. Applications of differential privacy in social network analysis: A survey. IEEE Transactions on Knowledge and Data Engineering. 2021;35(1):108-27.

[10] Hay M, Li C, Miklau G, Jensen D. Accurate estimation of the degree distribution of private networks. In: 2009 Ninth IEEE International Conference on Data Mining. IEEE; 2009. p. 169-78.

[11] Task C, Clifton C. What should we protect? Defining differential privacy for social network analysis. State of the Art Applications of Social Network Analysis. 2014:139-61.

[12] Paul S, Salas J, Torra V. Edge local differential privacy for dynamic graphs. In: Security and Privacy in Social Networks and Big Data. Springer; 2023. *to Appear*.

[13] Xu KS, Hero AO. Dynamic stochastic blockmodels for time-evolving social networks. IEEE Journal of Selected Topics in Signal Processing. 2014;8(4):552-62.

[14] Priebe CE, Conroy JM, Marchette DJ, Park Y. Scan statistics on Enron graphs. Computational & Mathematical Organization Theory. 2005;11:229-47.

[15] Salas J, González-Zelaya V, Torra V, Megías D. Differentially private graph publishing through noise-graph addition. In: International Conference on Modeling Decisions for Artificial Intelligence. Springer; 2023. p. 253-64.

[16] Zhang X, Moore C, Newman ME. Random graph models for dynamic networks. The European Physical Journal B. 2017;90(10):1-14.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230694

Reconstruction of the LHCb Calorimeter Using Machine Learning: Lessons Learned

Núria VALLS CANUDAS [a,1], Xavier VILASIS-CARDONA [a], Míriam CALVO GÓMEZ [a] and Elisabet GOLOBARDES RIBÉ [a]

[a] *Smart Society Research Group, Engineering Department, La Salle-Universitat Ramon Llull, Sant Joan de La Salle 42, 08022 Barcelona, Spain*

ORCiD ID: Núria Valls Canudas https://orcid.org/0000-0001-8748-8448, Xavier Vilasis-Cardona https://orcid.org/0000-0002-1915-9543, Míriam Calvo Gómez https://orcid.org/0000-0001-5588-1448, Elisabet Golobardes Ribé https://orcid.org/0000-0001-8080-0769

Abstract. In particle physics experiments, calorimeters are in charge of measuring the energy of incoming particles. In order to correctly estimate and evaluate the energy and other properties of these particles, a process, called reconstruction, is required. Because of the amount of collisions and the data-flow, reconstruction algorithms need to be time savvy. The nature of the problem seems appropriate for deep neural networks, yet the approach shows constraints. This paper presents the application to the calorimeter of LHCb in the first upgrade phase under the so-called Real Time Analysis framework, which is in charge of processing 30 MHz of data in real time, with its pros and cons.

Keywords. High Energy Physics, Calorimeter, Deep Learning, Optimization, Model Inference

1. The calorimeter reconstruction challenge

In the field of high energy physics (HEP) there are many computational challenges regarding data. As an example, the LHCb experiment at CERN currently processes 30 MHz of non-empty proton-proton collisions, which generate a data flow of 5 TB/s produced by the eight sub-detectors readout [2]. Far beyond the limits of storage technology, this rate is filtered by a factor 400 by the called *trigger system* [1] before its storage. This trigger relays on a full reconstruction of the collision events in order to make specific selections of physics signatures to decide which events are stored for further analysis. There are two stages in the trigger system: the High Level Trigger 1 (HLT1) [4], which processes the full detector readout in a GPU framework and performs a partial reconstruction of events and primarily selections; and the High Level Trigger 2 (HLT2) which makes a full reconstruction of events using a quasi-real-time alignment and calibration of the detector together with the order of a thousand specific selection algorithms. HLT2 has a throughput of 10 GB/s of reconstructed events and is executed in a CPU framework.

[1] Corresponding Author: Núria Valls Canudas, nuria.valls@salle.url.edu

With independence of the trigger stage, the reconstruction of events consists on a series of specific algorithms that process the data generated by each of the LHCb sub-detectors. By analysing the interaction of particles with the detector, the particles produced in the collision are identified.

One of the eight sub-detectors in LHCb is the electromagnetic calorimeter (ECAL) [10]. It is designed to measure the energy and position of particles as they interact with the detector material with high precision and is the only detector capable to identify neutral particles. It has a rectangular shape of 7.8×6.3 m and is placed perpendicular to the accelerator beam pipe. The ECAL structure is segmented into individual square-shaped modules that perform the energy measurement. Each module has a variable number of readout cells depending on the position, which conform three rectangular regions with increasing granularity around the central beam pipe. The output data obtained from the calorimeter are the values from the 6016 readout cells concerning the accumulated energy deposited by incident particles in a single collision. The energy deposits from a single collision are values from 0 to 10.240 called digits.

Regarding the reconstruction process, it consists on clustering together the energy deposits that belong to the same incident particle, which are then called *clusters*. Although the module shape is designed to contain the full energy of an incident particle, they may not impact at the center of a module. Therefore, the energy shower deposited by a particle is reconstructed as a 3×3 cell group. Given the current LHCb running conditions, ECAL has a high occupancy with the order of 250 clusters per event. This implies that many clusters are overlapping in the inner-most region, which adds more complexity to the reconstruction process.

Entering in the detail of a generic reconstruction for ECAL, it can be segmented into three processes. The first step is to identify the seeds of clusters, defined as the most energetic deposit in a local cell matrix of 3×3. Only deposits of more than 50 MeV are considered to be a seed as ECAL is focused in the reconstruction of photons and electrons that deposit all of its energy in the calorimeter cells. This also helps to filter out the background noise. The second step in the reconstruction chain is to tag the rest of the digits from a single collision to the closest seed. If a digit is in between two seeds, it is tagged to both of them and later processed as an overlapping cell. The final step consists in accumulating the energy of all the digits tagged to a seed and account it as a reconstructed cluster. In case of an overlapping cell, its energy is split according to the energy of the overlapping clusters and each part is linked to the respective cluster.

2. Evolution of methodologies

The process of reconstructing calorimeter data can be viewed as a clustering problem, where the primary objective is to group energy deposits from particles according to specific rules. Traditional unsupervised clustering algorithms use extensive recursive functions with distance or density metrics for creating clusters [6]. However, the approach taken for calorimeter data reconstruction in LHCb stands apart from classical clustering algorithms. This distinction arises due to the strong physics and execution time requirements.

Focusing on the field of calorimetry in High Energy Physics, the Cellular Automaton has been a benchmark solution for many years [3] in LHCb. A more recent optimization

of the latter, an algorithm named Graph Clustering [14] is the current solution for the ECAL reconstruction in HLT2. It is based in the use of graph data structures to store the energy deposits from an event, such that the related digits are closely linked in the graph. This allows for a fast and efficient processing of the overlapping clusters and a flexible representation of data. Although Graph Clustering clearly outperforms the previously used method in terms of execution time, it still has a quadratic complexity curve that grows with the number of energy deposits. This points out the vision of future LHCb upgrades in which the occupancy of the detector is expected to increase by a factor of 10 [8].

To keep up with such rates, there are two main approaches that can flatten the complexity curve. The first one is to take advantage of parallel architectures such as GPUs and FPGAs, which requires a non-trivial translation of the algorithms to take advantage of small local operations that are highly parallelizable. In this direction, some efforts are currently being made in order to adapt the logic of the Graph Clustering algorithm into the preliminary ECAL reconstruction algorithm in the HLT1 GPU framework [15].

On the other hand, the second approach to reduce the reconstruction complexity is through the use of deep learning models. If the ECAL readout data from a collision would be modeled as an image, it would have 6016 pixels with values from 0 to 10.240 from which the reconstructed clusters energy and positions need to be extracted.

It has long been demonstrated that artificial intelligence is well suited for many different HEP challenges [11,9]. Specifically, deep learning models have proven to solve many complex issues at very high speeds, only at the cost of increasing the time and complexity of the training. However, complex models tend to explode in number of parameters, which end up causing the inference to be slow. There is a current trend in HEP to make an efficient use of deep learning models by optimizing the networks itself. This optimization is achieved when the needs of the problem to solve are well understood and the network is modeled and trained accordingly.

On this same line, there is a proposed solution for the ECAL reconstruction that uses a sequence of two convolutional neural networks [13]. The main aspect of such approach is that segmenting the reconstruction process into steps allows to train neural networks on the rules that solve a general formulation of the problem. Furthermore, the understanding of the local nature of the problem leads to a simplification of the data-set, where thousands of training samples are extracted from only 2000 full ECAL simulations. Results regarding the inference time using simulation samples of LHCb collisions show a nearly constant behavior towards the number of digits in a collision compared to a version of the benchmark Cellular Automaton algorithm, as shown in Figure 1.

3. Limitations and constraints

In the latest years, there has been an increasing interest in machine learning and deep learning inference engines as a tool for fast deployment of AI models into production. One of the standards used in HEP is the TMVA tool for performing multivariate analysis in the ROOT framework [7]. It has long been used in LHCb for offline analysis of large data samples using mainly boosted decision trees (BDT) and multi layered perceptrons (MLP). But precisely because this tool is thought to be used in offline analysis, it priori-

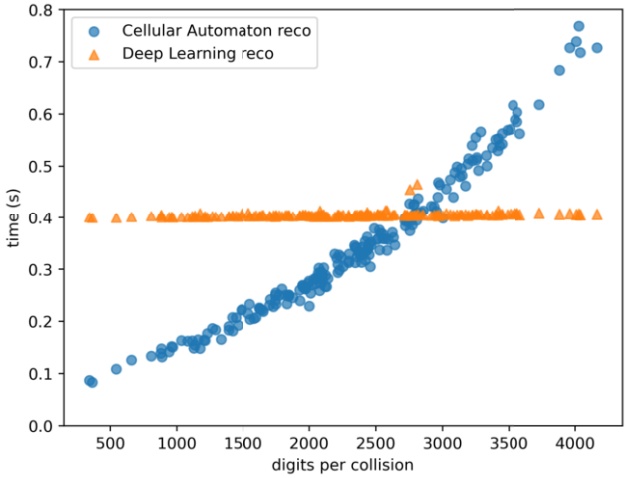


Figure 1. Scatter plot of the mean computational time over the number of digits per collision from LHCb simulations. Comparing a Python version of the LHCb benchmark Cellular Automaton and the simplified deep learning implementation [13].

tizes easy usability rather than inference performance. As an example, a fast neural network based algorithm was proposed in 2017 for LHCb [5]. However the use of TMVA tools was not sufficient to cope with the HLT2 throughput requirements and it required by-hand modifications of the code to allow auto-vectorization and further optimizations of the implementation.

Since this is not scalable and hard to maintain, other tools have started to come into play. Focusing on the model deployment, ONNX is an open format built to represent machine learning models. It defines the building blocks of ML and DL models as common operators and creates a common file format that allows to use the AI models in different frameworks. On the same line, TensorRT is a software development key from NVIDIA that provides high-performance deep learning inference for CUDA environments. It also allows to read and use ONNX files.

Taking advantage of the NVIDIA tools for CUDA, recent studies have used TensorRT to test the inference of two benchmark dense neural networks in the HLT1 GPU platform in LHCb [12]. The two networks tested are both MLP architectures using 17 input features from the LHCb tracking detectors. Looking at the overall results, the kernel overhead is the main bottleneck for throughput but large batch sizes minimize the throughput decrease. To give more detail, the first model tested is a dense neural network with two hidden layers. With the order of 1000 parameters, the HLT1 reconstruction sequence shows almost no throughput reduction using one instance of the network. However, compared to the second, larger approach, with six hidden layers of up to 128 neurons each, the throughput shows a decrease of almost 5%.

We can extrapolate those numbers to the simplified deep learning approach for the ECAL reconstruction. Recalling the network architectures of the approach, it consists of a first convolutional neural network with 1272 parameters for each of the three ECAL regions. The input data is represented as images of sizes 64×52, 64×40 and 48×36

pixels for the Outer, Middle and Inner regions respectively. As a second sequential step, the data from the first networks is processed in windows of 7×7 cells to isolate the relevant information for a single cluster. Then, this window is convoluted with a MLP kernel of size 5×5, which has $108,993$ parameters. Further details on the network chain can be found in the original paper [13]. Overall, the overhead of the data processing and the MLP network inference can be approximated to have the same cost as the six hidden layer model tested in HLT1. To have a broad estimation of the whole impact of the ECAL reconstruction approach, we need to add the inference cost of the first CNN. Although it is negligible for one instance, the approach has one instance per region, which has an impact of almost 6% to the total throughput. Therefore, with a broad estimation, we can say that the DL approach for ECAL reconstruction would imply a throughput reduction of 11% of the whole HLT1 sequence when executed inside the GPU framework of LHCb.

This high inference impact is further enhanced when compared with the cost of the current calorimeter reconstruction in HLT1, which uses a set of CUDA algorithms to find the cluster seeds and build 3×3 clusters around them. This simplified algorithm that does not take into account the overlap between clusters represents a 4% of the total HLT1 reconstruction throughput.

4. Discussion and conclusions

All the algorithms currently implemented for HLT1 and HLT2 have a strong dependency in the number of digits in a collision. Even tho the parallelization of CUDA algorithms can help to mitigate this effect, the future prospects for the LHC experiments include a major upgrade in which the objective is to increase the number of particles generated in every collision by a factor of 10. In such a scenario, a deep learning model with nearly constant inference time, regardless of data complexity, becomes highly valuable.

However, although the many application of AI models in HEP have demonstrated to be very effective in data analysis and reconstruction, the inference of such models to the real experiments frameworks is a clear bottleneck.

Through the study of the LHCb calorimeter reconstruction, we have seen that the standard HEP tools for the inference of models are not scalable and require an expert knowledge in advanced code optimization which sets a huge barrier for deploying or testing the models.

On the other hand, newer tools are starting to be mature enough to allow a generalized format for inferencing AI models that provide fast inferences in an optimized environment. However, the ECAL reconstruction process requires a set of complex operations that are non trivial for an AI application. Even with an optimized and simplified network architecture, the resulting algorithm is expected to have a non-negligible decreasing effect on the throughput even with parallel architectures.

The first outcome we want to raise is that deep learning models will only be suited for HEP trigger-like applications if they are small, simplified and well optimized. This can only be achieved when the insights of the problem are well understood and a network is modeled according to them, instead of expecting a model to learn the general rules of the problem by itself.

As a second message, it is key for the future development of deep learning applications in HEP to push the development of fast and optimized tools for inferencing AI models inside HEP frameworks either with CPUs or GPUs.

Acknowledgement

This work is supported by MICINN under Grant PID2019-106448GB-C32 and by Generalitat de Catalunya under Grant 2021 SGR 01398.

References

[1] Aaij, A.; Benson, S.; De Cian, M.; Dziurda, A; Fitzpatrick, C.; Govorkova, E.; Lupton, O.; Matev, R.; Neubert, S.; Pearce, A.; Schreiner, H.; Stahl, S. and Vesterinen, M. *A comprehensive real-time analysis model at the LHCb experiment*; *Journal of Instrumentation*, **2019**, *14(04)*.

[2] Bediaga, I.; Chanal, H.; Hopchev, P.; Cadeddu, S.; Stoica, S.; Calvo Gomez, M.; T'Jampens, S.; Machikhiliyan, I.V.; Guzik, Z.; Alves, A.A., Jr.; et al. *Framework TDR for the LHCb Upgrade: Technical Design Report*; Technical Report; LHCb-TDR-012; Geneva, CERN 2012.

[3] Breton, V.; Brun, N.; Perret, P. *A Clustering Algorithm for the LHCb Electromagnetic Calorimeter Using a Cellular Automaton*; Technical Report; CERN-LHCb-2001-123; Geneva, CERN 2001.

[4] The LHCb Collaboration. *LHCb Upgrade GPU High Level Trigger Technical Design Report*; Technical report, CERN-LHCC-2020-006, LHCb-TDR-021, CERN, Geneva, 2020.

[5] De Cian, M.; Farry, S.; Seyfert, P. and Stahl, S. *Fast neural-net based fake track rejection in the LHCb reconstruction*; Technical report, CERN-LHCb-PUB-2017-011, LHCb-PUB-2017-011, CERN, Geneva, 2017.

[6] Han, J.; Pei, J. and Kamber, M. *Data mining: concepts and techniques*; *Elsevier*, **2011**

[7] Hoecker, A.; Speckmayer, P.; et al. *TMVA - Toolkit for Multivariate Data Analysis*; *arXiv preprint physics/0703039*, 2009.

[8] The LHCb Collaboration. *Framework TDR for the LHCb Upgrade II - Opportunities in flavour physics, and beyond, in the HL-LHC era*; Technical report, CERN-LHCC-2021-012, LHCB-TDR-0231, CERN, Geneva, 2021.

[9] Mazurek, M. *Deep learning solutions for 2D calorimetric cluster reconstruction at LHCb.*; *4th Inter-experiment Machine Learning Workshop*; Talk; LHCb-TALK-2020-178; 2020.

[10] Omelaenko, O.; Dalpiaz, P.; Guzik, Z.; Spiridenkov, E.; Jarron, P.; Semenov, V.; Ocariz, J.; Khan, A.; Perret, P.; Schneider, O.; et al. *LHCb Calorimeters: Technical Design Report*; Technical Report; LHCb-TDR-002; Geneva, CERN 2000.

[11] Qasim, S.R.; Kieseler, J.; Iiyama, Y.; Pierini, M. *Learning representations of irregular particle-detector geometry with distance-weighted graph networks.*; *The European Physical Journal C*, **2019**, *79(7), 1-11*.

[12] Sclocco, A.; Ali, S.; van den Oord, G.; Cámpora Pérez, D.; de Vries, J.; Aaij, R. and van Veghel, M. *High-throughput machine learning inference with NVIDIA TensorRT*; Presented at *26th International Conference on Computing in High Energy and Nuclear Physics (CHEP23)*, Talk; Norfolk, Virginia, USA; 2023.

[13] Valls Canudas, N.; Calvo Gómez, M.; Golobardes Ribé, E.; Vilasis-Cardona, X. *Use of Deep Learning to Improve the Computational Complexity of Reconstruction Algorithms in High Energy Physics.*; *Applied Sciences*, **2021**, *11(23), 11467*

[14] Valls Canudas, N.; Calvo Gómez, M.; Vilasis-Cardona, X.; Golobardes Ribé, E. *Graph Clustering: a graph-based clustering algorithm for the electromagnetic calorimeter in LHCb.*; *European Physical Journal C*, **2023**, *83, 179*

[15] Valls Canudas, N.; Calvo Gómez, M.; Vilasis-Cardona, X.; Golobardes Ribé, E. *Preliminary performance study of an alternative calorimeter clustering solution for Allen in LHCb*; Presented at *26th International Conference on Computing in High Energy and Nuclear Physics (CHEP23)*, Talk; Norfolk, Virginia, USA; 2023.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230695

A Real-World Dataset for Benchmarking False Alarm Rate in Keyword Spotting

Sergi SÁNCHEZ DEUTSCH, Ivan HUERTA CASADO and Josep ESCRIG ESCRIG
i2CAT Foundation

Abstract. Over the past few years, Keyword Spotting (KWS) has emerged as a popular area of research. Although numerous open-source KWS datasets have been recently released, there is a general lack of realism in benchmarking the false alarm rate (FAR) in real environments. This can produce models that achieve great accuracies but are not able to work on real-world conditions due to a high number of false triggers. In this work, we demonstrate that two recent KWS models report state-of-the-art accuracies on Google Speech Command dataset but suffer from high false alarm rates in presence of noisy environments. To this end, we propose an extensive benchmark dataset comprising various real-world noises and sounds to evaluate specifically the FAR across different acoustic environments.

Keywords. keyword spotting, command understanding, benchmark, false alarm mitigation

1. Introduction

Keyword Spotting (KWS) is widely used nowadays for wake-up word detection and command recognition. One of the main challenges of KWS systems is to achieve a good detection accuracy of the target keywords while maintaining a low false alarm rate (FAR). An occurrence of a false alarm (FA) takes place when the model identifies a word that was not actually spoken, leading to a wrong action. It is crucial to reduce the number of false alarms to the minimum, as the impact of triggering the device when the keyword is not present is equal or even greater than the cost of not detecting it [1].

In recent years, several datasets have been released for training and benchmarking KWS models. The most well-known is Google Speech Commands (GSC) [2]. This corpus has 10 target keywords along with another 25 words that are used as negative samples to train the *unknown* class. Other less common datasets are Mobvoi [3], which has two target keywords and an extensive set of negative samples composed of other words and speech, and Multilingual Spoken Corpus [4], with 340,000 keywords in 50 different languages.

However, none of the abovementioned datasets include a set of negative samples that is representative enough of the background noises and sounds that can be found in a real environment. This can cause a model to report great accuracies on these datasets while they fail to work in a real environment due to a high number of false alarms per hour.

In this paper, we demonstrate that two state-of-the-art KWS models suffer from high false alarm rates on real acoustic conditions even though they report high accuracies on Google Speech Commands dataset. To this end, we propose an extensive dataset to

benchmark specifically the false alarm rate in real environments of any KWS model. This corpus contains more than 52,000 negative samples recorded in a variety of real environments that represent many acoustic situations, and can be used to evaluate FAR of any KWS model.

2. Methods

2.1. FAR Benchmark Dataset

Many open-source KWS benchmarks focus mainly on the ability to discern between the target keywords and other non-target words. However, there is a lack of realism in benchmarking FAR in presence of real-world noises and sounds, so that models are robust enough to work under real conditions. To alleviate this issue, we create an extensive dataset composed of 52,198 1-second audio samples to benchmark specifically the false alarm rate of any KWS model in real-world conditions. These samples contain sounds, speech, and background noises recorded in different real situations. All samples that comprise speech have been manually reviewed to ensure that they do not contain any of the target keywords of the Google Speech Commands dataset. The samples are recorded using a standard laptop microphone at 16 kHz, and are divided in multiple categories, as listed in Table 1: in the category *TV show*, we include samples with several TV contents playing in the background. To gather these samples, the microphone was placed in several locations around the room, and the TV volume was varied several times. The *podcast* category contains a loudspeaker playing several chapters of a BBC podcast. Similarly, the *street* category has a loudspeaker playing noises recorded from a real street in New York City. In the *people* category, we include samples with different crowds talking in the background and, finally, the *living room* category includes multiple daily sounds that can be found in a house, such as people eating at a table or the sound of a vacuum cleaner.

Table 1. Categories and number of samples in our proposed FAR Benchmark dataset

	TV Show	Podcast	Street	People	Living Room
#Samples	23,759	4,713	6,060	3,174	14,492
#Hours	6.6	1.3	1.68	4.03	4.03

2.2. Neural Network Architectures

We conduct the experiments using two state-of-the-art KWS architectures: the Multi-head attention RNN (MHAtt RNN) [5] and the Keyword Transformer (KWT) presented in [6]. The Multi-head attention RNN is based on the "Attention RNN" architecture [7], which combines a CNN, a bidirectional LSTM, and an attention layer. The model has the ability to capture local relations in a feature map and the ability of RNNs to focus on long-term dependencies. The MHAtt RNN model takes the same architecture as the Attention RNN with two main changes, which are a) replacing the LSTM with Gated Recurrent Units (GRU) and b) using the multi-head attention method presented in [8] to

focus on more than one relevant part of the input audio. The Keyword Transformer is the first Transformer-based architecture for keyword spotting that relies solely on attention layers. It takes as input the Mel spectrogram of the audio divided into several patches within the time domain. These patches are flattened and mapped to a higher dimension d using a linear projection matrix. Two learnable embeddings are then added to the input patches to provide the encoder layers with some extra information: a class embedding and a positional embedding. These updated inputs (the original spectrogram patches with the extra embeddings) are then fed into a set of L encoder layers, which perform multi-head attention with k different heads. We employ the code released by [9] to train the MHAtt RNN model. For the KWT, we use the implementation released by [6].

3. Experiments

3.1. Dataset

We train the two selected KWS architectures on the Google Speech Commands (GSC) dataset [2]. GSC is a large-scale open-source dataset that comprises over 100,000 utterances of 35 different words recorded by thousands of speakers. It is widely used in research and industry, and it has been key in advancing the field of speech recognition in recent years. From the 35 available words, 10 are used as target keywords and the rest of them are used to train the *unknown* class, as described in Table 2. In our experiments, we use the split files provided by [2] to divide the dataset into training, validation and testing set, with ratio 80:10:10. The unknown class is composed of a subset of 3,500 samples from the non-target utterances to maintain a balance among classes.

Table 2. Target and non-target keywords included in Google Speech Commands dataset.

Target keywords	yes	no	up	down	left
	right	on	off	stop	go
Non-target keywords	zero	one	two	three	four
	five	six	seven	eight	nine
	bed	bird	cat	dog	happy
	house	Marvin	Sheila	tree	wow
	backward	forward	follow	learn	visual

3.2. Results

In Table 3 we evaluate MHAtt RNN and KWT in terms of accuracy on GSC and FAR on our proposed FAR benchmark dataset. The two architectures report state-of-the-art accuracies on GSC, achieving 99.34% in the case of the MHAtt RNN and 99.72% in the case of KWT.

However, FAR results on our proposed dataset show that both models suffer from a serious false alarm problem when facing noisy samples recorded in multiple real environments. If we consider the case of a commercial voice assistant, it would be unacceptable that the system gets triggered by a non-target utterance more than 100 times every hour in the best case scenario (considering the MHAtt RNN). In the case of KWT, this

FAR increases up to 902 FA per hour. Overall, this demonstrates that the robustness of a KWS model can not be benchmarked with enough confidence using only the accuracy on GSC, as it may lead to a FA issue in real-life conditions.

Table 3. Accuracy on GSC and FAR results in our proposed FAR Benchmark dataset.

Model	Accuracy (GSC)	FAR (%)	FAR (FA/h)
MHAtt RNN	99.34	2.88	103.6
KWT	99.72	25.06	902.1

4. Conclusions and Future Work

In this paper, we demonstrated that current KWS benchmarks are not sufficient to evaluate the robustness of the models in terms of FAR in real-life conditions. To this end, we created an extensive dataset of real-world noises and sounds to benchmark FAR in a variety of acoustic environments. Using this dataset, we proved that two state-of-the-art models suffer from a serious false alarm issue when trained on Google Speech Commands dataset. In future work, we plan to further investigate this issue and propose a false alarm mitigation solution at the dataset level. Our next steps will be thus devoted to reduce FAR in real-world conditions with minor impact on the accuracy of the models.

Acknowledgements

This work has been partially funded by the RED.ES program under the 2021 Grant Program for the Funding of R&D Projects on Artificial Intelligence and other Digital Technologies, with grant number 2021/C005/00152763.

References

[1] Lopez-Espejo I, Tan ZH, Hansen J, Jensen J. Deep Spoken Keyword Spotting: An Overview. IEEE Access. 2021 nov:1-1.
[2] Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. arXiv; 2018.
[3] Hou J, Shi Y, Ostendorf M, Hwang MY, Xie L. Region Proposal Network Based Small-Footprint Keyword Spotting. IEEE Signal Processing Letters. 2019 8;26:1471-5.
[4] Mazumder M, Chitlangia S, Banbury C, Kang Y, Ciro J, Achorn K, et al. Multilingual Spoken Words Corpus. In: Vanschoren J, Yeung S, editors. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1. Curran; 2021. .
[5] Rybakov O, Kononenko N, Subrahmanya N, Visontai M, Laurenzo S. Streaming keyword spotting on mobile devices. INTERSPEECH. 2020 may;2020-Octob:2277-81.
[6] Berg A, O'Connor M, Cruz MT. Keyword Transformer: A Self-Attention Model for Keyword Spotting. In: Proc. Interspeech 2021; 2021. p. 4249-53.
[7] Coimbra de Andrade D, Leo S, Loesener Da Silva Viana M, Bernkopf C. A neural attention model for speech command recognition. ArXiv e-prints. 2018 Aug.
[8] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 2017-Decem. Neural information processing systems foundation; 2017. p. 5999-6009.
[9] Google Research. Streaming Aware neural network models [Source Code]; 2020. Available from: https://github.com/google-research/google-research/tree/master/kws_streaming.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230696

Predicting the Perceptual Rating of a Soundscape Using Artificial Intelligence

Daniel BONET-SOLÀ [a], Ester VIDAÑA-VILA [a,1], Rosa Ma. ALSINA-PAGÈS [a]
[a] *Human-Environment Research (HER), La Salle Campus Barcelona, Universitat Ramon Llull*
ORCiD ID: Daniel Bonet-Solà https://orcid.org/0000-0001-7362-199X, Ester Vidaña-Vila https://orcid.org/0000-0002-4265-2985, Rosa Ma. Alsina-Pagès https://orcid.org/0000-0003-2261-5471

Abstract. The Sons al Balcó project was a citizen science initiative that started during the lockdown consisting in collecting videos recorded in domestic balconies during 2020 and 2021, along with perceptual questionnaires about the soundscape. It has developed into a citizen science project that aims at analyzing the quality of urban quiet places. In previous work, we applied artificial intelligence techniques to automatically detect and classify acoustic events present on the videos. However, to better understand the impact of these events on people's lives, it is important to consider their perception of the soundscape. In this work, we propose a method to incorporate perceptual data into the acoustic event detection algorithm. Specifically, we aim to predict users' perception of the soundscape based on their responses to the questionnaires. Our target is to train a machine learning model to predict the perceptual rating of a soundscape based on the acoustic events and their acoustic features present in a location. When completed, our work will enable a more comprehensive and accurate assessment of the soundscape, taking into account both objective acoustic features and subjective perceptual responses.

Keywords. Soundscape, machine learning, AI application, neural networks, perceptual soundscape evaluation

1. Introduction

Acoustic event detection and classification is a research topic that the scientific community has widely explored in the last years [1]. Even though the field is still under development, automatic AI tools to detect acoustic events have become more available. These tools, together with an increment of awareness of the dangers of noise pollution, have enabled the emergence of new research areas related to soundscape evaluation. These evaluations can be carried out by means of quantitative data (e.g., equivalent noise level) or qualitative data (i.e., the perception of citizens). However, until now, the perceptual evaluations of acoustic environments have been usually carried out by questionnaires made to citizens, implying that the community still lacks an automatic tool to rate a soundscape depending on how pleasant it is for citizens.

[1] Corresponding Author: Ester Vidaña-Vila, ester.vidana@salle.url.edu

For this reason, this paper proposes an experimental set-up to develop an artificial intelligence model able to rate soundscapes trained with two different types of inputs: (1) acoustic event labels obtained using a deep neural network and (2) ratings of humans that have classified the soundscapes as pleasant or unpleasant using questionnaires.

2. Proposed experimental set-up

The data used to develop the automatic tool will come from the research project *Sons al Balcó*, which was a citizen science project that gathered more than 600 videos together with a perceptual questionnaire in different spots of Catalonia [2] during 2020 and 2021. In previous works [1], a neural network that detected acoustic events was trained using the sound from these videos. However, this network was not able to determine the quality of the soundscape of the video.

This work proposes a system capable of estimating the quality of the soundscape using as inputs the acoustic event labels together with the answers of the questionnaires (see Figure 1). Specifically, the proposed pipeline consists of two experiments:

Experiment 1: Rate the soundscape as good/bad depending on the perception of users. For this purpose, a binary classifier will be used (e.g., logistic regression).

Experiment 2: Estimate the soundscape quality using a 1-to-5 rating scale. In this case, a regression algorithm will be used.

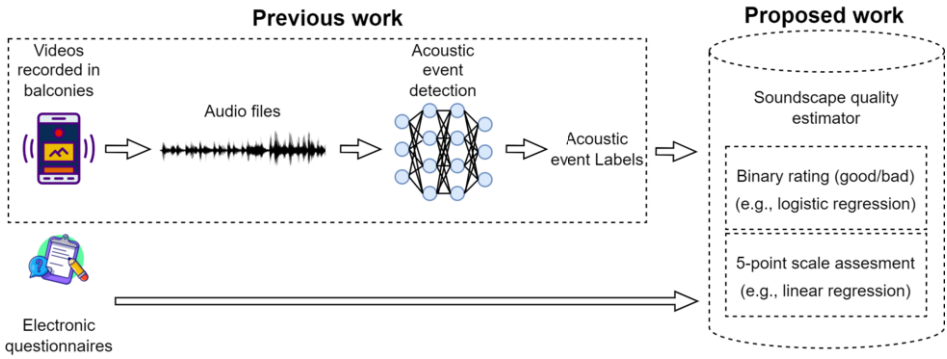


Figure 1. Proposed experimental set-up.

3. Conclusions

This paper proposes an experimental set-up to estimate the quality of a soundscape by merging acoustic event labels obtained using automatic classifiers and the answers of perceptual questionnaires.

References

[1] Bonet-Solà D, Vidaña-Vila E, Alsina-Pagès RM.Analysis and Acoustic Event Classification of Environmental Data Collected in a Citizen Science Project.International Journal of Environmental Research and Public Health. 2023 Feb; 29(4), doi: 10.3390/ijerph20043683

[2] Alsina-Pagès RM, Orga F, Mallol R, Freixes M, Baño X, Foraster M. (2020). Sons al balcó: Soundscape map of the Confinement in Catalonia. Engineering Proceedings, 2(1), 77.

Part 5

Consensus, Social Media Evaluation

This page intentionally left blank

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230698

An Application to Measure Consensus on Customers' Ratings Using Hesitant Fuzzy Linguistic Term Sets

Walaa ABUASAKER[a] , Jennifer NGUYEN[b], Núria AGELL[b,1], Francisco J. RUIZ[a], Mónica SÁNCHEZ[a]
[a] *UPC-BarcelonaTECH*
[b] *Esade Business School, Universitat Ramon Llull*

Walaa Abuasaker https://orcid.org/0009-0000-4630-6687;
Jennifer Nguyen https://orcid.org/ 0000−0002−7498−7536;
Núria Agell https://orcid.org/0000-0001-9264-2147;
Francisco J. Ruiz https://orcid.org/0000−0002−4314−101X;
Mónica Sánchez https://orcid.org/0000−0002−0201−345X

Abstract. In this study, we propose a methodology using hesitant fuzzy linguistic term sets to aggregate rating values. The main idea is to define an integrated methodology for collaborative platforms involving individuals that use the same set of linguistic terms with different semantics. The objective is to fuse the opinions of different profiles of reviewers or customers when assessing an alternative. The proposed approach allows considering previous assessments to give personalized meaning to a set of linguistic terms. To this end, the concepts of linguistic perceptual maps and projections among perceptual maps are introduced to model and aggregate reviewers' different profiles. Finally, we compute a measure of consensus to capture the agreement among reviewers' opinions. All these concepts are applied in a real case example of rating books from Amazon to demonstrate the potential of the methodology in facilitating the evaluation and comparison of different alternatives. Using our methodology, the degree of consensus among customers can be used to rate in a more precise way the books before recommendation.

Keywords. Hesitant Fuzzy Linguistic Terms Sets, Consensus measurement, Customers' profiles, Rating scales

1. Introduction

Ratings are usually used to assess alternatives and provide users with an insight into the shared opinions. Despite the widespread use of ratings as a means of evaluation, individuals may interpret them differently. Therefore, how can a system standardize the perception of diverse users to accurately interpret their rating scales?

In this paper, we present a new methodology to aggregate assessments and measure consensus coming from a rating context. In general, rating scales do not consider that the assessments may come from individuals with different profiles, and they may be considering differently the meaning of the rating scales [1]. Different linguistic scales are simultaneously considered in this study to represent and aggregate individuals' ratings taking into account not only their preferences but also their different perceptions [2, 3, 4]. In this way, the meaning of each linguistic term is not being considered the same for all reviewers, but different semantics for the same linguistic terms are considered depending on their profiles [5, 6].

In this study, we introduce the concept of *common perceptual map*, which provides a shared scale to represent different individuals' perceptions together [8]. Second, we propose the use of a *projection function* for the representation of each individual's linguistic assessment into this common perceptual map [8, 9]. Then, we propose the use of a *centroid* to aggregate the opinions of individuals. This centroid is obtained in the common perceptual map, where all opinions can be correctly represented. Finally, we consider a *measure of consensus* which considers the hesitancy and individual perceptions expressed in the common perceptual map.

Data from Amazon Books' recommender system is used to show the applicability of the proposed methodology in facilitating the evaluation and comparison of different books considering two different profiles of customers: *lenient* and *strict*. The data set considered for this application contains 3,000,000 reviews from more than 200,000 books.

The paper is structured as follows: Section 2 outlines the fundamental concepts and relations of hesitant fuzzy linguistic term sets (HFLTSs) to define the concept of *perceptual map*, including the notions of *centroid* and *degree of consensus*. Following this, Section 3, presents the methodology based on these concepts. In Section 4, an application to a real-world scenario based on Amazon Books' recommender system is developed. Lastly, Section 5 summarizes the conclusions, challenges, and potential directions for future research.

2. Preliminaries: Common perceptual map framework

The general theoretical framework considered in this paper is based on the concept of HFLTSs introduced in [10] that allow users to express hesitancy in their opinions. In addition, in this section, we consider the concept of *common perceptual map* as introduced in [8, 9], as the framework in which rating values from different users' profiles can be expressed and aggregated.

