On Learning Similarity Relations in Fuzzy Case-based Reasoning

Eva Armengol, Francesc Esteva, Lluís Godo, Vicenç Torra

Institut d'Investigació en Intel·ligència Artificial - CSIC
Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain)

Abstract. Case-based reasoning (CBR) is a problem solving technique that puts at work the general principle that similar problems have similar solutions. In particular it has been proved effective for classification problems. Fuzzy set-based approaches to CBR rely on the existence of a fuzzy similarity functions on the problem description and problem solution domains. In this chapter, we study the problem of learning a global similarity measure in the problem description domain as a weighted average of the attribute-based similarities and, therefore, the learning problem consists on finding the weighting vector that minimizes misclassification. The approach is validated comparing results with an application of case-based reasoning in a medical domain which uses a different model.

Keywords: Case Base Reasoning, Fuzzy Case Base Reasoning, Similarity relation, Aggregation.

1 Introduction

Case-based reasoning, CBR for short, amounts to inferring that what is true in some known cases might still be true, possibly up to some suitable adaptation, in a newly encountered situation which is similar enough to those already known cases (see e.g. [1,24]). In this way, case-based reasoning can be considered as a form of similarity-based or analogical reasoning since the basic principle implicitly assumed to apply in this kind of problem solving methodology is that *similar problems have similar solutions.*

Before going into more details, let us specify our working framework for classification-like case-based reasoning problems. 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 cl denote the class attribute. Moreover, let us denote by $D(a_i)$ and $D(cl)$ the domains of the attributes a_i and cl respectively (so $D(cl)$ is the set of solution classes). Then a case $c \in CB$ will be represented as a pair (d, r), where $d = (a_1(c), \ldots, a_n(c))$ is the set of the problem description values and $r = cl(c)$ is the solution class for the case c. If we write $\mathbf{D} = D(a_1) \times \ldots \times D(a_n)$ (D for description) and $\mathbf{R} = D(cl)$ (R for result), then a case base CB is just a subset of $\mathbf{D} \times \mathbf{R}$.

In this framework, given a case base $CB = \{c_i = (d_i, r_i)\}_{i \in I}$ and a new problem description d_0, the CBR task is to find (guess) a solution class r_0

for d_0, by applying the above general principle in some form, i.e. taking into account the possible similarity of d_0 with already solved cases $c_i \in CB$.

It is clear then that the notion of *similarity* plays a key role in CBR problems. In particular, the notion of graded similarity, which has been used in the framework of fuzzy sets theory for a long time [37], seems specially well suited for them. 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 more they resemble. In particular, $S(w, w') = 1$ means that w and w' are indistinguishable, 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 required to 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 separating, reflexive and symmetric fuzzy binary relations.

In some recent literature a fuzzy set-based approach to case-based reasoning has been developed, not only from a practical point of view (see e.g. [23,7,10,6,12,21,8]) but also from the formal modelling point of view [13,35,14,30,15,16,18,9]. In all these models it is assumed that fuzzy similarity relations S_D and S_R on the domains of problem description and solution attributes, $\mathbf{D}$ and $\mathbf{R}$ respectively, are known and given beforehand. In particular several models have been proposed corresponding to different interpretations of the above CBR principle in terms of constraints between the fuzzy similarity relations S_D and S_R. In this paper, within the class of the so-called non-deterministic models, we tackle the problem of learning a particular type of global similarity measure from the set of precedent cases stored in the base case. Namely, given similarity measures S_a defined on each attribute domain $D(a)$, we show how to determine a weighting vector for each case in order to define a global similarity on $\mathbf{D}$ as a weighted average of the S_a's that minimizes the misclassifications. We check their adequacy by comparing results with an application of case-based reasoning in a medical domain which uses a different model. In that model cases are not attribute-value tuples but they are represented in a relational way. A relational representation describe objects based on their components and the relations between these components. In Machine Learning there is a wide research field, Inductive Logic Programming, focused on relational representation of objects and methods to handle them (see e.g. [28]).

The chapter is organized as follows. In Section 2 we describe several fuzzy CBR models. Then, in Section 3 how to construct fuzzy similarity relations to be used in the fuzzy CBR models. Section 4 explains an alternative ap-

proach using symbolic similarity for comparing relational cases. The results of both approaches are compared in Section 5. The chapter finishes with the conclusions.

2 Fuzzy CBR models

A first class of models try to model a strong form of the case-based reasoning principle which reads:

> "the more similar are the description values (in the sense of S_D),
> the more similar are the class values (in the sense of S_R)".

In [13,14], this principle is modelled by viewing each case $c = (d, r)$ as a *fuzzy gradual rule* of the form

$$\text{If } X \text{ is } approximately\_d \text{ then } Y \text{ is } approximately\_r$$

Here X is a linguistic variable on the domain $\mathbf{D}$ of problem description tuples and Y is a linguistic variable on the domain $\mathbf{R}$ of solution classes, and $approximately\_d$ and $approximately\_r$ denote the fuzzy sets of attribute value tuples close or similar to d and r, and defined on $\mathbf{D}$ and $\mathbf{R}$ respectively by

$$\mu_{approximately\_d}(d') = S_D(d, d'), \quad \mu_{approximately\_r}(r') = S_R(r, r')$$

The semantics of fuzzy gradual rules (see [17]) capture the above intended meaning in the sense that the conditional possibility distribution they induce on $\mathbf{D} \times \mathbf{R}$ is

$$\pi(r' \mid d') \leq \left[\mu_{approximately\_d}(d') \Rightarrow_{\otimes} \mu_{approximately\_r}(r')\right]$$

where $\Rightarrow_{\otimes}$ is the residuum of a continuous t-norm $\otimes$. Such a binary operation on $[0, 1]$ has as fundamental property (independently of $\otimes$) that $\left[x \Rightarrow_{\otimes} y\right] = 1$ iff $x \leq y$. This means that, given a current problem d_0, the best solutions for d_0 which can be inferred from the above rule are those r_0 such that

$$\left[\mu_{approximately\_d}(d_0) \Rightarrow_{\otimes} \mu_{approximately\_r}(r_0)\right] = 1,$$

or equivalently such that

$$S_D(d, d_0) \leq S_R(r, r_0).$$

That is, the best solutions are those r_0 with a similarity to r that is at least as much as the similarity of d_0 to d. So, with this interpretation, neighborhoods around d are transferred to neighborhoods around r. Now, if we consider the whole case base $CB = \{c_i = (d_i, r_i)\}_{i \in I}$, and we perform the same kind

of inference for each case (d_i, r_i), then the global solution set will be the conjunctive aggregation of the individual sets

$$\bigcap_{i \in I} \{r_0 \in \mathbf{R} \mid S_D(d_i, d_0) \leq S_R(r_i, r_0)\}.$$

One problem here is that, in principle, nothing prevents that this intersection can be empty, if the case base CB is not fully consistent with the above principle. In particular this can happen when the case base CB contains cases with very similar description attribute values but with different class values. This is why such a model was called *deterministic* in [13,14].

