Classification of Schistosomiasis Prevalence Using Fuzzy Case-based Reasoning

Flávia T. Martins-Bedé[1], Lluís Godo[2], Sandra Sandri[1 2], Luciano V. Dutra[1], Corina C. Freitas[1], Omar S. Carvalho[3], Ricardo J. P. S. Guimarães[3], Ronaldo S. Amaral[4]

[1] Instituto Nacional de Pesquisas Espaciais - INPE
CP 515, 12201-970 - São José dos Campos - SP, Brazil
[2] Institut d'Investigació en Intel·ligència Artificial, IIIA - CSIC
Campus UAB s/n, 08193 Bellaterra, Spain
[3] Centro de Pesquisas René Rachou/FIOCRUZ, MG, Brazil
[4] Secretaria de Vigilância em Saúde/MS, Brazil

Abstract. In this work we propose the use of a similarity-based fuzzy CBR approach to classify the prevalence of Schistosomiasis in the state of Minas Gerais in Brazil.

1 Introduction

Case-based reasoning [1], CBR for short, can be considered as a form of similarity-based or analogical reasoning since the basic principle implicitly followed in this problem solving methodology is that *similar problems have similar solutions*. A weaker version of this principle stated is given by "it is only plausible (but not necessary) that similar problems have similar solutions". In [2] the authors propose a novel Fuzzy CBR algorithm for classification, in which this weaker principle is read, in the classification context [5] as:

"The more similar are the problem descriptions of two cases,
the more *possible* their classification values are similar"

Basically, in this context, a CBR base is composed of cases of the form $c = (d, cl)$, where d is the case description modeled as a vector of values for a set of attributes and cl its associated class. The assignment of a class to a new case description d_0 will depend on the similarity between d_0 and the descriptions of the cases in the CBR base. In the approach proposed in [2], this similarity is calculated as a weighted mean of the similarity between each attribute addressed in the description part of the cases in the learning data set. In this approach, the weight vectors are learnt in such a way as to minimize the misclassification of the cases already contained in the base.

Schistosomiasis mansoni is a disease with social and behavioral characteristics. Snails of the *Biomphalaria* species, the disease intermediate host, uses water as a vehicle to infect man, the disease main host. In Brazil, six million people are infected by it, mainly in poor regions of the country [10]. According to the data

presented at the Brazilian Information System for Notifiable Diseases (SINAN) of the Ministry of Health, from 1995 to 2005, more than a million positive cases were reported, 27% of them in the State of Minas Gerais.

In [7], the authors present a classification Schistosomiasis prevalence for the State of Minas Gerais, using remote sensing, climate, socioeconomic and neighborhood related variables. Two approaches were used, a global and a regional one. In the first approach, a unique regression model was generated and used to estimate the disease risk for the entire state. In the second approach, the state was divided in four regions, and a model was generated for each one of them. Imprecise classifications were also generated for both approaches, using the estimated standard deviation and several reliability levels as basis.

The aim of this paper is to check the usefulness of the fuzzy CBR approach to classification proposed in [2] in order to estimate and classify schistosomiasis prevalence, as an alternative to the linear regression model approach developed in [7]. To allow the comparison of the results, we present two approaches for solving the problem, a global and a regional one, following the guidelines in [7].

This work is organized as follows. In Section 2 we describe the similarity-based fuzzy CBR model we have used. In Section 3 we present our application context and discuss the experiments we have performed. Section 4 finally brings the conclusion.