Definition 1. Let S be a finite totally ordered set of *basic linguistic terms* $S = \{s_1, \dots, s_n\}$ with $s_1 < \cdots < s_n$. A *basic linguistic perceptual map* is a pair (S, μ) where μ is a normalized measure over S, that is, $\mu: S \to R^+$ such that $\sum_{i=1}^{n} \mu(s_i) = 1$.

Remark 1. Note that any basic linguistic perceptual map is associated to a partition P of the unit interval [0, 1] in n sub-intervals of lengths $\{\mu(s_1), \dots, \mu(s_n)\}$ and

landmarks $\lambda_0 = 0 < \lambda_1 < \cdots < \lambda_{n-1} < \lambda_n = 1$. The relation of the landmarks and the width of the basic linguistic labels is $\lambda_m = \sum_{k=1}^{m} \mu(s_k)$ and $\mu(s_m) = \lambda_m - \lambda_{m-1}$ for any $m = 1, \dots, n$.

The previous definition considers a universal set of basic linguistic terms but different interpretations of these terms, that is, all individuals use the same set of terms but each one has its own perceptual map.

Definition 2. ([10]) Given a set of basic linguistic terms $S = \{s_1, \dots, s_n\}$, a *hesitant fuzzy linguistic term set* (HFLTS) over S is a subset of consecutive elements of S, i.e. $H \equiv [s_i, s_j] = \{x \in S |\ s_i \leq x \leq s_j\}$, for some $i, j \in \{1, \dots, n\}$ with $i \leq j$.

The set of all HFLTS is denoted as $\mathcal{H}_S$. In [11] it is shown that this set has a lattice structure if the empty set is included.

When reviewers give a rating to assess an alternative, they normally use a precise rating (a basic term in the defined terminology). However, when this assessment is projected onto a common scale, we will see that this rating introduces a certain level of hesitancy. For this reason, the HFLTS framework is very suitable in our approach.

Remark 2. Any HFLTS $H = [s_i, s_j] \in \mathcal{H}_S$ is associated with the subinterval $[\lambda_{i-1}, \lambda_j] \subseteq [0,1]$, where λ_i, $i \in \{0, \dots, n\}$, are the landmarks corresponding to the partition P to which Remark 1 refers.

Definition 3. Given the set $\mathcal{H}_S$, a *linguistic perceptual map* is a pair $(\mathcal{H}_S, \mu)$ where μ is a normalized measure over S and $\mu([s_i, s_j]) = \sum_{k=i}^{j} \mu(s_k) = \lambda_j - \lambda_{i-1}$.

From now on, we call $\mu(H)$ the *width* of H. The width of H match the length of the real interval associated.

Definition 4. Let $(\mathcal{H}_{S^m}, \mu^m)$, $m \in \{1, \dots, k\}$ a set of k linguistic perceptual maps with $\#S^m = n_m$, the *common perceptual map,* $(\mathcal{H}_{S^U}, \mu^U)$, is the linguistic perceptual map associated to the partition of the unit interval P_U of landmarks $\bigcup_{m=1}^{k} \bigcup_{i=0}^{n_m} \{\lambda_i^m\}$.

Definition 5. If $H = [s_i^m, s_j^m] \in \mathcal{H}_{S^m}$, there exists a unique $H' \in \mathcal{H}_{S^U}$ such that it has the same subinterval associated, $[\lambda_{i-1}^m, \lambda_j^m] \subseteq [0,1]$. We call H' the *projection* of H into $\mathcal{H}_{S^U}$.

Definition 6. Let $(\mathcal{H}_S, \mu)$ be a linguistic perceptual map. Given $H_1, H_2 \in \mathcal{H}_S$, the *perceptual-based distance* between H_1 and H_2 is defined as:

$$D_\mu(H_1, H_2) = 2 \cdot \mu(H_1 \sqcup H_2) - \mu(H_1) - \mu(H_2) \qquad (1)$$

where $\sqcup$ denotes the connected union, that is, the less element of $\mathcal{H}_S$ to respect to the inclusion relation that contains both H_1 and H_2. It is easy to prove that this definition gives indeed a distance in $\mathcal{H}_S$.

Definition 7. Let $(\mathcal{H}_S, \mu)$ be a linguistic perceptual map. Let $\{H_1,\cdots, H_k\} \subset \mathcal{H}_S$, the *centroid* of this set, denoted as H^C, is defined as the element in $\mathcal{H}_S$ that minimizes the addition of distances to all the HFLTSs in this set:

$$H^c = arg\min_{H\in\mathcal{H}_S} \sum_{j=1}^{k} D_\mu(H, H_j) \tag{2}$$

When the set of $HFLTSs$ corresponds to the assessments provided by a group of DMs, the centroid is considered as the central opinion of the group of DMs. Note that, when k is an odd number, H^c is unique. Nevertheless, when k is an even number, there can be more than one HFLTS satisfying (2); then, in each application, an extra criterion must be used to select a specific centroid.

Finally, in this context, we define the degree of consensus of a set of opinions derived from the addition of all the distances between the opinions and the centroid.

Definition 8. Let Λ be an object to be assessed using HFLTS's from $(\mathcal{H}_S, \mu)$. Given a group of DMs, $G = \{d_1, d_2, \ldots, d_k\}$, let $\{H_1,\cdots, H_k\}$ be their assessments of Λ and H^c the centroid. The *degree of consensus* of the group is defined as:

$$\delta_\Lambda(G) = 1 - \frac{\sum_{j=1}^{k} D_\mu(H_j, H^c)}{\zeta} \tag{3}$$

where ζ is a normalization term assuring δ_Λ is between 0 and 1.

3. Methodology: Aggregating rating values from different customers' profiles

In this section, we present a methodology to fuse the opinions of different profiles of customers or reviewers when assessing alternatives in a certain context. The methodology is a reduced version of the methodology introduced in [13], and it is inspired by a previous work in which the fusion of ratings was considered in evaluations obtained from recommender systems [12].

The methodology follows four steps:

Step 1. *Define different profiles of customers:* We consider that different customers interpret the same linguistic terms in different ways. A profile is a set of customers that share the semantics for these set of terms. Each profile has its own linguistic perceptual map $(\mathcal{H}_S{}^m, \mu^m)$.

Step 2. *Obtain the linguistic perceptual map per each profile*: For the profile m, $\mu^m(s_i)$ is calculated, from $i = 1, \ldots, n_m$, as the relative frequency with which profile m assigns the basic label s_i in other assessments. It is reasonable that if the individuals of the profile use very often a determined basic label, the width of this basic label will be large. Once the set of widths has been determined, the landmarks of the associated partition $\{\lambda_0, \ldots, \lambda_{n^m}\}$ are computed as detailed in Remark 1.

Step 3. *Obtain the common perceptual map:* Once a linguistic perceptual map has been determined for each customers' profile, the common perceptual map, $(\mathcal{H}_{SU}, \mu^U)$, is calculated whereby all the profile perceptual maps are unified. Then, all assessments are projected from its own linguistic perceptual map to the common perceptual map.

Step 4. *Compute the centroid and the degree of consensus in the common perceptual map:* After the common perceptual map has been calculated, the centroid, H^c, and the degree of consensus, $\delta_A(G)$, are computed in this common space.

4. An application to on-line books rating

4.1. Data set description

The data[1] set used to illustrate the proposed methodology consists of two tables, one for books and another for reviews. Information about the books includes the title and category while the review information includes the reviewer id, rating, review text, and time it was written. To better exemplify the methodology, we focus on the ratings. After a cleaning process of dropping missing values and duplicates (we left the most recent review if the same reviewer reviewed the same book), there are 2,397,614 reviews, 216,023 books and 1,008,972 reviewers. Then, we filtered for reviewers who had written at least 50 reviews to obtain the profile of each one. There were 2,656 reviewers satisfying this constraint who reviewed 81,382 books. At the end, only 2,060 books reviewed by at least 25 reviewers have been considered.

4.2. Experimental approach

In this case, a book may be rated on a scale from 1 to 5 stars, with each star representing a particular level of satisfaction. The set of basic linguistic terms in the context of this case is $S = \{s_1, s_2, s_3, s_4, s_5\}$. We follow the four steps proposed in Section 3 to find the centroid for each book and to compute the degree of consensus among reviewers in the common perceptual map.

First, according to step 1, we divided reviewers into two groups or profiles based on the frequency with which they rated books with a rating of 5 stars. We have named the two groups of reviewers as: lenient and strict. Those who rated books with 5 stars more than 60% of the time were considered to be lenient and the remaining reviewers were considered strict. There were 1595 reviewers in the strict group who had written 191,610 reviews. In the lenient group there were 1061 reviewers who had written 133.552 reviews.

[1]The dataset is obtained from *Kaggle.com* and it is built from two sources:

- *https://jmcauley.ucsd.edu/data/amazon/* Citation Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering R. He, J. McAuley WWW, 2016 Image-based recommendations on styles and substitutes J. McAuley, C. Targett, J. Shi, A. van den Hengel SIGIR, 2015
- google books API.

Second, as presented in the step 2, we computed the relative frequency of the ratings for each group separately in order to obtain the perceptual map for each profile and its respective landmarks. These values are presented in Table 1. We also computed the perceptual map and its respective landmarks where different profiles are ignored.

Table 1: Relative frequencies and widths of basics and landmarks of the partitions for the two profiles: *lenient* and *strict* and also *without profiling.*

# scores	s_1	s_2	s_3	s_4	s_5
lenient	1431	1861	5872	21332	103056
strict	4961	10789	32424	71342	72094
total	**6392**	**12650**	**38296**	**92674**	**175150**
width of basics	μ_1	μ_2	μ_3	μ_4	μ_5
lenient	0.011	0.014	0.044	0.160	0.772
strict	0.026	0.056	0.169	0.372	0.376
no profiles	**0.020**	**0.039**	**0.118**	**0.285**	**0.539**
landmarks	λ_1	λ_2	λ_3	λ_4	λ_5
lenient	0.011	0.025	0.069	0.228	1.000
strict	0.026	0.082	0.251	0.624	1.000
no profiles	**0.020**	**0.059**	**0.176**	**0.461**	**1.000**

Third, from the group linguistic perceptual maps we obtained the common perceptual map considering the union of both sets of landmarks. Figure 1 shows the graphic representation of the perceptual maps corresponding to the two profiles and the common perceptual map. In addition, the perceptual map where different profiles are ignored is represented at the bottom of Figure 1.

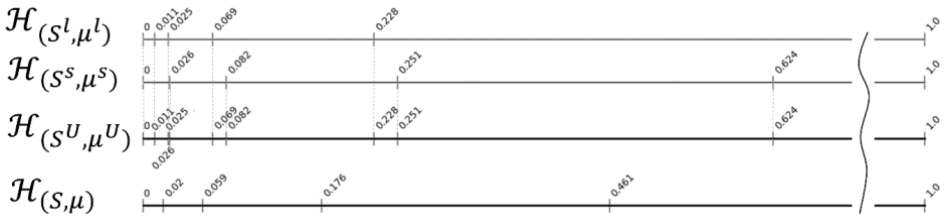


Figure 1. Graphic representation of perceptual maps for the two profiles: *lenient* and *strict*, the common perceptual map and, at the bottom, the perceptual map where no profiles are considered.

The cardinals of S^l for *lenient*, S^s for *strict* and S (where no profiles are considered) are 5, while the cardinal of S^U is N = 9 in this case.

Following the last step of the methodology, the reviews were projected to the common perceptual map and the centroid and degree of consensus were computed for each book using these projections. In addition, we calculated the centroids and degree of consensus in each original perceptual map before the projection. At the end, we calculated the centroids and degrees of consensus taking into account all reviewers without differentiating them by their profiles.

4.3. Results and discussion

First, before considering the projection of the reviews to the common perceptual map, we consider the distribution of centroids per each book separated by reviewer profile. This distribution is showed in Figure 2. Note that the centroid of a set of basics labels is always a basic label.

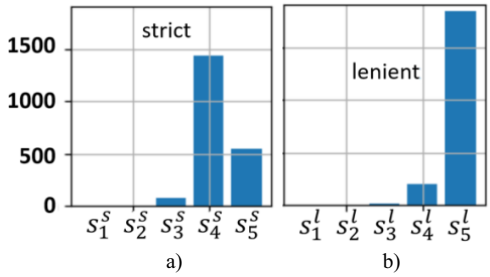


Figure 2. Distribution of the centroids separated by profile a) *strict*, b) *lenient*.

Figure 3 shows the distribution of centroids considering all reviewers together: on the left, the centroids in the common perceptual map (considering different profiles), on the right, the centroids where different profiles are ignored.

As we can observe in Figure 2, most of the book centroids in both the lenient and strict profiles have rating of 5 stars and some others 4 stars, making them very difficult to differentiate. In contrast, in the common perceptual map, the centroids are spread across multiple ratings and less concentrated in anyone rating (see Figure 3a). The comparison of the distribution of centroids in the common perceptual map and in the perceptual map where no different profiles are considered accentuates this difference (see Figure 3b).

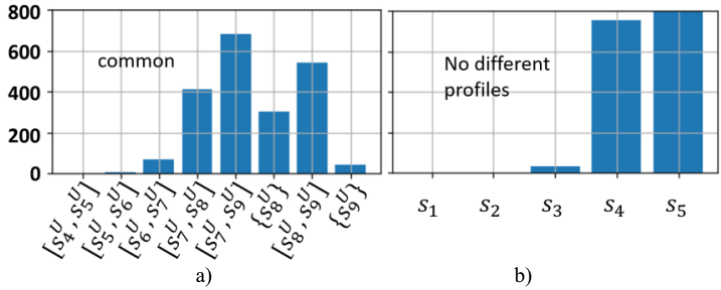


Figure 3. Distribution of the centroids a) in the common perceptual map, b) where no different profiles are taken into account.

Finally, for each book we compute the degree of consensus following (3), considering separately both profiles: *lenient* and *strict* (Figure 4) and considering both profiles together but using, on the one hand, the common perceptual map and, on the other hand, without using the profiles of the reviewers (Figure 5).

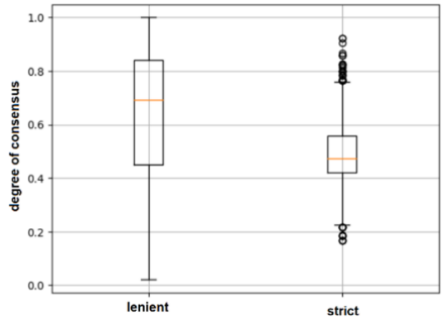


Figure 4. Box plot of distribution of degrees of consensus of reviewers separated by profile.

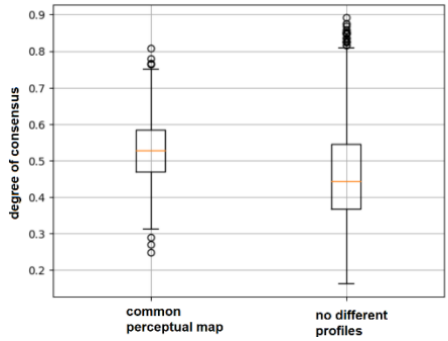


Figure 5. Box plot of distribution of degrees of consensus calculated on the common perceptual map and in the case where different profiles are ignored.

We can observe that the degree of consensus is greater in the lenient profile than in the strict one. On the other hand, the variability of this degree is also greater for the lenient profile.

More interesting is the comparison when all reviewers are considered together. Taking into account their different profiles and working in the common perceptual map, the degree of consensus is greater and has less variability than in the case where the different profiles are ignored.

5. Conclusions and future work

We present an application of the use of linguistic perceptual maps as a tool to represent ratings. We consider a methodology able to aggregate assessments coming from multiple linguistic scales. The obtained results demonstrate the suitability of the methodology in a rating context where individuals may assign different meanings to available linguistic labels.

A real case example with data from Amazon Books' recommender system is provided to show the efficiency of the proposed methodology in facilitating the evaluation and comparison of different books.

Using our methodology, results show that a more discriminating aggregated rating is obtained. On the one hand, the centroids distribution is more accurate (less biased). On the other hand, the presented methodology further refines the concept of consensus since agreement does not necessarily imply that reviewers have given the same linguistic evaluation and disagreement does not necessarily mean that reviewers have given different linguistic labels.

As future work, from a theoretical point of view, we will analyze the use of other aggregation functions to define the central opinion of a group of individuals' assessments, and we will study their mathematical properties. In addition, we plan to define other ways to deduce individuals' linguistic perceptual maps, such as text reviews or non-linguistic aspects such as the time spent to give the review.

Finally, from a practical and business point of view we intend to explore two key areas. Firstly, we will investigate in the development of an interpretability function, capable of translating the outcomes generated by the common perceptual map into each of the distinct initial linguistic perceptual maps. Secondly, we will focus on the practical application by considering a broader range of profiles and incorporating the sentiment analysis extracted from written reviews into the definition of the linguistic perceptual map for each profile.

Acknowledgement

This research has been partially supported by the PERCEPTIONS Research Project (PID2020-114247GB-I00), funded by the Spanish Ministry of Science and Information Technology.

References

[1] Agell, N., Sánchez, M., Prats, F., & Roselló, L. (2012). Ranking multi-attribute alternatives on the basis of linguistic labels in group decisions. Information Sciences 209, 49-60.
[2] Martínez, L., Ruan, D., Herrera, F. Computing with words in decision support systems: An overview on models and applications, International Journal of Computational Intelligence Systems 3 (2010) 382–395.
[3] Liao, H., Xu, Z., Herrera-Viedma, E., Hesitant fuzzy linguistic term set and its application in decision making: A state-of-the-art survey., Int. J. Fuzzy Syst. 20 (2018) 2084–2110.
[4] Wu, Z., Xu, J. A consensus model for large-scale group decision making with hesitant fuzzy information and changeable clusters, Information Fusion 41 (2018) 217–231.
[5] Chiclana, F., Mata, Pérez, L. G., Herrera-Viedma, E. Type-1 owa unbalanced fuzzy linguistic aggregation methodology: Application to Eurobonds credit risk evaluation, International Journal of Intelligent Systems 33 (2018) 1071–1088.
[6] Chen, Z.-S., Liu, X.-L., Rodríguez, R. M., Wang, X.-J., Chin, K.-S., Tsui, K.-L., Martínez, L. Identifying and prioritizing factors affecting in-cabin passenger comfort on high-speed rail in china: A fuzzy-based linguistic approach, Applied Soft Computing 95 (2020) 106558.
[7] Porro, O., Agell, N., Sánchez, M., Ruiz, F.J. (2022). A Multi-perceptual-Based Approach for Group Decision Aiding. In: Torra, V., Narukawa, Y. (eds) Modeling Decisions for Artificial Intelligence. MDAI 2022. Lecture Notes in Computer Science, vol 13408. Springer, Cham.
[8] Porro, O., Agell, N., Sánchez, M., & Ruiz, F. J. (2021). A multi-attribute group decision model based on unbalanced and multi-granular linguistic information: An application to assess entrepreneurial competencies in secondary schools. Applied Soft Computing, 111, 107662.
[9] Liao, H., Xu, Z., & Zeng, X. J. (2014). Distance and similarity measures for hesitant fuzzy linguistic term sets and their application in multi-criteria decision making. Information Sciences, 271, 125-142.

[10] Rodríguez, R. M., Martínez, L., & Herrera, F. (2011). Hesitant fuzzy linguistic term sets for decision making. IEEE Transactions on fuzzy systems, 20(1), 109-119
[11] Montserrat-Adell, J., Agell, N., Sánchez, M., Prats, F., & Ruiz, F. J. (2017). Modeling group assessments by means of hesitant fuzzy linguistic term sets. Journal of Applied Logic, 23, 40-50.
[12] Nguyen, J., Montserrat-Adell, J., Agell, N., Sánchez, M., & Ruiz, F. J. (2020) "Fusing hotel ratings and reviews with hesitant terms and consensus measures." Neural Computing and Applications, 32, 15301-15311.
[13] Abuasaker, W., Nguyen, J., Ruiz, F. J., Sánchez, M., Agell, N, Perceptual Maps to Aggregate Assessments from Different Rating Profiles: A Hesitant Fuzzy Linguistic Approach. Available at SSRN: https://ssrn.com/abstract=4457239

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230699

Mathematical and Computational Models for Crowdsourced Geolocation

Rocco BALLESTER [a] Yanis LABEYRIE [a,b] Mehmet Oguz MULAYIM [a]
Jose Luis FERNANDEZ-MARQUEZ [c] Jesus CERQUIDES [a]
[a] *IIIA-CSIC, Campus UAB, 08193 Cerdanyola, Spain*
[b] *Ecole Centrale de Marseille*
[c] *Centre Universitaire d'Informatique, University of Geneva*

Abstract. Social media generate large amounts of almost real-time data which can turn out extremely valuable in an emergency situation, specially for providing information within the first 72 hours after a disaster event. Despite there is abundant state-of-the-art machine learning techniques to automatically classify social media images and some work for geolocating them, the operational problem in the event of a new disaster remains unsolved. Currently the state-of-the-art approach for dealing with these first response mapping is first filtering and then submitting the images to be geolocated to a crowd of volunteers, assigning the images randomly to the volunteers. In this work we evaluate the potential of artificial intelligence (AI) to assist emergency responders and disaster relief organizations in geolocating social media images from a zone recently hit by a disaster by (i) building a simple model of the capabilities of each of the volunteers of the crowd; and (ii) intelligently assigning tasks to those volunteers which can perform them better. We present in this paper some new methods that outperform random allocation of tasks in a simplified, scalable scenario. Moreover, we show that for a given set of tasks and volunteers, we are able to process them with a significantly lower annotation budget, that is, we are able to make less volunteer solicitations without loosing any quality on the final consensus.

Keywords. social media, disaster response, machine learning, geolocation, crowdsourcing

1. Introduction

Photo geolocation is a hard task that has not been efficiently solved at the present time despite many efforts and plenty of different techniques and strategies. Currently, the state-of-the-art approach for geolocation is first filtering and then submitting the images to be geolocated to a crowd of volunteers (e.g., [4]), assigning the images randomly to the volunteers. In this work we aim to build a paradigm to learn important characteristics about the annotators and use this information to get a much more efficient consensus.

Our primary motivation for precise outdoor image geolocation is its application to the Disaster Management field, where it can have a critical impact because

a fast response is of paramount importance to help emergency aid. Our results show that we are able to learn volunteers' error profiles accurately enough so as to later use this information to increase the quality on our geolocations.

In the following section we provide a formalization of the main problems we will be dealing with, namely the *crowdsourced geolocation problem* and the *active crowdsourced geolocation problem*. Later, we define two statistical models, and several estimators and compare them on their quality when solving problems. We end up with a brief overview of the related work and some conclusions. To the best of our knowledge, this is the first time when the crowdsourced image geolocation problem has been mathematically formalized.

All experiments and analyses shown in this work can be found and reproduced at https://github.com/IIIA-ML/crowdsourced_geolocation.

2. Problem formalization

Definition 1 (Crowdsourced Geolocation Problem (CGP))**.** In the CGP we are interested in determining as accurately as possible the locations (in $\mathbb{R}^{\mathtt{d}}$) of $\mathtt{t}$ points given as a set of tasks $T = [1..\mathtt{t}]$. The information at our disposal to locate the points comes from a set of $\mathtt{a}$ *annotations* (i.e., reported locations) $A = [1..\mathtt{a}]$ by a set of $\mathtt{w}$ individuals $W = [1..\mathtt{w}]$ which we refer to as *workers* or *annotators*. For each reported location $a \in A$, we know (i) the point $t_a \in T$ for which the location is reported, (ii) the worker $w_a \in W$ that performed the annotation, and (iii) the value of the position reported $r_a \in \mathbb{R}^d$. So, the input to the CGP is a tuple $\langle t, w, r\rangle$ where $t \in T^{\mathtt{a}}, w \in W^{\mathtt{a}}$, and $r \in \mathbb{R}^{\mathtt{a}\times\mathtt{d}}$, and its output is a matrix $x \in \mathbb{R}^{\mathtt{t}\times\mathtt{d}}$ containing the estimated positions for the points in T.

In general, the real locations of the points are unknown. We will refer to the real location of point $k \in T$ as $p_k \in \mathbb{R}^{\mathtt{d}}$. We will use $w_j^{-1} = \{a \in A | w_a = j\}$ to denote the set of annotations performed by worker j and $t_k^{-1} = \{a \in A | t_a = k\}$ for the set of annotations of task k. In the following, we will assume $\mathtt{d} = 1$; that is, we will assume that the points have to be located in a line.

As in the CGP in the Active Crowdsourced Geolocation Problem (ACGP) we are interested in determining as accurately as possible the locations of $\mathtt{t}$ points. However, in this case the annotations are not provided to us, but instead we are entitled to select the annotators for each task.

Definition 2 (Active Crowdsourced Geolocation Problem (ACGP))**.** We are given a set of $\mathtt{w}$ individuals $W = [1..\mathtt{w}]$ which we refer to as *workers* and a set of $\mathtt{t}$ objects $T = [1..\mathtt{t}]$ which we need to geolocate and an annotation budget, which indicates how many annotations we can make. We are requested to define an annotation policy that at each point in time and based on the result so far, decides which is the next task to annotate and to which worker should the annotation be requested. The resulting annotation is incorporated as an input to the problem. Furthermore, after obtaining the set of $\mathtt{a}$ annotations $A = [1..\mathtt{a}]$, the output of the problem is a matrix $x \in \mathbb{R}^{\mathtt{t}\times\mathtt{d}}$ containing the estimated positions for the $\mathtt{t}$ points.

3. The Constant Normal Model

The simplest model we will introduce for crowdsourced geolocation assumes that the reported locations follow a normal distribution centered in the real value of the point and with a standard deviation which is different for each annotator.

Definition 3 (Constant Normal Model (CNM))**.** We say that the set of annotations A follow a CNM with parameter $\sigma = \langle \sigma_1, \ldots, \sigma_{\mathtt{W}} \rangle$ if for each annotation $a \in A$ we have that $r_a \sim N(p_{t_a}, \sigma_{w_a})$.

Next, in Sections 3.1-3.3 we propose different estimators for σ, provided the input of a CGP problem (pseudo-code algorithms for these estimators can be found in the supplementary material). Then, in Section 3.4, we compare the different estimators proposed and we evaluate how much benefit we can take of the constant normal model for the CGP in different scenarios.

3.1. Direct estimation

An initial estimation of σ can be obtained by (i) using the common estimator of the mean of a normal distribution for the position of each point, and then (ii) estimating sigma of each annotator by using the common estimator for the standard deviation. Formally, the direct estimation method proposes to estimate σ as $\hat{\sigma_j} = \sqrt{\frac{\sum_{i \in w_j^{-1}} (r_i - \hat{p}_{t_i})^2}{|w_j^{-1}|}}$ (1), where $\hat{p}_k = \frac{\sum_{i \in t_k^{-1}} r_i}{|t_k^{-1}|}$ (2). Thus, the final consensus can be computed by a weighted average.

3.2. Iterative estimation

In the former section we have proposed an estimator for σ which in turn relies on an estimator for p_k. There, we used the mean of the locations reported for task k as the estimator. This is the Maximum Likelihood Estimate (MLE) of the mean as long as the annotators have equal variances. However, when the annotators have different variances, the expression for the MLE of the mean is different, as shown below (proof provided in the supplementary material).

Proposition 1. *Provided a set of $x_1, \ldots, x_n$ values such that each x_i is a sample from $N(\mu, \sigma_i)$, the MLE for μ is* $\hat{\mu} = \frac{\sum_{i=1}^{n} \sigma_i^{-2} x_i}{\sum_{i=1}^{n} \sigma_i^{-2}}$ (3).

Note that Equation 3 can be understood as computing a weighted average of the observations. We can define the weight of observation x_i as $\omega_i = \frac{\sigma_i^{-2}}{\sum_{j=1}^{n} \sigma_j^{-2}}$ (4), with $0 < \omega_i < 1$ for each $i \in [1..n]$ and $\sum_{i=1}^{n} \omega_i = 1$, and then the MLE estimate can be expressed as $\hat{\mu} = \sum_{i=1}^{n} \omega_i x_i$ (5). The iterative estimation method takes it from where the direct estimation finished, and uses Proposition 1 to improve the estimate of the means, provided that we have an estimate for σ. Thus, provided we have an estimate $\hat{\sigma}$, it computes the means as $\hat{p}_k = \frac{\sum_{i \in t_k^{-1}} r_i \hat{\sigma}_{w_i}^{-2}}{\sum_{i \in t_k^{-1}} \hat{\sigma}_{w_i}^{-2}}$ (6). The resulting algorithm is described in Algorithm 1 in the supplementary material.

A reasonable criterion for convergence is that the norm of the difference between $\hat{\sigma}$ this iteration and the previous one is below a specific threshold. When analyzing the results of estimation, we will realize that the iterative estimator estimates σ_j to be 0 for the best annotator. Thus, its usage leads to erroneous estimates.

3.3. Conservative estimation

In this section, we try to tackle the problem of iterative estimation, namely that it overestimates the accuracy of one of the annotators. This objective is accomplished in the conservative estimation method by ensuring that the estimates for σ are *conservative*. By conservative we imply that, when used to compute the MLE of the means by Equation 3, the weights we assign to each observation (the ω_i's defined in Equation 4) are never too large or too small. In this sense, the most conservative estimate for σ will be the one which assigns equal value to every annotator. This will result in the weights ω_i being $\frac{1}{n}$, and the corresponding estimate (Equation 5) will be the average mean. Hence a conservative estimator for σ will result in variances larger than its true value for the more accurate annotators and variances smaller than the true value for the less accurate annotators.

The main idea to build a conservative estimator is to use different estimators for the mean for each of the annotators. In this way, instead of computing a single estimate $\hat{p}_k$ for each task $k \in T$, we compute an estimate $\hat{p}_{k,j}$ for each task $k \in T$, and for each annotator $j \in W$. $\hat{p}_{k,j}$ differs from $\hat{p}_k$ in that it excludes the annotations of annotator j. Hence given $\hat{\sigma}$, we compute $\hat{p}$ using $\hat{p}_{k,j} = \frac{\sum_{i \in t_k^{-1} \setminus w_j^{-1}} r_i \hat{\sigma}_{w_i}^{-2}}{\sum_{i \in t_k^{-1} \setminus w_j^{-1}} \hat{\sigma}_{w_i}^{-2}}$ (7), and given $\hat{p}$, we compute $\hat{\sigma}$ as $\hat{\sigma_j} = \sqrt{\frac{\sum_{i \in w_j^{-1}} (r_i - \hat{p}_{t_i,j})^2}{|w_j^{-1}|}}$ (8). The conservative estimator is described in Algorithm 2 in the supplementary material.

To improve our estimations, we modified our estimator by adding a check to ensure balanced weights and replacing the weight of the best annotator with that of the second-best. This upgraded approach will be called *conservative-2* estimation and aims to prevent overconfidence scenarios. We will show in section 3.4 that this modification results in better estimations.

3.4. Experimental setting and results

In this section, we explain the experimental setup[1] used in our study and compare different estimators based on results obtained from synthetic data. The experimental parameters, including number of points, number of annotators, number of annotators a point is given to annotate (the so-called redundancy), sigma distribution, and points distribution are varied to assess their impact on the methods' performance. Each experiment is repeated 50 times for statistical significance. Performance of each method is measured using mean location norm error and

[1]All experiments and analyses shown in this work can be found and reproduced at https://github.com/IIIA-ML/crowdsourced_geolocation

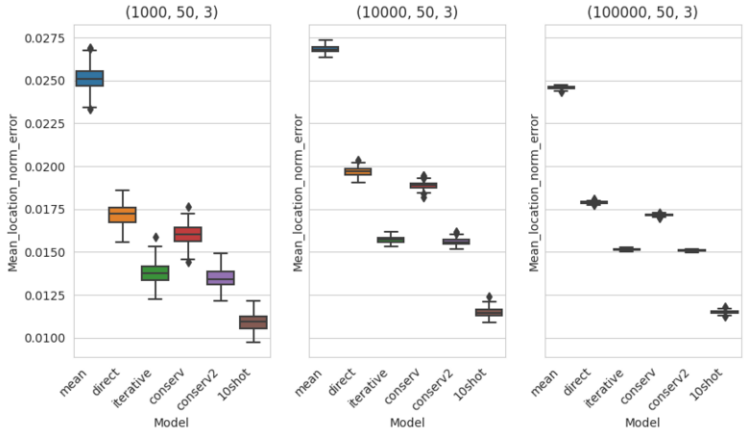


Figure 1. Location error depending on number of points annotated.

mean sigma norm error, which calculate the difference between predicted and true locations or sigmas.

Our baseline is the current state of the art practice, which consists in obtaining the locations by just computing the mean of the annotations at each point. The following results show the evaluation of this practice plus the regarding to the already explained methods. Furthermore, the active strategy is also shown (named *10Shot* in the plots) which will be explained later on.

The influence of the total number of points on the methods' precision is shown in Fig.1. We start with a redundancy value set to 3, and vary the number of points between 1000, 10000, and 100000. The points are requested at random from a pool of 50 annotators with variances sampled from a uniform distribution between 0 and 0.1.

More points do not significantly improve annotation quality, but they do reduce error variance. Working with more points leads to better mean and variance estimations. We then investigate how much less annotation budget is needed for a given error level when using a more advanced method. Fig. 2 shows the mean location error as a function of redundancy for different methods and error levels.

We can see that if we for instance set the desired error value to 0.01, the current practice would require around 20 annotators per image while the *direct* method would only need 5 and the *conservative-2* or *iterative* method would only need 3. Thus, looking at Fig.2 we could state that the best overall method for the CGP is the *conservative-2 estimation*, solving it with a significantly lower annotation budget having fixed the error.

3.5. Active strategy for the ACGP: the K-Shot method

In this section, we tackle the Active Crowdsourced Geolocation Problem (ACGP) where we are given an annotation budget and we actively decide which annotator is responsible for annotating each of the points.

We solve the ACGP implementing the K-Shot method, where we divide the points into K batches. We randomly select annotators to geolocate the first batch

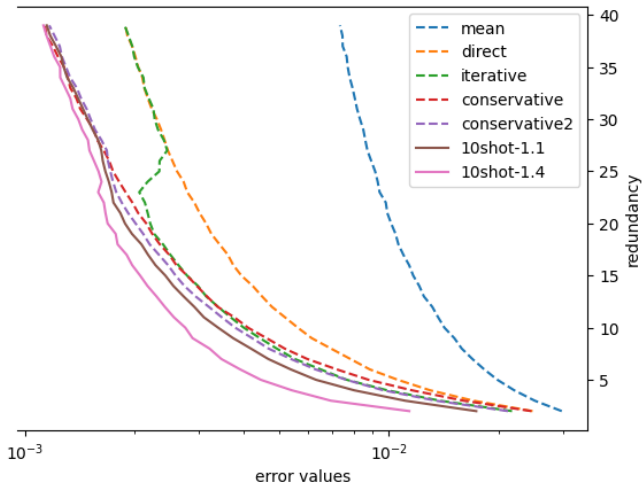


Figure 2. Mean location error depending on the redundancy value. CGP methods are shown in dotted lines while the ACGP method, 10shot, is shown in solid lines and using a greediness value of 1.1 and 1.4.

and estimate their variance profiles using *conservative-2* estimator. Then, we choose annotators with the best profiles for the following batches using the greediness parameter, properly explained later on in this section. We repeat this process for all batches until all points are annotated. Algorithm 3 in the supplementary material summarises this approach.

The greediness parameter helps control how many times certain points are requested to be annotated by annotators. A higher greediness value means that the best annotators are more likely to receive requests to annotate points. A greediness value of $\delta = 1$ distributes requests uniformly, while a value of $\delta = 2$ doubles the probability of the best annotator receiving requests compared to the worst annotator. In other words, this parameter balances precision and cost.

After several analyses, a greediness value of $\delta = 1.1$ ensures that the best annotator does not annotate more than twice the points it would have annotated if the points would have been evenly requested among all workers.

3.6. Comparison between CGP and ACGP

In this section, we aim to show that we can indeed benefit from actively assigning the points to the workers intelligently based on learning their variance profiles. We decided to arbitrary set $K = 10$ and $\delta = 1.1$ for the active method described in the previous section. As explained before, this greediness value guarantees that we do not overcharge some annotators.

K-Shot method outperforms other estimators in all experiment setups (Fig.1 and Fig.2), regardless of the number and distribution of points, and even with a small greediness. Moreover, and as it is expected, the precision of this method can be increased by raising the greediness parameter.

4. The Geographic Normal Model

In the previous section, we presented the constant normal model, where each annotator had a single parameter –namely, their standard deviation σ_j– and we assumed that the reported location for any point located by the annotator followed a normal distribution centered on the true position of the point with σ_j as standard deviation. In this section, we allow the standard deviation of an annotator to vary depending on the true position of the point. That is, instead of a single parameter $\sigma_j \in \mathbb{R}$, in the Geographic Normal Model (GNM) we assume that $\sigma_j : \mathbb{R}^{\mathtt{d}} \to (0, \infty)$ is a continuous function of location.

Definition 4 (Geographic Normal Model (GNM))**.** We say that annotations follow a GNM with standard deviation functions $\sigma = \langle \sigma_1, \ldots, \sigma_{\mathtt{w}} \rangle$ if for each annotation $a \in A$, we have that $r_a \sim N(p_{t_a}, \sigma_{w_a}(p_{t_a}))$.