In [18] alternative interpretations of the same principle were considered, namely by assuming the inequality

$$\text{if} \quad S_D(d_1, d_2) \leq S_D(d_1, d_3) \quad \text{then} \quad S_R(r_1, r_2) \leq S_R(r_1, r_3)$$

(and two other similar ones) to hold for any triple of cases $c_1 = (d_1, r_1), c_2 = (d_2, r_2), c_3 = (d_3, r_3) \in CB$. But still, these are deterministic models in the above sense, i.e $S_D(d_1, d_2) = 1$ implies $S_R(r_1, r_2) = 1$.

Deterministic models can be felt too strong in many real-world domains, since in some sense they assume that the description attributes completely characterize all possible situations. This is not often the case. In contrast the so-called *non-deterministic* models assume a weaker form of general CBR principle, in the sense that "it is only plausible (not necessary) that similar problems have similar solutions".

Expressed in terms of the fuzzy similarity relations S_D and S_R, this weaker Case-Based Reasoning principle can be expressed by the following rule [13,14]:

"The more similar are the problem descriptions in the sense of S_D, the more *possible* the solution classes are similar in the sense of S_R"

This amounts to state that for any case (d, r) in the case base CB, the closer is the current problem d_0 to d, the more plausible r is a solution class for d_0. This can be formalized again as a fuzzy rule

$$\text{If } X \text{ is } \mathit{approximately\_d} \text{ then } Y \text{ is } \mathit{approximately\_r}$$

but with a different semantics, corresponding to the so-called *possibility fuzzy rules* (see [17]). The semantics of such a possibility fuzzy rule is "the more X is $\mathit{approximately\_d}$ (i.e. the closer is a description to d), the more possible (plausible) is $\mathit{approximately\_r}$ a range for Y (i.e. the more plausible are those solutions close to r)". In terms of possibility distributions, the above rule induces the following constraint on the conditional possibility distribution on $\mathbf{D} \times \mathbf{R}$:

$$\pi(y \mid x) \geq \min(\mu_{\mathit{approximately\_d}}(x), \mu_{\mathit{approximately\_r}}(y)).$$

In other words, the plausibility degree of y being a solution for the problem x is lower bounded by the similarity degree between y and r, truncated by

the degree to which x is similar to s. When we consider the whole case base $CB = \{c_i = (d_i, r_i)\}_{i \in I}$, then the joint constraint induced by all fuzzy rules corresponding to the cases in CB is the disjunctive aggregation of the individual ones:

$$\pi(y \mid x) \geq \max_{(d_i, r_i) \in CB} \min(\mu_{approximately\_d_i}(x), \mu_{approximately\_r_i}(y)).$$

Then, given a current description problem d_0, such a joint constraint induces a *plausibility ordering* π_{d_0} of the possible solutions or classes for d_0, by defining

$$\pi_{d_0}(y) = \max_{(d_i, r_i) \in CB} \min(\mu_{approximately\_d_i}(d_0), \mu_{approximately\_r_i}(y)) \\ \max_{(d_i, r_i) \in CB} \min(S_D(d_i, d_0), S_R(r_i, y))$$

for all $y \in \mathbf{R}$, in sense that the greater is $\pi_{d_0}(y)$ the more plausible is y a solution for d_0. Sometimes, the plausibility values in themselves are not particularly important but only the ordering they induce on the set $\mathbf{R}$ of solution classes:

$$y \preceq_{d_0} y' \text{ iff } \pi_{d_0}(y) \leq \pi_{d_0}(y').$$

Finally, notice that if new cases are added to the case base CB, $\pi_{d_0}(y)$ can only increase, never decrease, according to the idea that new cases may incorporate new solutions but not discard old ones. Non-deterministic models have been also considered for instance in [22,10].

3 Learning similarity relations for a non-deterministic CBR model

As already mentioned in the Introduction, in this chapter we are concerned with the learning aspects of similarities within a non-deterministic fuzzy set-approach to CBR, with some particular choices. In this section we first set our working assumptions for the model, and then we describe how the relevant similarity relations can be constructed from the case base.

3.1 Working assumptions and relevant similarity relations

Background knowledge is assumed for each attribute $a \in \mathbf{A}$ under the form of a similarity relation S_a on $D(a)$, the domain of a. This similarity is a function $S_a : D(a) \times D(a) \to [0, 1]$. Additionally, we assume a similarity relation S_{cl} defined over the set of classes. This similarity was denoted by S_R in the previous section but from now on we use S_{cl} to stress the fact that is defined over the set of classes. This corresponds to a function $S_{cl} : D(Cl) \times D(Cl) \to [0, 1]$.

The goal is to define a similarity relation S_D between the cases 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. Of course, then we need additional information to assess the relevance of each attribute for retrieving a particular case. In particular, we shall assume there is a *weighting vector for each case* that evaluates the importance of each attribute when computing the similarity between this case and another arbitrary case description in the case base CB. This is formalized in the following definition.

Definition 1. Let $CB \subset \mathbf{D} \times \mathbf{R}$ be a case base, let $\mathbf{A}$ be the set of attributes considered in $\mathbf{D}$ and, for each $a \in \mathbf{A}$, let S_a be the corresponding similarity relation on $D(a)$. Finally, let w_c be the weighting vector[1] attached to a particular case $c \in CB$. Then, the similarity relation defined between an arbitrary case description $d \in \mathbf{D}$ and c is defined as follows:

$$S_D(c, d) = \sum_{a \in \mathbf{A}} w_c(a) \cdot S_a(a(c), a(d)) \tag{1}$$

As $S_a(\cdot, \cdot) \in [0, 1]$ and as w_{cb} is a weighting vector, it follows that $S(\cdot, \cdot)$ is a function into the $[0, 1]$ interval as well.

Note that, formally, S_D so defined is a function on $CB \times \mathbf{D}$, although from cases in CB we only make use of their description attribute values. Also, this definition is not commutative in the sense that it is the case c who determines the weighting vector (the weighting vectors may be different for two different cases), so the first argument in S_D plays a preeminent role.

