Decision Trees as a tool for Data analysis. Elections in Barcelona: A case Study

E. Armengol and À. García-Cerdaña

Artificial Intelligence Research Institute, (IIIA-CSIC)
Campus UAB, Camí de Can Planes, s/n, 08193 Bellaterra, Barcelona
email: eva@iiia.csic.es

Abstract. Decision trees are inductive learning methods that construct a domain model easy to understand from domain experts. For this reason, we claim that the description of a given data set using decision trees is an easy way to both discover patterns and compare the classes that form the domain at hand. It is also an easy way to compare different models of the same domain. In the current paper, we have used decision trees to analyze the vote of the Barcelona citizens in several electoral convocations. Thus, the comparison of the models we have obtained has let us know that the percentage of people with a university degree is the most important aspect to separe the neighbourhoods of Barcelona according to the most voted party in a neighbourhood. We also show that in some neighbourhoods has always won the same party independently of the kind of convocation (local or general).

Key words: Inductive Learning Methods, Decision Trees, Analysis of electoral results

1 Introduction

Decision trees are inductive machine learning algorithms useful to construct domain models. Commonly, such models are built having prediction in mind. The advantage of decision trees in front of other machine learning algorithms (i.e., support vector machines, neural networks, etc.) is that they are easily understandable by experts [8]. The most of our previous work has focused on building predictive models [2, 4] but, during the interaction with the experts, we observed that sometimes the expert was more interested on the attributes taken into account during the construction of the tree than in the predictivity of the final model.

In [1] we pointed out that, given a decision tree, the path from the root to a leaf can be interpreted as an explanation of the classification since it contains the pairs attribute-value relevant for the classification. In addition, in [3] we argued that the tree can be used to analyze a database. For instance, when a tree has a high depth, this means that all classes are very similar. In that case, we could conclude that the attributes used to describe the domain objects are not appropriated. Conversely, when two classes are separable using a few

attributes, this means that the classes are very different. We have used this kind of analysis to assess the life quality of people with intelectual disabilities [3] and caracterization of melanomas [3] [A2 = v2] and also to classify cows according their milk production [6].

In the present paper, we propose the use of decision trees to compare models of a domain or also to compare models of different domains. By the way they are built, decision trees let us know which attribute is the most important. Therefore, this allows making a first comparison to easily detect what is important for each model. As the tree is growing, the domain objects are distributed in the leaves and this also gives an idea of the similarities and differences between the classes.

To prove the feasibility of performing the comparison of models using decision trees, we analyzed several electoral results of the city of Barcelona. In particular, we have analyzed results of the elections of four convocations: Catalan Parliament held in 2017; Spanish Parliament held in April 2019; Council Hall held in May 2019; and Spanish Parliament held in November 2019. Our goal is to compare the results of these convocations and to check if the electoral behaviour of the voters has changed.

The paper is organized as follows. In Section 2 there is a brief explanation of decision trees. In Section 3 there is the description of the database used in the experiments. Section 4 contains a description and a discussion of the experiments carried on. Finally, Section 5 is devoted to conclusions and future work.

2 Decision Trees

A *Decision Tree* (DT) is a directed acyclic graph in the form of a tree. The root of the tree has not incoming edges and the remaining ones have exactly one incoming edge. Nodes without outgoing edges are called *leaf* nodes and the others are *internal* nodes. A DT is a classifier expressed as a recursive partition of the set of known examples of a domain [7]. The goal is to create a domain model predictive enough to classify future unseen domain objects.

Each node of a tree has associated a set of examples that are those satisfying the path from the root to that node. The leaves determine a partition of the original set of examples since each domain object only can be classified following one of the paths of the tree. The construction of a decision tree is performed by splitting the source set of examples into subsets based on an attribute-value test. This process is repeated on each derived subset in a recursive manner. Figure 1 shows the ID3 algorithm [9, 10] commonly used to grow decision trees. From a decision tree we can extract rules (i.e., patterns) giving descriptions of classes, since each path from the root to a leaf forms a classification rule. When all the examples of a leaf belong to the same class such description is *discriminant*. Otherwise, the description is *no discriminant*.