4.1. Parametric estimation

In this section, we propose to use a parametric expression to approximate the σ_j functions. Our approximation parameterizes each function by means of a number $\kappa \in (0, \infty)$, a set of $\mathtt{c}$ *reference points* $\mathbf{x} = \{x_1, \ldots, x_{\mathtt{c}}\}$, with $x_i \in \mathbb{R}^{\mathtt{d}}$, and a set of weights $\mathbf{v} = \{v_1, \ldots, v_{\mathtt{c}}\}$, with $v_i \in (0, \infty)$. Provided $\mathbf{x}$, $\mathbf{v}$, and a point p, the standard deviation at p is approximated as $s(p; \kappa, \mathbf{x}, \mathbf{v}) = \frac{\sum_{i=1}^{\mathtt{c}} v_i \exp(-\kappa d(p, x_i))}{\sum_{i=1}^{\mathtt{c}} \exp(-\kappa d(p, x_i))}$, where $d(p, x_i) = \|p - x_i\|$, that is, the Euclidean distance between p and x_i. We keep κ and $\mathbf{x}$ the same for all workers, but we have a different set of weights $\mathbf{v}^j$ for each worker $j \in W$. In the following we note $\hat{\sigma}_j(p) = s(p; \kappa, \mathbf{x}, \mathbf{v}^j)$ and so we have that for each $a \in A$, $r_a \sim N(\hat{p}_{t_a, w_a}, \hat{\sigma}_{w_a}(\hat{p}_{t_a}))$ where $\hat{p}_{k,j} = \frac{\sum_{i \in t_k^{-1} \setminus w_j^{-1}} r_i \hat{\sigma}_{w_i}(\hat{p}_k)^{-2}}{\sum_{i \in t_k^{-1} \setminus w_j^{-1}} \hat{\sigma}_{w_i}(\hat{p}_k)^{-2}}$, and $\hat{p}_k = \frac{\sum_{i \in t_k^{-1}} r_i}{|t_k^{-1}|}$ is the usual mean.

Now that the model is defined, our objective is to find the values for $\{\mathbf{v}^1, \ldots, \mathbf{v}^{\mathtt{w}}\}$ with maximum likelihood on that model. The problem can be effectively solved through quasi-Newton methods such as LBFGS [9] as long as the number of annotators and the number of reference points is not very large. Using MCMC simulation to obtain the mean of a set of posterior distribution samples with the help of the Stan software and by fixing the number of reference points to 15, we can get a proper estimation, as we will see in the following section.

When using this method to solve the CGP it will be referred as the *Spatial OneShot* method, as opposed to the name given to the active strategy.

4.2. Experimental setting and results

New experiments use continuous functions instead of single values to model annotators' profiles. Cubic polynomials are generated with a maximum height of 0.2 and a minimum of 0.001 using Scipy's CubicSpline library. Moreover, 10000 points, 10 annotators, a redundancy value of 5 and the same error metric are used. Again, each experiment is repeated 50 times for statistical significance.

Spatial OneShot is the best estimator according to Fig.3 (left), except for the *Spatial K-Shot* method. This is expected since the previous methods try to estimate constant variances.

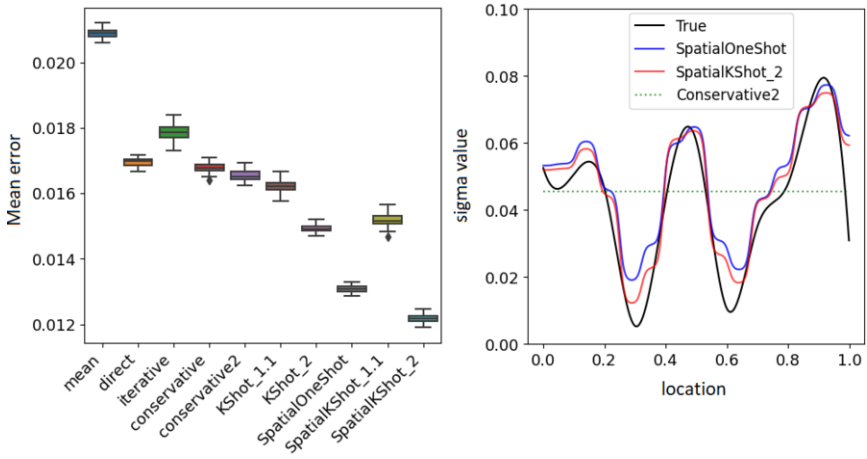


Figure 3. (Left) Mean norm error for each of the methods using $\delta = 1.1$ and $\delta = 2$ for the active methods. (Right) Annotator's variance profile (in black) and the variance profile learned by different methods using $\delta = 2$ for the active method.

4.3. Active strategy for the ACGP: the Spatial K-Shot method

The Spatial K-Shot method is similar to the K-Shot method. Still, it differs: it divides points into two groups, using a small percentage of the points to learn the variance profiles and iterating through the other points to recompute the variance functions. Initial annotations are randomly requested, while subsequent annotations are more likely to come from the best annotators, depending on the greediness parameter, as mentioned before. The algorithm described can be found in Algorithm 4 in the supplementary material.

4.3.1. Comparison between CGP and ACGP

In this section, we compare the methods regarding the CGP described previously and the active method exposed before. To do so, we again divided the K-Shot method in $K = 10$ equal batches and we used 10% of the points as the first batch in the Spatial K-Shot method. Moreover, we used a greediness value of $\delta = 1.1$ and $\delta = 2$ for both active methods. In Fig.3 (left) we can see the performance of the methods by using the mean error metric. We can see that the Spatial K-Shot method outperforms the rest only when the greediness parameter is high enough. Just as in our first model, the higher the greediness, the better the active method performs. However, a really high greediness value could have a detrimental effect since the best annotators would be overdemanded.

5. Related work

The difficulty in the automatic geolocation of photos without geotags has led to the use of crowdsourcing to compensate for the shortcomings of machine learning methods (e.g., [1, 6, 10, 13]). Geolocation precision is critical in emergencies to guide first responders. Studies show that leveraging the wisdom of the crowd

improves data quality, but task assignment is challenging and requires worker expertise and quality (e.g., [8]). Research on worker-task matching spans a broad spectrum of approaches, such as modelling worker skills (e.g., [11]), spatial crowdsourcing (e.g., [15]), and incorporating a trust model between workers to attain data quality (e.g., [14]). We contribute to creating computational models of worker geolocation performance, building on a probabilistic framework for annotation aggregation [3]. Familiarity with regions of interest, independent of worker location, may also benefit geolocation tasks. This paper presents a geolocation-specific extension to the mentioned probabilistic framework and will be integrated into the Crowdnalysis software library [2]. Refer to [5, 7, 12] for further perspectives on crowdsourcing and different assignment approaches.

6. Conclusions and future work

This paper presents two different models, the Constant Normal Model and the Geographic Normal Model, to tackle the Crowdsourced Geolocation Problem and the Active Crowdsourced Geolocation Problem, respectively. One key aspect of this work is the focus on learning annotators' variance profiles w.r.t. location of tasks, an innovative approach that surpasses current state-of-the-art methods. Several dedicated methods have been developed to achieve this, such as the *conservative-2* and *K-Shot* algorithms, which enhance consensus quality and reduce annotation costs. We believe these contributions advance the field of crowdsourced geolocation and hold potential for various domains of application, such as crowdsourced data classification or crowdsourced data evaluation.

There are several future research avenues for further enhancement of the proposed methods. Firstly, analysing a broader range of scenarios and determining the optimal method under different conditions would be beneficial. This could involve fixing variables at different values, exploring various combinations of the number of points and annotators, and even utilizing different data distributions. Optimizing specific hyperparameters that were arbitrarily chosen, such as the number of batches in active methods or the number of reference points in the parametric estimator, could lead to improved results.

Moreover, both models can be enhanced by increasing the dimensions used. For instance, expanding from one dimension to two dimensions would enable an effective representation of latitude and longitude. This enhancement would facilitate more accurate and context-aware predictions, empowering the models to better understand and interpret spatially related data.

Additionally, active learning approaches could be used to cease requesting annotations for instances that have already achieved a clear enough consensus. This effective strategy would reduce costs by eliminating the need for further annotations in such cases. Last but not least, it remains a future task to implement and validate the methods described in a real-world scenario, gathering and requesting data from actual citizens.

Acknowledgements. This work has received funding from the European Union's Horizon 2020 research and innovation programme HumanE-AI-Net and Crowd4SDG projects (N. 952026 and N. 872944, respectively).

.

References

[1] Carlo Bono, Mehmet Oğuz Mülâyim, Cinzia Cappiello, Mark James Carman, Jesus Cerquides, Jose Luis Fernandez-Marquez, Maria Rosa Mondardini, Edoardo Ramalli, and Barbara Pernici. A citizen science approach for analyzing social media with crowdsourcing. *IEEE Access*, 11:15329–15347, 2023.

[2] Jesus Cerquides and Mehmet Oğuz Mülâyim. crowdnalysis: A software library to help analyze crowdsourcing results. https://doi.org/10.5281/zenodo.5898579, January 2022.

[3] Jesus Cerquides, Mehmet Oğuz Mülâyim, Jerónimo Hernández-González, Amudha Ravi Shankar, and Jose Luis Fernandez-Marquez. A conceptual probabilistic framework for annotation aggregation of citizen science data. *Mathematics*, 9(8), 2021.

[4] Ramian Fathi, Dennis Thom, Steffen Koch, Thomas Ertl, and Frank Fiedrich. VOST: A case study in voluntary digital participation for collaborative emergency management. *Information Processing & Management*, 57(4):102174, July 2020.

[5] David Geiger, Stefan Seedorf, Thimo Schulze, Robert C. Nickerson, and Martin Schader. Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. *AMCIS 2011 Proceedings - All Submissions*, August 2011.

[6] Clemens Havas, Bernd Resch, Chiara Francalanci, Barbara Pernici, Gabriele Scalia, Jose Luis Fernandez-Marquez, Tim Van Achte, Gunter Zeug, Maria Rosa Mondardini, and Domenico Grandoni. E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. *Sensors*, 17(12):2766, 2017. Publisher: MDPI.

[7] Danula Hettiachchi, Vassilis Kostakos, and Jorge Goncalves. A survey on task assignment in crowdsourcing. *ACM Comput. Surv.*, 55(3), feb 2022.

[8] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, page 1301–1318, New York, NY, USA, 2013. Association for Computing Machinery.

[9] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. Publisher: Springer.

[10] Amudha Ravi Shankar, Jose Luis Fernandez-Marquez, Barbara Pernici, Gabriele Scalia, Maria Rosa Mondardini, and Giovanna Di Marzo Serugendo. Crowd4Ems: A crowdsourcing platform for gathering and geolocating social media content in disaster response. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:331–340, 2019.

[11] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 889–896, New York, NY, USA, 2009. Association for Computing Machinery.

[12] Eric Schenk and Claude Guittard. Crowdsourcing: What can be Outsourced to the Crowd, and Why ?, December 2009.

[13] Sukrit Venkatagiri, Jacob Thebault-Spieker, Rachel Kohler, John Purviance, Rifat Sabbir Mansur, and Kurt Luther. Groundtruth: Augmenting expert image geolocation with crowdsourcing and shared representations. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[14] Han Yu, Zhiqi Shen, Chunyan Miao, and Bo An. Challenges and opportunities for trust management in crowdsourcing. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 486–493, 2012.

[15] Yan Zhao, Yang Li, Yu Wang, Han Su, and Kai Zheng. Destination-aware task assignment in spatial crowdsourcing. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 297–306, New York, NY, USA, 2017. Association for Computing Machinery.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230700

Sentiment Analysis of Gastronomic Posts from Colour Palettes and Narrative Content

V.CASALES-GARCIA [a,1], I.SANZ [b] L.MUSEROS [b] Z.FALOMIR [b] and L.GONZALEZ-ABRIL [a]

[a] *Universidad de Sevilla, Dpto. de Economía Aplicada I, Avda. Ramón y Cajal, 1, Sevilla, E-41018, Spain*

[b] *University Jaume I, Engineering and Computer Science Dept., Av. Vicent Sos Baynat s/n, Castelló, E-12071, Spain*

Abstract. This paper presents an approach for analysing gastronomic images and also their related comments published by the *Getcookingcanada* Instagram account, which belongs to a cooking school. Our approach processes the published images to calculate the moods that the image can generate depending on its colour palette, and also analyses the comments related to each publication for determining the positive or negative sentiment associated to them. A dataset containing all these data has been built, and then the correlations among the data has been developed in order to explain the relation between the mood adjectives and the result of the sentiment analysis of the comments of the food images. 673 food images were analysed; the data analysis was carried out using the Kruskal-Wallis one-way ANOVA test on ranks and Jonckheere-Terpstra's test. Our results show that there is a significant difference between the different adjectives in terms of sentiment analysis.

Keywords. Sentiment analysis, Gastronomy, Instagram, Social media, image processing, colour palette, natural language processing, tourism, regression, ANOVA, deep learning

1. Introduction

Gastronomy is a key component in the promotion of tourist destinations. Gastronomic experiences influence human behaviour because they stimulate pleasure [1], and they can be used to understand local culture, its traditions, and its history [2,3,4]. These can also provoke feelings of involvement and attachment to a place [5]. Local food could influence destination choice and perception of experience both before and after a trip [6, 7,8]. According to [9], the internet provides positive destination image for pre-travelling as a marketing strategic to provide a holistic travel experience.

The most common channel for the gastronomic promotion of destinations is through official tourism websites and social media that display the most typical local dishes [10,11,12,9]. On these websites and social media, information, photos, and graphics are

[1] All authors contributed equally to this paper.

provided that create a gastronomic image [10]. According to [13] one of the ways to better attract tourists is by using the visual media-oriented Instagram social network, where posting gastronomy-related content is termed *foodstagramming*. Food images posted on social media help tourists meet their needs for entertainment, personal identification, and social interactions [14], resulting in better dining and travel experiences [15].

According to [13], in order to attract potential tourists, tourism authorities and marketers are leveraging touristic photos to pave the road for destination promotion and image development [16,17]. In other words, *tourists simultaneously become producers and consumers of locations and the visual culture connected with them* [18]. When a tourist travels away from home, taking and sharing travel images allows her to act as both a consumer of destination experiences and a narrator of these experiences to friends and family back home [18]. This combined role of actor and narrator has many advantages for vacation selfies.Furthermore, according to [19] the sentiment analysis can be used for the purpose of consolidating the pictorial and textual information and better understanding what foodstagrammers share about their culinary experiences. Therefore, being able to understand the impact of food images published in social media is important. And this is the aim of this research paper, specifically the paper analyses the food images published by the *Getcookingcanada* Instagram account in order to get knowledge about which sentiments produce, and the sentiment that their comments generates, and obtain relations among all of these. The rest of this paper is structured as follows. Section 2 presents how to analyse each image, which means how we can assign an adjective to an image following the methods presented in [20]. Also, in this section how the sentiment analysis done with the comments associates to each image is presented following the trends presented in [21]. Section 2.1 shows the methods used for our analysis. Results obtained are described and discussed in Section 3. Finally, conclusions and ideas for future work are presented in Section 4.

2. Methodology

2.1. Kobayashi's Model: Colour Image Scale

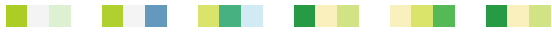
One of the aims of this paper is to analyse food colour and the moods that it can produce. To this end, we use Kobayashi's *Colour Image Scale* model, which was developed to define how a single colour and harmonic combinations of colours affect people's moods [22].

After a series of studies, Kobayashi created a system that: (i) defined a two-dimensional space based on the perceptual qualities of warmth/coldness and hardness/softness of colours, and defines coordinates for 130 colours on this space; (ii) assigns labels related to mood to particular areas in this space (for example, *happy* and *vigorous*); (iii) groups labels (i.e., areas in the perceptual space) that are related to different lifestyles (*causal, modern, romantic, natural, classic,* and *elegant*); and (iv) enables a predominant adjective to be assigned to a colour scheme.

To verify the adjectives in the dish's images, an automatic method was applied to associate a Kobayashi's adjective to the colour palette of the appearance of a dish.

In order to automatically assign an adjective/keyword based on Kobayashi's model to the colour palette corresponding to the image of a dish next process has been carried out:

Table 1. Examples of adjectives and corresponding Kobayashi's colour palettes

Adjective	Kobayashi's Colour palettes
Aromatic	
Citric	
Fresh	
Healthy-light	
Romantic	
Wholesome	

- Every picture is segmented, so that only the parts of the image showing food are kept. The segmentation was done automatically by applying the method described in [23].
- For each image of a dish the most representative Kobayashi's colours of the image are calculated. Then, for each combination of three of these colours, a Kobayashi's colour palette is searched. Not all the combinations result in a Kobayashi's colour palette.
- For each Kobayashi's colour palette obtained before, a keyword/adjective in the Kobayashi's model is assigned.

To calculating Kobayashi's colours, the Euclidean distance between the RGB colours in the images and the RGB colour representations of Kobayashi's model is used.

Some examples of adjectives and Kobayashi's palettes are shown in Table 1. This table also shows up to six Kobayashi's colour palettes associated to each adjective (note that some adjectives have less than six colour palettes defined in the model).

2.2. Sentiment Analysis of the comments associated to each image

In our study, the *polarity* of the comments associated to an image was obtained using a model based on DistilBERT [24], a deep learning model based on transformers. The score of this was discretised converting it into the qualitative labels *Positive*, *Neutral* and *Negative*. The *experience* by the Influencers or Instagrammers is measured as years pass by, since they are gaining experience progressively and their uploaded photos are increasingly more attractive to their followers.

3. Experimentation and results

This study is based on a dataset culinary images published in the Instagram account *@getcookingcanada*[2] by the *Get Cooking* online cooking school. These images and their associated comments were retrieved using the Instagram API.

To this end, the following steps have been carried out:

Data collection and preprocessing We downloaded images and kept only food images; other images containing people, such as a chef or a student, were discarded. We applied the methods described in the previous section to assign a sentiment analy-

[2] *Getcookingcanada* Instagram account: https://www.instagram.com/getcookingcanada/. We were granted permission to process their images.

sis score each image based on to the associated comments, and to assign up to five Kobayashi's adjectives to each image.

Data treatment: Processing and analysing the information obtained: Tabulation, comparative graphics, and analysis thereof, through research instruments. The data analysis to test the proposed research has been done using a Kruskal-Wallis one-way ANOVA by ranks test and Jonckheere-Terpstra test by using the version 26 SPSS program.

Hypothesis: There are differences between the adjectives in terms of sentimental analysis.

The initial dataset has 1523 culinary images, of which 958 had comments, with an average of 3.05 comments per image. Only the images with comments are used.

The first part of the experimentation is the application of the automatic label process of the Kobayashi's model to the images of the dataset, which involves extracting the colour palettes from the culinary images published in the Instagram account @*getcookingcanada*in order to associate a feeling or mood with them. Specifically, these images refer to dishes.

The main aim is to propose possible feelings that a user can have when seeing those images. Note that Kobayashi's model defines cognitive labels based on colour palettes only, and no extra information about the image is used. To be able to apply the proposed method to images, the following steps are necessary: first, colours in each image are discretised and the corresponding colour names (QC_{LAB}) are obtained. Then, all palettes based on the 5 most frequent chromatic colours in the image are considered, but only if the combined frequency of all colours is at least 30% of the total of colours (including also grey scale colours). This prevents assigning mood keywords to palettes that are not representative enough. It also prevents assigning no relevant adjectives to images where grey scale colours predominate. Table 2 shows the results of the experiment for some of the food images of the dataset (columns one, two and four).

Next, a study between the main adjective with the rate of sentiment analysis is carried out. In addition to removing images without comments the dataset was pre-processed to remove the adjectives that are not present in the dataset at least four times, because it is a condition to apply the Kruskal-Wallis test and the Jonckheere-Terpstra test. The 673 culinary image that fulfilled these requirements were used in the experimentation.

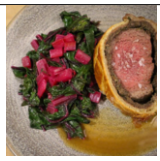
Table 2 shows (i) an example of the digital images used for extracting the colour palettes, (ii) the name of the dish, (iii) the sentiment analysis score obtained for the comments of the post, (iv) the semantic adjectives (representing a mood following Kobayashi' model) assigned, and (v) some of the comments analysed. The five most relevant semantic adjectives for each image are shown, in order of relevance (not all are equally relevant).

After applying the tests, the most important results are summarised in the following list (Result 1–6):

Result 1: The main adjectives that appear in food images are shown in Figure 1 and Table 3. It is worth noting that *Romantic* is the most repeated adjective (see Figure 1) with a percentage of 25.3%.

Result 2: A Kruskal-Wallis and one-way ANOVA by ranks test (see Table 4 and Figure 2) showed that there was a statistically significant difference between the score of the sentiment analysis and the Kobayashi's adjectives ($p = 0.00049$).

Table 2. Results: Sentiment analysis score (SA), five main adjectives given by Kobayashi's model, and comments (C# is a commentary and A# is the answer).

Images Digital	Name	SA	Adjectives and score	Comments
	Tempering Coconut Chutney	.999	Refreshing (2.22), Fresh (1.49), Amiable (0.90), Healthy (0.78), Masculine (0.78)	C1) I understood very little from all the words but looks delicious
	Dal Tadka, Indian cuisine	.996	Romantic (1.16), Amiable (0.76), Bright (0.67), Wholesome (0.62), Bold (0.61)	C1) Aloo Gobi had been one of the dishes I learned that I make the most since class. Love this dish!
	Salmon	.995	Romantic (1.34), Amiable (0.75), Supple (0.69), Soft (0.68), Masculine (0.67)	C1) looks fantastic C2) Looks amazing. I thing like a perfectly cooked salmon
	Risotto and Arancini	.964	Romantic (2.16), Light (1.43), Crystalline (1.43), Amiable (1.00), Charming (0.77)	C1) let's go!! A1) ummmmmmmmm YES PLEASE
	Soft Venison with Parsnip Pure	.741	Amiable (0.87), Bold (0.86), Innocent (0.86), Light (0.86), Bright (0.86), Festive (0.86)	C1) Omg this is incredible A1) thank you, We were pretty happy with this dish C2) This was sooooo yummy C3) Wow!!!!! Looks so amazing
	Beef Wellington Mushroom Sauce	.461	Romantic (1.16), Masculine (0.59), Supple (0.58), Intellectual (0.58), Charming (0.58)	C1) I would so happily do that!!, unfortunately we cannot take it this C2) It was an awesome class.
	Crostini with Roasted	−.761	Light (1.60), Romantic (1.59), Amiable (0.83), Wholesome (0.82), Healthy (0.81)	#Vegetarian #Cooking-Classes

Result 3: The Box-plot (see Figure 2) shows that the Wholesome adjective (15) has a minor score mean (0.256) in the sentiment analysis process from other adjectives. On the opposite side it is the Charming adjective (3) with the mayor score (0.443) although this one is not very representative as there are only four images.

Result 4: Since significant differences have been found, then multiple pairwise comparisons have been done. Table 5 shows the pairs where significant differences have

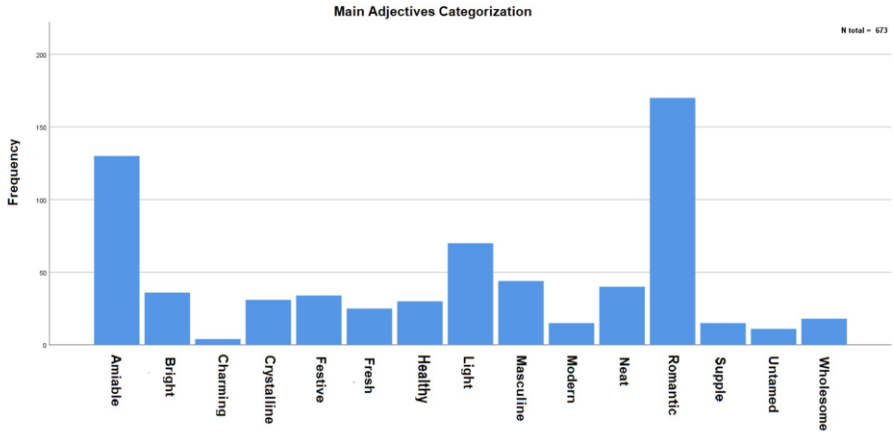


Figure 1. Frequency of adjectives appearance

Table 3. Frequency & Percentage of adjectives: Amiable (1), Bright (2), Charming (3), Crystalline (4), Festive (5), Fresh (6), Healthy (7), Light (8), Masculine (9), Modern (10), Neat (11), Romantic (12), Supple (13), Untamed (14), and Wholesome (15)

Adjectives	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
Frequency	130	36	4	31	34	25	30	70	44	15	40	170	15	11	18	673
Percentage (%)	19.3	5.3	0.6	4.6	5.1	3.7	4.5	10.4	6.5	2.2	5.9	25.3	2.2	1.6	2.7	100
Mean Score	.402	.368	.443	.293	.394	.348	.350	.316	.359	423	.351	.271	.317	377	256	–

Table 4. Kruskal-Wallis Test and Jonckheere-Terpstra Test

Kruskal-Wallis Test		Jonckheere-Terpstra Test	
N	673	N	673
Value of Statistic	49.61	Value of Statistic	82723.50
Grade of Freedom	14	Standard Error	2874.48
Asymptotic Sig. (2-sided test)	0.000	Standard Test Statistic	-5.388
		Asymptotic Sig. (2-sided test)	0.000

been found with a level of significance $\alpha = 0.05$ (21 of 105= $\binom{15}{2}$ pairwise).

Result 5: In order to confirm the results obtained with the Kruskal-Wallis test, the Jonckheere-Terpstra test for ordered alternatives is carried out. In Table 4 can be seen from the Jonckheere-Terpstra test that there was a statistically significant trend of sentiment scores with adjectives. That is, there is a difference between the adjectives in terms of sentiment analysis.

4. Discussion & Conclusion

Our main aim was to rate the users' preferences according to the colours displayed in the images and also to propose possible feelings that a user can have when seeing those

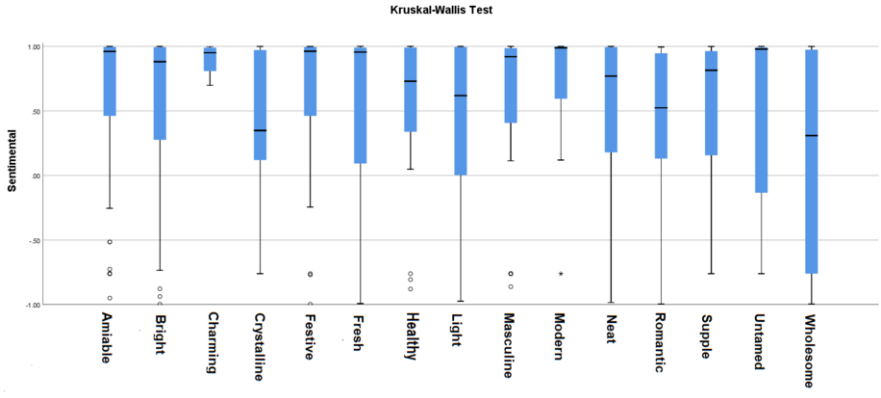


Figure 2. Box-plot of the score for each adjective

gastronomic images. With this information, restaurants would be able to design dishes using colours that may satisfy their future clients.

Our results show that there are statistically significant differences between the sentiment analysis scores assigned to images whose colours can be described using different adjectives using Kobayashi's model. This implies that it is possible to use a palette of colours to evoke a mood, and that this has detectable effects using sentiment analysis. Therefore, the evoked positive emotion may produce a good response for consumers. Moreover, this paper confirms the result by [20] that shows empirical evidence of the value of mood descriptors derived through colour.

We conclude that Kobayashi's model could be a good tool that can be used for marketing issues, considering restaurants and promoters of gastronomic tourism that advertise on social networks (e.g., Instagram, Facebook, TikTok). It is important to continue our studies into how colours influence people's behaviour when choosing a dish, and which sort of food is more attractive from the point of view of food colours in their gastronomy. In addition, this could give the chefs the possibility to categorise the adjectives into positive and negative moods. Furthermore, the moods that the chef wants to evoke in the dish are related to positive or negative Sentiment analysis. Businesses can display pictures of their food that have specific colour palettes to make it more appealing to potential customers or tourists. Therefore, if chefs or marketers want to evoke a sustainable or healthy mood through food images, they can use a recommendation system to help them.

Acknowledgments

This work has been partially funded by the Spanish Ministry of Science, Innovation and Universities through the project *Prueba de Concepto para la Plataforma de Análisis Social Dinámico en el Contexto del Turismo Sostenible* (PDC2021-121097-I00), by the Ministry of Economy and Commerce through the project TIN2016-88835-RET, the EDITH Research Project (PGC2018-102145-B-C21), and the project UJI-B2020-15 funded by Universitat Jaume I. Z. Falomir acknowledges the Ramon y Cajal fellowship

Table 5. Multiple pairwise comparisons. The columns shows Kobayashi's adjectives, the test statistic, the standard error, the test standard error and the *p*-value

	Multiple comparisons of pairwise			
Pairwise	Test Stat	Std. Error	Std. Test Stat	*p*-value
Wholesome ⟷ Neat	109.875	55.835	1.968	.049
Wholesome ⟷ Fresh	134.440	59.495	2.260	.024
Wholesome ⟷ Healthy	136.217	57.528	2.368	.018
Wholesome ⟷ Masculine	140.488	54.419	2.582	.010
Wholesome ⟷ Bright	154.222	55.835	2.762	.006
Wholesome ⟷ Amiable	162.737	49.611	3.280	.001
Wholesome ⟷ Untamed	162.955	72.785	2.239	.025
Wholesome ⟷ Festive	179.485	56.338	3.186	.001
Wholesome ⟷ Modern	193.964	68.007	2.852	.004
Wholesome ⟷ Charming	229.625	103.883	2.210	.027
Romantic ⟷ Healthy	74.593	36.800	2.027	.043
Romantic ⟷ Masculine	78.865	31.721	2.486	.013
Romantic ⟷ Bright	92.599	34.094	2.716	.007
Romantic ⟷ Amiable	101.113	22.496	4.495	.000
Romantic ⟷ Festive	117.862	34.912	3.376	.001
Romantic ⟷ Modern	132.341	51.670	2.561	.010
Crystalline ⟷ Amiable	82.463	37.642	2.191	.028
Crystalline ⟷ Festive	-99.211	46.148	-2.150	.032
Light ⟷ Amiable	77.934	28.743	2.711	.007
Light ⟷ Festive	94.682	39.229	2.414	.016
Light ⟷ Modern	-109.161	54.680	-1.996	.046

(RYC2019-027177-I / AEI / 10.13039/501100011033) awarded by the Spanish Ministry of Science, Innovation and Universities. All the authors acknowledge Getcookingcanada, and thank Chef Kathryn Joel in particular.

References

[1] Rozin P, Kurzer N, Cohen A. Free associations to food: effects of gender, generation, and culture. Journal of Research in Personality. 2002;36:419-41.
[2] Fields K. Demand for the gastronomy tourism product: motivational factors. Tourism and gastronomy. 2002;4(2):36-50.
[3] Henderson JC. Food tourism reviewed. British Food Journal. 2009;111(4):317-26.
[4] López-Guzmán T, Sánchez-Cañizares S. Culinary tourism in Córdoba (Spain). British Food Journal. 2012;114(2):168-79.
[5] Gross MJ, Brown G. An empirical structural model of tourists and places: Progressing involvement and place attachment into tourism. Tourism Management. 2008;29(6):1141-51.
[6] Duarte Alonso A. Olives, hospitality and tourism: a Western Australian perspective. British Food Journal. 2010;112(1):55-68.
[7] Quan S, Wang N. Towards a structural model of the tourist experience: An illustration from food experiences in tourism. Tourism Management. 2004;25(3):297-305.
[8] Smith S, Costello C. Segmenting Visitors to a Culinary Event: Motivations, Travel Behavior, and Expenditures. Journal of Hospitality Marketing & Management. 2009 jan;18(1):44-67.
[9] Björk P, Kauppinen-Räisänen H. Culinary-gastronomic tourism–a search for local food experiences. Nutrition & Food Science. 2014;44(4):294-309.

[10] Horng JS, Tsai CT. Government websites for promoting East Asian culinary tourism: A cross-national analysis. Tourism Management. 2010;31(1):74 85.
[11] Okumus B, Okumus F, McKercher B. Incorporating local and international cuisines in the marketing of tourism destinations: the cases of Hong Kong and Turkey. Tourism Management. 2007;28(1):253-61.
[12] Lepp A, Gibson H, Lane C. Image and perceived risk: A study of Uganda and its official tourism website. Tourism Management. 2011;32(3):675-84.
[13] Wong IA, Liu D, Li N, Wu S, Lu L, Law R. Foodstagramming in the travel encounter. Tourism Management. 2019;71:99-115.
[14] Liu B, Norman WC, Backman SJ, Cuneo K, Condrasky M. Shoot, taste and post: An exploratory study of food and tourism experiences in an online image-share community. e-Review of Tourism Research. 2012;10(3).
[15] Spence C, Okajima K, Cheok AD, Petit O, Michel C. Eating with our eyes: From visual hunger to digital satiation. Brain and Cognition. 2016;110:53-63. Food for thought: The functional and neural mechanisms of food perception and choice.
[16] Kim H, Stepchenkova S. Effect of tourist photographs on attitudes towards destination: Manifest and latent content. Tourism Management. 2015;49:29-41.
[17] Pan S, Lee J, Tsai H. Travel photos: Motivations, image dimensions, and affective qualities of places. Tourism Management. 2014;40:59-69.
[18] Dinhopl A, Gretzel U. Selfie-taking as touristic looking. Annals of Tourism Research. 2016;57:126-39.
[19] Chen Z, Chan ICC, Egger R. Gastronomic image in the foodstagrammer's eyes – A machine learning approach. Tourism Management. 2023;99:104784. Available from: https://www.sciencedirect.com/science/article/pii/S0261517723000663.
[20] Casales-Garcia V, Falomir Z, Museros L, Sanz I, Gonzalez-Abril L. A Study on How Food Colour May Determine the Categorization of a Dish: Predicting Meal Appeal from Colour Combinations. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2022;30(Supp02):293-308. Available from: https://doi.org/10.1142/S0218488522400232.
[21] Casales-Garcia V, Falomir Z, Museros L, Sanz I, Llido DM, Gonzalez-Abril L. Analysing Food-Porn Images for Users' Engagement in the Food Business. Frontiers in Artificial Intelligence and Applications. 2022 10:67-70. Available from: https://ebooks.iospress.nl/doi/10.3233/FAIA220316.
[22] Kobayashi S. Color image scale. Kosdansha International; 1991.
[23] Park D, Lee J, Lee J, Lee K. Deep Learning based Food Instance Segmentation using Synthetic Data. In: 2021 18th International Conference on Ubiquitous Robots (UR); 2021. p. 499-505.
[24] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108. 2019.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230701

Improving Disaster Response by Combining Automated Text Information Extraction from Images and Text on Social Media

Hafiz Budi FIRMANSYAH [a,1], Carlo Alberto BONO [b], Valerio LORINI [d], Jesus CERQUIDES [c], Jose Luis FERNANDEZ-MARQUEZ [a]
[a] *University of Geneva, Switzerland*
[b] *Politecnico di Milano, DEIB, Italy*
[c] *Artificial Intelligence Research Institute IIIA-CSIC, Spain*
[d] *European Comission - Joint Research Centre, Ispra*

Abstract. During disasters, social media can serve as a valuable source of real-time information about the impacts on people and infrastructure. However, due to the lack of geographical information in most social media posts, this information is often underutilized by first responders. Previous research has attempted to estimate the location of individual social media posts using text and image analysis, but limitations still exist in fine-grained disaster area mapping. To address this issue, this paper analyses the performance of combining text from social media post with textual information from the images on improving the geolocation of social media information during a disaster.

Keywords. Social Media, Disaster, Text Extraction, Geolocation

1. Introduction

Social media users share details about damaged infrastructure, injured victims, damage severity, and real-time situation of disaster scenes [1]. On the platforms, the social media posts may contain geolocation information. Such information is important for disaster responders to map the impacted area and take action to the situation quickly. However, only 3% of social media posts contain structured geolocation information [2,3], challenging the use of social media information in disaster management.

The potential of using social media data to improve disaster response attracted researchers to combine natural language processing (NLP) techniques and gazetteers to retrieve geolocation information. A large part of work contributions to geolocate social media posts focus on text analysis, for instance analyzing the text associated with images to find the location where the images were taken. These approaches have been proven to be able to geolocate a significant percentage of social media posts at country and city

[1] Corresponding Author: Hafiz Budi Firmansyah, hafiz.firmansyah@unige.ch

level. Text-based geolocation approaches have also been combined with crowdsourcing to increase the precision of the location. However, this hybrid approach, while performing well in terms of precision, can delay the production of a disaster area map due to the manual geolocation process. Therefore, it is necessary to increase the number of automatically geolocated images in order to reduce the time needed to produce the disaster area map.

The text contained in images, such as street names, traffic signals, and names of businesses or points of interest, serves a crucial role in crowdsourcing the geolocation of images. However, to the best of our knowledge, the automatic extraction of text from images and combining the extracted text with social media posts to retrieve geolocation has never been explored.