Once we have defined the similarity relation S_D, we can define how close is a problem description to a solution class just by comparing it to every case in CB that shares that solution class.

Definition 2. Let CB, $\mathbf{D}$, $\mathbf{A}$ and S_D be defined as in Definition 1. Then, the similarity between a case description $d \in \mathbf{D}$ and a solution class $r \in D(cl)$ is defined as follows:

$$SCL(d, r) = \max_{c \in \mathbf{CB}, cl(c) = r} S_D(c, d).$$

In turn, from this definition, we can define a similarity between any arbitrary pair of case descriptions as a kind of transitive closure of the SCL function as follows.

Definition 3. Let CB, $\mathbf{D}$, $\mathbf{A}$, S_D and SCL be defined as in Definition 2. Let $\otimes$ be a continuous t-norm. Then, the $\otimes$-similarity between two arbitrary set of case descriptions $d_1, d_2 \in \mathbf{D}$ is defined as follows:

$$SCC(d_1, d_2) = \sum_{r \in D(cl)} SCL(d_1, r) \otimes SCL(d_2, r).$$

Among others, possible choices for $\otimes$ are e.g. $\otimes = \cdot$ or $\otimes = \min$.

[1] $w = (w_1, \cdots, w_n)$ is a weighting vector of dimension n iff $w_i \geq 0$ for all $i = 1, ..., n$, and $\sum_{i=1,n} w_i = 1$.

3.2 Construction of the numerical similarity functions S_D

Definition 1 formalizes the similarity $S_D(c, d)$ between an arbitrary problem description d and a case c in the base case CB. Recall that this similarity has been defined as the weighted mean of the attribute similarities S_a for a in $\mathbf{A}$. However, expression (1) makes explicit that the function is not symmetric in the sense that the two arguments are not equally important. Instead, the role of the case c is preeminent because we make the weighting vector w_c depending on c. Next, we describe how to learn the appropriate weights for each $c \in CB$.

In the following we will assume that a case $cb \in CB$ is known and fixed along the learning process. In fact, the same process we describe below for cb will be applied for each case in CB. Naturally, for each case cb, the process would lead to its corresponding weighting vector w_{cb}.

To compute the vector, besides of the case cb, we need to fix a subset of the case CB, i.e. a collection of problem descriptions for which their solution class is known. This is the learning set and is denoted by LS. Of course we shall also make use of the similarities S_a for each attribute $a \in \mathbf{A}$.

Then, the weights determination can be formulated in the following way:

Problem 1 (Weight Determination Problem). Let LS be the learning set and let $cb \in CB$ be a case in the case base. Then the weight determination problem is to find the set of weights w_{cb} in Expression (1) such that, for each $c = (d, r) \in LS$, the similarity between cb and d, $S_D(cb, d)$, approximates as much as possible the similarity between the solution classes $cl(cb)$ and $cl(c) = r$, $S_{cl}(cl(cb), r)$.

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

Problem 2. Let $LS = \{c_j\}_{j \in J}$ be the learning set and let cb be a case in the case base. Then the weight determination problem relative to cb is to find the set of weights $w_{cb}(a_1), ..., w_{cb}(a_n)$ that minimizes the following expression:

$$\sum_{c_j \in LS} \Big(\sum_{a_i \in \mathbf{A}} w_{cb}(a_i) \cdot S_{a_i}(a_i(c_j), a_i(cb)) - S_{cl}(cl(c_j), cl(cb)) \Big)^2$$

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

Now, we introduce simplified terms for the similarities so that the problem can be further simplified. The notation we will use is as follows:

- $a_i^j = S_{a_i}(a_i(c_j), a_i(cb))$ is the similarity between the values corresponding to i-th attribute in $\mathbf{A}$ for the case cb and the j-th case in LS.

S_{a_1}	S_{a_2}	$\cdots$	$S_{a_{\lvert\mathbf{A}\rvert}}$	S_{Cl}
a_1^1	a_2^1	$\cdots$	$a_{\lvert\mathbf{A}\rvert}^1$	b^1
a_1^2	a_2^2	$\cdots$	$a_{\lvert\mathbf{A}\rvert}^2$	b^2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$a_1^{\lvert LS\rvert}$	$a_2^{\lvert LS\rvert}$	$\cdots$	$a_{\lvert\mathbf{A}\rvert}^{\lvert LS\rvert}$	$b^{\lvert LS\rvert}$

Table 1. Structure of the available data for finding the weighting vector

- $b^j = S_{cl}(cl(c_j), cl(cb))$ is the similarity between the solution classes of cb and the j-th case in LS.
- $w_i = w_{cb}(a_i)$ corresponds to the weight of the attribute a_i.

In this way, the data available to the problem has the form of Table 1 and the weights determination problem corresponds to:

Problem 3.

$$\begin{aligned} &Minimize\ \sum_{j=1}^{|LS|}\left(\sum_{i=1}^{|\mathbf{A}|} w_i a_i^j - b^j\right)^2 \\ &Subject\ to: \\ &\quad - \ \sum_{i=1}^{|A|} w_i = 1 \\ &\quad - \ w_i \geq 0, \text{ for all } i = 1, ..., |A| \end{aligned}$$

The minimization problem formulated above has been studied in [19], [32] and [33] in the general framework of parameter learning for aggregation operators. [19] introduced a method based on the gradient descent for learning the weights of the OWA operator [36]. The similarities between the OWA and the Weighted Mean make the method equally suitable for the Weighted Mean. In [32] an algorithm based on Active Set Methods was studied and applied to some medium-size data sets (problems up to 34 attributes were considered). Results reported were positively compared with the ones in [19]. More recently, [33] studied some issues left open in [32]. In particular, the work studied the situation with linearly dependent attributes. Then, an algorithm was introduced to deal with such situation. In this work we apply the algorithms introduced in [32] with the extension described in [33]. Related results about parameter learning for aggregation operators include [20], [26] and [31]. By the way, these latter results have been defined for learning parameters of more complex operators (e.g., the Choquet integral).

Methods based on active sets (see e.g. [29]) rely on the simplicity of computing the solution of quadratic problems with linear equality constraints. Iterative algorithms have been developed in which at each step inequality constraints are partitioned into two groups: those that are to be treated as active (considered as equality constraints) and those as inactive (essentially ignored). Once a partition is known, the algorithm proceeds, by moving on

the surface defined by the working set of constraints (the set of active constraints), to an improved point. In this movement some constraints are added to the working set and some others are removed. This process is repeated until the minimum is reached. When the function to minimize is convex (this is the case for the weighted mean) the method finds the minimum and, although the method is iterative, the final minimum is not influenced by the initial weight vector.