A key issue of the construction of decision trees is the selection of *the most relevant attribute* to split a node. There are different criteria to split a node and therefore, the selected attribute could be different depending on it and thus the whole tree could also be different. In our experiments we used the López

Algorithm 1 ID3 algorithm for growing a decision tree.

```
procedure ID3(E, A)                    ▷ E: Set of Examples; A: Set of attributes
    Create node
    if all e ∈ E belong to the same class then
        return class as the label for node
    else
        a_t ← best attribute
        for each value v_j of a_i do
            add a new tree branch below node
            E_{a_i} ← subset of examples of E such that a_i = v_j
            ID3(E_{a_i}, A − {a_i})
        end for
    end if
return node
end procedure
```

de Mántaras' distance [5], which is an entropy-based normalized metric defined in the set of partitions of a finite set. It compares the partition induced by an attribute, say a_i, with the *correct partition*, i.e., the partition that classifies correctly all the known examples. The best attribute is the one inducing the partition which is closest to the correct partition. Given a finite set X and a partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of X in n sets, the entropy of $\mathcal{P}$ is defined as ($|\cdot|$ is the cardinality function):

$$H(\mathcal{P}) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i, \text{ where } p_i = \frac{|P_i|}{|X|}$$

and where the function $x \cdot \log_2 x$ is defined to be 0 when $x = 0$. The *López de Mántaras*' distance (LM) between two partitions $\mathcal{P} = \{P_1, \ldots, P_n\}$ and $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ is defined as:

$$\mathrm{LM}(\mathcal{P}, \mathcal{Q}) = \frac{H(\mathcal{P}|\mathcal{Q}) + H(\mathcal{Q}|\mathcal{P})}{H(\mathcal{P} \cap \mathcal{Q})}, \tag{1}$$

where

$$H(\mathcal{P}|\mathcal{Q}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \cdot \log_2 \frac{r_{ij}}{q_j}, \quad H(\mathcal{Q}|\mathcal{P}) = -\sum_{j=1}^{m}\sum_{i=1}^{n} r_{ij} \cdot \log_2 \frac{r_{ij}}{p_i},$$

$$H(\mathcal{P} \cap \mathcal{Q}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \cdot \log_2 r_{ij},$$

$$\text{with } q_j = \frac{|Q_j|}{|X|}, \text{ and } r_{ij} = \frac{|P_i \cap Q_j|}{|X|}.$$

Decision trees can be useful for our purpose because their paths give us *patterns* describing classes of objects (electoral sections in our approach) in a

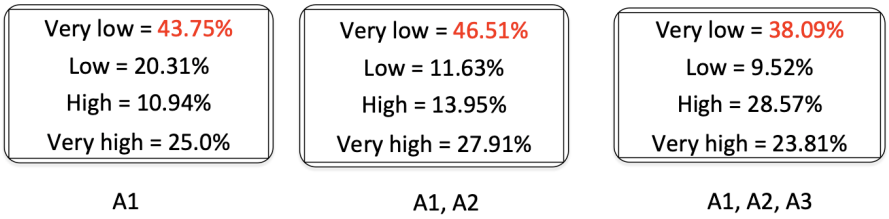

Fig. 1. Example of the stopping condition we used when growing a decision tree.

user-friendly manner. One shortcoming of decision trees is *overfitting*, meaning that there are few objects in most of the leaves of the tree. In other words, paths are actually descriptions that poorly represent the domain. The responsible of overfitting is the stoping condition of the algorithm: the set of examples has to be partitioned until all the examples of a node belong to the same solution class.

A way to either avoid or reduce overfitting is by pruning the tree, i.e., to expand all the nodes and then, with a post-process to merge two o more nodes; or, under some conditions, a node is no longer expanded. However, in both cases, this means that leaves can contain objects belonging to several classes and, therefore, paths do not represent discriminatory descriptions of classes. In other words, the descriptions or patterns represented by the branches of the tree are satisfied by objects of more than one class.