This paper presents a pipeline that utilizes text information extracted from social media images to improve the creation of disaster maps. The study systematically analyses the value of text embedded in images attached to social media posts for improving social media information during a disaster.

The remainder of the paper is structured as follows: Section 2 discusses related work. Section 3 presents the proposed approach. Section 4 shows the experimental results using real social media data from four disaster events. Section 5 describes the discussion. Finally, Section 6 reports the conclusion and future works.

2. Related works

The studies on geolocation in social media have attracted many researchers. Moreover, the research on text analysis for inferring geolocation information has been growing steadily recently. Several approaches have been presented, for instance, named entity recognition (NER), location-indicative word (LIW), user location profile (ULP), and social relations [4].

Some previous works attempted to develop methodologies based on machine learning, crowdsourcing, or hybrid approaches. Lorenzo et al. [2] proposed a text-based machine learning approach. They combined NER and network theory to extract relevant location information from tweets. The approach has been applied to nearly 3 million tweets, managing to extract country-level geolocation with 89.8% accuracy.

In another work, the authors exploited a naïve Bayes probabilistic method combined with location indicative words techniques. Focusing on feature selection, their work aimed at city-level geolocation. The results demonstrated how selected features outperformed the complete feature set [5].

Another popular approach to acquiring geolocation is extracting the location from the user's profile. In various works, authors processed the information in the user profile using a NER approach to obtain users' locations. However, the location quality could be limited, e.g., the users might write a fictitious location in their profile [6,7,8].

In [8] and [9], the authors proposed social relation techniques to obtain geolocation prediction. Specifically, they estimated the location based on the friendship network. However, also in this case limitations could arise. For example, users could be linked with someone they do not actually know (e.g., public figures). Furthermore, the approach demands high computational resources for processing.

In terms of text-based geolocation prediction algorithm, Scalia et al. proposed an approach called CIME, a technique based on NER and disambiguation of candidate lo-

cations [10]. An application of this methodology to large-scale flood characterization is proposed in [11]. Shankar et al. developed Crowd4EMS, a platform to geolocate a social media post, has successfully combined crowdsourcing techniques and an automatic approach for gathering information. Figure 1 illustrates how crowdsourcing can be leveraged to validate automated information extraction. The picture on the right, collected from Twitter, represents damaged buildings due to an earthquake. The picture on the left shows the same location on Google Street View. Using a crowdsourcing platform, crowd workers were asked to identify whether the image is located in a specific location. The approach was evaluated using data related to the Amatrice 2016 earthquake [12].

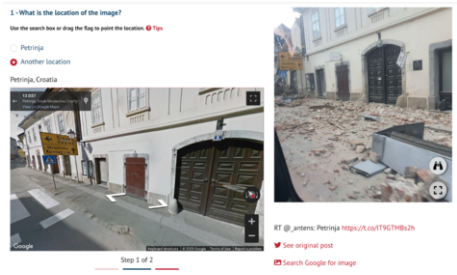


Figure 1. The crowdsourcing approach to validate geolocation information of a social media image

Besides text-based geolocation prediction, research on image-based geolocation prediction also exists. Murgese et al. address the problem of estimating images' locations by introducing Focal Modulation Network. They used geolocated images from Flickr and Mapillary [2] to evaluate their approach. However, the approach was not performing well with Flickr data which is more similar to the one used in disaster situations [13].

In a more recent study, Firmansyah et al. proposed a text extraction from images pipeline. This pipeline enables the integration of the extracted text into several geolocation prediction algorithms. The result shows that the geolocation prediction is reasonably good at the country level [14].

Despite extensive work on post-based and initial works on text extraction from image geolocation prediction, the approach proposed in this paper builds on the pipeline proposed by [15] and [14] which focuses on the geolocation of social media information by analyzing the text extracted from the image. Instead of focusing on a country level, this research aims to evaluate the number of more specific impacted disaster areas.

3. Our approach

This research evaluates the potential of using text extraction from social media images to find their locations. An image often comes with relevant text information such as the name of a street, a shop name, traffic signals directions, a hotel or a hospital name, which are key clues for deriving the location of an image.

The proposed approach aims to combine the extracted text from the images with social media post and predict the geolocation information using state-of-the-art algorithms from the literature.

[2] https://www.mapillary.com

Our approach proposes a pipeline for extracting relevant information from social media, related to a given disaster, and locating them in a disaster map to allow emergency responders to take better-informed decisions. Figure 2 illustrates the proposed pipeline, which is composed of five processes:

1. **Social media crawling** process fetches informative posts from social media by selecting appropriate keywords which depend on the type of disaster.
2. **Cleaning and Filtering** process retains posts that are relevant to the event, removing duplicates, memes, not safe to work, and images which do not provide information related to the disaster event (i.e. not informative images).
3. **Text extraction from images combined with text from post** process automatically extracts any text in the social media images. Then, we combined the extracted text with text from social media posts. They will be used later on by text-based geolocation algorithms.
4. **Geolocation prediction** process retrieves the candidate locations for each text fragment using text-based geolocation algorithms.
5. **Map visualization** process implements a disaster map where each dot represents a social media image.

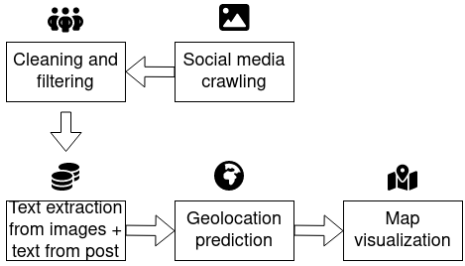


Figure 2. Pipeline using text extraction from images to geolocate social media data

4. Experimental settings and results

4.1. Dataset assessment

The dataset used for the validation of the presented approach is composed of Twitter data from two types of disasters, earthquakes and floods. The dataset consists of two modalities: text and images. The dataset has been created using the Social Media for Disaster Risk Management (SMDRM) Platform[16] gathered by Copernicus Emergency Management Service[3].

The text dataset contains 5,430 posts. From that number, we found that about 4,714 posts were unique. Furthermore, the images dataset contains 1,752 images from four disasters, Catania floods, 2021 (70 images), Central European floods, 2021 (214 images), Croatia earthquake, 2020 (549 images), and Haiti earthquake, 2010 (961 images).

[3] https://emergency.copernicus.eu/

To evaluate which images contain text, a crowdsourcing project was created using the Citizen Science project builder [4] as shown in Figure 3. This analysis is used later on as a ground truth to evaluate the automatic extraction of text. Each image is classified into three categories:(1) "yes", (2) "no", and (3) "yes, but not legible", as shown in Figure 3.

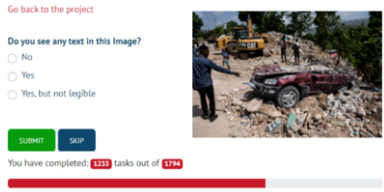


Figure 3. Classifying the images using Citizen Science Project Builder

From 1,752 images, we observed 852 images (49%) having text, 281 (16%) categorized as "yes but not legible" and 619 (35%) images not containing text. The class "yes but not legible" was defined as having small or blurry text which cannot be readable (see Figure 4).

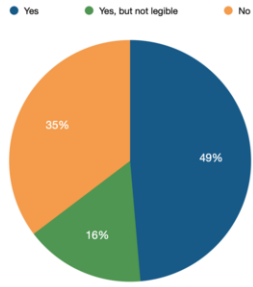


Figure 4. Images category distribution

Out of the 1,133 images containing text, we found 3 images were no longer accessible after these operations. Then, we finally considered 1,130 images.

4.2. Experimental settings

Recent advances in computer vision enable highly accurate extraction of text from images. In this research, we incorporated several existing technologies. We used the Amazon Rekognition service to extract text from images. The service implements a Convolutional Neural Network (CNN) pre-trained on a large labeled dataset. Previous research demonstrated that Convolutional Neural Networks show excellent performance when dealing with image data.

In terms of development environment, programming language, and library, we relied on Python and Google Colab as a development environment. The folium[5] library was used to visualize the predicted geolocation in a map.

[4] https://lab.citizenscience.ch/en/
[5] https://python-visualization.github.io/folium/

A geolocation algorithm, CIME, was implemented into the experiment. The CIME algorithm is deployed in two different configurations, with or without providing the detected language. The language was estimated directly from the text. The estimation was performed using the langdetect[6] library, which supports 55 different languages.

4.3. Experimental results

The Amazon Rekognition[7] service extracted the text from 1,130 images. From 1,130 images, Amazon Rekognition managed to detect the text of 1,029 images. From 1,029 images, we selected 532 images containing real text by excluding images with banners, news logos, images caption, and watermarks. To observe the improvement, we concatenated the extracted text from images into the post. Finally, we had 532 data that were concatenating extracted text from images and social media posts. The text from the post and text from the images were concatenated using a space as the separator. The text concatenated text was then fed to the CIME algorithm for inferring geolocation information. We aim to measure the delta improvement between text from the post and the combination of text from the post and text from images.

4.4. Assessing the relevance of texts from images for geolocating the image

This subsection shows the comparison between different settings in the experiment and measures their enhancement. The goal of this subsection is to contrast the results when incorporating the extracted text from images into social media posts and having only text from social media posts.

	CIME without language prediction			**CIME with language prediction**		
Macro average	Post	Post+Image	Delta	Post	Post+Image	Delta
Precision	0.91	0.96	0.05	0.91	0.96	0.05
Recall	0.97	0.97	0	0.97	0.97	0
F1 score	0.93	0.96	0.03	0.93	0.96	0.03

Table 1. Delta improvement in the two different settings of geolocation algorithm

Table 1 demonstrates the macro average score for each CIME algorithm setting. Macro average is the arithmetic mean of the scores on individual classes. The macro average includes precision, recall, and F1 score. The term post refers to text from post. The term post+image indicates to the concatenated text (text from post and text from images). Overall, both settings produce better performance for the combination of text from the post and text from images compared to text from the post. The improvement is ranging from 3 to 5 % for each performance metric. Furthermore, we found that the different settings (with or without language prediction) indicated no difference result.

In the CIME algorithm without language setting, the number of candidates were decreasing from 310 (only post) to 257 (post + extracted text). Despite the lower number on the post and extracted text, we found that this combination provided better precision, recall, and F1 score. In the CIME algorithm with language settings, the total number of

[6]https://pypi.org/project/langdetect/
[7]https://aws.amazon.com/id/rekognition/

candidates were equally declining from 307 (only post) to 258 (post + extracted text). Similar to previous settings, the result demonstrates better performance.

4.5. Evaluation of distribution of administrative level boundaries

This experiment analyses the administrative level boundaries obtained by the CIME algorithm using just posts and CIME algorithm using posts and texts extracted from the images. The administrative level represent the details of a feature within a government hierarchy. [8]. Intuitively, using texts from images and text from the post would better inform the CIME algorithm and the distribution of administrative level boundaries will be more precise, i.e. using the text extracted from images we are able to geolocate more images at the level of the street and number of the building.

Figure 5 illustrates the distribution of candidates based on the administrative level. In general, the x-axis represents the detail of the administrative level while the y-axis shows the candidate percentage. We can observe that the distribution is concentrated in administrative level 15 (equivalent to quarters, neighborhood). For the experiment without a language detection setting, the percentage is improving from 63% (see green bar) to 65% (see blue bar). For the experiment with the language detection setting, the percentage is increasing from 62 % (see orange bar) to 64 % (see grey bar).

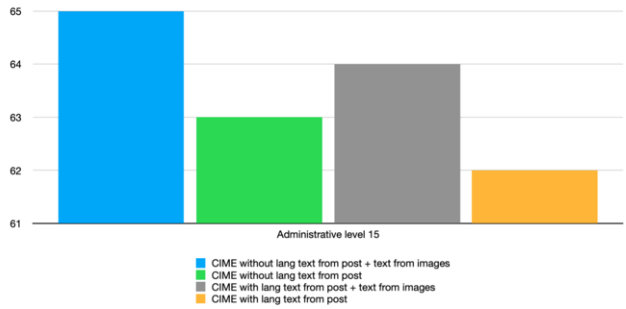


Figure 5. Admin level discrete distribution

4.6. Evaluation of candidate within bounding box and map visualization

This experiment analyses the ratio of geolocation candidates predicted within the affected area for each of the four disasters presented in the database. Each affected area is represented by a bounding box as shown in Figure 6. Each dot represents a geolocation candidate and the geolocation algorithm setting (without language or with language detection). The blue bounding box was manually created and represented the impacted area. The rationale for creating this bounding box is to provide a middle point and trade-off between ”too general” and ”too tight”. The term ”too general” refers to the wide area for instance country level, while ”too tight” means that it only concentrated on the impacted area based on Copernicus activation [9].

[8] https://wiki.openstreetmap.org/wiki/Key:admin_level
[9] https://emergency.copernicus.eu/mapping/list-of-activations-rapid

For CIME without language setting, about 42% (310/736) of text from posts were within the bounding box. While 43.2% (257/594) of concatenated text was inside the bounding box. There was a 1.2% improvement in the number of candidates predicted in the bounding box. For CIME with language setting, About (307/758) 40.8 % text from posts and (258/630) 40.9 % post + extracted text were within the bounding box.

Figure 6. Maps with predicted geolocation

5. Discussion

Results from the experiment show that adding extracted text from images to text from social media post increases the performance. The result demonstrates consistent improvement in both language prediction and without language prediction settings. Practically speaking, the operation team, public information team and the first reponders could benefit the practical contribution from the research as the end users [17].

As the additional practical contribution, we also proposed a pipeline that could be deployed as a feeder for a text-based geolocation prediction algorithm. The evaluation was done by measuring the performance of the extracted text from images into the contents of the posts with the text from posts.

However, we identified that there are four limitation on the use of social media data in the context of disaster response. First, the risk of having false geolocation information should be also considered carefully before adapting the approach to the production. Second, the issue of inclusiveness and accessibility needs to be anticipated for instance by providing data from various social media platforms (Facebook, Instagram, Flickr, and Foursquare). Third, the privacy aspect needs to be reviewed regularly in accordance with privacy policy. For instance, by removing personal information such as name, identification number, age, exact address in the social media data. Finally, the semantic meaning on Twitter might be more difficult to interpret compared to Facebook. This could happen because the character limitation in Twitter platform [18].

Even tough social media offers some advantages, it also comes up with the downside. As an example, the data might be irrelevant and noisy. One approach to address this problem is by incorporating the combination of machine learning and human (expert or crowd) to filter out the misinformation or noise.

6. Conclusion and Future Works

Geoinformation in social media posts serves a significant role for the first responders in the event of a disaster. However, most social media content does not contain geolocation information originally. Despite abundant research on text-based location prediction conducted by researchers in recent years, the research on combining text extraction-based geolocation from images and social media posts was still not discovered.

This paper analyses the potential of automated text extraction from images to enhance geolocation prediction. The proposed approach builds on the state-of-the-art text-based geolocation algorithms, by providing them a concatenation of extracted text from the images and the text from the post. The experimental result shows that the extracted text from images better infers the text-based geolocation algorithm thus improving the quality of geolocation candidate predictions.

As a future works, we plan to improve the geolocation prediction by filtering process by favoring the clear location indication in the social media images, might augment the quality of geolocation prediction.

Measuring multiple dimensions in order to establish a confidence score, for instance lightning condition, text direction in the images, extracted text confidence, number of times that the image appears on social media (also possibly in the past), geographical distribution of the predictions, and scoring derived from the disambiguation process, might be meaningful to investigate and corroborate the efficacy of the predictions. This measurement could be also considered as future direction of the works.

Acknowledgements

This work was funded by LPDP (Indonesia Endowment Fund for Education), the European Commission H2020 project Crowd4SDG “Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience”, #872944. This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work. The authors thank all the Crowd4SDG partners for the discussions and common work during the project.

References

[1] F. Alam, F. Ofli, and M. Imran, “CrisisMMD: Multimodal Twitter datasets from natural disasters,” *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*, 2018.

[2] J. Alonso-Lorenzo, E. Costa-Montenegro, and M. Fernández-Gavilanes, “Language independent big-data system for the prediction of user location on twitter,” in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 2437–2446.

[3] C. Francalanci, P. Guglielmino, M. Montalcini, G. Scalia, and B. Pernici, “IMEXT: A method and system to extract geolocated images from tweets — analysis of a case study,” *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, 2017.

[4] M. N. Utomo, T. B. Adji, and I. Ardiyanto, “Geolocation prediction in social media data using text analysis: A review,” *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018.

[5] H. Bo, E. Cook, and T. Baldwin, "Geolocation prediction in social media data by finding location indicative words," in *24th International Conference on Computational Linguistics (COLINGS 2012)*, 2012, pp. 1045–1062.

[6] M. Maghrebi, A. Abbasi, T. H. Rashidi, and S. T. Waller, "Complementing travel diary surveys with twitter data: Application of text mining techniques on activity location, type and time," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 2015, pp. 208–213.

[7] H. Shekhar, S. Setty, and U. Mudenagudi, "Vehicular traffic analysis from social media data," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 1628–1634.

[8] E. Williams, "Geocontext: Discovering geographical topics from social media," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 1342–1346.

[9] D. Xu, P. Cui, W. Zhu, and S. Yang, "Find you from your friends: Graph-based residence location prediction for users in social media," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 2014, pp. 1–6.

[10] G. Scalia, C. Francalanci, and B. Pernici, "CIME: Context-aware geolocation of emergency-related posts," *GeoInformatica*, vol. 26, no. 1, pp. 125—-157, 2022.

[11] C. Bono, B. Pernici, M. O. Mulayim, J. L. Fernandez-Marquez, A. R. Shankar, and E. Nemni, "TriggerCit: Early flood alerting using twitter and geolocation - a comparison with alternative sources," in *19th ISCRAM Conference*, 2022.

[12] A. Ravi Shankar, J. L. Fernandez-Marquez, B. Pernici, G. Scalia, M. R. Mondardini, and G. Serugendo, "Crowd4EMS: A crowdsourcing platform for gathering and geolocating social media content in disaster response," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-3/W8, p. 331–340, 2019.

[13] F. Murgese, G. Costa-Alcaina, M. O. Mulayim, J. Cerquides, and J. L. Fernandez-Marquez, "Automatic outdoor image geolocation with focal modulation networks," in *Artificial Intelligence Research and Development: Proceedings of the 24th International Conference of the Catalan Association for Artificial Intelligence*. IOS Press, 2022.

[14] H. B. Firmansyah, J.-L. Fernandez Marquez, J. Cerquides, C. Bono, V. Lorini, and B. Pernici, "Enhancing disaster response with automated text information extraction from social media images," in *IEEE Big Data Service*, 2023.

[15] C. Bono, M. O. Mülâyim, C. Cappiello, M. J. Carman, J. Cerquides, J. L. Fernandez-Marquez, M. R. Mondardini, E. Ramalli, and B. Pernici, "A citizen science approach for analyzing social media with crowdsourcing," *IEEE Access*, vol. 11, pp. 15 329–15 347, 2023.

[16] V. Lorini, E. Panizio, and C. Carlos, "Smdrm: A platform to analyze social media for disaster risk management in near real time," in *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media ICWSM*, 2022.

[17] D. P. Freitas, S. B. Marcos R., and d. C. Paulo Victor R., "A conceptual framework for developing solutions that organise social media information for emergency response teams," *Behaviour & Information Technology*, vol. 39, no. 3, pp. 360–378, 2020.

[18] C. Havas, B. Resch, C. Francalanci, B. Pernici, G. Scalia, J. L. Fernandez-Marquez, T. Van Achte, G. Zeug, M. R. R. Mondardini, D. Grandoni, B. Kirsch, M. Kalas, V. Lorini, and S. Rüping, "E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time," *Sensors*, vol. 17, no. 12, 2017.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230702

Text Characterisation for the Fight Against Political Misinformation

Elena Álvarez-García [a] and Daniel García-Costa [a] and Joan Vila-Francés [b] and Francisco Grimaldo [a,1]

[a] *Department of Computer Science, University of Valencia, Spain*
[b] *IDAL, Intelligent Data Analysis Laboratory, University of Valencia, Spain*

Abstract. Social media increase the transmission of false news, hoaxes and opinions, misleading, manipulating and misinforming the population due to the speed at which information spreads. Currently there are different websites known as fact-checkers that help to unmask news as Maldita or Newtral in Spain. They analyse certain news and publications in different media like Twitter, and report on their attempts to manipulate the population with accurate data. To do this, people in charge, analyse the news and look for reliable sources that help to contrast them, but as it is a tedious and manual job, it is not possible to deal with the large amount of information published on networks. The main objective of the work consists of analysing and understanding the emotional information of miss-informed texts in order to identify patterns and characteristics that could possibly be repeated in publications that seek to manipulate and provide false information. In this case we used tweets focused on the last political elections in different regions in Spain: Madrid, Catalunya, Castilla y León and Andalucía. The extracted characteristics, such as emotions, sentiments, toxicity and fakeness, have been obtained using pre-trained models. Finally, based on the characterisation of the texts and the veracity of each one, we train a model capable of approximating whether or not a tweet is a hoax based on the way it is written.

Keywords. Characterisation, Sentiment Analysis, Fake news, Politics, NLP

1. Introduction

The transmission of fake news, hoaxes and opinions is part of a generalized phenomenon that affects the population, misleading, manipulating and misinforming them. Although this problem has always been present in societies throughout the ages [1], it has increased due to the speed at which information spreads in media such as Facebook, Twitter or Tik Tok, and could lead to the amplification of the general lack of trust or the stigmatisation of vulnerable groups, increasing the difficulty to separate what is rumour from what is fact [2].

In order to unmask fake news and seek the truth of the information, fact-checkers nowadays have gained an important position in journalism [3], evaluating the veracity of claims generally by public figures such as politicians, journalists, etc. and trying to reveal this false claims to the population There are important companies or associations such as Chequeando [2] in Argentina, Verificado [3] in Mexico and Maldita[4] or Newtral[5] in Spain

[1]Corresponding Author: Francisco Grimaldo, Department of Computer Science, University of Valencia, Spain; E-mail: francisco.grimaldo@uv.es.
[2]https://chequeado.com/
[3]https://verificado.com.mx/
[4]https://maldita.es/
[5]https://www.newtral.es/

that try to analyse certain news and publications, written in Spanish, in different media like Twitter, and then report on their attempts to manipulate the population with accurate data. There are also well known fact-checkers for Catalan as Verificat [6]. Nevertheless, validating fact-checking is challenging, given the fact that to do this, people in charge have to analyse the news and look for reliable sources that help to contrast them manually. This tedious work becomes not possible to deal with the large amount of information published on networks.

The study in fake news detection on social media has grown enormously and has gained a lot of attention during the last few years, being a very attractive and interesting study for researchers from all over the world [4], developing and using computational techniques to solve this problem [5]. However, the detection of hoaxes or fake news in social networks poses many challenges and problems. Focusing on Twitter, there are many characteristics that we could look at to detect if a tweet is false, such as the way in which it is written, the credibility of the user, the content, the subject, the media ...

Although there are models such as NarrativaAI [7] that helps to detect potentially misleading news written in Spanish, depending on whether there is information about the news on reliable sites, we have focused on analysing the textual information of the tweet, as the sentiments, to identify patterns in hoaxes, with the objective of understanding the written differences between fake or real tweets. For this purpose, we extract and select the most representative sentiment and emotional characteristics from tweets proceeding from a set of Spanish political elections dated in the last years. We will analyse how the way hoaxes are written in Spanish and the malicious style of writing false news, so that people perceive them as true using NLP and Data Mining techniques as Sentiment analysis [6] to accelerate the process of identifying fake news.

2. Proposal and Methodology

This contribution proposes to analyse the written style of Spanish hoaxes by sentiment and emotional analysis using pre-trained models. In this case, the datasets used to carry out the work are about politically themed twitter texts extracted with the Twitter Academic API [8] during the election period in certain autonomous regions. We worked with these data because they can be pseudo-labelled more easily with fact-checkers webs. However, the technique we propose is agnostic to the subject matter of the texts. Finally, with the characterisation made, we trained a model with the text of the tweets together with the characterisation and a custom tag identifying manually the veracity of the information. The model will give an approximation of whether a tweet may be a hoax depending on the way it is written.

For creating the tweets datasets, we selected the last political elections in different regions in Spain: Madrid, Catalunya, Castilla y León and Andalucía. First of all, we made a search of the most propagated hoaxes in those election periods published in platforms against fake news such as maldita.es, newtral.es or verificat.cat [9]. Once the hoaxes have been identified, we proceed to extract the keywords of each fake to find other tweets that talk about this hoax, building with them different queries to the Twitter API using the

[6] https://www.verificat.cat/es
[7] https://huggingface.co/Narrativaai/fake-news-detection-spanish
[8] https://developer.twitter.com/en/products/twitter-api/academic-research
[9] https://verificat.cat/

Twarc2 Python library. For example, for the hoax "The company Indra is in charge of the electoral recount of the postal vote", keywords would be "(indra electoral recount)". In this way, and limiting the time period of the query to that of the election campaign, we obtain a set of tweets that talk about or spread the hoax in different ways. We label them as a fake news with the custom tag veracity false.

To obtain tweets that do not talk about such hoaxes and considered as real information, we made queries with generic election hashtags from each campaign (e.g.: *#EleccionesCatalanas*, *#14F*, *#EleccionesAndalucia2022*) to extract them with the API. Tweets previously classified as hoaxes with this hashtags are eliminated if any. They are labeled as not hoaxes with the custom tag veracity true. To obtain the best possible results the formed datasets containing hoaxes and non hoaxes tweets were balanced, having approximately the same number of hoaxes (veracity=false) as non-hoaxes (veracity=true).

Dataset	Captured period	Num hashtags	Num hoaxes	Veracity	Veracity Count
Andalucia	3 May 2022 - 19 July 2022	24	12	True	577
				False	561
Castilla y León	1 Dec. 2021 - 13 Mar. 2022	13	5	True	1537
				False	1530
Catalunya	1 Dec. 2020 - 14 Mar. 2021	6	45	True	1673
				False	1662
Madrid	18 Mar. 2021 - 4 June 2021	11	53	True	10826
				False	10078

Table 1. Size and number of dataset hoaxes and non-hoaxes (veracity).

Table 1 shows the four different datasets with the number of hoaxes and non hoaxes (Veracity Count). We also indicated the number of hoaxes reported in fact-checkers websites and the number of generic hashtags that we used to extract the tweets.

In order to make the sentiment analysis, we preprocess the tweets removing URLs and stop-words and lemmatizing the texts. Then, we apply different pre-trained models: Pysentimiento [10] for emotions or sentiments analysis, Newtral model [11] for obtaining the level of toxicity and NarrativaAI model to obtain an approximation about the potentially of being deceptive news. All extracted characteristics represent a percentage from 0 to 1 of the feature in the text.

Since the characteristics mentioned above are closely related to each other and this relationship could affect some models. We calculated Spearman correlations to denote which variables measure the same attribute and select the most independent ones in order to create the final model. Analysing it we decided to remove AGGRESSIVE (aggressive - hateful [0.95]), ANGER (disgust - anger [0.94]), NEGATIVE (negative - anger [0.71], negative - disgust [0.77], negative - neutral [-0.82], negative - positive [-0.85]) and VERY TOXIC (toxic - very toxic [0.87]) from the characteristics because they showed the highest correlation with another one.

Once removed, we calculated the AUC/ACC and the F1-score for each characteristic related to the veracity label. In this way we will be able to know which emotions are more related to the hoax class. The results showed a relation in tweets with the veracity label. In general, disgust, fear and sadness are the most predominant and repeated emotions in tweets tagged as hoax comparing the four datasets. In not fake tweets the most repeated emotions were joy, neutral and positive, which can lead to the thought that the way of writing hoaxes is usually more negative and tries to transmit something bad about some topic.

[10] https://github.com/pysentimiento/pysentimiento

[11] https://huggingface.co/Newtral/xlm-r-finetuned-toxic-political-tweets-es

Finally we created a non-parametric model (Random Forest) with the independent characteristics selected after comparing the results of the Spearman correlation to analyse and approach whether or not the tweet is a hoax according to the extracted characteristics. To verify if the trained model works in other contexts in addition to the training one, we created a cross validation test using the four datasets from the Spanish regional elections. Each dataset involves different politic parties, geographical contexts and in the case of the Catalonian datasets, language. We validated each model with the other three datasets. Table 2 shows AUC and F1-score obtained for each of the test models. In each test the Random Forest has been trained with the set of data coloured in grey and test with the rest. The media of AUC and (F1-score) of cells used for train in each model is 0.951 (0.877) while the media in test is 0.791 (0.709). The results show high AUC and F1-socre values in all tests, so we can say that the model is generalisable in different contexts.

	Madrid	Castilla y León	Catalunya	Andalucía
Model 1	0.907 (0.817)	0.902 (0.821)	0.717 (0.618)	0.819 (0.716)
Model 2	0.797 (0.723)	0.988 (0.937)	0.666 (0.628)	0.774 (0.715)
Model 3	0.811 (0.734)	0.849 (0.769)	0.934 (0.849)	0.772 (0.699)
Model 4	0.841 (0.705)	0.881 (0.789)	0.660 (0.588)	0.976 (0.905)

Table 2. Evaluation of Random Forest model generation. [AUC(F1-Score)]. Gray cell indicates the training dataset while white cells the test.

3. Discussions

Analysing the way tweets are written and the emotions they reflect and transmit can help us to approximate if we are dealing or not with a hoax. The results obtained in this work conducts us to understand that normally fake news are written in a more negative way, being disgust, fear and sadness the most representative emotions in tweets tagged as hoax in the four datasets of the regional elections that we used to carry out the work. We created a web-app to visualise the tweets, with the characteristics and the probability of a hoax extracted with the model, this can help to understand when it may be fake news.

References

[1] Taylor PM. Munitions of the Mind: A History of Propaganda from the Ancient World to the Present Day. Manchester: Politics Culture and Society in Early Modern Britain MUP. Manchester University Press; 2003.

[2] McKernon E. Fake news and the public. How the press combats rumor, the market rigger, and the propagandist. Harper's Magazine. 1925 Oct. Available from: https://harpers.org/archive/1925/10/fake-news-and-the-public/.

[3] Lim C. Checking how fact-checkers check. Research & Politics. 2018;5(3):2053168018786848. Available from: https://doi.org/10.1177/2053168018786848.

[4] Shu K, Sliva A, Wang S, Tang J, Liu H. Fake News Detection on Social Media: A Data Mining Perspective. SIGKDD Explor Newsl. 2017 sep;19(1):22–36. Available from: https://doi.org/10.1145/3137597.3137600.

[5] Sahoo SR, Gupta BB. Multiple features based approach for automatic fake news detection on social networks using deep learning. Applied Soft Computing. 2021;100:106983. Available from: https://www.sciencedirect.com/science/article/pii/S1568494620309224.

[6] Oshikawa R, Qian J, Wang WY. A Survey on Natural Language Processing for Fake News Detection; 2020.

This page intentionally left blank

Part 6

Boolean Satisfiability Problems

This page intentionally left blank

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230704

SAT-IT: The Interactive SAT Tracer

Marc CANÉ [a], Jordi COLL [b,1], Marc ROJO [a] and Mateu VILLARET [a,2]
[a] *Universitat de Girona, Girona, Spain*
[b] *Institut d'Investigació en Intel·ligència Artificial, CSIC, Bellaterra, Spain*
ORCiD ID: Jordi Coll https://orcid.org/0000-0002-9385-5723, Mateu Villaret https://orcid.org/0000-0002-8066-3458

Abstract. In this work we present the Interactive SAT Tracer (SAT-IT), a visual and interactive tool to monitor and illustrate the basic algorithms for solving the Boolean Satisfiability Problem (SAT). We consider three algorithms with progressively increasing sophistication: simple backtracking, its extension with unit-propagation, so called Davis-Putnam-Logemann-Loveland (DPLL), and its further extension with clause-learning and conflict driven, so called Conflict-Driven Clause Learning (CDCL). The motivation of this tool is to provide an environment where the user can see the full trace of a SAT solving process in a compact but detailed way and understand the reasons why each variable assignment, backtrack or backjump occurs. Moreover, we want the user to be able to control the evolution of the solving process at the desired pace, and let they choose what branchings (or decisions) must be done. Being able to control the solving process and having detailed information of this process, results in many pedagogical and research applications such as understanding algorithms or designing encodings.

Keywords. SAT, CDCL, DPLL, tool, tracer

1. Introduction

The Boolean Satisfiability problem (SAT) is the paradigmatic NP-complete problem [1,2]. This is the problem of deciding whether a Boolean propositional formula can be satisfied. This problem is not only relevant for being the first problem that was shown to be NP-complete, but his popularity and research interest have been continuously increasing during the last decades for its applicability as a problem-solving paradigm.

The key factors in the success of SAT are the facility to translate a plethora of hard constraint satisfaction and optimisation problems to SAT formulas, and the development of extremely efficient algorithms for solving such formulas. Despite the enormous theoretical computational complexity of SAT, nowadays we have efficient methods capable of solving huge formulas coming from problems of industrial interest such as circuit verification [3], planning [4], scheduling or timetabling [5,6], to name a few.

Except for very specific domains, there is one clearly predominating algorithm to solve SAT: the Conflict-Driven Clause Learning algorithm (CDCL). The essence of

[1]Corresponding Author: Jordi Coll, jcoll@iiia.csic.es, partially supported by Grants PID2019-111544GB-C21 and TED2021-129319B-I00 funded by MCIN/AEI /10.13039/501100011033.
[2]Corresponding Author: Mateu Villaret, mateu.villaret@udg.edu, partially supported by: Grant PID2021-122274OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe.

CDCL is a combination of search and inference. It consists of a (non-chronological) backtracking scheme that explores a search tree to find a solution if any exists, enhanced with the Unit Propagation (UP) rule and which uses the Resolution rule to learn new clauses from dead ends of the search tree. Therefore, understanding the basics of SAT-solving requires to get familiar with the previously mentioned inference rules and their integration into a SAT solving algorithm. In this regard, in [7] a very compact and precise rule-based framework to describe SAT solving algorithms was presented.

Inspired by the previously mentioned framework, in this work we present the Interactive SAT Tracer (SAT-IT), a visual and interactive tool to monitor and illustrate the basic algorithms for SAT solving. In order to facilitate the learning of SAT solving techniques to users that start without a background knowledge, we consider three algorithms with progressively increasing sophistication: simple backtracking, its extension with UP, so called Davis-Putnam-Logemann-Loveland (DPLL), and its further extension with conflict-driven clause learning, i.e. CDCL. In contrast to the teaching tool LearnSAT [8] our tool provides an interactive environment with some graphical support. The motivation of our tool is to provide an environment where the user can see the full trace of a SAT solving process in a compact but detailed way and understand the reasons why each variable assignment, backtrack or backjump occurs. Further information is included in the tool, such as the sequence of resolution rules involved in each conflict analysis, what clauses have been learnt, or what are the inspected literals involved in the 2-watched literals scheme for implementing UP. Moreover, the user is able to control the solving process evolution at the desired pace and choose what variables should be used for branchings (or decisions). The system displays a full log of the SAT solving process that allows to see the overall progress of the execution until the current point, and users can go back to any previous point of the solving process to reinspect or to try "what if" scenarios. In particular, one could see which variables are unit-propagated after some particular assignments. This could serve the user to get some practical insights of some properties of the encodings such as correctness or generalized arc consistency enforcement by UP.

In the rest of the paper, we review some preliminary concepts in Section 2, we present the SAT-IT rule-based notation in Section 3, we describe the considered algorithms for SAT solving in Section 4, we describe the SAT-IT tool in detail in Section 5 and we conclude in Section 6.

2. Preliminaries

A propositional *variable* x can take Boolean truth values, i.e. 0 (meaning *false*) or 1 (meaning *true*). A *literal* l is a variable x, or a negated variable $\neg x$. A *clause* is a disjunction of literals $l_1 \vee \cdots \vee l_k$. A *propositional formula in Conjunctive Normal Form* (CNF) is a conjunction of m clauses $c_1 \wedge \cdots \wedge c_m$. A CNF is usually seen as a set of clauses. An *assignment* is a mapping from propositional variables to $\{0, 1\}$. An assignment *satisfies* (respectively *falsifies*) literal x if $x = 1$ (respectively $x = 0$), and satisfies (respectively falsifies) literal $\neg x$ if $x = 0$ (respectively $x = 1$). An assignment is often expressed as a set of satisfied literals, e.g. $x = 1, y = 0, z = 1$ can be expressed as $\{x, \neg y, z\}$. A literal l is *defined* in an assignment M if $l \in M$ or if $\neg l \in M$, and otherwise is *undefined*. An assignment is *complete* if all the variables are assigned a value, otherwise it is *partial*.

Given an assignment, a clause is satisfied if at least one of its literals is satisfied, it is falsified if all its literals are falsified, and it is *unit* if all its literals are false but one which is unassigned. A CNF is satisfied if all its clauses are satisfied. An assignment satisfies a formula F if it satisfies all the clauses of F. We call an assignment satisfying a formula F a *model* of F. Given two formulas F and F', we say that F' is a *logical consequence* of F, denoted $F \models F'$, if every model of F is a model of F'. This notation can be extended to clauses and assignments: a clause C and a negation of a clause $\neg C$ are also formulas, and an assignment M can be seen as a conjunction of literals, i.e. also as a formula. Then, $M \models C$ means that assignment M satisfies clause C and $M \models \neg C$ means that M falsifies C.

