Classification and Learning of Similarity Measures*

Michael M. Richter
Fachbereich Informatik
Universität Kaiserslautern
6750 Kaiserslautern, FRG

Abstract

The background of this paper is the area of case-based reasoning. This is a reasoning technique where one tries to use the solution of some problem which has been solved earlier in order to obtain a solution of a given problem. As example of types of problems where this kind of reasoning occurs very often is the diagnosis of diseases or faults in technical systems. In abstract terms this reduces to a classification task. A difficulty arises when one has not just one solved problem but when there are very many. These are called "cases" and they are stored in the case-base. Then one has to select an appropriate case which means to find one which is "similar" to the actual problem. The notion of similarity has raised much interest in this context. We will first introduce a mathematical framework and define some basic concepts. Then we will study some abstract phenomena in this area and finally present some methods developed and realized in a system at the University of Kaiserslautern.

1 Introduction

We consider a universe U which is partitioned into a disjoint union of subsets called classes and we refer to the elements of U as objects. Each object has a structure; for simplicity we take this as a fixed number of n attribute-value pairs. This allows a twofold description of objects:

a) The objects are coded as vectors of length n of real numbers, each coordinate represents an attribute;

b) the objects are described as conjunctions of unary predicates $P(a)$ where P stands for an attribute and a for a value.

We call a) an *analytic* and b) a *logical* representation of the objects.

The task is to determine for a given object its class. The available information may be, however, incomplete in two respects:

1. The object itself is only partially known;
2. only for a restricted number of objects the class where its belongs to is known.

In order to predict the class of an object one assumes an underlying regularity in the formation of the classes which has to be determined or at least approximated on the basis of the available information. In machine learning one considers mainly two basic ways to achieve this:

a) The logical approach: Classes are described by formulas in predicate logic using the attributes. These may e.g. be rules which have conjunctions of attribute formulas or their negations as premises and class names as conclusions.

b) The analytic approach: There is a distance function d in $\mathbb{R}^n$ and the class of some presented vector a is the class of that particular vector b from the already classified vectors for which the distance $d(a,b)$ is minimal.

With both approaches a number of concepts are connected. In order to discuss the interrelations between them from a mathematical point of view we make use of a number of results in economics, in particular utility theory. This stems from the fact that the notion of similarity shares some mathematical properties with the notion of a preference order. In utility theory one studies objects which may be more or less preferrable; we will employ the mathematical analogy between the partial orderings coming from similarity and preference.
Both, the classifying rule system and the distance function have to be built up in the training phase. The algorithms for this task have some (sometimes hidden) common properties. A fundamental problem is to exhibit the the connections between the distance function and the classification problem. In a nutshell this reads as follows:

*This work was published in: *Proceedings der Jahrestagung der Gesellschaft für Klassifikation.* Opitz, Lausen, Klar (ed.), *Studies in Classification, Data Analysis and Knowledge Organisation*, Springer Verlag, 1992

How to construct a distance function d such that for sufficiently small $d(a,b)$ the objects a and b are in the same class?

This is essentially an a posteriori problem which can principally only be answered after the class of the objects is known. From this principal point of view this asks for an adaptive approach. Nevertheless one has first to explore the basic aspects and concepts of distance functions and the related similarity measures. This attempt focusses the attention on problems which should also be approached (at least presently) in an empirical way. The PATDEX-system discussed in section 6 realizes a number of essential tasks from a practical point of view.

2 Basic concepts

Each object is given by the values of a fixed number of attributes. If A is such an attribute with value a then this is denoted by $A(a)$. We describe objects alternatively as vectors where each coordinate corresponds to an attribute and the entry to its value. An object description is like an object except that instead of the value of an attribute a variable may occur (indicating that the value is unknown). The universe of our object descriptions is U. In general we do not distinguish between objects and object descriptions.
There are different ways to represent similarity which we will introduce now.