In our approach, we managed overfitting by controlling the percentage of elements of each class. Let S_N be the set of objects associated with an internal node N. The stopping condition in expanding N (the *if* of the ID3 algorithm) holds when the percentage of objects in S_N that belong to the majority class decreases in one of the children nodes. In such a situation, the node N is considered as a leaf.

As an example, let us suppose that the examples of a database can be classified in one of the following classes: *Very low*, *Low*, *High*, and *Very high*. Let us suppose now that when we grow a decision tree, we find that the most relevant attribute is A1. For the value $v1$ of such attribute we have the tree path (description) $D1 : [\mathsf{A1} = v1]$. The left hand side of Fig. 1 shows the distribution of the objects satifying $D1$ in each class. We see that the majority class with the 43.75% of examples is Very low. The next most relevant attribute is A2 and, for a value $v2$ we have the description $D2 : [[\mathsf{A1} = v1], [\mathsf{A2} = v2]]$. The centre of Fig. 1 shows the distribution of the objects satifying $D2$ in each class. Here the majority class is again Very low and the percentage is 46.51%; therefore the addition of A2 has improved the classification. Let us suppose that the next most relevant attribute is A3 and, for a value $v3$, we have the description $D3 : [[\mathsf{A1} = v1], [\mathsf{A2} = v2], [\mathsf{A3} = v3]]$. The right side of Fig. 1 shows that now the percentage of the majority class is 38.09%, i.e., lower than the one of $D2$. Therefore, now the procedure stops and the tree path we can use as description is $D2$.

Fig. 2. Administrative division of Barcelona in 10 districts.

3 A Case Study: An analysis of the electoral results in Barcelona

In Catalonia, there are four different kind of elections: Municipalities, Catalan Parliament, Spanish Parliament and European Parliament. In this study we want to use decision trees to compare the results of four electoral convocations: Catalan Parliament 2017, Spanish Parliament April 2019, Municipal elections 2019, and Spanish Parliament November 2019.

Previously to describe the database we have used in our experiments, we briefly explain the administrative organization in neighbourhoods of Barcelona and the political context.

3.1 Administrative organization of Barcelona

From the administrative point of view, Barcelona is composed of 10 districts (Fig. 2) each one in turn, composed of neighbourhoods. Thus, Barcelona is composed of 73 neighbourhoods. We focus our study in the political party that had won in each neighbourhood.

3.2 Electoral organization of Barcelona and Political context

Electoral landscape of Catalonia is formed by 5048 *electoral sections* each one of them composed of a minimum of 500 potential voters and a maximum of 2000.

Table 1. Division of Barcelona city in districts. For each district it is shown the number of neighbourhoods (#Neighb) and the number of electoral sections (#ES).

Number	District	#Neighb.	#ES
1	Ciutat Vella	4	55
2	Eixample	6	173
3	Sants-Monjuïc	8	117
4	Les Corts	3	57
5	Sarrià-Sant Gervasi	6	98
6	Gràcia	5	88
7	Horta-Guinardó	11	123
8	Nou Barris	13	117
9	Sant Andreu	7	96
10	Sant Martí	10	147

Following such criteria, Barcelona is formed by 1071 electoral sections distributed between the 10 districts as Table 1 shows. Notice that the number of electoral sections of each district gives an idea of its density of population.

From 2010 there is a complex political framework in Catalonia. In addition to the traditional ideologies left-right a new issue appears: the independence of Catalonia from Spain. Some of the historical Catalan parties already had the independence in their program, however it was not a main objective. Nevertheless, from a set of reasons that are out of the scope of this paper, the independence of Catalonia has become a priority for many population and for some parties, to the point that the choice independence/no independence has put the left-right dichotomy in a second term. It would be interesting to know how this factor has influenced the behaviour of the voters. Barcelona is a very populated city with many people having his origins in other Spanish regions or in other countries. For this reason we have considered interesting to study if the independence issue has some influence in the vote and which are the most reluctant neighbourhoods to independence.