2 A similarity-based fuzzy CBR model for classification

2.1 Working framework

Before going into more details, let us specify our working framework for classification-like case-based reasoning problems. Let us assume we have a base of cases CB consisting of an already solved set of cases, where a case is represented by a (complete) tuple of attribute values describing the situation or problem to solve together with a solution class or result. To fix ideas, let $\mathbf{A} = \{a_1, \ldots, a_n\}$ be the set of description attributes and let $class$ denote the class attribute. Moreover, let us denote by $D(a_i)$ and $D(class)$ the domains of the attributes a_i and $class$ respectively (so $D(class)$ is the set of solution classes). Then a case $c \in CB$ will be represented as a pair $c = (d, cl)$, where $d = (a_1(c), \ldots, a_n(c))$ is a n-tuple with the problem description values and $cl = class(c)$ is the solution class for the case c. If we write $\mathbf{D} = D(a_1) \times \ldots \times D(a_n)$ ($\mathbf{D}$ for descriptions) and $\mathbf{Cl} = D(class)$, then a case base CB is just a subset of $\mathbf{D} \times \mathbf{Cl}$. In the following, all definitions will use this classification-oriented notation.

In this framework, given a case base $CB = \{c_i = (d_i, cl_i)\}_{i \in I}$ and a new problem description d^*, the CBR task is to find (guess) a solution class cl^* for d^*, by applying the above general principle in some form, i.e. taking into account the similarity of d^* with already solved cases $c_i \in CB$.

In its most general sense, a fuzzy similarity relation on a domain Ω is a mapping $S : \Omega \times \Omega \to [0, 1]$ which assigns to every pair (w, w') of elements of Ω a number measuring how much w and w' resemble each other according to some

given criteria, in the sense that the higher $S(w, w')$, the larger their resemblance. In particular, $S(w, w') = 1$ means that w and w' are undistinguishable, while $S(w, w') = 0$ means that w and w' have nothing in common. One can also understand $\delta(w, w') = 1 - S(w, w')$ as a kind of distance between w and w'. Usual and reasonable properties (see e.g. [6]) required of such functions are reflexivity and symmetry, i.e. $S(w, w) = 1$ and $S(w, w') = S(w, w')$, for any $w, w' \in \Omega$. S is called *separating* if it verifies that $S(w, w') = 1$ iff $w = w'$. Sometimes they are also required to fulfill a weak form of transitivity, namely $S(w, w') \otimes S(w', w'') \leq S(w, w'')$, where $\otimes$ is a t-norm. For our purposes, and unless stated otherwise, we shall consider similarity relations as reflexive and symmetric fuzzy binary relations but neither necessarily transitive nor separating.

2.2 Main elements of the model proposed

The approach we will describe in the rest of this section requires that, for each attribute $a \in \mathbf{A}$, there is an available fuzzy similarity relation S_a on $D(a)$, as well as a fuzzy similarity relation S_{class} over the set of classes **Cl**, as defined in Section 2.1.

The main step in the method is to define a suitable similarity relation S_D between the case descriptions in the case base CB and an arbitrary problem description. Our working assumption is that such similarity will be defined as a weighted average of the existing similarity functions S_a for each attribute. Then, of course, we need additional information to assess the relevance (weight) of each attribute for retrieving a particular case. In particular, we will assume there is a best[5] *set of weighs vector for each case* that properly evaluates the importance of each attribute when computing the similarity between that case and another arbitrary case description in the case base CB.

Definition 1. *Let* $\mathbf{A} = \{a_1, \ldots, a_n\}$ *be the set of attributes considered in* **D** *and, for each* $a \in \mathbf{A}$*, let* S_a *be the corresponding similarity relation on* $D(a)$*, and let* $\mathbf{w} : \mathbf{A} \to [0, 1]$ *be a weight assignment to attributes, i.e. an assignment such that* $\sum_{a \in \mathbf{A}} \mathbf{w}(a) = 1$*. Then, we define the induced fuzzy similarity relation* $S_D^{\mathbf{w}} : \mathbf{D} \times \mathbf{D} \to [0, 1]$ *over case descriptions as follows:*

$$S_D^{\mathbf{w}}(d_1, d_2) = \sum_{a \in \mathbf{A}} \mathbf{w}(a) \cdot S_a(a(d_1), a(d_2)) \tag{1}$$

using the notation $d_1 = (a_1(d_1), \ldots, a_n(d_1))$ *and* $d_2 = (a_1(d_2), \ldots, a_n(d_2))$*.*

Note that, so defined, $S_D^{\mathbf{w}}$ is a indeed similarity relation in the sense of Section 2.1, i.e. it is reflexive and symmetric.