1. A binary predicate $SIM(x,y) \subseteq U^2$ meaning "x and y are similar";
2. a binary predicate $DISSIM(x,y) \subseteq U^2$ meaning "x and y are not similar";
3. a ternary relation $S(x,y,z) \subseteq U^3$ meaning "y is at least as similar to x than z is to x";
4. a quaternary relation $R(x,y,u,v) \subseteq U^4$ meaning "y is at least as similar to x than v is to u";
5. a function $sim(x,y) : U^2 \to [0,1]$ measuring the degree of similarity between x and y;
6. a function $d(x,y) : U^2 \to \mathbb{R}$ measuring the distance between x and y.

The obvious questions which arise here are:

(i) How to axiomatize these concepts, i.e. which laws govern them?

(ii) How are the concepts correlated, which are the basic ones and which can be defined in terms of others?

(iii) How useful are the concepts for the classification task?

There are split opinions about the properties of the various concepts. It is certainly agreed that SIM is reflexive. There are arguments that SIM should neither be symmetric nor transitive. A typical example to support the first claim is that one could say 'my neighbor looks similar to the president' but one would not use the reverse phrase. This argument, however, says nothing about the truth or falsity of the similarity relation; it is only concerned with its pragmatics. For this reason we will accept that SIM is symmetric. In order to reject the transitivity of SIM one gives examples like 'a small circle is similar to a large circle and a large circle is similar to large square but a small circle is not similar to a large square'. The reason for this effect is that one deals here with two different similarity relations, one concerning size and another concerning form. A basic problem is how one can amalgamate two different similarity relations into one. A second type of counter argument arises when the objects are partially unknown. Suppose we have three such objects a, x and b where $SIM(a,b)$ does not hold and x is partially unknown. An opportunistic view could then assume both $SIM(a,x)$ and $SIM(x,b)$, violating transitivity. As a consequence, we will not accept transitivity for SIM.

A next observation tells us that we should distinguish $DISSIM(x,y)$ from $\neg SIM(x,y)$. The latter means simply that there is not enough evidence for establishing $SIM(x,y)$ but that may not be sufficient to claim $DISSIM(x,y)$; we have here the same distinction as one has between the negation in classical and intuitionistic logic. The deeper reason for this is that similarity between objects is not given as a relation with truth values 0 and 1 but as something to which the terms 'more or less' apply. We will therefore not consider SIM and $DISSIM$ anymore but the arguments given above do also apply to the remaining concepts.

In the sequel we will encounter several preorderings. A preordering $\geq$ on a set U is a reflexive and transitive binary relation. $\geq$ is called complete if $y \geq z \vee z \geq y$ holds. Such a relation can always be decomposed into two parts:

(i) $y > z \leftrightarrow y \geq z \wedge \neg(z \geq y)$, this called the strict part of the relation;

(ii) $y \sim z \leftrightarrow y \geq z \wedge (z \geq y)$ (indifference).

'$>$' is always asymmetric and transitive and '$\sim$' is an equivalence relation.

The relation $S(x,y,z)$ induces for each x a binary relation $y \geq_x z$. We assume:

(i) $\geq_x$ is a complete preorder (with $>_x$ as its strict part and $\sim_x$ as the indifference relation);

(ii) $y >_x z$ implies $y >_x u$ or $u >_x z$;

(iii) $x \geq_x z$.

(iii) refers to the reflexivity of SIM; the symmetry of SIM has no counterpart here. A further axiom is often required where the structure of the objects is involved:

Monotonicity Law: If y' agrees at least on one more attribute with x than y does, then $y' \geq_x y$ holds.

We will not require this law in general because it includes a kind of independence between the values of the attributes. If the attributes depend on each other then the same value can have a different meaning in different contexts so that more agreement on the attribute values can mean less similarity.

The relation S allows to define the concept 'y is most similar to x' : For some set $M \subseteq U$ some $y \in M$ is called *most similar* to x with respect to M iff

$$(\forall z \in M) S(x, y, z)$$

This notion is essential in case-based reasoning.
For the relation R we assume the axioms

(i) $R(x, x, u, v)$;

(ii) $R(x, y, u, v) \leftrightarrow R(y, x, u, v) \leftrightarrow R(x, y, v, u)$.