However, an important problem arises for the practical application of these methods: some data can have linearly dependent columns (in our case, this corresponds to non-independent attributes). In this case, the algorithm fails to give a solution. In fact the problem arises only, as shown in [33], when there is a column/attribute a_i that can be written as a linear combination of the others of the form $a_i = \sum_j p_j a_j$ in such a way that $\sum_j p_j = 1$. In such 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.

4 A symbolic similarity approach to relational CBR

Reasoning and learning from cases is based on the concept of similarity, this is clear, but there exist several ways of modelling similarity. Numerical (or fuzzy) similarity-based approaches to case retrieval, as the ones we have considered in the previous sections, are mainly used for cases represented as attribute-value vectors. Instead, in this section we consider another approach where cases are represented in a scheme that uses relations among entities. We will call this setting *relational case-based learning*. In this approach, the similarity between two cases is understood as what they "share". But, in addition, we need to be able to evaluate whether what they share is "relevant" or "important" (or to which degree is relevant or important) for the problem at hand. In this section we first introduce the concepts of symbolic similarity and feature terms, the basis of the formalism we will use for representing relational cases. Then, we present LID (for Lazy Induction of Descriptions), a method for relational case-based learning. LID, introduced in [5], is based on two main notions: 1) the similarity is constructed as a symbolic description of what is shared between precedent cases and a specific *problem* to be classified, and 2) there is some assessment function to help the system decide which relations among entities are important or relevant to be shared with the precedent cases.

4.1 Symbolic similarity and feature terms

In real domains it can be useful to represent the knowledge in a structured way. As a means to do so, in [4] the authors introduced the so-called *feature terms* (also called feature structures or ψ-terms) that are a generalization of first order terms [2,11]. Formally, we defined a feature term as follows:

Definition 4. Given a signature $\Sigma = \langle \mathcal{S}, A, \leq \rangle$, where $\mathcal{S}$ is a set of sort symbols, $\mathbf{A}$ is a set of attribute symbols and $\leq$ is a decidable partial order on $\mathcal{S}$, and a set ϑ of variables, we define a *feature term* as an expression of the form:

$$\psi ::= X : s[a_1 \doteq \Psi_1 \ldots a_n \doteq \Psi_n] \tag{2}$$

where X is a variable in ϑ called the *root* of the feature term, s is a sort in $\mathcal{S}$, $a_1 \ldots a_n$ are attributes in $\mathbf{A}$, and each Ψ_i is a set of feature terms and variables. When $n = 0$ we are defining a variable without features. The set of variables occurring in ψ is noted as ϑ_ψ.

Figure 1 shows the description of a diabetic patient using feature terms. The patient *patient-371* is a feature term of sort *patient* (sorts are underlined in the figure). This patient is described by two attributes, *diabetes-data* and *consultation*. The value of *diabetes-data* is a feature term of sort *diab-data* that has, in turn two attributes, *dm-type* and *dm-year*.

Definition 5. A path $path(X, a_i)$ is defined as a sequence of attributes going from the variable X to the feature a_i.

Some examples of paths in the description of figure 1 are the following:

- $path(patient\text{-}371, retinopathy)$ = *patient-371*.consultation.eye-exam.visibility.retinopathy
- $path(patient\text{-}371, dyslipidaemia)$ = *patient-371*.consultation.additional-treatment.dyslipidaemia
- $path(consultation, visibility)$ = consultation.eye-exam.visibility

Sorts have an informational order relation ($\leq$) among them, where $s \leq s'$ means that s is more general than s' (or s' has more information than s), we also say then that s' is a subsort of s. When a feature has *unknown* value it is represented as having the sort *any*. All other sorts are more specific than *any*.

There is an ordering relation, called *subsumption*, among feature terms defined as follows:

Definition 6. A feature term ψ *subsumes* another feature term ψ', written $\psi \sqsubseteq \psi'$, when the following conditions are satisfied: 1) the sort of ψ' is either the same or a subsort of ψ, 2) if A_ψ is the set of attributes of ψ and $A_{\psi'}$ is the set of attributes of ψ' then $A_\psi \subseteq A_{\psi'}$ and 3) the feature terms of values of the attributes in A_ψ and $A_{\psi'}$ satisfy in turn the two conditions above.

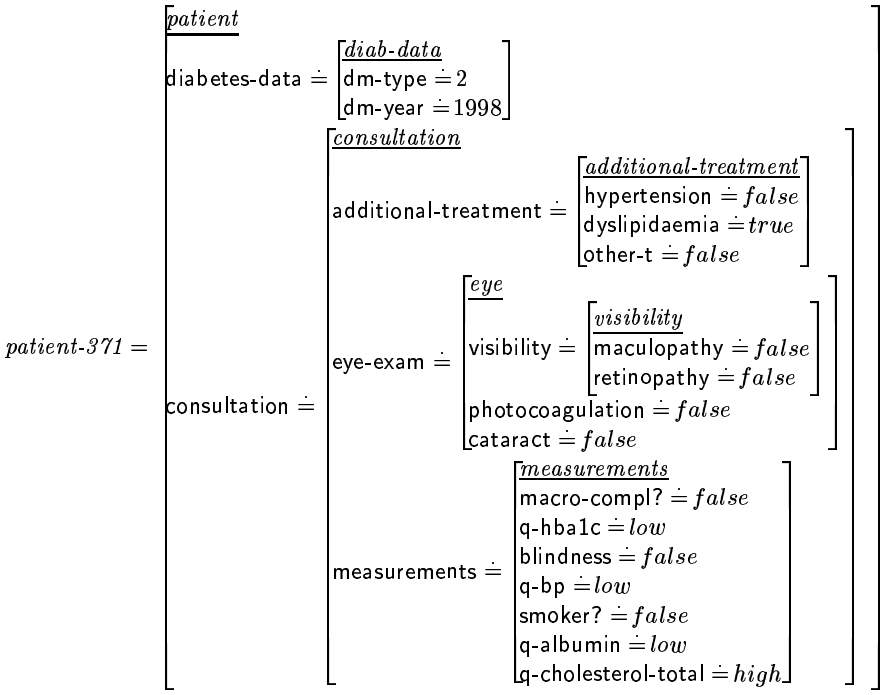

Fig. 1. Partial description of a diabetic patient. Here only 22 attributes are shown out of a total of more than 80 paths present in the complete description of the patient.