Once we have defined the similarity relation $S_D^{\mathbf{w}}$, we can define how adequate a solution class cl is for a problem description d^* just by comparing d^* to the descriptions of all those cases in CB sharing that solution class cl, and aggregating all these similarity values.

[5] In the sense explained in Section 2.3.

Definition 2. *Let be $CB \subseteq \mathbf{D} \times \mathbf{Cl}$ a case base. Given a set of weight assignments $\mathbf{W} = \{\mathbf{w}_c\}_{c \in CB}$ (one per each case in CB), and a suitable aggregation function F on $[0,1]$, the adequacy degree between a case description $d^* \in \mathbf{D}$ and a solution class $cl \in \mathbf{Cl}$ is defined as follows:*

$$\Pi_{\mathbf{W},F}(d^*, cl) = F(\{S_D^{\mathbf{w}_c}(d^*, d) \mid c \in CB, c = (d, cl\}).$$

where $S_D^{\mathbf{w}_c}$ is defined as in Definition 1.

Depending on the application, suitable aggregation functions [4] may be for example disjunctive functions, like the maximum or other t-conorms, or some kinds of average functions, like quasi-arithmetic means or even more sophisticated aggregation functions.

Finally, given case base CB, a set of weight assignments $\mathbf{W} = \{\mathbf{w}_c\}_{c \in CB}$ and a suitable aggregation function F on $[0,1]$, the last step in the fuzzy CBR approach consists in assigning to a case description $d^* \in \mathbf{D}$ the solution class cl^* such that

$$cl^* = \arg\max_{cl \in \mathbf{CL}} \Pi_{\mathbf{W},F}(d^*, cl).$$

2.3 Learning the weight assignments for each case

Next, we describe how to learn the appropriate weight assignment for each case in the base CB. So, in the following, we will assume that a case $c_0 \in CB$ is known and fixed along the learning process. In fact, the same process we describe below for c_0 will be applied for each case in CB. Naturally, for each case c_0, the process would lead to its corresponding weight assignment $\mathbf{w}_0$.

To do so, in addition to the case c_0, we need to fix a subset of cases $LS_0 \subseteq CB$, i.e. a collection of problem descriptions whose solution class is known. This is the learning set. Then, the weights determination can be formulated in the following way:

Problem 1 (Weight Determination Problem). Let $c_0 = (d_0, cl_0)$ be a case in base CB and let $LS_0 \subseteq CB$ be the learning set relative to c_0. Then the weight determination problem relative to c_0 is to determine a weight assignment $\mathbf{w}_0 : \mathbf{A} \to [0,1]$ such that, for each case $c = (d, cl) \in LS$, the similarity between d_0 and d, $S_D^{\mathbf{w}_0}(d_0, d)$, approximates as much as possible the similarity between the solution classes cl_0 and cl, $S_{Class}(cl_0, cl)$.

Using the square difference to measure the divergence between the two similarities (i.e., the similarity between the two case descriptions and the similarity between their classes), we can reformulate the problem as follows:

Problem 2. Let $c_0 = (d_0, cl_0)$ be a case in the base CB and let $LS_0 \subseteq CB$ be the learning set from which the weight assignment $\mathbf{w}_0 : \mathbf{A} \to [0,1]$ relative to c_0 will be determined. Then the weight determination problem relative to c_0 is to find the values $\mathbf{w}(a_1), ..., \mathbf{w}(a_n)$ that minimize the following expression:

$$\sum_{(d,cl)\in LS_0} \Big[S_D^{\mathbf{w}}(d, d_0) - S_{Class}(cl, cl_0)\Big]^2$$

subject to the following constraints over w_{c_0}:
(1) $\sum_{a\in\mathbf{A}} \mathbf{w}(a) = 1$, and
(2) $\mathbf{w}(a) \geq 0$ for all $a \in \mathbf{A}$.