(i) and (ii) are the counterparts of the reflexivity and of symmetry of SIM, resp.

The relation $R(x, y, u, v)$ induces a partial ordering $\geq$ on pairs of objects by $(x, y) \geq (u, v) \leftrightarrow R(x, y, u, v)$. $\geq$ can be decomposed as above and we assume the same axioms as for $>_x$. R also induces a relation S_R by $S_R(x, y, z) \leftrightarrow R(x, y, x, z)$.

The basic axioms for a similarity measure sim are:

(i) $sim(x, x) = 1$ (reflexivity);

(ii) $sim(x, y) = sim(y, x)$ (symmetry).

The dual notion is that of a *distance measure* $d(x, y)$ which may attain arbitrary nonnegative values. In the corresponding axioms reflexivity reads as $d(x, x) = 0$. One does not require, however, the triangle inequality and allows $d(x, y) = 0$ for $x \neq y$ which means that d is neither a metric nor even a pseudo-metric. The argument for skipping the triangle inequality is the same as the one for not requiring transitivity for SIM.

One says that d and sim correspond to each other iff there is an order reversing one-one mapping

$$f : range(d) \to range(sim)$$

such that $f(0) = 1$ and $sim(x, y) = f(d(x, y))$; we denote this by $d \equiv_f sim$.
Popular candidates for f are:
$f(z) = 1 - \frac{z}{1+z}$ for unbounded d or $f(z) = 1 - \frac{z}{max}$ if d attains a greatest element max.

Some interrelations between the introduced concepts are immediate. If d is a distance measure and sim a similarity measure then we define

$$\begin{aligned} R_d(x, y, u, v) : &\iff d(x, y) \leq d(u, v) \\ R_{sim}(x, y, u, v) : &\iff sim(x, y) \geq sim(u, v) \end{aligned}$$

and

$$\begin{aligned} S_d(x, y, z) : &\iff R_d(x, y, x, z) \\ S_{sim}(x, y, z) : &\iff R_{sim}(x, y, x, z) \end{aligned}$$

We say that d and sim are compatible , iff

$$R_d(x, y, u, v) \iff R_{sim}(x, y, u, v) \ ;$$

compatibility is ensured by $d \equiv_f sim$ for some f.
As usual in topology the measures also define a neighborhood concept. For $\in > 0$ we put

$$V_\in(x) := V_{d,\in}(x) := \{y | d(x, y) \leq \in\},$$

and analogously $V_{sim,\in}(x)$ is defined; if d is a metric then these sets are ordinary closed neighborhoods. $S_d(x, y, z)$ expresses the fact that each neighborhood of x which contains z also contains y. In order to be useful for the classification task the neighborhood system has to be compatible with the partition into classes in the sense that the neighborhood should group the elements of the classes 'closely together'

3 Ordinals and cardinals

The concepts presented in (1) to (6) of Section 2 contain in an increasing order more and more information about the similarity of object descriptions. Least informative are SIM and $DISSIM$ and most informative are the measures and distance functions. The latter ones define the relations R_d and R_{sim} as indicated above in such a way that their axioms are satisfied. From R we obtain the relation S; again the axioms for S follow from those for R. S finally can, using some threshold, define relations SIM and $DISSIM$.

Comparing first $y \geq_x z$ and $sim(x, y)$, $sim(x, z)$ the additional information provided by sim is that it tells us *how* more similar y is to x than z is to x. S contains only an ordinal information while sim has also a cardinal aspect.

In the application to classification the main use of this cardinal aspect is that one forms differences like $|sim(x, y) - sim(x, z)|$. Such a difference is of interest

when one searches the object y most similar to x. If $|sim(x,y) - sim(x,z)|$ is small, then one could choose z instead of y with a small error only; for the classification task this may be sufficient. From this point of view $R(x,y,u,v)$ contains some cardinality information. Another type of implicit cardinality information is contained in the *sensibility potential*, cf. Wagener (1983).