Definition 1 *The* Boolean satisfiability *(SAT) problem consists in determining whether there exists a model for a given CNF formula.*

The unit propagation and the resolution rules play a central role in SAT solvers.

Definition 2 *Given an assignment M and a clause $C \vee l$ such that $M \models \neg C$ and l is undefined in M, (i.e., $C \vee l$ is a unit clause given M), the* unit propagation *(UP) rule extends M to $M \cup \{l\}$.*

Definition 3 *Given two clauses $C \vee l$, $C' \vee \neg l$, the* resolution *rule derives the clause $C \vee C'$. It holds that $(C \vee l) \wedge (C' \vee \neg l) \models C \vee C'$.*

Modern SAT solvers implement the Conflict-Driven Clause Learning (CDCL) algorithm. It is an algorithm that combines search and inference. Moreover, CDCL solvers include a series of techniques which are crucial to achieve the best performance, such as *restarts*, the *variable-elimination inprocessing* and the *learnt clause removal* [9]. However, for now SAT-IT only focuses on the basic search and inference components.

In particular, SAT-IT is designed to explain the search component of SAT-solving algorithms by starting with a naive backtracking scheme which is incrementally enriched, obtaining the Davis-Putnam-Logemann-Loveland (DPLL) algorithm at a first step, and the CDCL algorithm at a second step. The most basic backtracking scheme tries to find a model by successively assigning arbitrary truth values to the variables (it makes *decisions*). If a clause is falsified (it finds a *conflict*), it undoes a decision (it does a *backtrack*) and continues the search, thus forming a search tree. If this basic backtracking scheme is improved by applying UP whenever possible, we obtain the DPLL algorithm. The *decision level* of an assigned literal l is the number of decisions made before being assigned (including l), and the decision level of an assignment (or of a point of the search tree) is the maximum decision level of its literals. If a conflict is found at decision level 0, the formula is unsatisfiable. Finally, CDCL improves DPLL with clause learning. When it reaches a conflict, *conflict analysis* is performed to derive a new clause explaining the conflict, and it does a non-chronological backtrack (or *backjump*). This new clause is inferred with a sequence of applications of the resolution rule, and is added to the formula (it is *learnt*). We refer the reader not familiar with SAT to [10,11,12,7,13] for more detailed explanations about DPLL and CDCL.

3. SAT-IT Rules

In [7] there was presented a rule-based framework to describe the DPLL algorithm and some extensions such as CDCL, or DPLL(T) for solving Satisfiability Modulo Theories

DECIDE:		
$M \parallel F$	$\Longrightarrow M\ l^d \parallel F$	**if** $\begin{cases} l \text{ or } \neg l \text{ occurs in a clause of } F \\ l \text{ is undefined in } M \end{cases}$
BACKTRACK:		
$M\ l^d\ N \parallel F \cup \{C_i\}$	$\Longrightarrow M\ \neg l^k \parallel F \cup \{C_i\}$	**if** $\begin{cases} M\ l^d\ N \models \neg C_i \\ N \text{ contains no decision literals} \end{cases}$
FAIL:		
$M \parallel F \cup \{C_i\}$	$\Longrightarrow FailState$	**if** $\begin{cases} M \models \neg C_i \\ M \text{ contains no decision literals} \end{cases}$
UNITPROPAGATE :		
$M \parallel F \cup \{C_i\}$	$\Longrightarrow M\ l^i \parallel F \cup \{C_i\}$	**if** $\begin{cases} C_i \text{ has the form } C' \vee l \\ M \models \neg C' \\ l \text{ is undefined in } M \end{cases}$
BACKJUMP:		
$M\ l^d\ N \parallel F \cup \{C_i\}$	$\Longrightarrow M\ l'^j \parallel F \cup \{C_i\}$	**if** $\begin{cases} M\ l^d\ N \models \neg C_i, \text{ and} \\ \text{there is some clause } C_j \text{ such that:} \\ \quad C_j \text{ has the form } C' \vee l', \\ \quad F \cup \{C_i\} \models C_j \text{ and } M \models \neg C', \\ \quad l' \text{ is undefined in } M, \text{ and} \\ \quad l' \text{ or } \neg l' \text{ occurs in } F \text{ or in } M\ l^d\ N \end{cases}$
LEARN:		
$M \parallel F$	$\Longrightarrow M \parallel F \cup \{C_j\}$	**if** $\begin{cases} \text{each literal of } C_j \text{ occurs in } F \text{ or in } M \\ F \models C_j \end{cases}$

Figure 1. Rule-based representation of SAT-IT.

(SMT). This framework allows to formally reason about the algorithms using a simple representation. SAT-IT works with (a slight modification of) a subset of the rules presented in [7], that is listed in Figure 1, and described after providing some notation.

A *state* is defined as a pair of the form $M \| F$, where M is a (partial) assignment and F is a formula in CNF, i.e. a finite set of clauses. Every rule, of the form $M \| F \Longrightarrow M' \| F'$, can only be applied if a set of conditions over $M \| F$ are satisfied, and its application changes the state from $M \| F$ to $M' \| F'$. Then, the process of solving a SAT formula F is represented as a sequence of states $S_0 \Longrightarrow S_1 \Longrightarrow S_2 \Longrightarrow \cdots \Longrightarrow S_n$ obtained by a sequence of rule applications, where $S_0 = \{\} \| F$. If F is satisfiable, then S_n is of the form $M \| F'$, where M is a model of F. Otherwise (i.e., F is unsatisfiable), S_n is the special state $FailState$. A partial assignment is not represented as a set of literals but as a sequence M of literals tagged with superscripts. We write together assignments and literals to denote concatenation, e.g. $M\ l$ is the sequence M followed by literal l.

The rules used by SAT-IT are the following:

DECIDE: Extends the partial assignment with an undefined literal l (it *decides* l).

BACKTRACK: This rule can be applied when an existing clause C_i is falsified by the partial assignment, and the partial assignment contains some decided literal. The last decided literal l is negated, and all literals after l are removed from the assignment.

FAIL: This rule can be applied when an existing clause C_i is falsified by the partial assignment, and the partial assignment does not contain any decided literal. The result is the *FailState* (the formula is unsatisfiable).

UNITPROPAGATE: This rule applies UP on the unit clause $C_i = C' \vee l$.

BACKJUMP: This rule can be applied in the same case as the BACKTRACK rule. However, instead of negating the last decision, the rule removes the partial assignment until (and including) a decision l^d, which is not necessarily the last decision (i.e., it performs non-chronological backtracking). The backjump decision level d is determined by a clause $C_j = C' \vee l'$ which is falsified by $M\ l\ N$. More precisely, d is one plus the largest decision level of literals in C'. Note that C_j does not necessarily exist in F, C_i, but it must be a logical consequence of F, C_i. Note also that clause C' is unit with assignment M, and therefore we extend M with l'^j due to UP on clause C_j.

LEARN: This rule adds a new clause to the formula. The added clause must be logical consequence of the formula.

The only rules that add a literal to the partial assignment, possibly after removing other literals, are DECIDE, UNITPROPAGATE, BACKTRACK and BACKJUMP. Like in [7], a literal l that has been added to an assignment M with the DECIDE rule will have the symbol d as a superscript, meaning that l has been assigned with the application of the DECIDE rule. However, in [7] no superscript is added in the remaining cases. SAT-IT rules introduce a small difference: literals assigned by BACKJUMP and UNITPROPAGATE contain an integer superscript i, meaning that l has been assigned by UP on the unit clause C_i; literals assigned by BACKTRACK contain the symbol k as a superscript. This representation is closer to the real implementation of CDCL SAT solvers [12], where each literal in the sequence of assignments (a.k.a. *trail*), has a *reason* to be assigned, i.e. it has been either decided or unit-propagated due to a particular case (no backtrack k reason exists in CDCL). Moreover, this extra information allows a better understanding of the development of the solving process, as well as a better identification of the clauses involved in conflict analysis. Example 1 explains how to read the trail of a given state.

Example 1 *Consider the trail* 1^d -5^5 2^d -3^k 4^0 *corresponding to the second line of Figure 3, left (where* x_i *is denoted just i and* $\neg x_i$ *as -i, as we explain later). Literals* 1 *and* 2 *have been decided (*DECIDE *rule), literal* -3 *is the negation of a previous decision (*BACKTRACK *rule), and literals* -5 *and* 4 *have been unit-propagated due to clauses* C_5 *and* C_0 *respectively, either by* UNITPROPAGATE *rule or by* BACKJUMP *rule. Literals* 1 *and* -5 *have decision level 1, and literals* 2, -3 *and* 4 *have decision level 2.*

4. SAT-IT Algorithms

SAT-IT supports three SAT-solving algorithms with increasing sophistication: BACKTRACKING, DPLL and CDCL. These three algorithms can be described as applications of different sets of rules from Figure 1. In fact, using these rules one could design other solving algorithms. We consider those three since they follow a natural evolution form a naive backtracking one to a full (though basic) CDCL solver. In the provided pseudocode for these algorithms, we use the auxiliary function *canApply*(R,M,F), which checks the shape of M and the conditions of rule R on the state $M \| F$ to determine the applicabil-

Algorithm 1: BACKTRACKING algorithm

Input: $F = \{C_1, \ldots, C_{|F|}\}$, a set of clauses.
Output: If F is SAT: (SAT, M), where M is a model of F. Otherwise: $UNSAT$.

```
1 M ← {}
2 while True do
3     if CanApply(BACKTRACK,M,F) then (M,F) ←Backtrack(M,F) ;
4     else if CanApply(FAIL,M,F) then return UNSAT ;
5     else if M is a complete assignment then return (SAT,M) ;
6     else (M,F) ←Decide(M,F) ;
```

Algorithm 2: DPLL algorithm

Input: $F = \{C_1, \ldots, C_{|F|}\}$, a set of clauses.
Output: If F is SAT: (SAT, M), where M is a model of F. Otherwise: $UNSAT$.

```
1 M ← {}
2 while True do
3     if CanApply(BACKTRACK,M,F) then (M,F) ←Backtrack(M,F) ;
4     else if CanApply(FAIL,M,F) then return UNSAT ;
5     else if M is a complete assignment then return (SAT,M) ;
6     else if CanApply(UNITPROPAGATE,M,F) then (M,F) ←UnitPropagate(M,F) ;
7     else (M,F) ←Decide(M,F) ;
```

Algorithm 3: CDCL algorithm

Input: $F = \{C_1, \ldots, C_{|F|}\}$, a set of clauses.
Output: If F is SAT: (SAT, M), where M is a model of F. Otherwise: $UNSAT$.

```
1  M ← {}
2  while True do
3      if CanApply(BACKJUMP,M,F) then
4          C_j ←ConflictAnalysis(M,F)
5          (M,F) ←Backjump(M,F,C_j)
6          (M,F) ←Learn(M,F,C_j)
7      else if CanApply(FAIL,M,F) then return UNSAT ;
8      else if M is a complete assignment then return (SAT,M) ;
9      else if CanApply(UNITPROPAGATE,M,F) then (M,F) ←UnitPropagate(M,F) ;
10     else (M,F) ←Decide(M,F) ;
```

ity of the rule. The auxiliary functions *Decide*, *UnitPropagate*, *Backtrack*, *Backjump* and *Learn* perform one application of the corresponding rule and return the resulting state.

The simplest algorithm is BACKTRACKING, described in Algorithm 1, and can be defined in terms of the rules DECIDE, BACKTRACK and FAIL. The DPLL algorithm, described in Algorithm 2, is obtained by adding to BACKTRACKING the UNITPROPAGATE rule. Finally, the CDCL algorithm, described in Algorithm 3, replaces the BACKTRACK rule with the BACKJUMP rule, and adds the LEARN rule. Both rules are parametric to a clause C_j that we provide as an argument. In order to obtain C_j, we define the *ConflictAnalysis* function that requires as input a state $M \| F$ where we can apply the

Algorithm 4: Conflict Analysis

Input: F, M: A state where F contains some clause falsified by M.
Output: C: A new clause derived with resolution from a conflict until the 1UIP.

```
C ← a clause in F such that M ⊨ ¬C
while C contains more than one literal of the last decision level do
    l^i ← the rightmost literal of M s.t. ¬l ∈ C
    C ← Resolution(C, C_i)
return C
```

BACKJUMP rule, and therefore where there exists a clause $C_i \in F$ which is falsified by the current assignment M. The conflict analysis procedure is described in Algorithm 4. Starting from an unsatisfied clause C, and by means of the resolution rule, the conflict analysis transforms C until it contains only one literal of the last decision level. Such literal is usually referred to as the First Unique Implication Point (1UIP) [10]. After conflict analysis, the returned clause C_j is learnt and the 1UIP is unit-propagated due to clause C_j in the BACKJUMP rule.

5. SAT-IT GUI and Functionalities

SAT-IT is publicly available[3]. This tool allows the user to work with CNFs in DIMACS format using any of the three given algorithms. We find an example execution of CDCL in Figure 2, and of DPLL and BACKTRACKING in Figure 3. These figures consider the same example input formula:

$$C_0 : x_3 \vee x_4 \vee \neg x_1 \vee x_5 \quad C_2 : x_3 \vee \neg x_4 \vee \neg x_1 \quad C_4 : x_1 \vee \neg x_2 \quad C_6 : \neg x_3 \vee \neg x_4 \vee x_5$$
$$C_1 : \neg x_3 \vee x_4 \vee x_5 \quad C_3 : x_1 \vee x_2 \quad C_5 : \neg x_1 \vee \neg x_5$$

Moreover, when a decision is done, the provided examples always choose the first unassigned variable in the order $1, 2, \ldots, 5$, and always the positive literal.

SAT-IT illustrates the solving process and lets the user control the evolution of the process using the views and functionalities that we describe now. The main window of SAT-IT can be seen in Figure 2. We start by explaining the four main views of the main window of SAT-IT, that we refer as to *trail's history*, *clause list*, *event viewer* and *control buttons*. After this, we describe further functionalities.

Trail's history: The trail's history is shown in the enclosed area at the top-left part in Figure 2. This view shows the evolution of the trail (i.e. of M) as the algorithm progresses. In this window, a tagged literal x_i^t is denoted i^t, and a literal $\neg x_i^t$ is denoted $-i^t$. This representation corresponds to the standard DIMACS format. At the beginning of the execution, it starts with an empty trail. The new literals added to the trail with the rules DECIDE and UNITPROPAGATE are appended to the end of the current trail. Whenever a rule that removes literals from the trail is applied, namely BACKTRACK or BACKJUMP, the current trail is left in its form before the rule application, and the trail resulting of the rule application is added in a new line. See for instance that in Figure 2 there are 3 lines,

[3] https://imae.udg.edu/Recerca/LAI/

since we applied 2 times the BACKJUMP rule. In order to visually illustrate the backtrack point, the part of the trail that is kept after BACKTRACK or BACKJUMP is blurred out. The application of a BACKTRACK or BACKJUMP, which is written at the right of the current trail, is always preceded by the message *CONFLICT i*, meaning that C_i is the conflicting clause that triggered the rule. Moreover, the BACKJUMP rule is also followed by the message *LEARNED j*, meaning that clause C_j is learnt and is the clause that establish the backjump point (recall Algorithm 3). Finally, the last trail will either finish with *SAT* or *UNSAT*. The former case means that the formula is satisfiable and the last trail is a model, while the latter means that unsatisfiability is proven. Note that the content of this history once the solving process has finished shows in full detail what has been the process that the algorithm has followed to find a solution.

Clause list: The bottom-left enclosed part lists all the clauses of the formula. Every row, of the form $i : l_1\ l_2\ l_3\ \ldots\ l_n$ shows a different clause, where i is the name given to clause C_i, and the right part is the list of literals of the clause. Again, the minus sign means negated literal, hence following the standard DIMACS format. At the beginning, the list contains the original clauses. Every time a clause C_j is learned, the corresponding message *LEARNED j* is shown after the trail, and a new clause is added to the list. For instance, in Figure 2 clauses 7 and 8 are learned. Note that the ids of learnt clauses are painted in orange. The colouring of the literals during the solving process is modified. The colouring code of the literals, for a current trail (assignment M), is the following: red means falsified, green means satisfied, and black means unassigned.

Event viewer: This is the top-right part of the window and it shows, in chronological order, some events that the solver has processed together with extra information. The events that we consider are: (i) a decision, specifying the decided literal and its decision level; (ii) a maximal chain of consecutive unit propagations; (iii) a backtrack followed by the backtrack resulting literal; (iv) or a backjump followed by the decision level at which it "backjumps". This view provides therefore a summary of the whole solving process. Moreover, if the user double-clicks a backjump entry, a popup will show the chain of resolution rule applications followed in the conflict analysis in order to derive the clause to be learnt (see Figure 2 and recall Algorithm 4).

Control buttons: These buttons allow the user to control the solving process at the desired pace. The *Decision* button performs a literal decision. Depending on the selected configuration the literal is either picked automatically, or chosen by the user. In case of automatic picking, we can configure SAT-IT to pick the literals either based on input order, or following the VSIDS heuristic [11]. The latter case is only enabled if we use CDCL, and in that case the scores of the variables will be shown. The *Unit Prop.* button advances the execution step by step, doing only one propagation each time, or a decision if needed. The *Conflict* button advances the execution until a conflict is found. The *Resolve C.* button is enabled only when a conflict is found and is used to resolve the conflict. The *End* button will trigger the automatic completion of the solving process.

Other Functionalities: SAT-IT uses and illustrates the two-watched literals (2WL) scheme (see [11]) to apply UP. The literals that are added to the trail are processed in the order they are added since the last decision, in order to find unit clauses where to apply UP. That is, we will consider all unit clauses containing the literal at position i of the trail (starting from the last decision), before continuing with the unit clauses containing the

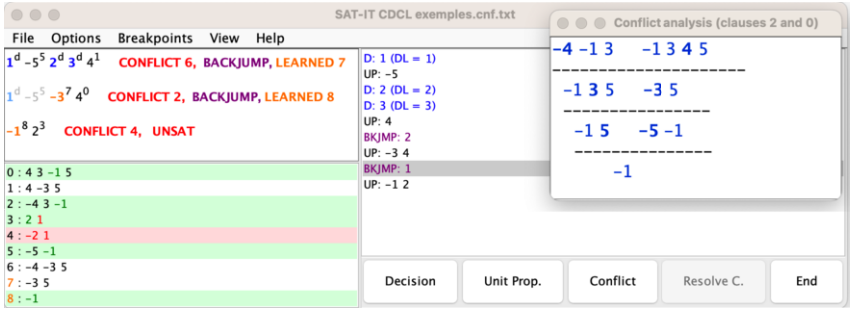


Figure 2. Main view of SAT-IT, with a completed execution of the CDCL algorithm, and the pop-up window showing the conflict analysis for the second backjump.

Figure 3. Left: trail history of a DPLL run. Right: three first and three last lines of the trail history of a BACKTRACKING run (the total number of lines of the example is 16).

literal at position i+1. 2WL benefits from the fact that inspecting two unassigned literals in a clause is enough to check the moment when the clause becomes unit. In SAT-IT, each clause in the clause list view is dynamically reordered to put the two watched literals in the two first positions of the list of literals. The user can thus appreciate how a chain of unit propagations updates the watched literals. SAT-IT also implements a *breakpoint system*, that lets the user indicate variables to be tracked. Then, a message appears every time one of these variables is assigned. In order to try the consequences of different decisions, SAT-IT implements *undo and redo* functionality. This allows to undo the execution process decision by decision, and pick another path (by manual decision selection). Finally, SAT-IT incorporates a preprocess of the input formulas. If the input CNF contains unit clauses, UP is applied until fix point. The formula is simplified accordingly, meaning that satisfied clauses are removed and falsified literals are removed from the remaining clauses. The literals propagated by this preprocess are listed in a line at the beginning of the trail history, with the superscript symbol p.

6. Conclusions an Future Work

We have presented SAT-IT, a visual and interactive tool to monitor and illustrate the basic search-based exact algorithms for SAT solving in a compact but detailed and very informative way. Whereas the tool can rapidly solve fairly large formulas, it has been designed to deal with formulas of few tens or hundreds of variables, so that the evolution of the solving process can be visually inspected. We believe that this tool can be really helpful to assist the teaching basics on SAT solving for different reasons. First, it is a very intuitive graphical tool that shows a good selection of the most relevant information involved in a SAT solving process. Second, the three included algorithms with increasing

sophistication allow the users to approach gently to the full CDCL algorithm. The used language and the trace of the algorithms are completely deterministic and the notation is also suitable for hand writing resolution of SAT formulas. These facts favours the use of the tool as an assistant to correct practical exercises and assist autonomous study.

The practical utility of the tool goes much beyond the learning of the CDCL algorithm or even teaching. For instance, SAT-IT can be used as an assistant for creating and validating examples for research reports. In particular, in papers on SAT modelling and solving, we can frequently find illustrative examples where an initial CNF formula is provided, and the inferences made by UP on the formula are then analysed. Similarly, SAT-IT can be used to analyse the properties of some SAT formulas or to find counterexamples. As an illustrative example we consider the analysis of the propagation strength of a SAT encoding. We are given a SAT formula consisting of the encoding of the cardinality constraint $x_1+x_2+x_3+x_4\geq 3$, and this encoding is supposed to maintain Generalized Arc Consistency (GAC) by UP [14]. Satisfying the GAC property implies, for instance, that assigning *false* to one of the three variables must make UP assign *true* to the other three. This scenario can be easily verified with SAT-IT by means of a manual decision followed by UP, and the user can inspect which clauses are involved in this process.

We plan to improve SAT-IT by including more features of CDCL SAT solvers such as restarts, learnt clause removal and bounded variable elimination, by providing graphic representation of unsatisfiability proofs, and by supporting MaxSAT solving algorithms.

References

[1] Cook SA. The Complexity of Theorem-Proving Procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing. STOC '71; 1971. p. 151-158.
[2] Levin LA. Universal sequential search problems. Problemy peredachi informatsii. 1973;9(3):115-6.
[3] Kaufmann D, Biere A, Kauers M. Verifying Large Multipliers by Combining SAT and Computer Algebra. In: 2019 Formal Methods in Computer Aided Design (FMCAD); 2019. p. 28-36.
[4] Kautz HA, Selman B, et al. Planning as Satisfiability. In: ECAI. vol. 92. Citeseer; 1992. p. 359-63.
[5] Demirovic E, Musliu N, Winter F. Modeling and solving staff scheduling with partial weighted maxSAT. Ann Oper Res. 2019;275(1):79-99.
[6] Bofill M, Coll J, Garcia M, Giráldez-Cru J, Pesant G, Suy J, Villaret M. Constraint Solving Approaches to the Business-to-Business Meeting Scheduling Problem. J Artif Intell Res. 2022;74:263-301.
[7] Nieuwenhuis R, Oliveras A, Tinelli C. Solving SAT and SAT Modulo Theories: From an abstract Davis–Putnam–Logemann–Loveland procedure to DPLL(T). Journal of the ACM. 2006;53(6):937-77.
[8] Ben-Ari MM. LearnSAT: a SAT solver for education. In: International Conference on Theory and Applications of Satisfiability Testing. Springer; 2013. p. 403-7.
[9] Elffers J, Giráldez-Cru J, Gocht S, Nordström J, Simon L. Seeking Practical CDCL Insights from Theoretical SAT Benchmarks. In: Lang J, editor. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI; 2018. p. 1300-8.
[10] Marques-Silva JP, Sakallah KA. GRASP: A search algorithm for propositional satisfiability. IEEE Transactions on Computers. 1999;48(5):506-21.
[11] Moskewicz MW, Madigan CF, Zhao Y, Zhang L, Malik S. Chaff: Engineering an efficient SAT solver. In: Proceedings of the 38th annual Design Automation Conference; 2001. p. 530-5.
[12] Eén N, Sörensson N. An extensible SAT-solver. In: Theory and Applications of Satisfiability Testing: 6th International Conference, SAT 2003, Santa Margherita Ligure, Italy, May 5-8, 2003, Selected Revised Papers 6. Springer; 2004. p. 502-18.
[13] Biere A, Heule M, van Maaren H, Walsh T, editors. Handbook of Satisfiability - Second Edition. vol. 336 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2021.
[14] Bofill M, Coll J, Nightingale P, Suy J, Ulrich-Oltean F, Villaret M. SAT encodings for Pseudo-Boolean constraints together with at-most-one constraints. Artif Intell. 2022;302:103604.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230705

General Boolean Formula Minimization with QBF Solvers

Eduardo CALÒ [a,1] and Jordi LEVY [b]
[a] *Utrecht University, The Netherlands*
[b] *IIIA, CSIC, Spain*
ORCiD ID: Eduardo Calò https://orcid.org/0000-0003-3881-8994, Jordi Levy https://orcid.org/0000-0001-5883-5746

Abstract. The minimization of propositional formulae is a classical problem in logic, whose first algorithms date back at least to the 1950s with the works of Quine and Karnaugh. Most previous work in the area has focused on obtaining minimal, or quasi-minimal, formulae in conjunctive normal form (CNF) or disjunctive normal form (DNF), with applications in hardware design. In this paper, we are interested in the problem of obtaining an equivalent formula in any format, also allowing connectives that are not present in the original formula. We are primarily motivated in applying minimization algorithms to generate natural language translations of the original formula, where using shorter equivalents as input may result in better translations. Buchfuhrer and Umans have proved that the (decisional version of the) problem is Σ_2^p-complete. We analyze three possible (practical) approaches to solving the problem. First, using brute force, generating all possible formulae in increasing size and checking if they are equivalent to the original formula by testing all possible variable assignments. Second, generating the Tseitin coding of all the formulae and checking equivalence with the original using a SAT solver. Third, encoding the problem as a Quantified Boolean Formula (QBF), and using a QBF solver. Our results show that the QBF approach largely outperforms the other two.

Keywords. SAT Solvers, QBF Solvers, Boolean Formula Minimization, Natural Language Processing

1. Introduction

The minimization of complex Boolean expressions is a longstanding problem in logic. The first algorithms developed in the 1950s, e.g., the works of Quine, McCluskey [1,2,3], and Karnaugh [4] paved the way for extensions and optimizations in the following years (e.g., the Petrick's method [5], and the Espresso heuristic logic minimizer [6], *i.a.*). These works have focused on obtaining minimal equivalent representations in specific canonical forms (e.g., conjunctive normal form (CNF) or disjunctive normal form (DNF)), and confined the studies to a limited set of connectives. Here, we are interested in the *general* Boolean formula minimization, where no assumptions are made in the form of the input

[1] Corresponding Author: Eduardo Calò, e.calo@uu.nl. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860621 and the MICINN project PROOFS (PID2019-109137GB-C21).

formula or the output. In fact, our minimization methods allow us to use distinct sets of connectives for the input and the output.

We frame Boolean minimization (i.e., finding the logically equivalent shortest formula(e) to a given one) as a Quantified Boolean Formulae (QBF) satisfiability problem and design an algorithm that consistently finds the shortest equivalents of a given formula. We compare this algorithm with a brute force baseline, and an approach based on SAT.

Motivation We have two main motivations behind our work. The first one is shared in [7], where the authors present `qbf2epr`, a tool that translates QBF to formulae in effective propositional logic (EPR). Their aim is to generate benchmarks for EPR and compare solvers for QBF and EPR. Similarly, our formula minimization problem, encoded as QBF, generates benchmarks for QBF solvers and allows us to compare SAT and QBF techniques. The automated deduction community is divided into sub-communities (e.g., SAT, QBF, SMT, MaxSAT, EPR), which try to solve distinct classes of problems, from SAT that is NP-complete, to EPR that is NEXPTIME-complete, passing by QBF that is PSPACE-complete, and each one has its own competition and set of benchmarks. However, many ideas that proved effective in one area (like learning in SAT) have been exported to others. In this sense, problems that could be solved with two distinct technologies, like ours, contribute to comparing the level of maturity reached in each area.

Our second motivation relates to a use case of minimization algorithms in natural language processing. Grasping the meaning of logical formalisms is a crucial task for many scholars, yet sometimes even experienced logicians might have trouble deciphering a complex formula. This problem is exacerbated with students of logic, particularly when they encounter unfamiliar formal systems [8]. Techniques from natural language generation [9,10], and in particular logic-to-text generation methodologies [11,12], can be used for simplifying and translating logical formulae into optimally intelligible text in natural languages (NLs) (such as English, Mandarin, or Korean), which can effectively explain formulae to systems' users. For example, given the following first-order formula:

$$\exists x(Problem(x) \wedge \forall y(Researcher(y) \rightarrow Interested(y,x)))$$

we want a system that can automatically generate a faithful and comprehensible explanation, via the following (or another semantically equivalent) text:

There is a problem that every researcher finds interesting.

What are the characteristics that a formula should have to become a suitable input for a logic-to-text translation system? One aspect that one might want to look at is length. Brevity has surrounded linguistic debate at least since [13]. Arguably, shorter utterances should be preferred over longer ones and unnecessary prolixity should be avoided. This principle might also apply to logical formulae. Intuitively, a short formula, rather than a longer logical equivalent, should be better suited to be translated into NL. In this paper, we tackle exclusively the logical aspect of the problem. For the application of the algorithm developed in this paper to the linguistic aspect of the problem, see [14]. We focus on propositional logic, a formalism in which equivalence is decidable, and limit

our examination to formulae's length,[2] aiming only for the shortest equivalents to a given formula.

1.1. Related Work

Formula Minimization Boolean formula minimization is a natural optimization problem in the second level of the Polynomial-Time Hierarchy Σ_2^p. Indeed, the problem is used by [15] to motivate the definition of the Polynomial Hierarchy. Its decisional version can be formulated as the following problem: Given a Boolean formula, prove the existence of a (smaller) formula (in the same set of variables) that gets the same evaluation of the given formula, for all possible assignments of the variables. The fact that both sets of quantified variables, as well as the time to evaluate the formulae, are bounded in the input proves its inclusion in Σ_2^p. As we will see in Section 2, this corresponds to our brute-force algorithm. It is assumed that both the given formula and its minimization are circuits or formulae of the same form. However, in our implementation, we leave open the possibility to use distinct sets of connectives. Apart from some completeness proofs for some particular forms of the input and output [16], the proof for the general form had eluded researchers until [17] proved Σ_2^p-completeness of the problem.

The optimization of complex Boolean expressions has been studied extensively in electronic circuits, where practical matters (i.e., complex circuits take up physical space and costs more resources in their implementation) make it crucial to find optimal circuit representations. Well-known minimization methods include the Quine-McCluskey algorithm [1,2,3] and the Karnaugh map [4]. In the Karnaugh map, Boolean results are transferred from a truth table onto a two-dimensional grid, where each cell position represents one combination of input conditions, while each cell value is the corresponding output value. Optimal groups of 0s and 1s are identified, which represent the terms of a canonical form that can be used to write a minimal expression. The Quine–McCluskey algorithm finds all the prime implicants of a function and uses them in a chart to find (i) the essential prime implicants of the function, and (ii) other prime implicants that are necessary to cover the function. The method is functionally identical to the Karnaugh map, but its tabular form makes it more efficient to employ in computer systems.

However, despite this long history of research and attempts to extend well-established methods (e.g., [18] tries to implement the `XOr` operator in the Quine-McCluskey algorithm), most work has focused on a limited set of connectives and canonical forms (e.g., CNF or DNF). For our scope, we need a more general approach where all connectives could be, in principle, taken into account.

Quantified Boolean Formulae Quantified Boolean Formulae (QBFs) are an extension of propositional logic, where universal and existential quantifications are allowed [19]. The use of quantifiers results in a greater expressive power than classic propositional logic. If all variables occurring in a QBF ϕ are bound, then ϕ is called *closed*. QBFs often assume a canonical prenex conjunctive normal form (PCNF) $\phi = \exists\vec{x}\forall\vec{y}\exists\vec{z}\cdots\psi$, where the portion containing only quantifiers and bound variables is called the *prefix*, followed by ψ that is a quantifier-free Boolean formula with conjunctions over clauses, called the *matrix*.

[2] We define length as the number of symbols (i.e., predicates and connectives, parentheses excluded) contained in a formula.

Algorithm 1: Brute-force algorithm

```
Input: φ
Output: a minimal equivalent formula ψ
1  Function equivalent(φ, ψ):
2      foreach Assignment I : Var(φ) → {0,1} do
3          if I(φ) ≠ I(ψ) then
4              return false
5      return true
6  Function main(φ):
7      foreach i = 1,...,|φ| do
8          foreach formula ψ s.t. |ψ| = i and Var(ψ) ⊆ Var(φ) do
9              if equivalent(φ, ψ) then
10                 return ψ
```

The QBF satisfiability problem [20] consists of determining, for a given QBF ϕ, the existence of an assignment for the free variables, such that ϕ evaluates to true under this assignment. Hence, ϕ is true iff, there exists a truth assignment to $\vec{x}$, such that, for all truth assignments to $\vec{y}$, there exists a truth assignment to $\vec{z}$,... such that ψ is true. Several QBF solvers have been developed over time,[3] and applications of QBFs technologies range from AI to planning [21,22,23]. QBF solvers only use to provide the instantiation of most externally existentially-quantified variables $\vec{x}$, since for the other ones, instantiation depends on previous universal variables $\vec{z} = f(\vec{y})$. In this work, we exploit QBFs to encode and solve the Boolean minimization problem.

1.2. Structure of the Paper

The rest of the paper is structured as follows. Section 2 introduces the algorithms that we employ in our experiments and the QBF encoding we develop. Section 3 illustrates the experiments we carry out, comparing the three aforementioned approaches, and shows the results. We present some reflections on possible future directions in Section 4.

2. Algorithms

In our experimentation, we analyze three algorithms that we will call brute-force, SAT-based, and QBF-based.

Brute-force Algorithm The brute-force algorithm (see Alg. 1) is the algorithm that we mention in Section 1.1 as proof that formula minimization is in Σ_2^p. Two formulae ϕ and ψ are equivalent iff $\phi \leftrightarrow \psi$ is a tautology. Like TAUT, the formula equivalence problem is CoNP-complete. Considering that we test the equivalence for all formulae ψ smaller than ϕ, the average time required by `equivalent(`ϕ, ψ`)` is the same as considering ψ

[3] http://www.qbflib.org

a random formula smaller than ϕ. Assuming that the probability of $I(\phi) = 1$ is $1/2$, this average time would be $\mathcal{O}(|\phi|)$ (notice that in this situation, half of the calls finish after checking one assignment, $1/4$ after checking two assignments, etc. hence, on average we would check $\sum_{i=1}^{2^{|\phi|}} i\frac{1}{2^i} < 2$ assignments in every call). However, the assumption $P(I(\phi) = 1) = 1/2$ is, in general, false.

We can also estimate the number of calls to this function as follows. The number of distinct complete trees of size n that we can construct with k binary symbols is k^n. If the trees can have any form, then the computation is more complicated. Let $\mathscr{C}$ be the set of possible binary symbols (hence, we are not considering `Not`) and $\mathscr{V}$ be the set of possible leaves. The number of forms of trees with m binary nodes and $m+1$ leaves is given by the recurrence $f(m) = \sum_{i=0}^{m-1} f(i) f(m-i-1)$ that define the Catalan numbers C_m. The number of distinct trees will be $C_m |\mathscr{C}|^m |\mathscr{V}|^{m+1}$. Using Stirling approximation, this can be approximated as $\frac{4^m}{\sqrt{\pi} m^{3/2}} |\mathscr{C}|^m |\mathscr{V}|^{m+1}$. As a function of the tree size $n = 2m+1$, this is $\mathcal{O}((4|\mathscr{C}||\mathscr{V}|)^{n/2}/n^{3/2})$ calls to the `equivalent` function.

Algorithm 2: SAT-based algorithm

```
Input: φ
Output: a minimal equivalent formula ψ
Function tseitin(φ, x):
    if φ = φ1 ∧ φ2 then
        y1, y2 := freshvars()
        return tseitin(φ1, y1) ∪ tseitin(φ2, y2) ∪ CNF({x ↔ y1 ∧ y2})
    ··· /* Similarly for other connectives or variables */
Function equivalent(φ, ψ):
    x1, x2 := freshvars()
    Γ := tseitin(φ, x1) ∪ tseitin(ψ, x2) ∪ CNF({¬(x1 ↔ x2)})
    return SAT(Γ) ≠ satisfiable
Function main(φ):
    foreach i = 1, ..., |φ| do
        foreach formula ψ s.t. |ψ| = i and Var(ψ) ⊆ Var(φ) do
            if equivalent(φ, ψ) then
                return ψ
```

SAT-based Algorithm The second algorithm (see Alg. 2) is based on the use of a SAT solver and the Tseitin encoding of the two formulae that we want to prove equivalent. Given two formulae ϕ, ψ, we can find, in linear time $|\phi| + |\psi|$, a CNF formula Γ such that the two formulae are equivalent iff Γ *is not* satisfiable. Experiments show that, in practice, we still can get some gain with respect to the brute-force algorithm (see Section 3).