Intuitively, a feature term ψ subsumes ψ' when all information in ψ is also contained in ψ'.

Definition 7. A feature term S_t is a *similarity term* of two cases c_1 and c_2 if and only if $S_t \sqsubseteq c_1$ and $S_t \sqsubseteq c_2$, i.e. iff the similarity term of two cases subsumes both cases. The set Δ_S of cases subsumed by the similarity term S_t is called the *discriminatory set* associated with S_t.

As we will see in Section 4.2, a similarity term can be seen as a symbolic similarity among cases. In short, given a problem description d, LID classifies it as belonging to the solution class $r \in \mathbf{R}$ when all the cases in the discriminatory set Δ_S associated with the similarity term S_t have r as solution class. This assumption can be made because LID builds the similarity term taking into account the attributes that are more relevant for classifying the problem d. Intuitively, LID considers that if the problem description d shares a set of relevant attributes (those in the similarity term S_t) with a (sub)set of cases belonging all them to the same solution class r, then d can

```
Function LID(d, S_t, Δ_S, R)
    if stopping-condition(Δ_S)
        then return Class(Δ_S)
        else a_d := Select-attribute (d, Δ_S, R)
             S'_t := Add-path(path(d, a_d), S_t)
             Δ_S' := Discriminatory-set (S'_t, Δ_S)
             LID(d, S'_t, Δ_S', R)
    end-if
end-function
```

Fig. 2. The LID algorithm. S_t is the similarity term, Δ_S is the discriminatory set associated with S_t, R is the set of solution classes, $class(\Delta_S)$ is the class $r_i \in R$ to which belong all the elements in Δ_S.

also be classified as belonging to r. In Section 4.2 we explain how to select the relevant attributes of a case to form the similarity term.

4.2 The LID Method

In this section we describe the method LID, useful for relational case-based learning. LID combines the notion of symbolic similarity with an heuristic that measures the importance of an attribute as the distance between the partition induced by the values of this attribute and the correct one. The criterion used in LID is based on the discrimination power of the similarity term which is evaluated using the so-called RLM distance [25].

The main steps of the LID algorithm are shown in Figure 2. Input parameters of the LID algorithm are a problem description d to be classified, a similarity term S_t, the discriminatory set Δ_S associated with S_t and the set of solution classes $\mathbf{R} = D(cl)$ where the problem d can be classified. In the first call $LID(d, S_t, \Delta_S, \mathbf{R})$, the similarity term S_t is initialized to the most general description (the one having no features) and the set Δ_S is initialized to CB, the whole case base.

The stopping condition of LID is when all the cases in the discriminatory set Δ_S belong to only one solution class $r_0 \in R$. In such a situation LID gives the similarity term S_t as an explanation of the classification of d in r_0 and Δ_S is the support set justifying that result[2]. The similarity term S_t can be viewed as a *partial* description of the solution class r_0 because it contains a subset of attributes that are discriminating enough to classify a case as belonging to r_0. Notice however that S_t is not the most general "generalization" of r_0,

[2] There is an *abnormal* stopping condition when there cases in Δ_S belonging to different solution classes but the similarity term S_t cannot be further specialized. In such a case, the output of LID is not a single class but the set of solution classes of the cases in Δ_S.

since in general S_t does not subsume all the cases belonging to r_0 but only a subset of them (those sharing the attributes of S_t with the new problem). The similarity term S_t depends on the current problem, for this reason there are several partial descriptions (i.e. similarity terms) for the same solution class.

In the first call of LID the stopping condition is not satisfied since Δ_S contains all the cases in CB. This means that the similarity term S_t is satisfied by cases belonging to several solution classes, therefore S_t has to be specialized.

The specialization of a similarity term S_t is achieved by adding attributes to it. LID considers as candidates to specialize S_t only those attributes that are present in the problem description d to be classified. Let A_l be the set of attributes in d with no *unknown* value.

The next step of LID is the selection of an attribute $a_d \in A_l$ to specialize the similarity term S_t. The selection of the most discriminatory feature in the set A_l is heuristically done using the RLM distance over the features in A_l.

The RLM distance assesses how similar are two partitions (in the sense that the smaller the distance is the more similar they are). Each attribute $a_i \in A$ induces a partition $\mathcal{P}_i = \{sp_{i1}, \ldots, sp_{in}\}$ of the case-base, where the classes sp_{ij} are formed by those cases having the same value for the attribute a_i. The *correct partition* is the partition $\mathcal{P}_c = \{\bar{r}_1, \ldots, \bar{r}_m\}$, where the classes correspond to the solution classes, i.e. the cases belonging to $\bar{r}_i$ are those with the same solution class r_i. For each partition $\mathcal{P}_i$ induced by an attribute a_i, LID computes its RLM distance to the correct partition $\mathcal{P}_c$. The proximity of the partition $\mathcal{P}_i$ to $\mathcal{P}_c$ estimates the relevance of feature a_i.

Definition 8. Given two partitions $\mathcal{P}_i = \{sp_{i1}, \ldots, sp_{in}\}$ and $\mathcal{P}_c = \{\bar{r}_1 \ldots \bar{r}_m\}$ of the case base CB, the RLM distance between them is computed as follows:

$$RLM(\mathcal{P}_i, \mathcal{P}_c) = 2 - \frac{I(\mathcal{P}_i) + I(\mathcal{P}_c)}{I(\mathcal{P}_i \cap \mathcal{P}_c)}$$

with

$$I(\mathcal{P}_i) = -\sum_{j=1}^{n_i} p_j \cdot log_2 p_j \quad ; \quad p_j = \frac{|CB \cap sp_{ij}|}{|CB|}$$

$$I(\mathcal{P}_c) = -\sum_{k=1}^{m} p_k \cdot log_2 p_k \quad ; \quad p_k = \frac{|CB \cap \bar{r}_k|}{|CB|}$$

$$I(\mathcal{P}_i \cap \mathcal{P}_c) = -\sum_{j=1}^{n_i} \sum_{k=1}^{m} p_{jk} \cdot log_2 p_{jk} \quad ; \quad p_{jk} = \frac{|CB \cap \bar{r}_k \cap sp_{ij}|}{|CB|}$$

where $I(\mathcal{P}_i)$ measures the information contained in the partition $\mathcal{P}_i$; n_i is the number of possible values of the attribute inducing $\mathcal{P}_i$; p_j (p_k resp.) is the probability of occurrence of the class sp_{ij} ($\bar{r}_k$ resp.), i.e. the proportion of examples in CB that belong to sp_{ij} ($\bar{r}_k$ resp.); $m = |\mathcal{P}_c|$, i.e. the number of solution classes; $I(\mathcal{P}_i \cap \mathcal{P}_c)$ is the mutual information of the two partitions; and p_{jk} is the probability of the intersection $\bar{r}_j \cap sp_{ik}$, i.e. the proportion of examples in CB that belong to $\bar{r}_j$ and to sp_{ik}. In this definition, as it is common in the case of entropy, $0 \log 0$ is defined as zero.