In the experiments described in Section 3 below, we have divided case base CB in two parts, a training set CBt and a validation set $CBv = CB - CBt$, and for each case $c_0 \in CBt$ we obtain a weight assignment $\mathbf{w}_0$ by solving the above minimization problem in a leave-one-out fashion, by taking as learning set the whole CBt after removing case c_0, i.e., for each $c_0 \in CBt$, $LS_0 = CBt - \{c_0\}$.

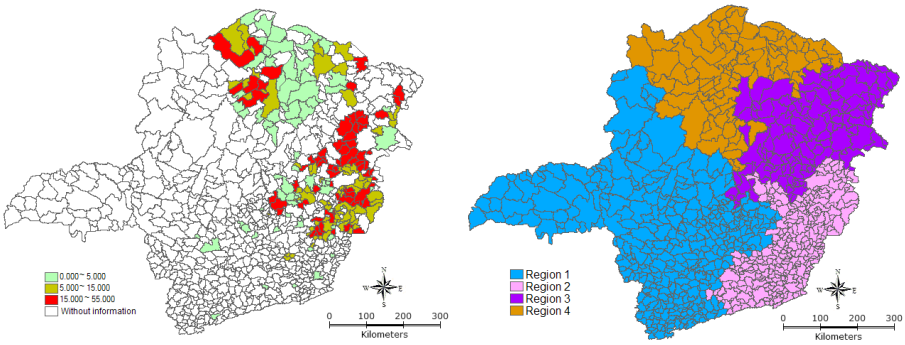
In order to solve this minimization problem, we apply the algorithm introduced in [8] with the extension described in [9]. It is worth pointing out that this minimization algorithm, as similar ones existing in the literature, fails to give a (unique) solution when there exists a particular kind of linear dependence among the columns in the data matrix of the problem. In our case, this would refer above to the matrix of similarity values $\mathbf{S} = \{s_{i,j}\}$ with $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where $s_{ij} = S_{a_i}(a_i(d_j), a_i(d_0))$, assuming that $\mathbf{A} = \{a_1, \ldots, a_n\}$ and $LS_0 = \{(d_1, cl_1), \ldots, (d_m, cl_m)\}$. In fact, as shown in [9], the problem only arises when there is a column $\mathbf{s}_i = (s_{1i}, \ldots, s_{ni})$ that can be written as a linear combination of the others in the form $\mathbf{s}_i = \sum_{j\neq i} p_j \cdot \mathbf{s}_j$ with $p_j \geq 0$ and such that $\sum_j p_j = 1$. In such a case, when removing one of the linearly dependent columns we get the same minimum we would get when considering all the attributes. Therefore, an alternative approach is to consider as many subproblems as dependent attributes, where each subproblem corresponds to the original one after removing one of the dependent attributes. The solution with a minimum error would correspond to the solution of the original problem.

3 Classification of Schistosomiasis prevalence

3.1 The original experiments: materials and methods

In the original experiments presented in [7], the disease prevalence data was provided by the Health Secretary of the State of Minas Gerais state. The prevalence is known for 197 municipalities out of the 853 composing the state (see Figure 3.1.a). In the original experiments, 86 independent variables of various types were used to classify prevalence: Remote Sensing (22), climatic (6), socioeconomic (34) and neighborhood characterization (24). The Remote Sensing variables were derived from sensors MODIS (Moderate Resolution Imaging Spectroradiometer) and SRTM (Shuttle Radar Topography Mission), and are supposedly related to the snail habitat type. The climatic variables were obtained from the Weather Forecast and Climate Studies Center (CPTEC) from the National Institute for Space Research (INPE) and reflects the conditions of survival of the snail and the various forms of the larvae of Schistosoma mansoni. The socioeconomic variables were obtained from SNIU (National System of Urban Indicators) such as

the water accessing means and sanitation condition aspects. The neighborhood characterization variables measure the disparity between neighboring municipalities with relation to variables of income, education, sewerage, water access and water accumulation.