Table 2 shows the political parties that concurred to the electoral convocations we analyzed and their ideology. For the sake of simplicity we call *ECP* a party that has concurred with different names in all the elections: *Barcelona en Comú, En Comú Podem, Unidas Podemos,* among others. In that table we only show those parties that have been winners in some of the neigborhoods, therefore it is not an exhaustive table of all the parties that were eligible.

3.3 The database

The database we have is composed of 73 records, each one of them corresponds to one neighbourhood of Barcelona. Each record has socio-demographic information and the party that had won in each one of the electoral convocations. Socio-demographic data has been obtained from the files of the official web of the Barcelona City Hall (https://www.bcn.cat/estadistica/angles/dades/inf/barris/a2018/index.htm) that contains socio-demographic information about each neigh-

bourhood of Barcelona. The most voted party for each neighborhood has been obtained from the public results in https://www.bcn.cat/estadistica/angles/dades/inf/ele/index.htm.

Socio-demographic attributes are the following: density, women, men, age0-14, age15-25, age25-64, age+65, Barcelona, Catalunya, Spain, other, university degree, birth rate, alone+65, over-aging rate, unemployed and income. The over-aging rate has been calculated as the rate of $\frac{people+75}{people+65} * 100$ and the income is a percentage that has been calculated taking 100 as the index of the whole city.

The majority of the data above are percentages. We discretized them by dividing the whole range of an attribute in intervals of equal length. We have used the elbow method to determine the better number of intervals for each attribute. For the attributes discretized in three intervals, we have associated the labels *L (low), M (medium)*, and *H (high)*. For the attributes discretized in four intervals, we have associated the labels *VL (very low), L (low), H (high)*, and *VH (very high)*. The attribute income has been discretized in four intervals, but we have used the labels *L (low), M (medium), H (high)*, and *VVH (very very high)*, where *VVH* corresponds to those neighbourhood having an income greater than 100%, the other intervals have been calculated using the equal length width and the elbow method.

4 Experiments

Our goal is to analyze how the most voted party changes for each electoral convocation at each neighborhood of Barcelona. We focused on the results of four electoral convocations: Catalan Parliament 2017, Spanish Parliament April 2019, Municipal elections 2019, and Spanish Parliament November 2019. In all the experiments we have considered all the socio-demographic attributes and, as solution class, the winner of each neighbourhood. We performed four independent experiments, one for each electoral convocation. Figure 3 shows the decision trees we have obtained.

Notice that, for all of them, the relevant attribute is the percentage of people having a university degree. Concerning the results of the convocation of 2017, seems clear that the voters were polarized according independentist/no indepen-

Table 2. Political parties that have won in some neighbourhoods in some of the electoral convocations we analyzed. For each party we show its ideology in terms of right.left and independentist-no independentist.

Party	Ideology	Independentist?
Cs	right	no
ERC	center-left	yes
JxC	center-right	yes
ECP	left	no defined
PP	right	no
PSC	cener-left	no