QBF-based Algorithm The third algorithm (see Alg. 3) is based on the use of a QBF solver. Here, instead of testing every possible minimal formula ψ, we test every possible depth δ. This supposes a significant improvement since there is a linear number of depths to try, instead of an exponential number of formula candidates. Second, instead of

Algorithm 3: QBF-based algorithm

Input: ϕ
Output: a minimal equivalent formula ψ

Function `scheme(`δ, z`):`
  $x_{false} := freshvar(\exists^{(1)})$
  $\Gamma := CNF(\{x_{false} \rightarrow \neg z\})$
  foreach $y \in \forall^{(1)}$ **do**
    $x_y := freshvar(\exists^{(1)})$
    $\Gamma := \Gamma \cup CNF(\{x_y \rightarrow (z \leftrightarrow y\})$
  if $\delta > 0$ **then**
    $z_1, z_2 := freshvar(\exists^{(2)})$
    $\Gamma := \Gamma \cup scheme(\delta - 1, z_1) \cup scheme(\delta - 1, z_2)$
    foreach $c \in \mathscr{C}$ **do**
      $x_c := freshvar(\exists^{(1)})$
      $\Gamma := \Gamma \cup CNF(\{x_c \rightarrow (z \leftrightarrow z_1\ c\ z_2)\})$
  return $\Gamma \cup CNF(\{x_{false} + \sum_{y \in \forall^{(1)}} x_y + \sum_{c \in \mathscr{C}} x_c = 1\})$
Function `equivalent(`ϕ, δ`):`
  $z_1, z_2 := freshvars(\exists^{(2)})$
  $\forall^{(1)} := Vars(\phi)$
  $\Gamma := tseitin(\phi, z_1) \cup scheme(\delta, z_2) \cup CNF(\{z_1 \leftrightarrow z_2\})$
  return $QBF(\Gamma) = true$
Function `main(`ϕ`):`
  foreach $\delta = 1, \ldots, depth(\phi)$ **do**
    if $equivalent(\phi, \delta)$ **then**
      return ψ *extracted from* Γ

a Tseitin encoding of the candidate, we compute a *scheme of the candidate*. This means that given a depth δ, we consider all terms of depth δ as possible candidates, without fixing the content of each node of the candidate. The equivalence between the original formula and this *scheme* can be encoded as a QBF formula with three quantifier alternations: $\exists \vec{x}.\forall \vec{y}.\exists \vec{z}.\Gamma$. In Alg. 3, these three sets of variables are represented as $\exists^{(1)}$, $\forall^{(1)}$, and $\exists^{(2)}$ and individual variables are named x, y, and z, respectively. If the QBF formula is true, the values we got for variables $x \in \exists^{(1)}$ will encode the minimal formula. Notice that QBF solvers only provide the instantiations of the most external existentially-quantified variables, since the values of the other existentially-quantified variables depend on more externally universally-quantified variables.[4]

These *schemes* are defined as follows. We assume that there is a maximal arity for all connectives; in our case 2 (although it could be generalized for any set of fixed-arity connectives). A scheme of depth δ is basically a complete tree of depth δ, therefore containing $2^{\delta+1} - 1$ nodes. For every node i of the scheme, and for every truth constant

[4] In the case of using a QBF solver unable to provide these instantiations, we cannot compute the minimal equivalent formula.

(we only consider the constant $false$), or variable y of the original formula, or connective c, we have a variable in x^i_{false}, x^i_y, $x^i_c \in \exists^{(1)}$. When $x^i_c \in \exists^{(1)}$ gets the value true, then at position i of the scheme, we have the connective c. Respectively, when variable $x^i_y \in \exists^{(1)}$ is true, we have variable y at position i, or constant $false$ when x^i_{false} is true. Additionally, we will have a variable x^i_{dummy} that gets the value true when at position i of the scheme there is not any content. The constraints:

$$x^i_{dummy} + x^i_{false} + \sum_{y \in \forall^{(1)}} x^i_y + \sum_{c \in \mathscr{C}} x^i_c = 1 \tag{1}$$

in the QBF formula will ensure that one, and only one, of them get the value true. Variables in $\forall^{(1)}$ are just the set of variables in the original formula. The original formula and the scheme are equivalent if *for all* assignments to these variables, both the original formula and scheme get the same evaluation. Variables in $z^i \in \exists^{(2)}$ encode the truth values for every possible subformula at position i of the scheme or of the Tseitin encoding of the original formula. The QBF formula will also contain restrictions of the form:

$$\begin{aligned} &x^i_{false} \rightarrow \neg z^i \\ &x^i_y \rightarrow (z^i \leftrightarrow y) \\ &x^i_{\texttt{Not}} \rightarrow (z^i \leftrightarrow \neg z^{i \cdot 1}) \\ &x^i_c \rightarrow (z^i \leftrightarrow z^{i \cdot 1}\, c\, z^{i \cdot 2}) \end{aligned} \tag{2}$$

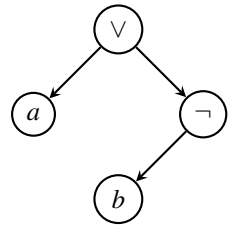
The intended meanings of these constraints are: If at position i of the scheme we have the constant $false$, the sub-scheme is evaluated to false, if there is an original variable $y \in \forall^{(1)}$, it is evaluated to y, and if there is a connective $c \in \mathscr{C}$, then the sub-scheme gets the same value as the connective c operated on the evaluations $z^{i \cdot 1}$ of the left-child of i and the evaluation $z^{i \cdot 2}$ of the right-child.

When in a node of the scheme we put a `Not`, we only use one of the children, and in the case of putting a variable, we do not use any of the children. To avoid useless search in the QBF solver, we can force all these useless nodes to be fixed to the dummy value by adding the following constraints to the QBF formula:

$$\begin{aligned} &x^i_{\neg} \rightarrow x^{i \cdot 2}_{dummy} \\ &(x^i_y \vee x^i_{dummy} \vee x^i_{false}) \rightarrow (x^{i \cdot 1}_{dummy} \wedge x^{i \cdot 2}_{dummy}) \end{aligned} \tag{3}$$

In this encoding of schemes, from the assignment computed by the QBF solver for the variables in the most-externally existentially-quantified $\exists^{(1)}$ of the formula Γ, we can obtain the minimized formula ψ. For instance, if the variables set to true are the ones on the left, the formula ψ is the one on the right:

$x^{\varepsilon}_{\texttt{Or}}$
x^1_a $\quad$ $x^2_{\texttt{Not}}$
$x^{1 \cdot 1}_{dummy}$ $\quad$ $x^{2 \cdot 1}_b$
$x^{1 \cdot 2}_{dummy}$ $\quad$ $x^{2 \cdot 2}_{dummy}$



Notice that the size of the QBF formula we get is $\mathcal{O}(2^{depth(\phi)} \cdot (|\mathscr{C}| + |Vars(\phi)|))$. Assuming that the original formula is balanced, in practice $2^{depth(\phi)} \approx |\phi|$. Therefore, we could consider it a polynomial encoding. Notice also that we do not make a profit from the commutativity and associativity of most connectives.

In this approach, we only bound the depth of the scheme. If we also want to limit its size, we can add the encoding of some cardinality constraint that bound the number of nodes in the schema that are distinct from the dummy:

$$\sum_i \neg x^i_{dummy} \leq size_bound \tag{4}$$

Notice that this QBF-based algorithm may be used to minimize the depth of an equivalent formula, without using Eq. (4). If instead, we want to minimize the size of an equivalent formula, it is not so simple as adding Eq. (4). It is possible that, given an original formula ϕ, there exists an equivalent formula ψ, satisfying $size(\psi) < size(\phi)$, but where $depth(\psi) > depth(\phi)$, and it would not be found by the algorithm. Stricto sensu, we should try all schemes of depth $depth(\psi) \leq size(\phi)$ to avoid this situation and ensure that all smaller-sized equivalent candidates are considered. However, this would lead us to obtain QBF formulae of size $\mathcal{O}(2^{size(\phi)})$. In the experiments (see Section 3), we have observed that it is enough to consider all schemes of size $size(\psi) \leq size(\phi)$ and imposing a bound $depth(\psi) \leq \log_2 size(\psi) + 1$, to get the same results as with the other algorithms.

3. Experiments and Results

We conduct some experimentation with our algorithms. The three algorithms are implemented in Python 3 and are publicly available at `https://gitlab.nl4xai.eu/eduardo.calo/QBF-boolean-minimization`. In the case of the SAT-based algorithm, we use `Glucose 4` [24], a state-of-the-art SAT solver, via the Python module `python-sat`.[5] In the case of the QBF-based algorithm, we use the state-of-the-art `CAQE` [25,26] QBF solver, although any other QBF solver that accepts QDIMACS standard[6] input and output may be used.

For every size in $s = 1, \ldots, 20$, we generate 100 random formulae of size s and minimize them using the three algorithms. We make sure that all syntactically distinct formulae are generated with the same probability. We do it carefully to avoid any bias. However, we do not consider commutativity and associativity of connectives or other formula equivalences. We generate formulae of size s over a set of $\sqrt{s}$ variables[7] and connectives $\mathscr{C} = \{\texttt{Not}, \texttt{And}, \texttt{Or}\}$, and minimizations are searched among formulae with connectives $\mathscr{C}' = \{\texttt{Not}, \texttt{And}, \texttt{Or}, \texttt{Implies}\}$.[8] We also check that our conclusions are the same for another number of variables (e.g., half of the desired size, etc.) and that the implementations agree on the results in terms of formula size using any of the three methods.

[5] `https://github.com/pysathq/pysat`

[6] `http://www.qbflib.org/qdimacs.html`

[7] Using s/c or s^c variables does not seem to affect substantially the results.

[8] We include `Implies` in $\mathscr{C}'$ to remark that the set of input and output connectives may be different (contrarily to other approaches).

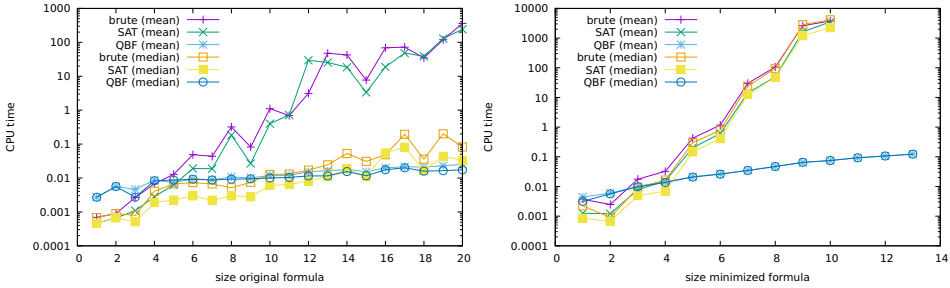


Figure 1. Average and median time required by the three algorithms with respect to the size of the original formula (left) and the resulting minimized formula (right).

We use a cluster with 11 calculating nodes with 2 Intel Xeon CPUs at $2.2GHz$ with 10 cores/CPU and $92GB$ of RAM. We set a time-out of $20,000s$. The brute-force and the SAT-based algorithms reach the time-out in some instances for $s = 15, 18, 19, 20$. These values are not considered in the computation of the mean and median times. Therefore, these mean and median values are abnormally low.

In Figure 1 (left), we show the average and median (logarithm of) CPU time required by each one of the algorithms as a function of the size of the input formula. We clearly observe that the QBF-based algorithm outperforms the other two algorithms, which seem to require exponential time on the size of the input. We also observe that the SAT-based is consistently better than the brute-force algorithm (a constant distance between the functions, in logarithmic axes, means an improvement of constant factor). This is quite surprising since, as we mention in Section 2, the computation of the formula equivalence can be done in linear average time.[9] It is also remarkable that, in the case of brute-force and SAT-based, there is a significant difference between the average and median time. The reason, as we comment in detail below, is the significant variability in the times required by each instance. The same effect produces a fluctuation in the values of the average time. We can conclude that, although in most of the instances (attending to the median), the three algorithms minimize the formula in less than one second, for sizes smaller than 20, just a few instances make brute-force and SAT-based require around $1h$ on average when the size is around 20.

In Figure 1 (right), we show the average and median (logarithm of) CPU time as a function of the size of the obtained minimal formula. Here the differences between the mean and median times are smaller. Hence, we can conclude that the size of the output determines the time required by the algorithms. Again, it is clear that the QBF-based algorithm outperforms the other two. We still observe that the median time is smaller than the average time, which indicates that significant variability still exists. Curiously, the times depend on the parity of the formula sizes: Even-size formulae are easier than odd-size formulae. The reason could be that, except in the case of negation, the rest of the connectives are binary.

In Figure 2 (left), we show how the average size of the minimized formula grows with respect to the size of the original formula. We observe that the growth is close to the square root of the original size. Recall that we generate random formulae of size s

[9] Equivalence checking of two propositional formulae is CoNP-complete in the worse case. However, for random formulae, and random assignments, the average time is decent.

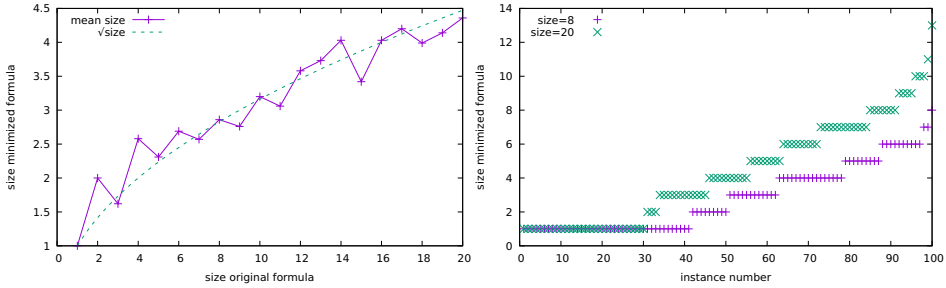


Figure 2. Average size of the minimized formula w.r.t. the size of the original formula (left) and distribution of minimized sizes for formulae of original size 8 and 20 (right).

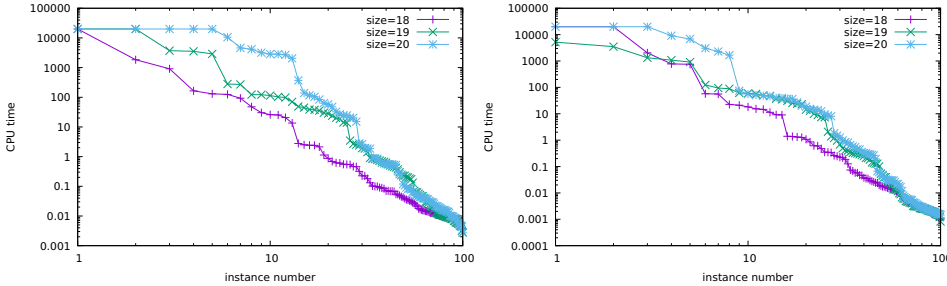


Figure 3. Distribution of times for the instances solved with the brute-force algorithm (left) and SAT-based algorithm (right).

and with $\sqrt{s}$ variables. Curiously, we observe that odd-size formulae are simplified more than even-size formulae, although the reason for this is not clear. In Figure 2 (right), we show the distribution of sizes of the minimized formulae (for original formulae of sizes 20 and 8).

As mentioned above, we observe significant variability in CPU times, for the brute-force and SAT-based algorithms. In Figure 3, we sort the instances in decreasing order of CPU time and represent, in double logarithmic axes, these times for the 100 instances. We observe that this representation is close to a line (truncated on the top due to time-outs) with an increasing (negative) slope when the size increases. This implies that the CPU time in these algorithms follows a power-law probability distribution, where the time required by a few instances is responsible for most of the average time. The standard solution in these situations is to use some kind of restart policy or some randomization of the algorithm. In our case, we could randomize the order of the candidates to minimal formulae. However, since we want to obtain the minimal equivalent formula, we cannot randomize the order of the sizes of formulae that we try.

4. Conclusion and Future Work

In this paper, we have analyzed the practical use of three algorithms for general Boolean formula minimization: a simple one that proves that the problem is in Σ_2^p, one based on the use of a SAT solver to check formula equivalences, and one that uses a Tseitin

encoding of a formula's scheme and a QBF solver. We show that the third one clearly outperforms the other two. The use of QBF solvers represents thus the state-of-the-art for the Boolean minimization problem.

We have done the experimentation using random formulae. We plan to expand the work using formulae that are not randomly generated, e.g., deriving from natural language, or realistic random formulae [27,28,29]. Moreover, our experiments have been limited to Boolean formulae. A natural extension of this work would be to see if this or similar methods could scale up to other (more expressive) formalisms, e.g., first-order logic (FOL). This would open up a range of interesting research questions, as in FOL, equivalence is undecidable. Adapting the QBF approach would probably not be feasible, yet, a semi-brute force approach, e.g., using a first-order theorem prover, could prove successful.

References

[1] Quine WV. The Problem of Simplifying Truth Functions. The American Mathematical Monthly. 1952;59(8):521-31.

[2] Quine WV. A Way to Simplify Truth Functions. The American Mathematical Monthly. 1955;62(9):627-31.

[3] McCluskey EJ. Minimization of Boolean functions. The Bell System Technical Journal. 1956;35(6):1417-44.

[4] Karnaugh M. The map method for synthesis of combinational logic circuits. Transactions of the American Institute of Electrical Engineers, Part I: Communication and Electronics. 1953;72(5):593-9.

[5] Petrick SR. A direct determination of the irredundant forms of a Boolean function from the set of prime implicants. Air Force Cambridge Res Center Tech Report. 1956:56-110.

[6] Brayton RK, Hachtel GD, Hemachandra LA, Newton AR, Sangiovanni-Vincentelli ALM. A comparison of logic minimization strategies using ESPRESSO: An APL program package for partitioned logic minimization. In: Proceedings of the International Symposium on Circuits and Systems; 1982. p. 42-8.

[7] Seidl M, Lonsing F, Biere A. qbf2epr: A Tool for Generating EPR Formulas from QBF. In: Fontaine P, Schmidt RA, Schulz S, editors. PAAR-2012. Third Workshop on Practical Aspects of Automated Reasoning. vol. 21 of EPiC Series in Computing; 2013. p. 139-48.

[8] Rector A, Drummond N, Horridge M, Rogers J, Knublauch H, Stevens R, et al. OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In: Engineering Knowledge in the Age of the Semantic Web: 14th International Conference, EKAW 2004, Whittlebury Hall, UK, October 5-8, 2004. Proceedings 14. Springer; 2004. p. 63-81.

[9] Reiter E, Dale R. Building Natural Language Generation Systems. Studies in Natural Language Processing. Cambridge University Press; 2000.

[10] Gatt A, Krahmer E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research. 2018;61:65-170.

[11] Ranta A. Translating between language and logic: what is easy and what is difficult. In: Proceedings of the International Conference on Automated Deduction. Springer; 2011. p. 5-25.

[12] Calò E, van der Werf E, Gatt A, van Deemter K. Enhancing and Evaluating the Grammatical Framework Approach to Logic-to-Text Generation. In: Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics; 2022. p. 148-71.

[13] Grice HP. Logic and conversation. In: Speech acts. Brill; 1975. p. 41-58.

[14] Calò E, Levy J, Gatt A, Van Deemter K. Is Shortest Always Best? The Role of Brevity in Logic-to-Text Generation. In: Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023). Toronto, Canada: Association for Computational Linguistics; 2023. p. 180-92. Available from: https://aclanthology.org/2023.starsem-1.17.

[15] Garey MR, Johnson DS. Computers and intractability. vol. 174. freeman San Francisco; 1979.

[16] Umans C, Villa T, Sangiovanni-Vincentelli AL. Complexity of two-level logic minimization. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. 2006;25(7):1230-46.

[17] Buchfuhrer D, Umans C. The complexity of Boolean formula minimization. Journal of Computer and System Sciences. 2011;77(1):142-53.
[18] Turton BCH. Extending Quine-McCluskey for Exclusive-Or logic synthesis. IEEE Transactions on Education. 1996 Feb;39(1):81-5.
[19] Beyersdorff O, Janota M, Lonsing F, Seidl M. Quantified Boolean Formulas. In: Biere A, Heule M, van Maaren H, Walsh T, editors. Handbook of Satisfiability - Second Edition. vol. 336 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2021. p. 1177-221.
[20] Giunchiglia E, Marin P, Narizzano M. Reasoning with quantified boolean formulas. In: Handbook of satisfiability. IOS Press; 2009. p. 761-80.
[21] Cashmore M, Fox M. Planning as QBF. International Conference on Automated Planning and Scheduling Doctoral Consortium (ICAPS 2010). 2010.
[22] Diptarama RY, Shinohara A. QBF encoding of generalized tic-tac-toe. In: 4th International Workshop on Quantified Boolean Formulas (QBF) Co-located with 19th International Conference on Theory and Applications of Satisfiability Testing (SAT), Bordeaux, France; 2016. p. 14-26.
[23] Shukla A, Biere A, Pulina L, Seidl M. A Survey on Applications of Quantified Boolean Formulas. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). Portland, OR, USA: IEEE; 2019. p. 78-84.
[24] Audemard G, Simon L. On the glucose SAT solver. International Journal on Artificial Intelligence Tools. 2018;27(01):1840001.
[25] Rabe MN, Tentrup L. CAQE: A Certifying QBF Solver. In: 2015 Formal Methods in Computer-Aided Design (FMCAD); 2015. p. 136-43.
[26] Tentrup L. CAQE and QuAbS: Abstraction Based QBF Solvers. Journal on Satisfiability, Boolean Modeling and Computation. 2019 Sep;11:155-210.
[27] Ansótegui C, Bonet ML, Levy J. Random SAT Instances à la Carte. In: Artificial Intelligence Research and Development, Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence, CCIA 2008. vol. 184 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2008. p. 109-17. Available from: https://doi.org/10.3233/978-1-58603-925-7-109.
[28] Ansótegui C, Bonet ML, Levy J. Towards Industrial-Like Random SAT Instances. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI 2009; 2009. p. 387-92. Available from: http://ijcai.org/Proceedings/09/Papers/072.pdf.
[29] Ansótegui C, Bonet ML, Levy J. Scale-Free Random SAT Instances. Algorithms. 2022;15(6):219. Available from: https://doi.org/10.3390/a15060219.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230706

A Complete Tableau Calculus for the Regular MaxSAT Problem

Jordi COLL, [a] Chu-Min LI, [b,c] Felip MANYÀ [a] and Elifnaz YANGIN [a]
[a] *Artificial Intelligence Research Institute (IIIA), CSIC, Bellaterra, Spain*
[b] *Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France*
[c] *MIS, Université de Picardie, Amiens, France*

Abstract. We define a tableau calculus for solving the Maximum Satisfiability problem of regular propositional logic (Regular MaxSAT). Given a multiset of regular clauses Φ, we prove that the calculus is sound in the sense that if the minimum number of contradictions derived among the branches of a completed tableau for Φ is m, then the minimum number of unsatisfied clauses in Φ is m. We also prove that it is complete in the sense that if the minimum number of unsatisfied clauses in Φ is m, then the minimum number of contradictions among the branches of any completed tableau for Φ is m. Furthermore, we describe how to extend the proposed calculus to solve Regular MaxSAT in the case where we consider weighted formulas.

Keywords. regular propositional logic, maximum satisfiability, semantic tableaux, completeness.

1. Introduction

Regular propositional logic is a multiple-valued logical formalism for knowledge representation that lies in the intersection of the areas of constraint programming, many-valued logics and annotated logic programming [6].

Regular propositional formulas are multiple-valued propositional formulas equipped with a regular sign. Given a finite truth value set N equipped with a total order $\leq$, a regular sign is a subset of N either of the form $\{j \in N \mid j \geq i\}$, denoted by $\geq i$, or $\{j \in N \mid j \leq i\}$, denoted by $\leq i$, for some $i \in N$. For simplicity, we assume that N takes rational values between 0 and 1. Given a multiple-valued propositional formula ϕ, a regular propositional formula is an expression either of the form $\geq i : \phi$ or $\leq i : \phi$. In Regular logic, a truth assignment v maps every propositional variable to a value of N. An assignment v is extended to regular propositional formulas by interpreting conjunction as the minimum function, disjunction as the maximum function and negation of a propositional formula ϕ as $1 - v(\phi)$. Then, an assignment v satisfies a regular propositional formula $\geq i : \phi$ iff $v(\phi) \geq i$, and v satisfies a regular propositional formula $\leq i : \phi$ iff $v(\phi) \leq i$.

In contrast to signed propositional formulas [6], where a sign can be any subset of N, regular propositional formulas offer advantages such as the distinction between positive and negative literals. In this paper, we assume that N is finite, but in the case that N is infinite, regular signs allow us to deal with infinite subsets of truth values.

In the area of satisfiability testing, one of the problems that has attracted more interest in recent years is the Maximum Satisfiability problem (MaxSAT) [5,15,16]. SAT is usually considered in the particular case of formulas in Conjunctive Normal Form (CNF), i.e. a problem instance consists of a conjunctions of clauses, where a clause is a disjunction of literals. Whereas SAT is the problem of deciding if there exists a truth assignment for a given Boolean CNF formula that satisfies all clauses, MaxSAT is the problem of finding a truth assignment that minimizes the number of unsatisfied clauses in a Boolean CNF formula.

In practice, SAT is used as a generic problem solving formalism for decision problems and MaxSAT for optimization problems. The development of highly competitive MaxSAT solvers (e.g. [5,14]) has allowed to apply MaxSAT to solve challenging optimization problems in various fields such as bioinformatics [22], circuit design and debugging [23], combinatorial testing [3], diagnosis [10], planning [24], scheduling [7] and team formation [21].

In this paper we focus on the MaxSAT problem for regular propositional formulas in CNF, or Regular MaxSAT. Namely, the problem is to find a truth assignment that satisfies the largest possible number of regular clauses from a given multiset. Our aim is to define a complete tableau-style proof system for Regular MaxSAT.

Our work is motivated by the fact that the logic machinery defined for Regular SAT is not valid for Regular MaxSAT. Unfortunately, the inference rules for Regular SAT are unsound in Regular MaxSAT because they preserve satisfiability but do not preserve the minimum number of unsatisfied clauses between the premises and the conclusions. As a consequence, in the Boolean case, new complete resolution and tableau-style proof systems for MaxSAT have had to be defined (see e.g. [4,8,9,11,12,13,17,18]). In the multiple-valued case, there exist also resolution and tableau-style proof systems for MaxSAT (see e.g. [1,2,19,20]). Nevertheless, specific proof systems for Regular MaxSAT have not been investigated so far.

The main contributions of the present paper can be summarized as follows:

- The definition of the first tableau-style proof system for Regular MaxSAT and the corresponding proofs of soundness and completeness.
- The extension of the proposed proof system for dealing with clauses that have an associated weight, or in other words, a sound and complete proof system for Weighted Regular MaxSAT.

The paper is structured as follows. Section 2 defines basic concepts. Section 3 defines a Regular MaxSAT tableau calculus and proves its soundness and completeness. Section 4 defines an extension of the proposed calculus to Weighted MaxSAT. Section 5 ends the paper with some concluding remarks.

2. Preliminaries

Given a countable set of propositional variables $\mathscr{V} = \{x_i\}_{i\in\mathbb{N}}$, we define a propositional clause as a set of literals connected by the disjunction ($\vee$) operator, where a literal is either a variable x_i or its negation $\neg x_i$. Let $N = \{0, \frac{1}{n-1}, \ldots, \frac{n-1}{n-1}\}$ be a finite set of truth values. The *$|N|$-valued propositional logic* considered in this paper is established by a semantics comprehending an assignment $v : \mathscr{V} \to N$, which is extended to literals and clauses as follows:

$$v(\neg x_i) = 1 - v(x_i);$$
$$v(l_1 \vee l_2 \vee \cdots \vee l_n) = \max\{v(l_1), v(l_2), \ldots, v(l_n)\}.$$

Given $N = \{0, \frac{1}{n-1}, \ldots, \frac{n-1}{n-1}\}$ equipped with the natural total order $\leq$ on rational numbers, a regular sign is a subset of N either of the form $\{j \in N \mid j \geq i\}$, denoted by $\geq i$, or $\{j \in N \mid j \leq i\}$, denoted by $\leq i$, for some $i \in N$. Given a clause C, a regular clause is an expression either of the form $\geq i : C$ or $\leq i : C$. An assignment v satisfies $\geq i : C$ iff $v(C) \geq i$, and satisfies $\leq i : C$ iff $v(C) \leq i$.

Given a multiset of regular clauses Φ, the Regular Maximum Satisfiability Problem, or Regular MaxSAT, is to find an assignment that minimizes the number of unsatisfied clauses in Φ.

A weighted regular clause is a pair (C, w), where C is a regular clause and w, its weight, is a positive number. Given a multiset of weighted regular clauses Φ, Weighted Regular MaxSAT is to find an assignment that minimizes the sum of weights of unsatisfied clauses in Φ.

3. A Regular MaxSAT Tableau Calculus

We define a Regular MaxSAT tableau calculus and prove its soundness and completeness. The expansion rules we define in Definition 3.2 are graphically represented in Figure 1.

Definition 3.1. *A tableau is a tree with a finite number of branches whose nodes are labelled by either a regular clause or a box ($\Box$). A box in a tableau denotes a contradiction. A branch is a maximal path in a tree, and we assume that branches have a finite number of nodes.*

Definition 3.2. *Let $\Phi = \{C_1, \ldots, C_m\}$ be a multiset of regular clauses. A tableau for Φ is constructed by a sequence of applications of the following rules:*

Initialize *We start by generating a tree with a single branch with m nodes such that each node is labelled with a clause of Φ. Such a tableau is a tableau for Φ, called the initial tableau. All its clauses are declared to be active.*

Given a tableau T for Φ and a branch b of T, the result of applying any of the following rules results in a tableau for Φ with new branches extending b. All the following rules declare the premises to be inactive in the new branches, and declare the new nodes to be active:

$\vee$-rule ($\leq$ sign): *If b contains an active regular clause $\leq i : l_1 \vee l_2 \vee \cdots \vee l_n$, where, for $1 \leq j \leq n$, l_j is a literal and $n \geq 2$, the resulting tableau contains a new (left) branch with a node below b labelled with $\Box$, and a new (right) branch with one node for every literal l_j below b labelled with $\leq i : l_j$.*

$\vee$-rule ($\geq$ sign): *If b contains an active regular clause $\geq i : l_1 \vee l_2 \vee \cdots \vee l_n$, where, for $1 \leq j \leq n$, l_j is a literal and $n \geq 2$, the resulting tableau contains, for every literal l_j, a new branch with a node below b labelled $\geq i : l_j$.*

$$\frac{\leq i : l_1 \vee l_2 \vee \cdots \vee l_n}{\begin{array}{c|c} \Box & \leq i : l_1 \\ & \leq i : l_2 \\ & \cdots \\ & \leq i : l_n \end{array}} \qquad \frac{\geq i : l_1 \vee l_2 \vee \cdots \vee l_n}{\geq i : l_1 \mid \geq i : l_2 \mid \ldots \mid \geq i : l_n}$$

provided that $n \geq 2$

$\vee$-rule

$$\frac{\leq i : \neg x}{\geq (1-i) : x} \qquad \frac{\geq i : \neg x}{\leq (1-i) : x}$$

$\neg$-rule

$$\frac{\begin{array}{c} \leq i : x \\ \geq j : x \end{array}}{\begin{array}{c|c} \Box & \Box \\ \leq i : x & \geq j : x \end{array}}$$

provided that $i < j$

$\Box$-rule

Figure 1. Tableau expansion rules for Regular MaxSAT

$\neg$-rule: *If b contains an active regular clause $\leq i : \neg x$ (resp. $\geq i : \neg x$), where x is a variable, the resulting tableau appends a node below b labelled with $\geq (1-i) : x$ (resp. $\leq (1-i) : x$).*

$\Box$-rule: *If b contains two active regular clauses $\leq i : x$ and $\geq j : x$ such that $i < j$, where x is a variable, the resulting tableau contains a new (left) branch with two nodes below b labelled with $\Box$ and $\leq i : x$, and a new (right) branch with two nodes below b labelled with $\Box$ and $\geq j : x$.*

Definition 3.3. *Let T be a tableau for a multiset of regular propositional clauses Φ. A branch b of T is saturated when no further expansion rules can be applied on b, and T is completed when all its branches are saturated. The cost of a saturated branch is the number of boxes in the branch. The cost of a completed tableau is the minimum cost among all its branches.*

We show below that the minimum number of clauses that can be unsatisfied in a multiset of regular propositional clauses Φ is m iff the cost of a completed tableau for Φ is m. Thus, the systematic construction of a completed tableau for Φ provides an exact method for solving Regular MaxSAT.

Example 3.4. *Figure 2 shows how to create a tableau, with the previous calculus, to prove that the minimum number of unsatisfied regular propositional clauses in the multiset $\Phi = \{\leq \frac{1}{3} : \neg x_3, \geq \frac{2}{3} : x_1 \vee x_2, \leq \frac{1}{3} : x_2 \vee x_3\}$ is one, assuming $N = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$. The first tableau (the leftmost tableau on top) is the initial tableau, which contains one node for each regular propositional clause in Φ. The second tableau shows the application of the $\neg$-rule to $\leq \frac{1}{3} : \neg x_3$ and the $\vee$-rule to $\geq \frac{2}{3} : x_1 \vee x_2$. The third tableau shows the application of the $\vee$-rule to $\leq \frac{1}{3} : x_3 \vee x_2$. The fourth tableau shows the application of the $\Box$-rule to $\geq \frac{2}{3} : x_3$ and $\leq \frac{1}{3} : x_3$, and to $\geq \frac{2}{3} : x_2$ and $\leq \frac{1}{3} : x_2$. Finally, the fifth tableau ap-*

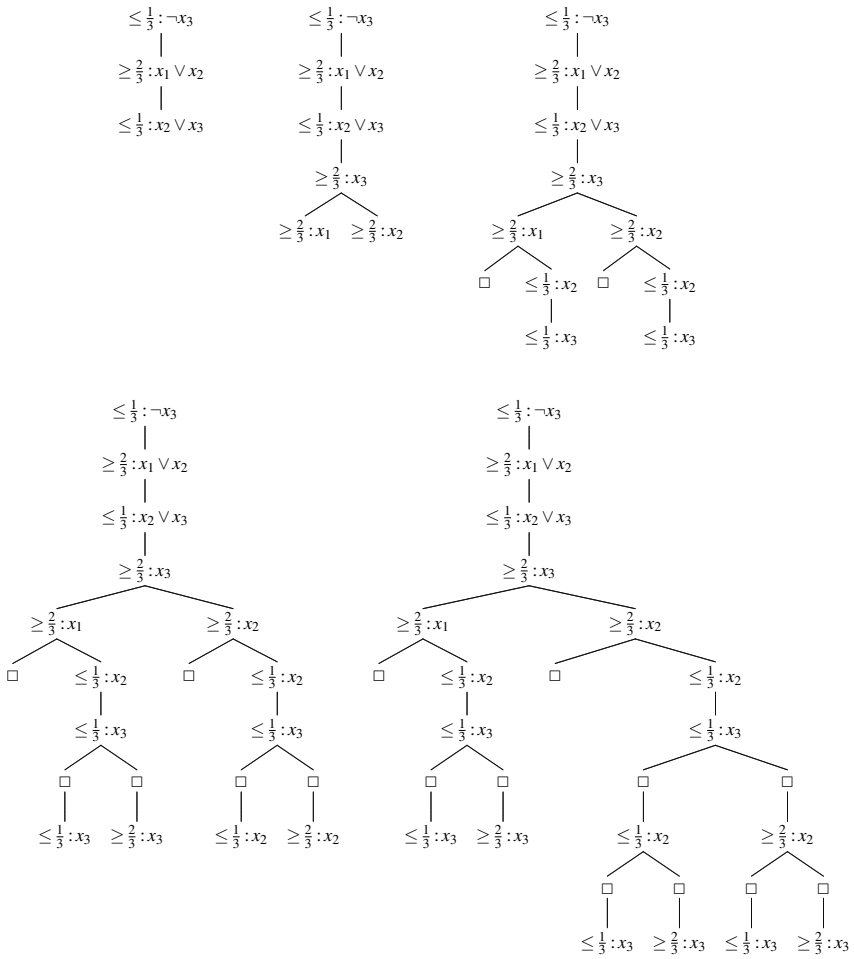


Figure 2. A tableaux for the Regular MaxSAT instance $\leq \frac{1}{3} : \neg x_3, \geq \frac{2}{3} : x_1 \vee x_2, \leq \frac{1}{3} : x_2 \vee x_3$.

plies the $\square$-rule to $\geq \frac{2}{3} : x_3$ and $\leq \frac{1}{3} : x_3$ on the two rightmost branches. Since the tableau is completed and the minimum number of boxes among its branches is one, the minimum number of regular propositional clauses that can be unsatisfied in Φ is also one.

3.1. Soundness and completeness

We prove the soundness and completeness of the proposed calculus. Before presenting the completeness theorem, we prove termination and the soundness of the expansion rules.