The reverse way from the ordinal to the cardinal view is more involved. First, the relations SIM and $DISSIM$ carry very little information about the relation S. Given S, one has for every object description x the preorder $\geq_x$. In order to obtain R from S we proceed in several steps:

1. define: $R_1(x,y,x,z) \leftrightarrow S(x,y,z)$;
2. obtain R_2 from R_1 by adding the tuples (x,x,y,z);
3. define $\geq_3$ as the transitive closure of $\geq_2$;
4. obtain $\geq$ from $\geq_3$ by extending it to a complete preorder in such a way that $y >_3 z$ implies $y > z$ (this is always possible).
5. Define $R(x,y,u,v) \leftrightarrow (x,y) > (u,v)$.

If we define from this R as above the relation S_R the strict parts of the preorders may, however, be different. This is due to the fact that in step 1) where essentially the join $\cup \geq_x$ of the preorders $\geq_x$ was formed some cycles in the strict parts of the join may occur which means that elements are now indifferent which were strictly ordered before. Therefore we require that this cannot happen and call it the *compatibility condition* on S.

The step from R (or $\geq$) to a measure or distance function is done by embedding $\geq$ into $\mathbb{R}^2$. This is possible because our universe is finite.

We emphasize again that for our classification task the relation S is the one which is used. To be of interest the compatibility condition has to be satisfied. This is essentially the step to the relation R which, as remarked above, has additional benefits. In our learning process below we will learn the measure directly but will essentially use information about relation S.

4 The amalgamation of similarity measures

Suppose we are given different experts E_i who are confronted with a fixed object x and a number of objects which may be more or less similar to x. The task for these experts is to arrange the objects according to their similarity to a, i.e. to establish an ordering $\geq_a^i$. Each expert is supposed to represent a certain aspect and will come up with his individual arrangement. Furthermore, there is a general manager who takes these individual ratings and whose task is to amalgamate the different ratings into a general ordering of the objects under consideration.

A very simple method for integrating such orderings is to use a number assignment according to the orderings and sum up these numbers. This is Borda's method which he invented in 1781. We give an example with 5 participating objects t,y,z,u and v and 5 experts (representing 5 aspects) :

	t	y	z	u	v
1	4	3	2	1	0
2	2	4	3	0	1
3	3	2	1	0	4
4	4	3	0	2	1
5	1	4	2	3	0
Sum	14	16	8	6	6

The winner, i.e. the object most similar to x is y, followed by t, z etc. Suppose now that we want to remove the objects z and u from the database because they are perhaps not of great interest anyway. Then we are left with three objects and we apply the same method to rank them. We get the following table:

	t	y	v
1	2	1	0
2	1	2	0
3	1	0	2
4	2	1	0
5	1	2	0
Sum	7	6	2

The result is that the final ordering of the remaining objects is changed and that now t is the winner. This effect is very undesirable because the elimination of uninteresting objects leads to a change of the ordering of the remaining objects; the whole data base is subject to a global analysis in order to recompute the similarity relation. We will explain now that this is not an accident which is due to the special method but that there is an underlying deeper phenomenon.

We start with a set U of object descriptions s.t. $|U| \geq 3$ and an index set $M \neq \emptyset$. We consider partial orderings as introduced in section 2. Let S be the set of such orders on U and $F = \{f | f : M \rightarrow S\}$. M represents the different aspects and F the orderings (i.e. the strict part) with respect to similarity to the reference object according to these aspects. What one looks for is a mapping $\sigma : F \rightarrow S$ which amalgamates the individual orderings into a universal one. The function σ has to satisfy certain very plausible conditions:

(a) If $y\ f(m)z$ for all $m \in M$, then $y\ s(f)z$;

(b) if f and g coincide on y and z, then $\sigma(f)$ and $\sigma(g)$ coincide on y and z too.

(c) There is no $m \in M$ such that for all y and z in U we have:
If $y\ f(m)z$, then $y\sigma(f)z$.

These conditions have a clear motivation. (a) says that the universal ordering should not contradict all aspects. (b) was discussed above and (c) says that one cannot reduce the problem to one aspect.