Definition 9. Let $\mathcal{P}_i$ and $\mathcal{P}_j$ the partitions induced by the attributes a_i and a_j respectively. We say that a_i is *more discriminatory than* a_j iff $RLM(\mathcal{P}_i, \mathcal{P}_c) < RLM(\mathcal{P}_j, \mathcal{P}_c)$, i.e. when the partition induced by a_i is closer to the correct partition $\mathcal{P}_c$ than the partition induced by a_j.

LID uses the *more discriminatory than* relationship to estimate the attributes that are more relevant for the purpose of classifying a current problem. Let us call a_d the most discriminatory attribute in $\mathbf{A}$, i.e. a_d induces the partition of the discriminatory set closest to the correct partition.

The specialization step of LID defines a new similarity term S'_t by adding to the current similarity term S_t the sequence of attributes specified by $path(d, a_d)$. After this addition S'_t has a new path with all the attributes in the path taking the same value they take in d. After adding the path $path(d, a_d)$ to S_t, the new similarity term $S'_t = S_t + path(d, a_d)$ subsumes a subset of cases in Δ_S, namely the discriminatory set $\Delta_{S'}$.

Next, LID is recursively called with the similarity term S'_t and the discriminatory set $\Delta_{S'}$. The recursive call of LID has $\Delta_{S'}$ as parameter (instead of Δ_S) because the cases that are not subsumed by S'_t will not be subsumed by any further specialization. The process of specialization reduces the discriminatory set $\Delta_{S^n} \subseteq \Delta_{S^{n-1}} \subseteq \ldots \subseteq \Delta_{S'} \subseteq \Delta_S$ at each step. This specialization of the discriminatory set will provoke the algorithm to terminate.

5 Application

In this section we describe the results obtained when applying the non-deterministic fuzzy CBR approach, with the fuzzy similarity learning method described in Section 3, into a medical domain (in a Diabetis related problem). Similarity relations have been built for a set of cases and results are reported and compared with those obtained by the LID method. Additionally, we describe the classification results obtained when using the *leave one out* approach. The section starts with the description of the problem and follows with the description and analysis of the experiments.

5.1 The Diabetes Domain

Diabetes Mellitus is one of the most common human chronic diseases. There are two major types of diabetes: diabetes type I (or *insulin-dependent*) and

diabetes type II (or *noninsulin-dependent*). The diabetes type I is usually found in people under 40 years and is the consequence of a pancreatic malfunction. Instead, diabetes type II is more frequent in aged people and it could be explained as a loss of effectivity of the insulin on the body cells. In fact, both forms of diabetes produce the same short-term symptoms (i.e. frequent urination, increase of thirst and high blood glucose values) and long-term complications (i.e. blindness, renal failure, gangrene, coronary heart disease and stroke). The main concern in the management of diabetes is to reduce the risk of the patient to develop long-term complications. In 1989, representatives of Government Health Departments, patient organisations and diabetes experts celebrated a meeting in Saint Vincent (Italy) (http://www.show.scot.nhs.uk/crag/topics/diabetes/vincent.htm) to elaborate some recommendations to be followed in the diabetes management with the goal of minimizing the diabetic complications of the patients.

Since then, experts have analyzed a lot of data that allowed to define the main parameters (attributes) that prevent the development of complications. In particular, they found that keeping the analytical data of the patient as close as possible to the "normal" ranges, clearly reduces the complication risk. In other words, the patient has to modify some of his life habits (like diet, physical exercise, alcohol ingestion or smoking habits) in order to maintain analytical parameters, such as the cholesterol, the blood pressure or the creatinine, within the same ranges than those of a healthy person. Furthermore, it is necessary a strict eye and foot control to prevent the development of a retinopathy and of a polyneuropathy respectively.

In this work we have considered the assessment of infarct risk of a diabetic patient. Four solution classes are considered: *low, moderate, high* and *very-high*. The assessment of the risk can be seen as a classification task where the goal is to identify the solution class to which the new problem belongs. In the next section we describe the process of building a similarity function to determine the attributes that are more important in order to assess the infarct risk of a diabetic patient. An evaluation of the built similarity function is also considered.

5.2 Empirical description and analysis

As briefly described above, the procedure explained in Section 3.2 has been applied to find similarities between cases. We have considered a case base CB with 30 cases describing patient analytical data and patient's risk assessment (real analytical data was supplied by the Mataró Hospital and the risk of each patient was assessed by a physician). Each case was described in terms of 84 attributes.

Then, the procedure outlined in Section 3.2 for the determination of the weighting vectors has been applied. In this way, we have learned a weighting vector for each case $cb \in CB$. To do so, we have used as learning set the whole CB after removing the case cb. This is, for each $cb \in CB$, $LS_{cb} = CB - \{cb\}$.

The learning stage lead to 30 weighting vectors, each one of dimension 84. This process required first the definition of similarity functions S_{cl} and S_a for all $a \in \mathbf{A}$.

Following the process explained in Section 3.2 we built a matrix similar, in structure, to the one in Table 1. It has to be mentioned that at that point, and before applying the learning algorithm, we considered to run the experiment with both the original data and normalized data. In the second case, each variable was normalized in [0,1] defining

$$a'^i_j = \frac{(a^i_j - min_k a^i_k)}{(max_k a^i_k - min_k a^i_k)} .$$

The normalization was considered because the similarity range of some variables never reached 0 or 1 for the cases considered in the learning set.

Linear dependency problems raised in the computation of the weights for 15 of the cases in CB (for both normalized and non-normalized data). In these cases, we proceeded, as explained in Section 3.2, to remove one dependent attribute at a time and computing the weights for all these subproblems. For all of these subproblems, the best weighting vector was selected. Nevertheless, some of the problems still had linear dependent attributes. For instance, the 6-th and 7-th cases had more than two dependent columns when normalized data was considered, and the same occurred for the 7-th case with the non-normalized data. No additional treatment was applied to these cases and thus the best weighting vector selected was in fact not relevant.