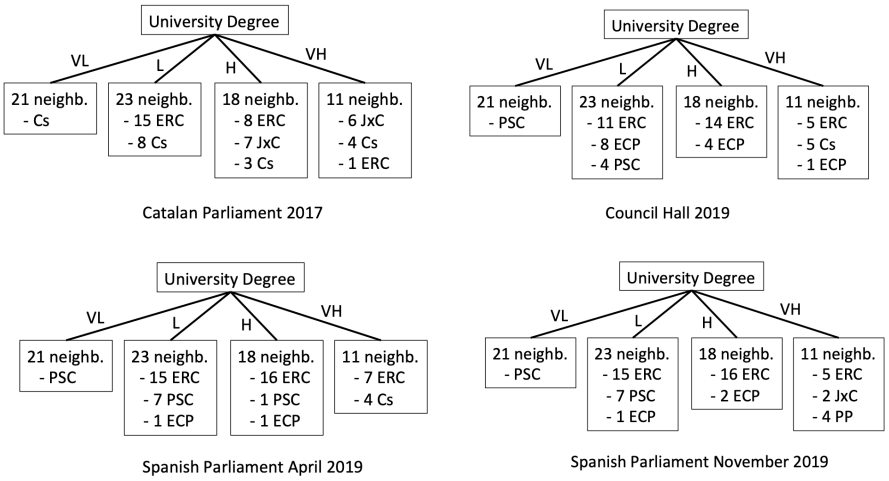

Fig. 3. Models of each one of the electoral convocations. Each node shows the number of neighbourhoods (*neigb.*) satisfying it, and also how many of them have voted to the parties.

dentist options since all the votes went to ERC and JxC (both independentists) and Cs (no independentist). The majority for the independentist option is clear, however notice that all the neighbourhoods having very low percentage of people with university degree had voted to Cs. As the percentage of university degrees increases, the percentage of votes to Cs decreases. In the rest of convocations these neighbourhoods changed the vote to PSC (center left). In any case, no voters had moved to independentist options. Also, it is interesting to remark that elections of 2017 were specially polarized by the independence/no independence option, and in this particular aspect, the political party Cs was more beligerant against independence than PSC. This can also be shown in the 18 neighbourhoods with high percentage of people with university degree. In 2017, 15 of them had voted independentist (8 to ERC and 7 to JxC) whereas the remaining 3 voted to Cs. In April 2019 the no independentist vote was divided between PSC and ECP; and in the remaining two convocations the vote went to ECP (this party has not a defined position about Catalonia independence). For the other percentages of university degrees there is not a clear separation using only that attribute. It is also interesting to see that the 21 neighbourhoods with low percentage of university degree that in 2017 had voted to Cs, in the rest of convocations have changed their vote in favour to PSC (a more moderated option). Also notice that the votes to Cs of the neighbourhoods with very high percentage of university degrees had change to PP (right, no independentist) in the last elections.

The complete model for the convocation held in April 2019 can be seen at the left hand side of Fig. 4. Notice that for neighbourhoods with high or very high

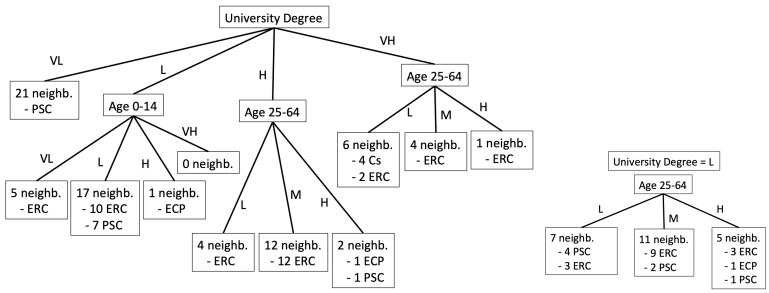

Fig. 4. Model for the Spanish elections held in April 2019. Each node shows the number of neighbourhoods (*neigb.*) satisfying it, and also how many of them have voted to the parties.

percentage of people with university degree, the most relevant attribute is the range of age between 25 and 64, whereas when the percentage is low, the most relevant attribute is the range of age between 0 and 14. When the percentage is very low no additional attributes are necessary. Notice that neighbourhoods with low percentage of young people (from 0 to 14) could be interpreted as neighbourhoods with adult people who almost all of them can vote (in Spain the minimum age to vote is 18). In fact, this node could be considered as expressing the same as the one corresponding to [age25-64=H], since this last means that are neighbourhoods with low people under 24. For this reason, we have forced to use the attribute age25-64 for all the values of university degree. The result is the subtree show at the right hand side of Fig. 4.

The analysis of this result shows that for those neighbourhoods with either low or high percentage of university degrees, the vote is divided between independentist (ERC) and moderately no independentist (PSC) and ECP (undefined about independence). In neighbourhoods where the percentage of university degrees is high and the population with age between 25 and 64 is high, the vote goes to no independentist or undefined options (PSC or ECP) whereas in other situations, this means, in neighbourhoods with low or medium percentage of people between 24 and 64 years (i.e., mostly young people between 18 and 24 or people over 65) the vote goes to independentist options. Notice that the neighbourhoods with very high percentage of university degrees and low percentage of people between 25 and 64 years are the only ones that vote strong no independentist options (Cs).