Theorem: There is no function f satisfying (a), (b) and (c).

This theorem is due to Arrow (cf. Arrow (1963)) and well known in the area of social choice functions. There the partial orderings are preference orderings, M is the set of voters, (a) is the principle of democracy and (c) excludes dictatorship. The function σ combines the individual votes. Arrow's impossibility theorem is also called the theorem of the dictator and was considered as somewhat paradoxical. Slight variations of the condition do not change the validity of the theorem. The crucial and most discussed condition is (b). It is also important for our situation; according to the theorem changes in the data base have other consequences. The most we can hope for is that these consequences have a local character.

5 General forms of distance functions and similarity measures

We consider objects which are defined in terms of boolean valued attributes and study their relations using distance functions only. There is a great variety of distance functions and an enormous amount of literature. When distance functions are used for classification purposes they cluster the objects in such a way that the cluster coincide with the given classes as much as possible. If this is the case then one can say that the function contains some knowledge about the classes. Different applications lead to different types of classes and therefore to different kinds of distance functions; this explains mainly the richness of this area. In our approach we are not so much interested in our introducing a particular clever distance function but rather in showing how some general knowledge can be improved by an adaptive process. The type of functions we introduce is general enough to study these techniques but many other distance functions would have worked as well. We will restrict ourselves here to Boolean attributes, i.e. we have values 0 and 1 only. The most simple distance measure is the Hamming distance. A generalization of the Hamming distance is given by the *Tversky-Contrast model* (cf. Tversky (1977)). For two objects x and y we put
$A :=$ The set of all attributes which have equal values for x and y ;
$B :=$ the set of all attributes which have value 1 for x and 0 for y ;
$C :=$ the set of all attributes which have value 1 for y and 0 for x ;
The general form of a Tversky distance is

$$T(x,y) = \alpha \cdot f(A) - \beta \cdot f(B) - \gamma \cdot f(C)$$

where α, β and γ are positive real numbers. Most of the other possible distance functions are located between the Hamming and the Tversky measure with respect to the information which they can contain. In PATDEX (see below) we start out with a measure for which we need some notation. An object description from the case base is denoted by x and an arbitrary one by x_{act} (indicating that this is the actual description for which we want a similar one from the base). We put
$x_{act} = (w_{i_1}, ..., w_{i_k})$, $x = (v_{r_1}, ..., v_{r_j})$; here we list only the coordinates with a known value.

$\mathrm{H} = \{i_1, ..., i_k\}$,

$\mathrm{K} = \{r_1, ..., r_j\}$;

$\mathrm{E} = \{i | i \in \mathrm{H} \cap \mathrm{K}, w_i = v_i\}$,
the set of attributes where the values agree;

$\mathrm{C} = \{i | i \in \mathrm{H} \cap \mathrm{K}, w_i \neq v_i\}$,
the set of attributes with conflicting values;

$\mathrm{U} = \mathrm{H} \setminus \mathrm{K}$,
the set of attributes with a known value for x but unknown value for the actual object;

$\mathrm{R} = \mathrm{K} \setminus \mathrm{H}$,
the set of attributes with a redundant value for x_{act}.

The measure used is of the form:

$$\mathrm{sim_{PAT}(x_{act}, x)} = \frac{\alpha \cdot |\mathrm{E}|}{\alpha \cdot |\mathrm{E}| + \beta \cdot |\mathrm{C}| + \gamma \cdot |\mathrm{U}| + \eta \cdot |\mathrm{R}|}$$

The parameters α, β, γ and δ can be chosen; presently we use:

$$\alpha = 1, \qquad \beta = 2, \qquad \eta = 1/2, \qquad \gamma = 1/2;$$

which gives:

$$\mathrm{sim_{PAT}(x_{act}, x)} = \frac{|\mathrm{E}|}{|\mathrm{E}| + 2 \cdot |\mathrm{C}| + 1/2 \cdot |\mathrm{U}| + 1/2 \cdot |\mathrm{R}|}$$

This measure pays special attention to attributes with missing values. On the other hand, it abstracts from the Tversky measure in so far that it sees only the cardinality of sets instead the sets themselves.