Results show that most weights are almost zero and that only a few of them are significantly greater than zero. This is analogous to the procedure adopted by the physician, who only considers a few number of attributes. Therefore, the method is appropriate for selecting relevant attributes and for disregarding the other ones. Table 2 displays the number of relevant weights for all the cases. Two thresholds have been used in this table for considering an attribute to be relevant: 0.1 and 0.05. The table shows that the number of relevant attributes found are similar with normalized and non-normalized data.

A detailed analysis of the relevant attributes shows that, for several cases, the set of relevant attributes is similar in both CBR approaches, the numerical fuzzy similarity approach and the LID method. For example, LID considers the attribute *smoke* as relevant for 12 cases in the case base. For 8 out of these 12 cases the corresponding weighting vectors assign relevant weights to $w_{cb}(smoke)$. Table 3 reviews these cases and provides the relative position of the corresponding weights in the weighting vectors obtained for the same cases. Similarly, LID finds as relevant the attributes $Chol - total$ (10 cases), $HbA1c$ (4 cases), TG (1 case) and $Edu - member$ (1 case). These attributes have also non-zero values in the corresponding weights (5 cases for the $Chol - total$, 4 for $HbA1c$, 1 case for both the TG and $Edu - member$).

The study also shows that there are some cases when this correspondence between attributes in both approaches is not straightforward. Nevertheless,

some of the divergences between both methods are due to the presence of similar or related attributes in both descriptions. For example, the attribute $food - education$ is considered related with another attribute that informs whether the patient goes on some diet or belongs to an educational association.

Threshold	Number of relevant weights $w_{cb}(a_i)$ for all 30 cases in CB
0.05	9 7 7 6 7 8 0 7 8 7 6 7 8 8 8 8 6 8 8 8 4 8 8 7 6 8 6 8 8 8
0.1	2 3 4 4 2 3 0 2 3 5 2 1 1 3 3 2 3 2 3 3 3 2 2 3 3 3 3 1 2 2
0.05	7 7 7 7 7 0 0 7 6 8 6 6 7 8 7 8 5 7 7 5 4 8 7 6 6 7 6 7 8 8
0.1	2 3 4 4 4 0 0 2 3 3 1 2 2 3 3 3 4 2 2 3 3 2 2 3 3 3 3 3 3 2

Table 2. Number of relevant weights considering two thresholds (0.05 and 0.1): non normalized data (top) and normalized data (bottom). Each column corresponds to one case from CB.

Case where smoke was relevant	Relative position of the weight in normalized data	Relative position of the weight in non-normalized data
1	4-th	3-rd
2	4-th	4-th
6	4-th	n.r.
7	n.r.	n.r.
9	n.r.	n.r.
16	4-th	4-th
17	4-th	4-th
18	4-th	4-th
19	n.r.	n.r.
26	5-th	n.r.

Table 3. Cases where attribute "smoke" was relevant according to the LID method and the relative position of the corresponding weight in the weighting vectors obtained for the same cases. (n.r. stands for "not relevant" weight - i.e., too small weight)

Additionally, to further evaluate the fuzzy similarity CBR approach, we have considered the classification problem using the previous "diabetes" case base CB. The evaluation has been based on the *leave one out* approach. This is, each time we have removed one case from the case base and we have computed the similarity of this case with all the remaining cases in the case base. Table 4 displays the number of correct classifications using this approach. As it can be seen, a little less than half of the cases were correctly classified and for the misclassified ones, the method assigned the

nearest solution class. Here, the nearest solution class is understood in terms of the similarity S_{cl} on the set $\mathbf{R}$ of possible classes. Although the totally correct results in the classification process are not as good as expected, they are indeed quite acceptable as soon as we take into account that, as a matter of fact, we introduced a similarity relation on the solution classes, so not all classes different from the correct one are equally bad. In terms of the plausibility orderings mentioned in Section 2, Table 4 shows that the one of two more plausible solutions is (almost) always is the correct one.

From the case base perspective, the difficulty on correctly classifying the cases can be explained in terms of a lack of redundancy in the case base. It is important to remember that the case base consisted of only 30 cases in a 84-dimensional space (84 attributes). Therefore, the space of possible problem descriptions is very large, and according to the results, most cases are highly unique in the case base. In fact, LID yields similar results when evaluated using also the *leave one out* approach. For the same case base, 4 cases were incorrectly classified, while for the rest of the cases LID always produced two answers, the correct class and one of its nearest classes.

	Correct class	Nearest class	Other classes
no normalization	13	16	1
normalization	13	17	0

Table 4. Assignments obtained in the classification problem

6 Conclusions and future work

In this chapter, within a fuzzy CBR model, we have described a method to construct a particular type of global similarity measure in the problem description domain. Namely, given a case base CB, the similarity between a problem description d and a case $c \in CB$ is defined in terms of a weighted average of the attribute similarities, where the weights are particular to each c. We have applied the approach in a medical domain and compared with another case-based reasoning method, LID, which uses a relational approach with feature terms as a kind of symbolic similarity between cases. The comparison results shows the suitability of the approach.

In particular, results show that only a few attributes are relevant for each case (this is analogous to the procedure adopted by physicians) and that these attributes correspond, in general, to the ones selected by the alternative case-based reasoning method. Classification performance of both approaches also lead to similar results.

Future extensions of this work include the study of additional types of global similarity measures (e.g., using other aggregation operators instead of

the weighted mean in Equation 1). The consideration of non-linear models for defining the similarity relations relate this work with Kernel functions and non-linear Support Vector Machines [34] that, as it is known, is a classification method that operates mapping data into an alternative and higher dimensional space so that classes are linearly separable. In our case, the consideration of such transformations would probably increase the performance of the system when remaining stick in the use of weighted average based similarity. This process would be analogue to the one described in [27] for using the Fuzzy C-Means clustering method in the new high dimensional space.

Acknowledgments

The authors are partially supported by the EU project IBROW (IST-1999-2005) and the CICYT projects STREAMOBILE (TIC2001-0633-C03-02) and e-INSTITUTOR (TIC2000-1414). The authors thank Dr. Albert Palaudàries for his assistance in developing the diabetes application.