If we let expand the tree, for the elections to the Council Hall (see Fig. 5) the next most relevant attribute is Other, i.e., the percentage of people of a neighbourhood that has born in a country different of Spain. In other words, this attribute represents the immigration percentage. We can see that in neighbourhoods with very high percentage of university degrees there are not neighbourhoods having high or very high percentage of immigration. The majority of the neighbourhoods with very low percentage of immigration vote to Cs (strongly no independentist) and the vote of the neighbourhoods with low percentage of immi-

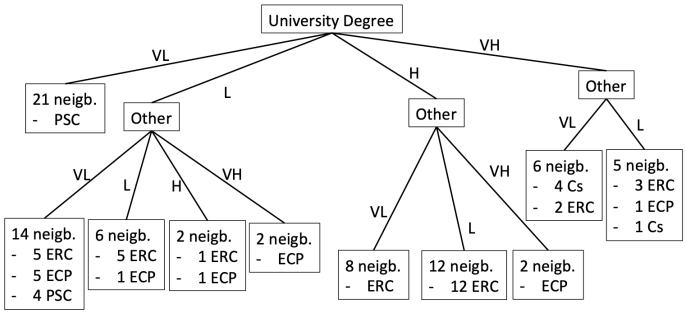

Fig. 5. Model for the Council Hall elections held in May 2019. Each node shows the number of neighbourhoods (*neigb.*) satisfying it, and also how many of them have voted to the parties.

gration is divided between ERC, ECP and Cs. In fact, the party Cs has win only in some neighbourhoods with [university degree=VH] . When the percentage of university degrees is high, only in the neighbourhoods with very high percentage of immigration ([Other=VH]) have voted the no independentist option of ECP. When the percentage of university degree is low, in the two neighbourhoods with very high immigration, the winner was ECP and in the 6 neighbourhoods with low percentage of immigration in 5 of them had won ERC. For other percentages there is not a clear winner and there is no way to expand the tree to separate the neighbourhoods.

Finally, the model for the elections held in November 2019 (see Fig. 6) show that the most relevant attribute is also age25-64 as in the elections of April 2019. Again, for neighbourhoods where the percentage of university degrees is low, the most relevant attribute is age0-14. As we explained, such atribute could be seen as complementary to age25-64, for this reason in Fig. 6 we used this last attribute for all the values of university-degree.

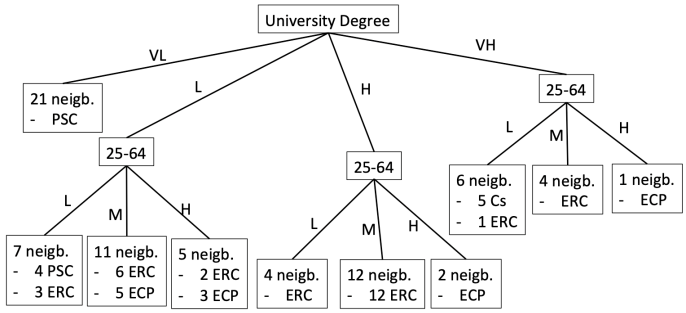

Fig. 6. Model for the Spanish elections held in November 2019. Each node shows the number of neighbourhoods (*neigb.*) satisfying it, and also how many of them have voted to the parties.

Comparing the subtrees of [university-degree=L] of the models of the Spanish Elections (Figs. 4 and 6), we can see that in essence they are not so different. For [age25-64=L] the distribution of votes is the same; for [age25-64=M] three neighbours of ERC (indepemdentist) in April have changed to ECP (undefined) in November, and all the votes of PSC have went to ECP; and when [age25-64=H], ERC has lost one neighbour with respect to the results of April, and ECP has won all the no independentist neighbour that in April had voted to PSC. For [university-degree=H] the only change is that when [age25-64 = H] in both neighbourhoods the winner has been ECP. For [university-degree = VH] Cs has won in one more neighbourhood in November than in April; and also when [age25-64 = H] in the neighbourhood that in April had won ERC, in November has won ECP.