(a) (b)

Fig. 1. The state of Minas Gerais in Brazil: a) known prevalence of Schistosomiasis, b) regionalization obtained through algorithm SKATER.

From the original 86 variables, a smaller set was selected, according to tests using multiple linear regression [7]; the independent variables chosen were those that had high correlation with the dependent variable and low correlation with other independent variables. Two main approaches were used: i) a global one, in which all the municipalities with known disease prevalence were used, for either constructing or validating a linear regression model, and ii) a regional one, in which the state was divided in four homogeneous regions and a linear regression model was created for each one of them. The number of independent variables used in the experiment varied; in the global approach 5 variables were used, while in the regional approach 2 variables were used for region R1, 5 for region R2, 4 for region R3 and 3 for region R4 (see details in [7]). In both the global and regional approaches, approximately 2/3 of the samples were used as training set, and the remainder 1/3 as the test set. Algorithm SKATER [3] was used to obtain the homogeneous regions in the regional model; this algorithm creates areas such that neighboring areas with similar characteristics belong to the same region (see Figure 3.1.b).

3.2 The fuzzy CBR experiments: materials and methods

This work uses the same data, variables and regionalization than those used in [7]. As already mentioned, to construct the similarities $S_D^{\mathbf{w}^c}$, our approach needs a similarity relation S_a for each independent attribute $a \in \mathbf{A}$ to be given. In our experiments all considered attributes are real-valued attributes. Then, for each attribute a we have taken S_a to be of the form $S_a = S_{\lambda_a}$, where S_{λ_a} is a

parametrized similarity relation on $[0,1]$ defined as

$$S_{\lambda_a}(x,y) = \max(0, 1 - \frac{|x-y|)}{\lambda_a \cdot length(a)})$$

where $\lambda_a \in (0,1]$ and $length(a) = \max_{c \in CB} a(c) - \min_{c \in CB} a(c)$ is the maximal variation of a in the whole case base.

On the following, we will simply denote a similarity relation for a description attribute as S_λ, $\lambda \in (0,1]$, and we will synthesize the notation of a set of such similarity relations as $S_{(\lambda_1,\ldots,\lambda_n)}$, meaning that S_{λ_1} is applied to attribute a_1, S_{λ_2} to a_2, etc.

The dependent attribute $Class$ is defined in the domain $\mathbf{Cl} = \{L, M, H\}$, for *low*, *medium* and *high* disease prevalence, respectively. In the experiments, we have taken the similarity relation S_{Class} on $\mathbf{Cl}$ to be of the form $S_{Class} = T_\lambda$ for some $\lambda \in [0,1]$, where T_λ is defined as $T_\lambda(w,w) = 1$ and $T_\lambda(w,w') = T_\lambda(w',w)$, for all $w, w' \in \mathbf{Cl}$, $T_\lambda(H,M) = T_\lambda(M,L) = \lambda$ and $T_\lambda(H,L) = 0$.

3.3 Experimental results and analysis

Table 1 brings the best results obtained from experiments made with the data, for a) the regression models employed in [7] and b) by the fuzzy CBR method. On Table 1.b, besides the CBR approach accuracy value, we have indicated the similarity relations used in the description and solution variables.

We have applied the fuzzy CBR method using various parameters sets for the description and class similarity relations. As aggregation function F in Definition 2, we have used the maximum (as proposed in [2]) and other operators; the best results were obtained with the arithmetic means (see Table 1). Notice that for region R1, we have obtained the same accuracy (.56) using similarities $(S_{(.3,.4)}, T_{.5})$ and $(S_{(.1,.1)}, T_0)$ for the regional learning approach.