Proposition 3.5. *A tableau for a multiset of regular propositional clauses Φ is completed in a finite number of steps.*

Proof. We first create an initial tableau and then apply expansion rules in the newly created branches until they become saturated. The $\vee$- and $\neg$-rule reduce the number of

connectives. Since we began with a finite number of connectives, these rules can only be applied a finite number of times. The $\Box$-rule has two regular propositional clauses as premises and each conclusion contains exactly one clause in the conclusion that could be a premise of another rule and, therefore, it can only be applied a finite number of times. Hence, the construction of any completed tableau terminates in a finite number of steps. $\Box$

Proposition 3.6. *The $\vee$-, $\neg$- and $\Box$-rule preserve the number of premises unsatisfied by an assignment v in at least one branch and do not decrease that number in the other branches (if any).*

Proof. We show that the proposition holds for each rule separately:

- **$\vee$-rule ($\leq$ sign):** If v satisfies $\leq i : l_1 \vee l_2 \vee \cdots \vee l_n$, it holds that $v(l_j) \leq i$ for all $1 \leq j \leq n$, because of the maximum function. Hence, v satisfies all the clauses in at least one of the branches of the rule, specifically in the right branch. If v unsatisfies $\leq i : l_1 \vee l_2 \vee \cdots \vee l_n$, it holds that there is at least one l_j such that $v(l_j) > i$. Hence, v unsatisfies one clause of the left branch, and at least one clause of the right branch.
- **$\vee$-rule ($\geq$ sign):** If v satisfies $\geq i : l_1 \vee l_2 \vee \cdots \vee l_n$, no branch of the rule can decrease the number of unsatsified clauses in the premise, which is 0. Moreover, there is at least one l_j such that $v(l_j) \geq i$, and therefore there is at least one branch without unsatisfied clauses. If v unsatisfies $\geq i : l_1 \vee l_2 \vee \cdots \vee l_n$, it holds that $v(l_j) < i$ for all $1 \leq j \leq n$, and therefore v unsatisfies one clause of every branch.
- **$\neg$-rule:** When the premise is $\leq i : \neg x$, v satisfies $\leq i : \neg x$ iff $v(\neg x) = 1 - v(x) \leq i$ and, therefore, v satisfies $\leq i : \neg x$ iff it satisfies $\geq (1-i) : x$. The case where the premise is $\geq i : \neg x$ holds analogously.
- **$\Box$-rule:** Since $i < j$, at least one premise is unsatisfied. If $v(x) \leq i$, exactly one premise and one clause in the left conclusion are unsatisfied, and two in the right conclusion. If $i < v(x) < j$, two premises and two clauses in each conclusion are unsatisfied. If $v(x) \geq j$, exactly one premise and one clause in the right conclusion are unsatisfied, and two in the left conclusion.

$\Box$

Theorem 3.7. Soundness & completeness. *The cost of a completed tableau T for a multiset of clauses Φ is m iff the minimum number of unsatisfied clauses in Φ is m.*

Proof. (*Soundness:*) T was derived by creating a sequence of tableaux $T_0, \ldots, T_n$ ($n \geq 0$) such that T_0 is an initial tableau for Φ, $T_n = T$, and T_i was obtained by a single application of one of the $\vee$-, $\neg$- and $\Box$-rules on an branch of T_{i-1}, for $i = 1, \ldots, n$. By Proposition 3.5, we know that such a sequence is finite. Since T has cost m, T_n contains one branch b with exactly m boxes and the rest of the branches contain at least m boxes. Moreover, the active clauses in every branch of T_n are regular literals of the form $\leq i : x$ or $\geq i : x$ such that, for each propositional variable x, the intersection of all the signs is non-empty; otherwise, we could yet apply expansion rules and T_n would not be completed. Given a saturated branch with m boxes, the assignment that sets each variable x to one value of the intersection of all the signs of active regular unit clauses containing x only unsatisfies

the m boxes. Moreover, there cannot be any assignment satisfying less than m clauses in a branch of T_n, because each branch contains at least m boxes. Therefore, the minimum number of active regular propositional clauses that can be unsatisfied among the branches of T_n is m.

Proposition 3.6 guarantees that the minimum number of unsatisfied active regular clauses is preserved in the sequence of tableaux $T_0, \ldots, T_n$. Thus, the minimum number of unsatisfied regular clauses in T_0 is also m. Since T_0 is formed by a single branch that only contains the clauses in Φ and all these clauses are active, the minimum number of clauses that can be unsatisfied in Φ is m.

(*Completeness:*) Assume that there is a completed tableau T for Φ that does not have cost m. We distinguish two cases:

(i) T has a branch b of cost k, where $k < m$. Then, T has a branch with k boxes and a satisfiable multiset of active clauses because T is completed. This implies that the minimum number of unsatisfied active clauses among the branches of T is at most k. By Proposition 3.6, this also holds for T_0, but this is in contradiction with m being the minimum number of unsatisfied clauses in Φ because $k < m$. Thus, any branch of T has at least cost m.

(ii) T has no branch of cost m. This is in contradiction with m being the minimum number of unsatisfied clauses in Φ. Since the tableau expansion rules preserve the minimum number of unsatisfied clauses and the branches of any completed tableau only contain active clauses that are boxes or regular literals from which we cannot derive contradictions, T must have a saturated branch with m boxes. Thus, T has a branch of cost m.

Hence, each completed tableau T for a multiset of clauses Φ has cost m if the minimum number of clauses that can be unsatisfied in Φ is m. □

4. A Tableau Calculus for Weighted Regular MaxSAT

Figure 3 displays the expansion rules of a complete tableau calculus for Weighted Regular MaxSAT. If there is one premise (C, w), the weighted expansion rules are identical except for the fact that we propagate the weight from the premise to the conclusions. In fact, dealing with weighted regular propositional clauses can be understood as collapsing several unweighted Regular MaxSAT inferences into a single inference, because a weighted clause (C, w) can be replaced by w copies of the unweighted clause C. Therefore, the soundness and completeness proofs for unweighted Regular MaxSAT are trivially extensible to Weighted Regular MaxSAT. If there are two premises (C_1, w_1) and (C_2, w_2) with different weights ($w_1 \neq w_2$), (C_1, w_1) and (C_2, w_2) become inactive but $(C_1, w_1 - w)$ and $(C_2, w_2 - w)$, where $w = \min(w_1, w_2)$, are added as active clauses (clauses with weight 0 are not added). In this case, the conclusions of the inference have weight w. Note that, in the left branch of the □-rule, the conclusion $(\leq i{:}x, w_1)$ is the result of merging $(\leq i{:}x, w)$ and $(\leq i{:}x, w_1 - w)$, by accumulating the weight $w_1 = w + w_1 - w$ (and the analogous case occurs in the right branch).

Example 4.1. *Figure 4 displays a completed tableau for the Weighted Partial Regular MaxSAT instance* $\{(\geq 1 : x_1 \vee x_3, 3), (\geq 1 : \neg x_3, 2), (\leq \frac{1}{3} : x_1 \vee x_2, 5)\}$*, assuming* $N = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$*. The first tableau (top left) is the initial tableau. The second tableau results*

$$\frac{(\leq i : l_1 \vee l_2 \vee \cdots \vee l_n, w)}{(\Box, w) \mid \begin{array}{c}(\leq i : l_1, w)\\ (\leq i : l_2, w)\\ \cdots\\ (\leq i : l_n, w)\end{array}} \qquad \frac{(\geq i : l_1 \vee l_2 \vee \cdots \vee l_n, w)}{(\geq i : l_1, w) \mid (\geq i : l_2, w) \mid \ldots \mid (\geq i : l_n, w)}$$

provided that $n \geq 2$

$\vee$-rule

$$\frac{(\leq i : \neg x, w)}{(\geq (1-i) : x, w)} \qquad \frac{(\geq i : \neg x, w)}{(\leq (1-i) : x, w)}$$

$\neg$-rule

$$\frac{\begin{array}{c}(\leq i : x, w_1)\\ (\geq j : x, w_2)\end{array}}{\begin{array}{c|c}(\Box, w) & (\Box, w)\\ (\leq i : x, w_1) & (\geq j : x, w_2)\\ (\geq j : x, w_2 - w) & (\leq i : x, w_1 - w)\end{array}}$$

provided that $i < j$

where $w = \min(w_1, w_2)$

$\Box$-rule

Figure 3. Tableau expansion rules for Weighted Regular MaxSAT

of the application of the $\vee$-rule to $(\geq 1 : x_1 \vee x_3, 3)$, the $\neg$-rule to $(\geq 1 : \neg x_3, 2)$, and the $\vee$-rule to $(\geq 1 : x_1 \vee x_3, 3)$. The third tableau results of the application of the $\Box$-rule to $(\leq \frac{1}{3} : x_1, 5)$ and $(\geq 1 : x_1, 3)$. The last tableau results of the application of the $\Box$-rule to $(\geq 1 : x_3, 3)$ and $(\leq 0 : x_3, 2)$ on both branches of the right subtree. Since the minimum cost among all the branches is 10, the minimum sum of weights of the unsatisfied clauses is 10.

5. Conclusions

We presented a complete tableau calculus for Regular MaxSAT, proved its soundness and completeness and defined its extension to Weighted Regular MaxSAT. As future work, we plan to extend the calculus to other formalisms that may be better suited to encode specific optimization problems. For instance, in Weighted Partial Regular MaxSAT, we have hard and weighted soft clauses; in Regular MinSAT, the goal is to maximize the number of unsatisfied clauses; and in first-order logic, we can make use of quantifiers and predicates.

Acknowledgements: This work has been supported by the French Agence Nationale de la Recherche, reference ANR-19-CHIA-0013-01, and grants PID2019-111544GB-C21, PID2022-139835NB-C21 and TED2021-129319B-I00 funded by MCIN/AEI/10.13039/501100011033.

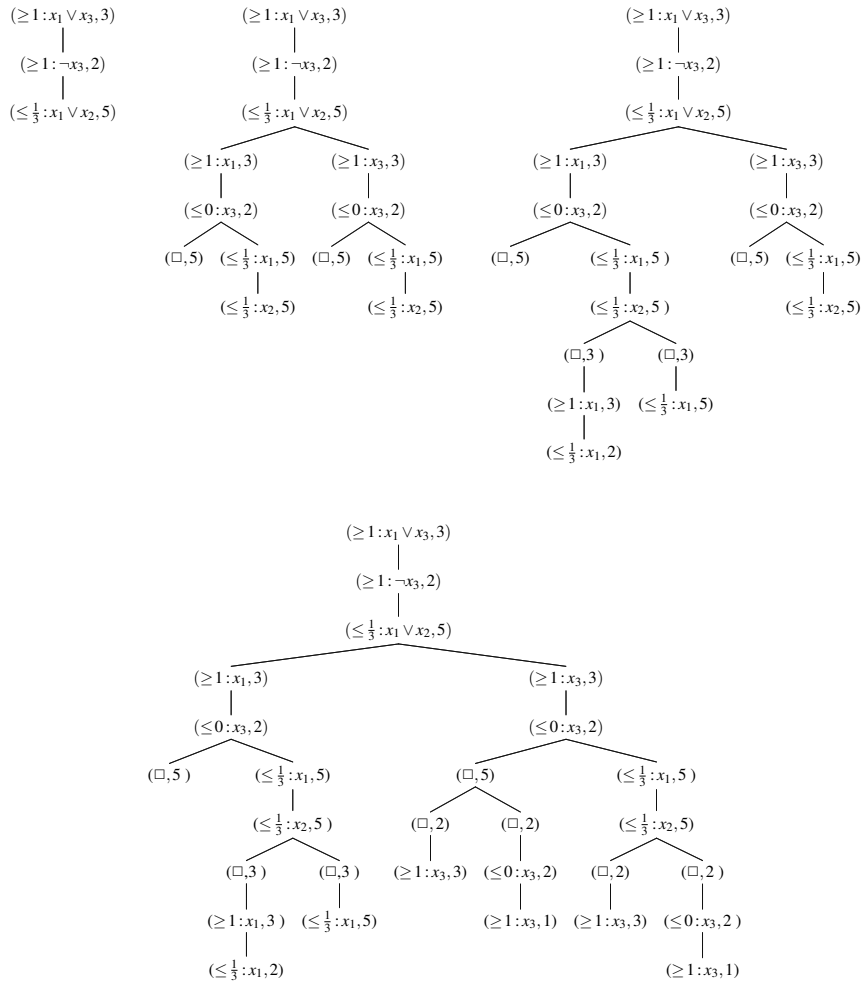


Figure 4. A tableau for the Regular MaxSAT instance $\{(\geq 1 : x_1 \vee x_3, 3), (\leq \frac{1}{3} : x_1 \vee x_2, 5), (\geq 1 : \neg x_3, 2)\}$.

References

[1] C. Ansótegui, M. L. Bonet, J. Levy, and F. Manyà. The logic behind weighted CSP. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI-2007, Hyderabad, India*, pages 32–37, 2007.

[2] C. Ansótegui, M. L. Bonet, J. Levy, and F. Manyà. Resolution procedures for multiple-valued optimization. *Information Sciences*, 227:43–59, 2013.

[3] C. Ansótegui, F. Manyà, J. Ojeda, J. M. Salvia, and E. Torres. Incomplete MaxSAT approaches for combinatorial testing. *Journal of Heuristics*, 28(4):377–431, 2022.

[4] J. Argelich, C. M. Li, F. Manyà, and J. R. Soler. Clause tableaux for maximum and minimum satisfiability. *Logic Journal of the IGPL*, 29(1):7–27, 2021.

[5] F. Bacchus, M. Järvisalo, and M. Ruben. Maximum satisfiability. In A. Biere, M. Heule, H. van Maaren, and T. Walsh, editors, *Handbook of Satisfiability*, pages 929–991. IOS Press, 2021.

[6] B. Beckert, R. Hähnle, and F. Manyà. The SAT problem of signed CNF formulas. In D. Basin, M. D'Agostino, D. Gabbay, S. Matthews, and L. Viganò, editors, *Labelled Deduction*, volume 17 of

Applied Logic Series, pages 61–82. Kluwer, Dordrecht, 2000.

[7] M. Bofill, J. Coll, M. Garcia, J. Giráldez-Cru, G. Pesant, J. Suy, and M. Villaret. Constraint solving approaches to the business-to-business meeting scheduling problem. *Journal of Artificial Intelligence Research*, 74:263–301, 2022.

[8] M. L. Bonet, J. Levy, and F. Manyà. A complete calculus for Max-SAT. In *Proceedings of the 9th International Conference on Theory and Applications of Satisfiability Testing, SAT-2006, Seattle, USA*, pages 240–251. Springer LNCS 4121, 2006.

[9] M. L. Bonet, J. Levy, and F. Manyà. Resolution for Max-SAT. *Artificial Intelligence*, 171(8–9):240–251, 2007.

[10] D. D'Almeida and É. Grégoire. Model-based diagnosis with default information implemented through MAX-SAT technology. In *Proceedings of the IEEE 13th International Conference on Information Reuse & Integration, IRI, Las Vegas, NV, USA*, pages 33–36, 2012.

[11] G. Fiorino. New tableau characterizations for non-clausal MaxSAT problem. *Logic Journal of the IGPL*, 30(3):422–436, 2022.

[12] J. Larrosa and F. Heras. Resolution in Max-SAT and its relation to local consistency in weighted CSPs. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-2005, Edinburgh, Scotland*, pages 193–198. Morgan Kaufmann, 2005.

[13] J. Larrosa, F. Heras, and S. de Givry. A logical approach to efficient Max-SAT solving. *Artificial Intelligence*, 172(2–3):204–233, 2008.

[14] C. Li, Z. Xu, J. Coll, F. Manyà, D. Habet, and K. He. Combining clause learning and branch and bound for MaxSAT. In *Proceedings of the 27th International Conference on Principles and Practice of Constraint Programming, CP, Montpellier, France*, volume 210 of *LIPIcs*, pages 38:1–38:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[15] C. M. Li and F. Manyà. MaxSAT, hard and soft constraints. In A. Biere, M. Heule, H. van Maaren, and T. Walsh, editors, *Handbook of Satisfiability*, pages 903–927. IOS Press, 2021.

[16] C. M. Li and F. Manyà. Inference in MaxSAT and MinSAT. In W. Ahrendt, B. Beckert, R. Bubel, and E. B. Johnsen, editors, *The Logic of Software. A Tasting Menu of Formal Methods - Essays Dedicated to Reiner Hähnle on the Occasion of His 60th Birthday*, volume 13360 of *Lecture Notes in Computer Science*, pages 350–369. Springer, 2022.

[17] C. M. Li, F. Manyà, and J. R. Soler. A clause tableaux calculus for MaxSAT. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI-2016, New York, USA*, pages 766–772, 2016.

[18] C. M. Li, F. Manyà, and J. R. Soler. A tableau calculus for non-clausal maximum satisfiability. In *Proceedings of the 28th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods, TABLEAUX, London, UK*, pages 58–73. Springer LNCS 11714, 2019.

[19] C. M. Li, F. Manyà, and A. Vidal. Tableaux for maximum satisfiability in Łukasiewicz logic. In *Proceedings of the 50th IEEE International Symposium on Multiple-Valued Logic, ISMVL, Miyazaki, Japan*, pages 243–248, 2020.

[20] S. Li, J. Coll, D. Habet, C. M. Li, and F. Manyà. A tableau calculus for signed maximum satisfiability. In *Proceedings of the 53rd IEEE International Symposium on Multiple-Valued Logics (ISMVL), Matsue, Japan*, pages 170–175, 2023.

[21] F. Manyà, S. Negrete, C. Roig, and J. R. Soler. Solving the team composition problem in a classroom. *Fundamamenta Informaticae*, 174(1):83–101, 2020.

[22] J. Marques-Silva, J. Argelich, A. Graça, and I. Lynce. Boolean lexicographic optimization: algorithms & applications. *Annals of Matematics and Artificial Intelligence*, 62(3-4):317–343, 2011.

[23] S. Safarpour, H. Mangassarian, A. G. Veneris, M. H. Liffiton, and K. A. Sakallah. Improved design debugging using maximum satisfiability. In *Proceedings of 7th International Conference on Formal Methods in Computer-Aided Design, FMCAD, Austin, Texas, USA*, pages 13–19, 2007.

[24] L. Zhang and F. Bacchus. MAXSAT heuristics for cost optimal planning. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Ontario, Canada*, pages 1846–1852, 2012.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230707

On the Density of States of Boolean Formulas [1]

Carlos ANSÓTEGUI [a] María Luisa BONET [b] Jordi LEVY [c]
[a] *Universitat de Lleida (DIEI, UdL), Lleida, Spain*
[b] *Universitat Politècnica de Catalunya (CS, UPC), Barcelona, Spain*
[c] *Artificial Intelligence Research Institute (IIIA, CSIC), Bellaterra, Spain*
ORCiD ID: Carlos Ansótegui https://orcid.org/0000-0001-7727-2766, María Luisa Bonet https://orcid.org/0000-0003-1646-7177, Jordi Levy https://orcid.org/0000-0001-5883-5746

Abstract. Given a CNF formula with m clauses, and an integer k, where $0 \leq k \leq m$, a *density of states* procedure will count the number of truth assignments that falsify exactly k clauses of the formula. We present the first approach to compute the density of states of Boolean formulas exactly. This is the first non-trivial result on this known hard problem. There are previous approaches for computing approximately the density of states of Boolean Formulas [1,2], where the authors point out they are not aware of any complete solver that is able to compute the exact density of states. The present work is the first step to fill this gap.

The idea is, given a formula φ and a parameter c, construct a formula φ_c such that, the number of models of φ_c is the number of assignments that falsify exactly c clauses of φ. Then, a #SAT solver is used as a black box to count the models of φ_c. This approach can be also used to compute approximately the density of states by using an approximate #SAT solver. Finally, the method can be extended trivially to deal with Weighted Boolean formulas.

Keywords. SAT, #SAT, Model Counting, Density of States

1. Introduction

Given a Boolean formula in CNF with n variables and m clauses, we can calculate the number of truth assignments that falsify exactly k clauses, where k is any number between 0 and m. Therefore, given the set of 2^n possible truth assignments, we can partition this set into $m+1$ parts. Computing the size of all subsets in the partition is the problem known as *density of states*, a term borrowed from statistical physics.

The problem of the density of states is a generalization of *model counting* for Boolean formulas (abbreviated #SAT). A model counting procedure outputs the number of assignments that satisfy a given CNF formula. There are several procedures, that either count the number of models exactly (CDP [3],Relsat [4], Cachet [5], sharpSAT [6] and c2d [7]), or simply approximate the number [8,9,10,11,12]. See [13] for a complete

[1]This research is part of the projects PROOFS (PID2019-109137GB-C21) and PROOFS BEYOND (PID2022-138506NB-C21) funded by the AEI and MCIN.

presentation on model counting (exact or approximate) for SAT. #SAT is a harder problem than assessing satisfiability (SAT), and in fact, it is #P-Complete. As a consequence, the software solving this problem is orders of magnitude slower than a SAT solver. There are a number of applications of model counting in areas like planning or probabilistic reasoning. For instance, performing Bayesian inference can be translated into model counting [14,15,16].

The density of states is an even more ambitious problem. As we said, for any number k (between 0 and the total number of clauses), we want to know how many assignments falsify exactly k clauses. For $k = 0$, a procedure solving the density of states will give us the number of satisfying assignments. On the other hand, a density of states procedure can also be used to find the minimal k such that the number of assignments falsifying k clauses is not zero, hence solving MaxSAT. Therefore, solving the density of states subsumes both #SAT and MaxSAT.

As a generalization of model counting, density of states has applications in areas like planning and probabilistic reasoning. Also, a detailed study of the search space of all truth assignments can provide insight into the development of new algorithms (for SAT or MaxSAT), as well as a better understanding of current approaches. For instance, [17] study local search dynamics using density of states. They observe that independently of the starting point, local search executions eventually stagnate around particular cost (number of falsified clauses) intervals. These intervals correspond to costs where the formula has many assignments falsifying that number of clauses.

There isn't much previous work developed around density of states. In fact, we are the first to propose a way to compute density of states exactly, and the method can be easily adapted to an approximate method. The rest of the literature presents methods to approximate density of states [17,1,2] using different sampling schemes. These estimation methods can be faster than our exact computation method, but many times it is hard to asses the accuracy of the estimates. Therefore our method can also be used to evaluate the performance of the estimates produced by other algorithms.

The key idea behind our work is the following. Given a CNF formula φ and a parameter c, we obtain a formula φ_c such that the number of assignments that satisfy φ_c is the same as the number of assignments that falsify exactly c clauses of φ. After that, we use a model counter on φ_c to calculate (or estimate) the number of satisfying assignments. The method can be extended trivially to deal with weighted Boolean formulas. Therefore to state the method in full generality we will describe it using weighted Boolean formulas.

This method requires encoding cardinality or pseudo-Boolean constraints as CNF formulas, possibly introducing additional variables, with the property that once we instantiate the original variables, the remaining formula has either 0 or 1 models. The well-known Tseitin encoding of a circuit fulfills this property. However, several state-of-the-art encodings, mostly for efficiency issues, do not satisfy it. We revisit state-of-the-art encodings of cardinality constraints and we present them guaranteeing that once we instantiate the original variables, the remaining formula has either 0 or 1 models. For efficiency reasons, our encodings should also have the property of arc-consistency. We are not the first in studying encodings with these two properties. To be able to do model counting for Integer Linear Programming, [18] study encodings with the same properties.

We conduct an experimental investigation on a selection of industrial Partial MaxSAT instances. The #SAT solvers we use are: the exact model counter sharpSAT [6], and the approximate model counters satss [8] and sampleCount [19].

The paper proceeds as follows. After some preliminary definitions in Section 2, we describe a simple method for computing the density of states based on successive calls to a #SAT solver in Section 3. This method requires encoding cardinality and pseudo-Boolean constraints as CNF formulas without introducing redundant models. In Section 4, we explain how this can be done for several state-of-the-art encodings. In Section 5, we present some experimental results on industrial instances used in the MaxSAT evaluation. Finally, in Section 6, we show how to compute a good approximation of the density of states function by interpolation.

2. Preliminaries

We consider an infinite countable set of *Boolean variables* $\mathscr{X}$. A *literal* l is either a variable $x_i \in \mathscr{X}$ or its negation $\overline{x_i}$. A *clause* C is a finite set of literals, denoted as $C = l_1 \vee \cdots \vee l_r$, or as $\square$ for the empty clause. A *(CNF) formula* φ is a finite set of clauses.

A *weighted clause* is a pair (C,w), where C is a clause and w is a natural number or infinity, indicating the penalty for falsifying the clause C. A clause is *hard* if the corresponding weight is infinity, otherwise, the clause is *soft*. A *weighted formula* is a multiset of weighted clauses $\varphi = \{(C_1,w_1),\ldots,(C_m,w_m),(C_{m+1},\infty),\ldots,(C_{m+m'},\infty)\}$ where the first m clauses are soft and the last m' clauses are hard.

The set of variables occurring in a (weighted) formula φ is noted as $\mathrm{var}(\varphi)$.

A *(total) truth assignment* for a formula φ is a function $I : \mathrm{var}(\varphi) \to \{0,1\}$, that can be extended to literals, clauses, and formulas, in the standard way. This definition of total assignment can be extended to partial assignments $I : V \to \{0,1\}$. Then, $I(\varphi)$ is another formula with variables in $\mathrm{var}(\varphi) \setminus V$.

For a weighted formula $\varphi = \{(C_1,w_1),\ldots,(C_m,w_m),(C_{m+1},\infty),\ldots,(C_{m+m'},\infty)\}$ and truth assignment $I : \mathrm{var}(\varphi) \to \{0,1\}$, we define the *cost* as $\mathrm{cost}(\varphi,I) = \sum_{i=1}^{m} w_i(1 - I(C_i))$, if I satisfies all hard clauses, i.e. $I(\{C_{m+1},\ldots,C_{m+m'}\}) = 1$. Otherwise, we say that the formula is unsatisfiable, and $\mathrm{cost}(\varphi,I) = \infty$.

We define the *optimal cost* of φ as $\mathrm{cost}(\varphi) = \min\{\mathrm{cost}(\varphi,I) \mid I : \mathrm{var}(\varphi) \to \{0,1\}\}$ and an *optimal assignment* of φ as any assignment I such that $\mathrm{cost}(\varphi) = \mathrm{cost}(\varphi,I) \neq \infty$.

We define *MaxSAT* as the problem of finding a truth assignment that maximizes the number of satisfied clauses of a CNF formula. We define *Weighted Partial MaxSAT* as the problem of finding a truth assignment that satisfies all the hard clauses and minimizes the cost of a weighted formula, i.e. finding an optimal assignment. Similarly, we define *MinSAT* as the problem of finding a truth assignment that minimizes the number of satisfied clauses, and *Weighted Partial MinSAT* as the problem of finding a truth assignment I that satisfies all the hard clauses and maximizes the number of falsified clauses.

A *cardinality constraint* is a restriction of the form $\sum_{i=1}^{m} x_i \leq k$, where $k \in \mathbb{Z}$ and variables x_i may get values in $x_i \in \{0,1\}$. A *pseudo-Boolean constraint* is a restriction of the form $\sum_{i=1}^{m} w_i x_i \leq k$, where $w_i, k \in \mathbb{Z}$ and variables x_i also may get values in $x_i \in \{0,1\}$.

3. Density of States via Model Counting

In this section, we describe our basic approach to exactly compute the density of states of a Boolean formula. For the sake of clarity, we will first describe how to solve the

Weighted Partial MaxSAT problem. The density of states approach will be based on these ideas.

Given a weighted formula $\varphi = \{(C_1, w_1), \ldots, (C_m, w_m), (C_{m+1}, \infty), \ldots, (C_{m+m'}, \infty)\}$, first, we check that all hard clauses are satisfiable. Then, we generate the formula $\varphi_k = \{C_1 \vee b_1, \ldots, C_m \vee b_m, C_{m+1}, \ldots, C_{m+m'}\} \cup CNF\left(\sum_{i=1}^{m} w_i b_i \leq k\right)$, for any possible value of $k \in \{0..\sum_{i=1}^{m} w_i\}$ and find the smallest value of k such that φ_k is satisfiable. This solves the Weighted Partial MaxSAT problem, using the fact:

$$\text{cost}(\varphi) = \min\{k \mid \varphi_k \text{ is SAT and } \varphi_{k-1} \text{ is UNSAT}\}.$$

The search for such k can be done using a dichotomy search and calling a SAT solver. A more efficient way would be to use modern SAT-based state-of-the-art MaxSAT solvers [20].

Now we are ready to state the more general problem. Counting the number of models with a cost equal to a given k can be defined as computing the following function.

Definition 1 *Given a weighted formula* $\varphi = \{(C_1, w_1), \ldots, (C_m, w_m), (C_{m+1}, \infty), \ldots, (C_{m+m'}, \infty)\}$ *and a value* $c \in \mathbb{N} \cup \{\infty\}$ *we define*

$$\text{dens}(\varphi, c) = \left|\{I : \text{var}(\varphi) \to \{0,1\} \mid \text{cost}(\varphi, I) = c\}\right|$$

Computing the density of states consists in obtaining $\text{dens}(\varphi, c)$, for any possible value of $c \in \{0, \ldots, \sum_{i=1}^{m} w_i\}$. In what follows we will describe how we can obtain each value $\text{dens}(\varphi, c)$.

First, we generate the formula

$$\varphi_c = CNF\left(\left\{C_1 \leftrightarrow \overline{b}_1, \ldots, C_m \leftrightarrow \overline{b}_m, C_{m+1}, \ldots, C_{m+m'}, \sum_{i=1}^{m} w_i b_i = c\right\}\right) \tag{1}$$

where given a formula ψ, $CNF(\psi)$ is a translation of ψ into Conjunctive Normal Form. Here we need to take into account that such translations usually use additional variables, and they will have to fulfill certain properties.

Second, we use a SAT model counter to obtain the number of models φ_c has. The idea behind these two steps is that:

$$\text{dens}(\varphi, c) = \#SAT(\varphi_c), \qquad \text{for any } c \in \mathbb{N} \tag{2}$$

To justify the correctness of equation (2), we need to see that any interpretation I, such that $I(\varphi) = c$, can be extended to a unique interpretation of the b_i variables, and can also be extended to a unique interpretation of the additional variables of the CNF encoding that satisfies φ_c; also, we need to see that any interpretation that satisfies φ_c, can be restricted to an interpretation of the original variables of φ, say I', such that $I'(\varphi) = c$. The second part of the statement is clearly true, but to ensure the first part, we need a CNF encoding of the pseudo-Boolean constraints that satisfies the *counting safety* property. The notion of counting safe encodings was first defined in [18] in the context of counting models in integer domains. They propose to encode pseudo-Boolean and integer linear programs into SAT, and then to use #SAT solvers, as an effective method for counting models in these domains. Their *counting safety* definition stated that the pseudo-Boolean

constraint and the Boolean encoding had to have the same number of models. Here we strengthen the definition.

Definition 2 *We say that a CNF encoding of a pseudo-Boolean (or cardinality) constraint*[2] $CNF(\sum_{i=1}^{m} w_i x_i \leq k)$ *is* **counting safe** *if, for any assignment* $I : \{x_1, \ldots, x_m\} \to \{0,1\}$*, the formula* $I(CNF(\sum_{i=1}^{m} w_i x_i \leq k))$ *has either* **zero** *models (when* $\sum_{i=1}^{m} w_i I(x_i) \not\leq k$*) or just* **one** *model (when* $\sum_{i=1}^{m} w_i I(x_i) \leq k$*).*

Notice that this property is not trivial since state-of-the-art CNF encodings of pseudo-Boolean constraints usually contain additional auxiliary variables. In order to make encodings more efficient to calculate, the counting safety property is dropped in many of them. Therefore, in many encodings, the property that is preserved is that if under an assignment the constraint is true, the encoding is satisfiable, and vice versa. This is true about the encodings of [21,22,23,22,24,25]. There are a few encodings that preserve counting safety (see [18] for details on two of them).

In order to obtain an efficient encoding we are also interested in other properties like having few new clauses, few additional variables, and also that, simple consequences of a partial assignment applied to a pseudo-Boolean constraint can be obtained by performing unit propagation to the CNF encoding. This last property is called *arc-consistency*. In the case of pseudo-Boolean (and cardinality) constraints, the property can be defined as follows.

Definition 3 *We say that a CNF encoding* φ *of a pseudo-Boolean (or cardinality) constraint* $CNF(\sum_{i=1}^{m} w_i x_i \leq k)$ **preserves arc-consistency** *if, for any partial assignment* $I : V \to \{1\}$*, where* $V \subseteq \{x_1, \ldots, x_n\}$*, and variable instantiation* $I' : \{x_i\} \to 1$*, if* $I(\sum_{i=1}^{m} w_i x_i \leq k)$ *is satisfiable and* $I'(I(\sum_{i=1}^{m} w_i x_i \leq k))$ *is unsatisfiable, then* $\bar{x}_i$ *may be inferred from* $I(\varphi)$ *by unit propagation.*

In section 4, we will talk about encodings of cardinality and pseudo-Boolean constraints that have both properties: safe counting and arc-consistency. One of these encodings will be used in our implementation.

In what follows, we propose various improvements to gain efficiency. In our implementation, we solve the MaxSAT and MinSAT problems for the formula φ. This gives us two values say k_1 and k_2, respectively, such that $\text{dens}(\varphi, c) = 0$ for $c < k_1$ and for $c > k_2$. So, we restrict the computation of the density to $c \in \{k_1, \ldots, k_2\}$.

As we will see, the time needed by an exact model counter on a formula φ_c increases with the number of models. Therefore, we use an approximate model counter when the number of models of φ_c is big.

It is also possible that for a number c, $k_1 \leq c \leq k_2$, that the number of models of φ_c is equal to 0. For instance, if φ is a formula with all even weights, then $\text{dens}(\varphi, c) = 0$ for c odd. Therefore, it is desirable to check the satisfiability of φ_c in that range $k_1 \leq c \leq k_2$, using a state-of-the-art SAT solver, before calling a model counter. This is especially important if we use an approximated model counter, that may have problems when the number of models is very small.

[2] Here we only consider less-equal constraints, with the understanding that the less, greater, greater-equal and equal constraints can easily be expressed in terms of least-equal.

4. Encodings of Cardinality and Pseudo-Boolean Constraints

There are several (standard) ways to encode a cardinality constraint $\sum_{i=1}^{m} x_i \leq k$ or a pseudo-Boolean constraint $\sum_{i=1}^{m} w_i x_i \leq k$ as a CNF formula preserving arc-consistency [21,22,23,24,25]. In [18], it is proved that the encoding in [26], based on binary adders, and the [24] encoding, based on BDDs, both satisfy the counting safety property. The first one, however, does not preserve arc-consistency.

In the next subsections, we review two more encodings for cardinality constraints, one based on sequential counters, and another based on sorting networks. We also analyze how to translate BDDs into CNF formulas in order to ensure the counting safety property.

4.1. Sequential Counters

A simple approach to encoding a sequential counter is to use a variable s_i^j that gets the value true when $x_1 + \ldots + x_i = j$, and a pair of clauses of the form $s_i^j \wedge \bar{x}_{i+1} \rightarrow s_{i+1}^j$ and $s_i^j \wedge x_{i+1} \rightarrow s_{i+1}^{j+1}$ that define the value of s_i^j in terms of the values of variables with smaller indexes. Then, the encoding of the cardinality is given by the following set of clauses.

$$
\begin{array}{ll}
\bar{s}_i^j \vee x_{i+1} \vee s_{i+1}^j & \text{for } j = 0 \ldots k,\ i = j \ldots n+j-k-2 \\
\bar{s}_i^j \vee \bar{x}_{i+1} \vee s_{i+1}^{j+1} & \text{for } j = 0 \ldots k,\ i = j \ldots n+j-k-1 \\
s_0^0 & \\
\bar{s}_i^{k+1} & \text{for all } i = k+1, \ldots, n
\end{array}
$$

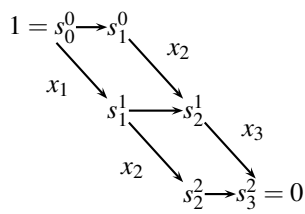
This encoding has the counting safety property, but not arc-consistency.

In order to have arc-consistency, we can implement a monotonic variant of this Boolean function. Sinz [21] proposes the use of variables s_i^j that are true if $x_1 + \ldots + x_i \geq j$, and the following set of clauses.

$$
\begin{array}{ll}
\bar{s}_i^j \vee s_{i+1}^j & \text{for } j = 0 \ldots k+1,\ i = j \ldots n+j-k-2 \\
\bar{s}_i^j \vee \bar{x}_{i+1} \vee s_{i+1}^{j+1} & \text{for } j = 0 \ldots k,\ i = j \ldots n+j-k-1 \\
s_0^0 & \\
\bar{s}_n^{k+1} &
\end{array}
$$

This formula can be simplified by unit propagation, leading to a formula with $k(n-1)$ auxiliary variables and size $\mathcal{O}(nk)$. The encoding is arc-consistent, but does not have the counting safety property, as the following counter-example shows.

Example 4 For the cardinality constraint $x_1 + x_2 + x_3 \leq 1$, the sequential counter implementation described in [21] generates the following set of clauses, where literals in red may be removed by unit propagation.



$$
\begin{array}{ll}
\bar{s}_0^0 \vee \bar{x}_1 \vee s_1^1 & s_0^0 \\
\bar{s}_1^0 \vee \bar{x}_2 \vee s_2^1 & \bar{s}_0^0 \vee s_1^0 \\
\bar{s}_1^1 \vee \bar{x}_2 \vee s_2^2 & \bar{s}_1^1 \vee s_2^1 \\
\bar{s}_2^1 \vee \bar{x}_3 \vee s_3^2 & \bar{s}_2^2 \vee s_3^2 \\
\bar{s}_3^2 &
\end{array}
$$

For $x_1 = x_2 = x_3 = 0$ the formula has three models. In all them $s_0^0 = s_1^0 = 1$ and $s_2^2 = s_3^2 = 0$, but $\langle s_1^1, s_2^1 \rangle \in \{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,1\rangle\}$.