5 Conclusions

In this paper we have introduced a new approach to analyze electoral data: the decision trees. This kind of methods are commonly used to construct domain models useful for prediction. Our focus has been the political party that has won in each Barcelona neighbourhood.

Thanks to the representation as a tree we can see that: 1) the most relevant attribute to characterize the neighbourhoods of Barcelona according the winner party is university-degree; 2) from the point of view independentist/no independentist, the results are not substantially different in the four analyzed convocations; 3) the 21 neighbourhoods having a very low percentage of people with university degree, always have voted no independentist options, although in 2017 the winner was Cs (strongly against independence) and in the next convocations changed to PSC (more moderated); 4) The Cs party has had and important decrement of votes, however, the neighbourhoods with [university-degreeVH] and [age25-64=L] have a great fidelity to this party since there are the ones in which this party has won in all the convocations; and 5) the percentage of immigration is only important for the elections to the Council Hall, and the ECP party has won in more neighbourhoods than in other convocations.

The conclusion is that the citizens of Barcelona have not substantially changed their vote according to the kind of elections, since we have found very similar models for each one of them. In addition, a very important conclusion is that the percentage of university degree is the most important factor influencing the electoral result of a neighbourhood. Clearly, the independentist/no independentist dichotomy has had an influence in the result since in convocations previous to 2010 the sense of vote changed according to the elections: in convocations to Spanish Parliament tend to win parties that are delegations of national parties (for instance PSC) whereas in Catalan elections tend to win Catalan parties.

In the future we plan to analyze in the same way other Catalan cities as, for instance Girona, Lleida and Tarragona (the main cities of each one of the Catalonia concurrencies) and compare the similarities and differences.

Acknowledgments

This research is funded by the project RPREF (CSIC Intramural 201650E044); and the grant 2014-SGR-118 from the Generalitat de Catalunya.

References

1. E. Armengol. Usages of generalization in CBR. In R.O. Weber and M. M. Richter, editors, *ICCBR-2007. Case-based Reasoning and Development*, number 4626 in Lecture Notes in Artificial Intelligence, pages 31–45. Springer-Verlag, 2007.
2. E. Armengol. Building partial domain theories from explanations. *Knowledge Intelligence*, 2/08:19–24, 2008.
3. E. Armengol, À. García-Cerdaña, and P. Dellunde. Experiences using decision trees for knowledge discovery. In *Fuzzy Sets, Rough Sets, Multisets and Clustering*, pages 169–191. Springer, 2017.
4. E. Armengol and E. Plaza. Discovery of toxicological patterns with lazy learning. In V. Palade, R.J. Howlett, and L. Jain, editors, *KES-2003*, number 2774 in Lecture Notes in Artificial Intelligence, pages 919–926. Springer, 2003.
5. R. López de Mántaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6:81–92, 1991.
6. M. López-Suárez, E. Armengol, S. Calsamiglia, and L. Castillejos. Using decision trees to extract patterns for dairy culling management. In Vassilis Plagianakos Lazaros Iliadis, Ilias Maglogiannis, editor, *Proceedings of the 14th International Conference on Artificial Intelligence Applications and Innovations. AIAI-2018*, page In Press, 2018.
7. O. Maimon and L. Rokach, editors. *Data Mining and Knowledge Discovery Handbook, 2nd ed.* Springer, 2010.
8. M. J. Pazzani. Knowledge discovery from data? *IEEE Intelligent Systems*, 15(2):10–13, 2000.
9. J. R. Quinlan. Discovering rules by induction from large collection of examples. In *Expert Systems in the Microelectronic Age. D. Michie (Ed.)*, pages 168–201. Edimburg Eniversity Press, 1979.
10. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.