Region	regional	global
R1	0.56	0.50
R2	0.51	0.40
R3	0.72	0.48
R4	0.76	0.59

(a)

Region	regional	global
R1	0.56 $(S_{(.3,.4)}, T_{.5})$	0.56 $(V_{(.2,.2)}, T_0)$
R2	0.56 $(S_{(.2,.2,.2,.2)}, T_{.5})$	0.49 $(V_{(.1,.1,.1,.1)}, T_0)$
R3	0.62 $(S_{(.2,.4,.3,.3)}, T_0)$	0.71 $(V_{(.2,.2,.2,.2)}, T_0$
R4	0.38 $(S_{(.2,.2,.2)}, T_{.5})$	0.65 $(V_{(.1,.1,.1)}, T_{.5})$

(b)

Table 1. Classification accuracy, with learning made on either a global or a regional basis: (a) regression models and (b) fuzzy CBR models.

The results obtained with the fuzzy CBR approach are comparable to those obtained with the regression models and, in one case, the fuzzy CBR approach is better than regression (region R_2). It is interesting to note that in the fuzzy CBR approach, contrary to what happened with the regression models, the global learning approach have very often outperformed the regional one. As a matter of fact, the global fuzzy CBR approach obtained invariably better results than regression in the global approach.

4 Conclusions

In this work we have described the use of a similarity based fuzzy CBR approach to classify the prevalence of Schistosomiasis in the state of Minas Gerais in Brazil. We have compared our results to the ones obtained from the literature that uses linear regression. The comparison results shows the suitability of the approach.

The classification method is such that at the end of an experiment, we obtain a similarity degree between the description of a new case and each one of those in the case base. Then we derive the compatibility of the new case with each class value, by aggregating all the similarity degrees obtained from the cases in the case base that are classified with that value. Here we have verified that the usual operator, max, is not always the most suitable for a given application. In particular, in our problem, the best aggregation operator from those tested was found to be the arithmetic means.

Acknowledgments The auhors acknowledge partial support of the CSIC-CNPq bilateral project, Ref. 2007BR0053. Godo and Sandri are also partially supported by the Spanish projects IEA (TIN2006-15662-C02-01), MULOG2 (TIN2007-68005-C04-01) and "Agreement Technologies" (CONSOLIDER CSD2007-0022, INGENIO 2010). The authors thank Vicenç Torra for unvaluable help in realizing the experiments and to an anonymous reviewer for helpful comments.

References

1. Aamodt, A., Plaza, E., (1994), Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, AI Communications 7 39–59.
2. Armengol, E., Esteva, F., Godo, L., Torra, V., (1994) On learning similarity relations in fuzzy case-based reasoning, Lecture Notes in Computer Science 3135 14–32.
3. Assuno, R.M., Neves, M.C., Cmara, G., Freitas, C.C., (2006) Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees, International Journal of Geographical Information Science, 20 797–811.
4. Calvo, T., Kolesarova, A., Komornikova, M., Mesiar, R., (2002) Aggregation operators: Properties, classes and construction methods. In T. Calvo, G. Mayor, and R. Mesiar, editors, Studies in Fuzziness and Soft Computing, Vol. 97, pages 3?104.
5. Dubois, D., Esteva, F., Garcia, P., Godo, L., Lòpez de Mantaras, R., Prade, H., (1998), Fuzzy Set Modelling in Case-based Reasoning. International Journal of Intelligent Systems 13:4 345–373.
6. Dubois, D., Prade, H., (eds.) (2000), Fundamentals of Fuzzy Sets, The Handbooks of Fuzzy Sets Series, Kluwer Academic, Dordrecht, 2000.
7. Martins, F., Freitas, C., D., Dutra, L. Sandri, S., Drummond, I., Fonseca, F., Guimarães, R, Amaral, R., Carvalho, O., Risk mapping of Schistosomiasis in the state of Minas Gerais, Brazil, using MODIS and socioeconomic spatial data, submitted.
8. Torra, V., (2000), On the learning of weights in some aggregation operators: the weighted mean and OWA operators, Math. and Soft Comp. 6 249–265.
9. Torra, V., (2002), Learning weights for the quasi-weighted means, IEEE Trans. on Fuzzy Systems 10:5 653–666.
10. World Health Organization, (1993), "The Control of Schistosomiasis. Second Report of the WHO Expert Committee". Technical Report Series no. 830, Geneva.