References

1. Aamodt, A., Plaza, E., (1994), Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, AI Communications 7 39–59
2. Aït-Kaci, H., Podelski, A., (1993), Towards a Meaning of LIFE, J. Logic Programming 16 195–234.
3. Armengol, E., Palaudaries, A., Plaza, E., (2001), Individual Prognosis of diabetes long-term risks: A CBR approach, Methods of Information in Medicine (special issue on prognosis models in medicine: AI and Statistics) 40 46–51.
4. Armengol, E., Plaza, E., (2000), Bottom-up Induction of Feature Terms, Machine Learning 41 259–294.
5. Armengol, E., Plaza, E., (2001), Lazy Induction of Descriptions for Relational Case-based Learning, Machine Learning: ECML-2001, Springer-Verlag, Lecture Notes in Artificial Intelligence 2167 13-24.
6. Bonissone, P., Ayub, S., (1994), Representing cases and rules in plausible reasoning systems, Proc. of the 1994 ARPA/RL Planning initiative, (Tucson, AZ), 305–316.
7. Bonissone P., Cheetman W. (1997), Applications of fuzzy case-based reasoning to residential property valuation. Proc of the 6th IEEE Int. Conference on Fuzzy Systems, FUZZ-IEEE'97, Barcelona (Spain), pp. 37–44.
8. Bonissone, P., López de Mántaras, Fuzzy Case-Based Reasoning Systems, in Handbook of Fuzzy Computing, Ruspini, E., Bonissone, P. P., Pedrycz, W., Eds. IOS Publishing Ltd., F 4.3: 1–17.
9. Burkhard, H.-D., Richter, M., (2001), On the notion of similarity of Case Based Reasoning and fuzzy theory, in Soft Computing in Case Base Reasoning, S. K. Pal, T. S. Dillon, D. S. Yeung, Eds., Springer, pp . 29-46.
10. de Calmès, M., Dubois, D., Hüllermeier, E., Prade, H., Sèdes, F., (2002), Case-based querying and prediction: A fuzzy set approach. Proc of the IEEE International Conference on Fuzzy Systems, Hawaii (USA), pp. 735–740.

11. Carpenter, B., (1992), The Logic of typed Feature Structures, Tracts in theoretical Computer Science, Cambridge University Press, Cambridge, UK.
12. Cheetham, B., Cuddihy, P., Goebel, K., Applications of Soft CBR at General Electric, in Soft Computing in Case Base Reasoning, S. K. Pal, T. S. Dillon, D. S. Yeung, Eds., Springer, pp 335-365.
13. Dubois, D., Esteva, F., Garcia, P., Godo, L., Lòpez de Màntaras, R., Prade, H., (1997), Fuzzy modelling of case-based reasoning and decision. Proc. of the 2nd International Conference on Case-Based Reasoning (ICCBR-97), Lecture Notes in Artificial Intelligence 1266 599–610.
14. 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.
15. Dubois D., Hüllermeier E., Prade H. (2000), Formalizing case-based inference using fuzzy rules. In S.K. Pal, D.Y. So, and T. Dillon, editors, Soft Computing in Case-Based Reasoning. Springer-Verlag, 47–72.
16. Dubois D., Hüllermeier E., Prade H. (2000), Flexible Control of Case-Based Prediction in the Framework of Possibility Theory. Proc. of EWCBR'2000, (E. Blanzieri et L. Portinali eds.), Lecture Notes in Computer Science 1898 61–73
17. Dubois D., Prade H. (1996), What are fuzzy rules and how to use them, Fuzzy Sets and Systems 84 169-185.
18. Esteva, F., Garcia, P., Godo, L., (2002), Fuzzy Similarity-based Models in Case-based Reasoning. Proc. of 11th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2002), Hawaii (USA), pp. 1348–1353.
19. Filev. D. P., Yager, R. R., (1998), On the issue of obtaining OWA operator weights, FSS 94 157-169.
20. Grabisch, M., Nguyen, H. T., Walker, E. A., (1995), Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference, Kluwer Academic Publishers, Dordreht, The Netherlands.
21. Hansem, B., Riordan, D., (1998), Fuzzy case-based prediction of ceiling and visibility, 1st Conference on Artificial Intelligence, American Metereological Society, pp. 118-123.
22. Hüllermeier E., Dubois D., Prade H. (2000). Knowledge based extrapolation of cases: a possibilistic approach. Proc. IPMU'2000, pp. 1575–1582.
23. Jaczynski M., Trousse B., (1994), Fuzzy logic for the retrieval step of a case-based reasoner. Proc. of the EWCBR'94, 313–321.
24. Kolodner J. (1993) Case-Based Reasoning. Morgan Kaufmann, San Mateo, CA.
25. López de Mántaras, R., (1991), A distance-based attribute selection measure for decision tree induction, Machine Learning 6 81–92.
26. Marichal, J.-L., Roubens, M., (1999), Determination of weights of Interacting Criteria from a Reference Set, Papiers de Recherche, Faculté d'Economie de Gestion et de Sciences Sociales, Groupe d'Etude des Mathematiques du Management et de l'Economie, N. 9909.
27. Miyamoto, S., Suizu, D., (2002), Fuzzy c-Means clustering using transformations into high dimensional spaces, Proc. SCIS&ISIS conference (CD-ROM), Tsukuba, Japan.
28. Muggleton, S., De Raedt, L., (1994), Inductive Logic Programming: Theory and Methods, Journal of Logic Programming 19-20 629–679.
29. Luenberger, D. G., (1973), Introduction to Linear and Nonlinear Programming, Addison-Wesley, Menlo Park, California.

30. Plaza E., Esteva F., Garcia P., Godo L., López de Màntaras R. (1998) A logical approach to case-based reasoning using fuzzy similarity relations. Information Sciences 106 105-122.
31. Tanaka, A., Murofushi, T., (1989), A learning model using fuzzy measure and the Choquet integral, Proc. of the 5th Fuzzy System Symposium, Kobe, Japan, 213-217 (in Japanese).
32. 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.
33. Torra, V., (2002), Learning weights for the Quaasi-weighted means, IEEE Trans. on Fuzzy Systems 10:5 653–666.
34. Vapnik, V. N., (2000), The Nature of the Statistical Learning Theory, 2nd Ed., Springer, New York.
35. Yager R. R., (1987). Case-based reasoning, fuzzy systems modelling and solution composition. Proc of ICCBR-97, pp. 633-643.
36. Yager, R. R., (1988), On ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Trans. on SMC 18 183–190.
37. Zadeh L. A. (1971). Similarity relations and fuzzy orderings. Journal of Information Sciences 177–200.