The next lemma presents an encoding that preserves arc-consistency and has the counting safety property. Like in [21], the variable s_i^j is true when $x_1 + \ldots + x_i \geq j$.

Lemma 5 *For any cardinality constraint $x_1 + \ldots + x_n \leq k$, the following set of clauses*

$$
\begin{array}{rl}
\overline{s}_i^j \vee s_{i+1}^j & \text{for } j=0...k+1 \text{ and } i=j...n+j-k-2 \\
s_i^j \vee x_{i+1} \vee \overline{s}_{i+1}^j & \text{for } j=0...k+1 \text{ and } i=j...n+j-k-2 \\
\overline{s}_i^j \vee \overline{x}_{i+1} \vee s_{i+1}^{j+1} & \text{for } j=0...k \text{ and } i=j...n+j-k-1 \\
s_i^j \vee \overline{x}_{i+1} \vee \overline{s}_{i+1}^{j+1} & \text{for } j=0...k \text{ and } i=j...n+j-k-1 \\
\overline{s}_i^i \vee x_i & \text{for } i=1...k \\
s_0^0 & \\
\overline{s}_n^{k+1} &
\end{array}
$$

defines an equivalent formula with both the arc-consistency and the counting safety properties.

PROOF: All clauses are sound, in the sense that, any interpretation of the x_i variables satisfying the cardinality constraint can be extended to an interpretation of the s_i^j variables satisfying the formula.

The formula is an extension of the formula proposed in [21] without additional variables. Therefore, this formula is also arc-consistent because it has at least the same inference power.

Now, we will prove that it has the counting safety property, i.e. that there exists at most one model once we have instantiated the x_i variables. The proof is by induction. The value of s_0^0 is 1 in all models. For the induction step, we prove that the values of $s_{i+1}^0, \ldots, s_{i+1}^{r+1}$, for $r = \min\{i,k\}$, are determined by the values of $s_i^0, \ldots, s_i^r$, for $r = \min\{i,k+1\}$. Notice that the value of s_i^0 is 1, and the value of s_i^{k+1} is 0, in all models. There are two cases. If $x_{i+1} = 0$, then the clauses $\overline{s}_i^j \vee s_{i+1}^j$ and $s_i^j \vee x_{i+1} \vee \overline{s}_{i+1}^j$ ensure that s_i^j and s_{i+1}^j have the same value, for $j = 0, \ldots, \min\{i,k+1\}$. When $i \leq k$, the value of s_{i+1}^{i+1} is determined by the clause $\overline{s}_{i+1}^{i+1} \vee x_{i+1}$ that sets it to 0. If $x_{i+1} = 1$, then the clauses $\overline{s}_i^j \vee \overline{x}_{i+1} \vee s_{i+1}^{j+1}$ and $s_i^j \vee \overline{x}_{i+1} \vee \overline{s}_{i+1}^{j+1}$ ensure that s_{i+1}^{j+1} and s_i^j have the same value, for $j = 0, \ldots, \min\{i,k\}$. Notice that s_{i+1}^0 is always 1 and the value of s_{i+1}^{k+1}, when $i \geq k$, is 0. ■

4.2. Sorting and Cardinality Networks

Sorting networks can be used to encode cardinality constraints [22,23]. A sorting network is a circuit $\langle y_1, \ldots, y_n \rangle = SORT(\langle x_1, \ldots, x_n \rangle)$ that given a sequence, returns a sorted permutation of the input such that $y_i \geq y_j$, for all $1 \leq i < j \leq n$. A merging network is a circuit $\langle z_1, \ldots, z_{2n} \rangle = MERGE(\langle x_1, \ldots, x_n \rangle, \langle y_1, \ldots, y_n \rangle)$ that given two sorted sequences, returns a sorted permutation of the concatenation of both sequences.

We can implement a sorting network using the recurrence

$$SORT(\langle x_1,\ldots,x_{2n}\rangle) = MERGE(SORT(\langle x_1,\ldots,x_n\rangle), SORT(\langle x_{n+1},\ldots,x_{2n}\rangle))$$

We can also implement a merge-network

$$\langle z_1,\ldots,z_{4n}\rangle = MERGE(\langle x_1,\ldots,x_{2n}\rangle, \langle y_1,\ldots,y_{2n}\rangle)$$

using two merge networks (one for the even position and the other one for the odd positions)

$$\langle t_1,\ldots,t_{2n}\rangle = MERGE(\langle x_1,x_3,\ldots,x_{2n-1}\rangle, \langle y_1,y_3,\ldots,y_{2n-1}\rangle)$$
$$\langle t'_1,\ldots,t'_{2n}\rangle = MERGE(\langle x_2,x_4,\ldots,x_{2n}\rangle, \langle y_2,y_4,\ldots,y_{2n}\rangle)$$

and adding the following gates

$$z_1 = t_1$$
$$z_3 = AND(t_2,t'_1) \ldots z_{4n-1} = AND(t_{2n},t'_{2n-1})$$
$$z_2 = OR(t_2,t'_1) \quad \ldots z_{4n-2} = OR(t_{2n},t'_{2n-1})$$
$$z_{4n} = t'_{2n}$$

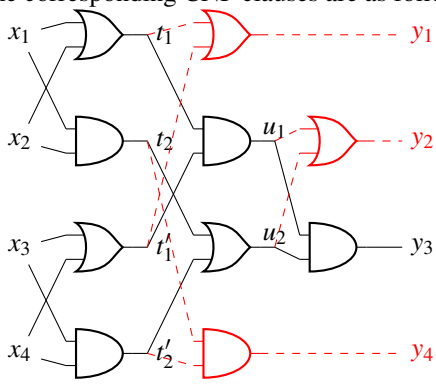
These sort and merge circuits have size $\mathcal{O}(n\log^2 n)$ and $\mathcal{O}(n\log n)$, respectively.

We can implement a circuit that checks $x_1+\ldots+x_n \le k$, implementing $\langle y_1,\ldots,y_n\rangle = SORT(\langle x_1,\ldots,x_n\rangle)$ and checking if $y_{k+1}=0$. Then, by encoding gates as clauses, we get a CNF encoding of the cardinality constraint of size $\mathcal{O}(n\log^2 n)$. This encoding also preserves arc-consistency [22].

Notice that we only check one of the outputs of the circuit. Therefore, we can remove all gates that are only needed to compute the value of the other outputs y_i for $i \neq k+1$. By iteratively removing these useless gates, we obtain an encoding of size $\mathcal{O}(n\log^2 k)$ [23]. This encoding is (asymptotically) better than the encoding based on sequential counters (where the size was $\mathcal{O}(nk)$). However, this is not significant when k is small.

Gates are encoded as clauses that only ensure that the output is bigger or equal to the circuit output. More precisely, gates $y = AND(x_1,x_2)$ are encoded as $x_1 \wedge x_2 \rightarrow y$, and gates $y = OR(x_1,x_2)$ as $\{x_1 \rightarrow y,\ x_2 \rightarrow y\}$. This makes the resulting formula not have the counting safety property.

Example 6 Consider the cardinality constraint $x_1+x_2+x_3+x_4 \le 2$. The circuit encoding the cardinality constraint is drawn below, where gates in red are unused gates (removed in [23]), and the corresponding CNF clauses are as follows.



$$x_1 \rightarrow t_1$$
$$x_2 \rightarrow t_1$$
$$x_1 \wedge x_2 \rightarrow t_2$$
$$x_3 \rightarrow t'_1$$
$$x_4 \rightarrow t'_1$$
$$x_3 \wedge x_4 \rightarrow t'_2$$
$$t_1 \wedge t'_1 \rightarrow u_1$$
$$t_2 \rightarrow u_2$$
$$t'_2 \rightarrow u_2$$
$$u_1 \wedge u_2 \rightarrow y_3$$
$$\bar{y}_3$$

For $x_1 = 1$ and $x_2 = x_3 = x_4 = 0$ (satisfying the constraint), we have six models:

t_1	t_2	t'_1	t'_2	u_1	u_2	y_3
1	0	0	0	0	0	0
1	0	0	0	0	1	0
1	0	0	0	1	0	0
1	0	0	1	0	1	0
1	0	1	0	1	0	0
1	1	0	0	0	1	0

The problem in the previous example is that the Boolean function implemented with the clauses is not the function implemented in the circuit, but a *monotone* variation. Notice that when $x_1 = x_2 = 1$, both the gate $y = AND(x_1, x_2)$ and the clause $x_1 \wedge x_2 \rightarrow y$ force $y = 1$. However, when $x_1 = 0$ and $x_2 = 1$, the gate $y = AND(x_1, x_2)$ forces $y = 0$, but the clause $x_1 \wedge x_2 \rightarrow y$ allows y to be 0 or 1. We can construct a counting safety formula by adding clauses ensuring that the gates and the clauses have the same behavior.

Lemma 7 *If we encode sorting networks, adding gates $y = AND(x_1, x_2)$ the clauses $\{x_1 \wedge x_2 \rightarrow y,\ y \rightarrow x_1,\ y \rightarrow x_2\}$, and for gates $y = OR(x_1, x_2)$ the clauses $\{x_1 \rightarrow y,\ x_2 \rightarrow y,\ y \rightarrow x_1 \vee x_2\}$, then the resulting formula has the counting safety property and the encoding preserves arc-consistency.*

PROOF: The formula is an extension of the formula proposed in [22,23], hence arc-consistent. The additional clauses force the formula to behave exactly like the circuit, where all intermediate points have a determined value. ■

4.3. Reduced Ordered BDDs

Reduced Ordered BDDs are a canonical representation of Boolean functions [27], hence of pseudo-Boolean constraints. These BDDs may also be translated into arc-consistent CNF formulas [22]. Although ROBBDs may be exponentially large [24,25], we have obtained good (practical) results using them in our MaxSAT solvers [28,29,20,30].

A BDD may be encoded in CNF [22] by introducing an auxiliary variable y_i for every decision variable x_i. Then, if the false and true child of x_i are respectively x_j and x_k, we add the following clauses:

$$\begin{array}{lll} \bar{x}_i \wedge \bar{y}_j \rightarrow \bar{y}_i & x_i \wedge \bar{y}_k \rightarrow \bar{y}_i & \bar{y}_j \wedge \bar{y}_k \rightarrow \bar{y}_i \\ \bar{x}_i \wedge y_j \rightarrow y_i & x_i \wedge y_k \rightarrow y_i & y_j \wedge y_k \rightarrow y_i \end{array}$$

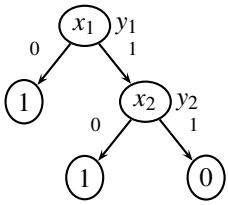
It is trivial to prove that this CNF encoding has the counting safety property, because auxiliary variables in leaves are always true or false, and the values of other auxiliary variables y_i are determined by the values of their child's auxiliary variables y_j and y_k (once we have instantiated the decision variable x_i).

In [25] it is proposed an alternative translation from monotonic BDDs to CNF that only uses two clauses per node and that is also arc-consistent. In this case, we also have an auxiliary variable per node, and the two clauses are as follows:

$$\overline{y}_j \to \overline{y}_i$$
$$x_i \wedge \overline{y}_k \to \overline{y}_i$$

In this case, the resulting formula does not have the counting safety property, as the following example shows.

Example 8

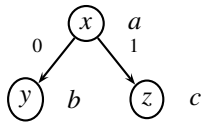


The monotonic BDD on the left, resulting from translating the pseudo-Boolean constraint $2x_1 + x_2 \leq 2$, is translated as the following CNF by [25]

$$\varphi = \{y_1,\ x_1 \wedge \overline{y}_2 \to \overline{y}_1,\ x_2 \to \overline{y}_2\}$$

For the interpretation $I(\langle x_1, x_2\rangle) = \langle 0,0\rangle$ the formula $I(\varphi)$ has two models $\langle y_1, y_2\rangle \in \{\langle 1,0\rangle,\ \langle 1,1\rangle\}$.

Lemma 9 *If we translate a pseudo-Boolean constraint into a BDD, and then, for every BDD node of the following form we add the clauses on the right,*



$$\overline{b} \to \overline{a}$$
$$x \wedge \overline{c} \to \overline{a}$$
$$\overline{x} \wedge b \to a$$
$$c \to a$$

then the resulting formula is arc-consistent and has the counting safety property.

PROOF: The formula is an extension of the one proposed in [25], where we add two new clauses for each node, hence arc-consistent. To prove that the formula has the counting safety property, assume that the decision variables in the BDD are all assigned. We will prove by induction that all auxiliary variables get a unique determined value. By induction hypothesis we have that values of decision variables (the x in the picture) and values of auxiliary variables in the sons (b and c) are determined. Now, we will prove that they determine the value of the auxiliary variable in the father (a). To prove this, notice that by construction the BDD is monotonic and $c \to b$ is always satisfied, thus the case $c = 1$ and $b = 0$ does not need to be considered. For the rest of cases, the following table gives the unique value for a allowed by the four clauses:

x	b	c	a
0	0	0	0
0	1	0	1
0	1	1	1
1	0	0	0
1	1	0	0
1	1	1	1

■

5. Experimental Results

We have conducted an experimental investigation to explore the density of states of industrial formulas and test our method. We have run our experiments on machines with 2.27GHz CPU and 1G RAM.

We have selected 10 instances from the Partial MaxSAT industrial category in the MaxSAT evaluation 2011. We have also added two pigeonhole formulas PHP_9^{10} and PHP_5^{10} where "each pigeon has to go to one hole" is a soft clause, and "two pigeons cannot go to the same hole" is a hard clause.

As we have said, our approach is parametric on any #SAT solver. Here, we used the exact #SAT solver sharpSAT [6], and the approximate #SAT solvers satss [8] and sampleCount [19]. We will refer to the resulting solvers as DOS(sharpSAT), DOS(satss) and DOS(sampleCount). We encode cardinality constraints as cardinality networks as it is described in Subsection 4.2. Initially, we compute the optimal cost of the MaxSAT (k_1) and MinSAT problems (k_2) as described in Section 3. We will only show the results for the interval $c \in [k_1, k_2]$, since for the rest of the values of c, the number of truth assignments violating c clauses is 0. In the figures, we show the number of truth assignments that violate exactly $c \in [k_1, k_2]$ soft clauses in the Partial MaxSAT formula. We do not consider models falsifying the hard clauses.

The timeout for the DOS(sharpSAT) solver is a global limit for the whole instance. We use 20 hours. Therefore, we compute the points following the sequence $c = k_1, k_2, k_1+1, k_2-1, k_1+2, \ldots$. When the cutoff is applied, we obtain an incomplete DOS where middle points (the costliest to compute) are missing. Approximate #SAT solvers may run forever, trying to improve the accuracy of the solution. In this case, we have to provide a timeout for each value c. We use 30 minutes.

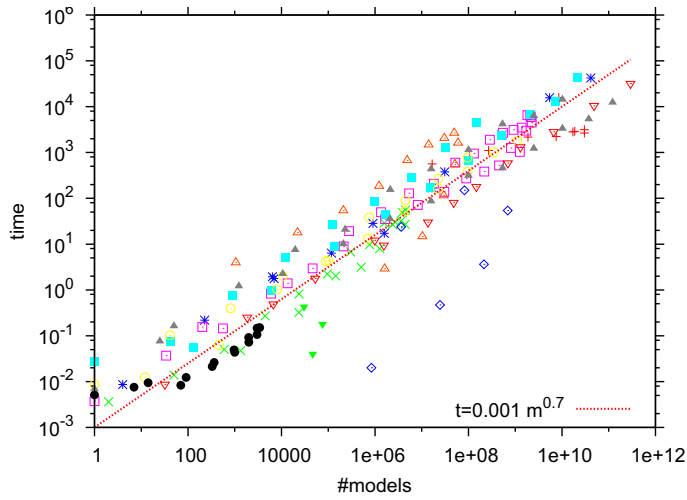


Figure 1. Time needed by sharpSAT on φ_c as a function of the number of models of φ_c for distinct industrial MaxSAT formulas φ (each family of formulas in a distinct color).

We analyze the running time for DOS(sharpSAT) on each data point. We can see that time is almost proportional to the number of assignments (see Figure 1). In fact, it seems to follow a relation $t = 0.001m^{0.7}$, where t is the time in seconds and m the number of

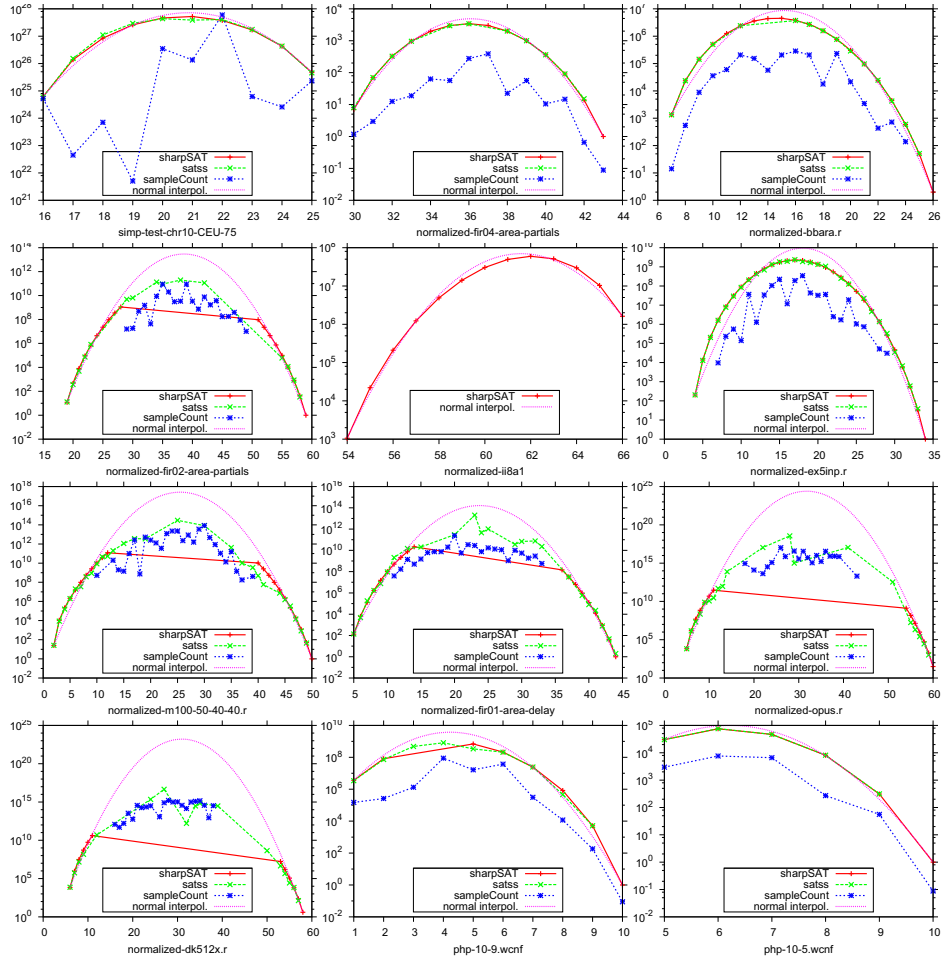


Figure 2. Density of states, $\log(\mathrm{dens}(\varphi, c))$ as a function of c, computed by DOS(sharpSAT), DOS(satss) and DOS(sampleCount) on some industrial Partial MaxSAT instances and two pigeonhole formulas. We also include a (truncated) normal distribution interpolation (parabolic interpolation in this logarithmic representation) computed from the values of $\mathrm{dens}(\varphi, k_1)$, $\mathrm{dens}(\varphi, k_2)$ and $\sum_{c=k_1}^{k_2} \mathrm{dens}(\varphi, c)$.

models. Therefore, by predicting the c's with maximum number of assignments we can also produce the hardest instances for #SAT solvers. Notice that #SAT solvers are tested on SAT instances, while with our approach we can extend the benchmarks to contain DOS queries on Partial MaxSAT instances.

Figure 2 shows the density of states $\mathrm{dens}(\varphi, c)$ computed by DOS(sharpSAT), DOS(satss) and DOS(sampleCount). As we can see, the number of truth assignments that violate c clauses follows approximately a (truncated) normal distribution, since representing the results in logarithmic scale we obtain an inverted parabolic representation. In general, approximate #SAT solvers compute approximate values smaller than the real values. Results of DOS(satss) seem more accurate than results of DOS(sampleCount). However, we have to take into account that these solvers are parametrizable. By adjusting parameters differently, we may obtain better results.

6. Approximating DOS by Interpolation

SAT-based MaxSAT solvers may be (easily) modified in order to produce a SAT formula whose models are the optimal assignments. Running a #SAT solver on this formula, we compute $\text{dens}(\varphi, k_1)$. In a similar way, we can easily compute $\text{dens}(\varphi, k_2)$. In a formula (without hard clauses), the number of possible assignments is 2^n, being n the number of variables. However, in a weighted formula, this number is smaller because we do not consider assignments violating hard clauses. We can easily compute the number of assignments satisfying hard clauses as $\sum_{c=k_1}^{k_2} \text{dens}(\varphi, c) = \#SAT(\varphi_{hard})$. These three values represent two points and the area below $\text{dens}(\varphi, c)$. They are enough to interpolate a parabola in the representation of $\log(\text{dens}(\varphi, c))$, i.e. to approximate a (truncate) normal distribution for $\text{dens}(\varphi, c)$. We represent this interpolation in Figure 2. As we can see, in our examples, and probably in most industrial MaxSAT instances, this interpolation is better than the results obtained by DOS(satss) and DOS(sampleCount). Moreover, it can be efficiently computed with one call to a #SAT solver to get $\#SAT(\varphi_{hard})$, one call to a MaxSAT solver to get k_1, and then a call to a #SAT solver to get $\text{dens}(\varphi, k_1)$, one call to a MinSAT solver to get k_2, and then a call to a #SAT solver to get $\text{dens}(\varphi, k_2)$.

7. Conclusions

In this pioneering work, we present the first method to compute *exactly* the density of states of a MaxSAT formula. The idea is to reduce the problem to a sequence of computations of #SAT. The use of an approximate #SAT solver results into an approximate DOS solver.

While there exists previous approximate DOS solvers [1,2] for MaxSAT, this work describes the first *exact* DOS solver. Moreover, it is also the first implementation able to manage *Partial* and *Weighted* MaxSAT formulas.

Finally, our work produces sets of test instances for #SAT solvers from over-constrained instances. Given a formula φ, the set of instances φ_c for $c \in \{k_1, \ldots, k_2\}$ (k_1 and k_2 are the MaxSAT and MinSAT values respectively), produce a set of benchmarks of different hardness for a model counter.

References

[1] Ermon S, Gomes CP, Selman B. Computing the Density of States of Boolean Formulas. In: CP'10. vol. 6308 of LNCS; 2010. p. 38-52.

[2] Ermon S, Gomes CP, Selman B. A Flat Histogram Method for Computing the Density of States of Combinatorial Problems. In: IJCAI'11; 2011. p. 2608-13.

[3] Birnbaum E, Lozinskii EL. The Good Old Davis-Putnam Procedure Helps Counting Models. Journal of Artificial Intelligence Research. 1999;10:457-77.

[4] Bayardo RJ, Pehoushek JD. Counting Models Using Connected Components. In: AAAI'00; 2000. p. 157-62.

[5] Sang T, Bacchus F, Beame P, Kautz HA, Pitassi T. Combining Component Caching and Clause Learning for Effective Model Counting. In: SAT'04; 2004. p. 20-8.

[6] Thurley M. sharpSAT - Counting Models with Advanced Component Caching and Implicit BCP. In: SAT'06. vol. 4121 of LNCS; 2006. p. 424-9.

[7] Darwiche A. New Advances in Compiling CNF into Decomposable Negation Normal Form. In: ECAI'04; 2004. p. 328-32.

[8] Gogate V, Dechter R. Approximate Counting by Sampling the Backtrack-free Search Space. In: AAAI'07; 2007. p. 198-203.
[9] Gomes CP, Sabharwal A, Selman B. Model Counting: A New Strategy for Obtaining Good Bounds. In: AAAI'06; 2006. p. 54-61.
[10] Kroc L, Sabharwal A, Selman B. Leveraging belief propagation, backtrack search, and statistics for model counting. Annals of Operations Research. 2011;184(1).
[11] Wei W, Erenrich J, Selman B. Towards Efficient Sampling: Exploiting Random Walk Strategies. In: AAAI'04; 2004. p. 670-6.
[12] Wei W, Selman B. A New Approach to Model Counting. In: SAT'05; 2005. p. 324-39.
[13] Gomes CP, Sabharwal A, Selman B. 20. In: Model Counting. vol. 185 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2009. p. 633-54.
[14] Bacchus F, Dalmao S, Pitassi T. Algorithms and Complexity Results for #SAT and Bayesian Inference. In: FOCS'03; 2003. p. 340-51.
[15] Sang T, Beame P, Kautz HA. Performing Bayesian Inference by Weighted Model Counting. In: AAAI'05; 2005. p. 475-82.
[16] Darwiche A. Modeling and Reasoning with Bayesian Networks. Cambridge University Press; 2009.
[17] Belaidouni M, Hao JK. SAT, Local Search Dynamics and Density of States. In: Artificial Evolution; 2001. p. 192-204.
[18] Morgado A, Matos PJ, Manquinho VM, Marques-Silva JP. Counting Models in Integer Domains. In: SAT'06; 2006. p. 410-23.
[19] Gomes CP, Hoffmann J, Sabharwal A, Selman B. From Sampling to Model Counting. In: IJCAI'07; 2007. p. 2293-9.
[20] Ansótegui C, Bonet ML, Levy J. SAT-based MaxSAT algorithms. Artificial Intelligence. 2013;196:77-105.
[21] Sinz C. Towards an Optimal CNF Encoding of Boolean Cardinality Constraints. In: CP'05. vol. 3709 of LNCS; 2005. p. 827-31.
[22] Eén N, Sörensson N. Translating Pseudo-Boolean Constraints into SAT. JSAT. 2006;2(1-4):1-26.
[23] Asín R, Nieuwenhuis R, Oliveras A, Rodríguez-Carbonell E. Cardinality Networks: a theoretical and empirical study. Constraints. 2011;16(2):195-221.
[24] Bailleux O, Boufkhad Y, Roussel O. A Translation of Pseudo Boolean Constraints to SAT. JSAT. 2006;2(1-4):191-200.
[25] Abío I, Nieuwenhuis R, Oliveras A, Rodríguez-Carbonell E. BDDs for Pseudo-Boolean Constraints - Revisited. In: SAT'11; 2011. p. 61-75.
[26] Warners JP. A Linear-Time Transformation of Linear Inequalities into Conjunctive Normal Form. Information Processing Letters. 1998;68(2):63-9.
[27] Bryant RE. Graph-Based Algorithms for Boolean Function Manipulation. IEEE Trans Computers. 1986;35(8):677-91.
[28] Ansótegui C, Bonet ML, Levy J. Solving (Weighted) Partial MaxSAT through Satisfiability Testing. In: SAT'09; 2009. p. 427-40.
[29] Ansótegui C, Bonet ML, Levy J. A New Algorithm for Weighted Partial MaxSAT. In: AAAI'10; 2010. p. 3-8.
[30] Ansótegui C, Gabàs J, Levy J. Exploiting subproblem optimization in SAT-based MaxSAT algorithms. J Heuristics. 2016;22(1):1-53.

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

Subject Index

3D orientation 90
3D printers 70
6G 7
additive kernels 143
AI 153
AI application(s) 110, 287
AI ethics 153, 221
AI for good 241
AI impact 3
AI problem solving 102
AI real-time application 132
algorithmic trading 110
allocation 173
ANOVA 311
application to biomedical data 183
artificial haptic system 48
artificial vision 98, 102
author bibliographic coupling 80
authorship 80
automotive industry 106
Bayesian networks 128
Bayesian optimization 143
benchmark 283
bibliographic coupling 80
bibliometric analysis 80
bioethics and human rights 221
black swan events 110
boolean formula minimization 347
bootstrap clustering 70
brain landmarking 94
calorimeter 277
cancer research 143
capsule endoscopy 134
car configurator 106
cascade network 38
causal inference 16
CBCT image segmentation 136
CDCL 337
characterisation 330
citation networks 80
citizen science 59
classification 204
cleansing 134
clinical decision support systems 204
clustering 70
cognitive similarity 80
colour palette 311
command understanding 283
completeness 359
conceptual model 211
consensus measurement 291
constrained optimization 143
correlation 106
cosine similarity 80
cost-effectiveness models 143
crowdsourcing 301
customers' profiles 291
data generation 128
decision support (system) 4, 70
deep learning 38, 94, 98, 124, 136, 267, 277, 311
deep spectrum features 138
density of states 369
diabetic retinopathy 204
digestive content 134
disaster response 301, 320
distributed space systems 7
domain adaptation 183
DPLL 337
drug dose titration 4
drug prescription 4
dynamic graph 273
dynamic stochastic block model 273
ECG 28
edge privacy 273
education 241, 250
embryo selection 153
emotion recognition 138
energy microgrids 120
evidence-based medicine 4
experimental data 256
expert finding 59
explainability 153
explainable AI 262
fake news 330
false alarm mitigation 283

financial trading 110
food waste 256
forecasting 193
gastronomy 311
Gaussian processes 143
generalization 48
generative adversarial networks 183
genetic algorithm 106
genetic programming 120
geolocation 301, 320
head pose estimation 90
healthcare 173
healthcare demand 193
hesitant fuzzy linguistic terms sets 291
high energy physics 277
human autonomy 221
image processing 98, 102, 311
image-similarity 102
industry 3
INSESS 211
instagram 311
inter-satellite links 7
internal representation 48
IVF 153
Kepler 7
keyword spotting 283
knowledge representation 4
learning-to-rank 59
lesions 163
LHCb 115
LiDAR 124
local differential privacy 273
longitudinal 163
low-earth orbits 7
machine learning 98, 115, 132, 221, 262, 287, 301
magnetic resonance imaging 94
maximum satisfiability 359
medical imaging 136
medication errors 4
metainformation 211
MML 3
model counting 369
model inference 277
MRI 163
multi-behaviour navigation 267
multi-valued treatment effect estimation 16
multi-view projections 90
multiple linear regression 193
multiple sclerosis 163
natural language processing 59, 311, 330, 347
neural networks 16, 98, 102, 132, 287
noise filtering 128
noise modeling 128
non-terrestrial networks 7
online discussion 59
optimization 277
paralinguistics 138
particle detectors 115
pattern discovery 70
pattern mining 231
peer-review 80
perceptual soundscape evaluation 287
Petri nets 120
physiological signals 28
point clouds 124
politics 330
portfolio management 110
preprocessing 211
QBF solvers 347
qualitative reasoning 256
qualitative representations 241
questionnaire 211
rating scales 291
reasoning 241
recommender systems 250
refereeing 80
regression 311
regular propositional logic 359
reinforcement learning 110, 173
relative size 241
remote sensing 124
reversal error 241
road cracks 38
SAT 337, 347, 369
segmentation 124, 134, 163
semantic segmentation 38, 98
semantic tableaux 359
sentiment analysis 311, 330
sequence learning 231
serious games 241, 241
SGP4 7
Shepard's Law 48
sign language recognition 132
similarity measures 256
simplified perturbations models 7

simulation models 143
SNN 102
social comfort 267
social games 231
social media 301, 311, 320
social navigation 267
social network 273
socially aware robotics 267
soundscape 287
SP-type problems 241
Spanish sign language 132
speech technology 3
stock market 110
stop words 80
supervised learning 7
sustainable decision-making 256
synthetic data 28
synthetic data generation 183
text extraction 320
time series 28, 70
time series classification 204
tool 337
tourism 311
tracer 337
transfer learning 28
transformers 98
triplet-loss 102
urban services 173
video quality 134
wastewater-based epidemiology 262
wide range 90
yaw 90

This page intentionally left blank

Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

Author Index

Abuasaker, W. 291
Agell, N. 256, 291
Alavi, A. 59
Alsina-Pagès, R.Ma. 287
Álvarez-García, E. 330
Álvarez-Napagao, S. 153
Angerri, X. 211
Angulo, C. 120
Ansótegui, C. 369
Aramendia, P. 250
Arcos, J.L. 143
Arjona, J. 28
Ballester, R. 301
Bandera Rubio, J.P. 267
Bandera, A. 267
Bech, S. 183
Béjar, J. 28
Béjar, R. 262
Benito-Altamirano, I. 138
Bernet Andrés, S. 115
Betorz, F. 7
Bonet, M.L. 369
Bonet-Solà, D. 287
Bono, C.A. 320
Burgos, X.P. 136
Calò, E. 347
Calvo Gómez, M. 115, 277
Camacho, A. 120
Campdepadros, R. 231
Cané, M. 337
Carós, M. 124
Casales-Garcia, V. 311
Casas, J. 136
Casas-Roma, J. 163
Caselles, V. 136
Cerquides, J. 16, 59, 143, 301, 320
Chan, J. 59
Cívico, S. 153
Coll, J. 337, 359
Corominas, J. 250
Corominas, L. 262
Cruz-Ramírez, N. 128
De Lecuona, I. 221
Díaz, M. 143
Diaz-Hurtado, M. 163
Escrig Escrig, J. 7, 283
Estrada, O. 193
Falomir, Z. 241, 311
Fernandez-Marquez, J.L. 301, 320
Ferrer, E. 7
Firmansyah, H.B. 320
Formiga, L. 3
Fortea, J. 94
García Piquer, À. 115
García, D. 28
Garcia-Calvés, P. 250
García-Costa, D. 80, 330
Garcia-Gasulla, D. 153
García-Piquer, Á. 106
Garcia-Sanchez, C.A. 120
García-Sánchez, J.M. 106
Garcia-Segarra, P. 241
Giacalone, R. 98
Gibert, K. 70, 211
Gilabert, P. 134
Golobardes Ribé, E. 115, 277
Gómez-Guillén, D. 143
González, M. 153
Gonzalez-Abril, L. 311
Grimaldo, F. 80, 128, 330
Guerra-Hernández, A. 128
Guerrero-Mosquera, C. 98, 102, 110
Guillén Ruiz, S. 267
Heredia-Lidón, Á. 90
Hernández Guillamet, G. 193
Heuzé, Y. 94
Hidalgo-Paniagua, A. 267
Hinjos, D. 153
Huerta Casado, I. 283
Just, A. 124
Kamišalić, A. 4
Kanber, B. 163
Karna, A. 70
Komander, B. 59
Labeyrie, Y. 301
Lerma-Martín, A. 106

Levy, J. 347, 369
Li, C.-M. 359
Limón, X. 128
Llufriu, S. 163
Lobato, B. 183
López Seguí, F. 193
López, B. 193, 231
Lorini, V. 320
Malagelada, C. 134
Malé, J. 94
Manyà, F. 359
Marco-Tordera, L. 80
Martí, P. 120
Martínez-Abadías, N. 90, 94
Martínez-Galicia, D. 128
Martínez-Heras, E. 163
Martin-Torres, P. 153
Mateu, C. 262
Méndez, M. 153
Miralles, D. 48
Moreno, A. 204
Moreno-Rus, A. 241
Mulayim, M.O. 301
Muñoz, A. 183
Museros, L. 311
Nguyen, J. 256, 291
Nin, J. 173
Núñez, J.F. 28
Okran, A.M. 38
Ortega Beltrán, E. 138
Pascual-Fontanilles, J. 204
Paul, S. 273
Perez Ruchat, B.M. 250
Piedra, W. 136
Piera, J. 59
Prados, F. 136, 163
Prat-I-Pubill, Q. 256
Pueyo-Ros, J. 262
Puig, D. 38
Pujol, O. 183, 221
Rashwan, H.A. 38
Ribalta, M. 262
Ribas Ripoll, V. 221
Rojo, M. 337
Romero-Aroca, P. 204
Rubión, E. 262
Ruiz, F.J. 291
Ruiz-De-Azua, J.A. 7
Sabater, A. 231
Sabater-Mir, J. 250
Salas, J. 273
Salbanya, B. 173
Saleh, A. 38
Sánchez Deutsch, S. 283
Sánchez, M. 291
Santamarta, V. 241
Sanz, I. 311
Seguí, S. 124, 134
Serrano, C. 48
Sevillano, X. 90, 94, 98
Soria, J.M. 183
Subías-Beltrán, P. 221
Svensson, K.F. 102
Tormos, A. 28
Torra, V. 273
Unceta, I. 173
Urcelay, L. 153
Vallés, V. 132
Valls Canudas, N. 115, 277
Valls, A. 204
Velasco, M. 120
Velasco-Regulez, B. 16
Ventura, C. 138
Ventura-Campos, N. 241
Vicente-Gomez, A. 163
Vidaña-Vila, E. 132, 287
Vila-Francés, J. 330
Vilasís-Cardona, X. 106, 115, 277
Villamarín Díaz, F. 110
Villaret, M. 337
Vitrià, J. 124, 134
Vivó, F. 163
Wenzek, H. 134
Yangin, E. 359
Zacharaki, K. 256