6 The Patdex system

The difficulty with the similarity measure is that its quality is related to the final success of the whole reasoning procedure; this is an *a posteriori* criterion. A priori it is not clear what the criteria for similarity of objects should be; they do not only depend on the objects themselves but also on the pragmatics of reasoning. In case-based reasoning it is usually clear whether a solution for a given problem (in our situation a classification problem) is correct but is far from clear what it means that two problems are similar enough so that the solution for one problem also works for the other one. Looking at the object descriptions only one neither knows a suitable general form of the measure nor has one an indication how the parameters should be determined. An even more serious difficulty arises when the world of problems is continuously changing. This suggests that the similarity should not be defined in some fixed way but instead be the result of an adaptive learning process. This will be carried out in the PATDEX-System.

PATDEX is a part of the MOLTKE-System (cf. Althoff (1992)) which was developed in the past years at the University of Kaiserslautern. Its domain is the fault diagnosis of technical systems. Here we are only concerned with the aspect that diagnosis can be regarded as a classification task and we will suppress the other aspects. For this reason we modify the present terminology of PATDEX. The system accepts a description of an object as an input; this description may be partial, some attribute values may be unknown. The basic instrument for the classification is the *case base*; a *case* is a pair (Object x, class(x)) where class(x) is the class to which x belongs.

The first version of PATDEX is PATDEX/1. It contains the basic structures which have been extended later on. It is convenient to describe it first. As basic techniques, PATDEX/1 applies learning by memory adaptation and analogical reasoning. The toplevel algorithm of PATDEX reads as follows:

Input: The actual object description x
Output: a class C or failure

1. Find a case in the case base with an object x' most similar to x. If there is no case with an object at least '*minimally similar*' to x then stop with failure.
2. If x and x' are '*sufficiently similar*' then accept the class C of x' also for x and goto 4).
   3) Otherwise select an attribute with unknown value and determine its value in order to obtain an improved situation and goto 1).
3. If the class is correct then add the case (x, C) to the case base and stop with success.
4. If the class is not correct then cancel temporarily (i.e. for the actual problem) all cases with class C and goto 3).

Here we need an external teacher who says whether a class is correctly chosen or not. We also have to explain *'minimally similar'* and *'sufficiently similar'*. For this we need a partition of the case base which is given after the introduction of the similarity measure.

For object descriptions PATDEX we introduced as a first proposal the similarity measure sim_{PAT} in section 5 with parameters $\alpha = 1,\ \beta = -2,\ \gamma = \eta = -1/2$. This special choice of the parameters is at the moment mainly motivated by experimental results. It has a defensive, pessimistic character. A high negative contribution to the measure is given for conflicting attribute values, i.e. we strongly wish to avoid false classification.

For the partition of the case base we choose real numbers $\in$ and δ such that $0 < \in < \delta < 1$ and define:

Def.: The object descriptions x_1 and x_2 are called:

(i) *indistinguishable*
$\Leftrightarrow sim(x_1, x_2) = 1$;

(ii) *sufficiently similar*
$\Leftrightarrow \delta \leq sim(x_1, x_2) < 1$;

(iii) *at least minimally similar*
$\Leftrightarrow \in \leq sim(x_1, x_2) < \delta$;

(iv) *not minimally similar*
$\Leftrightarrow 0 \leq sim(x_1, x_2) < \in$;

The lower bound $\in$ is called the hypothesis threshold, a case succeeding here is said to be qualified for further processing. If the value exceeds an upper bound δ it is even qualified as providing the classification (classification threshold). If, for a given case, the similarity value equals 1 this case is said to be proven. The thresholds are locally defined for each case of the case base, i.e. we have the possibility to make the numbers $\in$ and δ dependent on the respective cases.

It is an important feature of PATDEX that it supports for an object description the selection of an attribute with an unknown value. An optimal or at least good choice of such an attribute is crucial for an efficient

classification procedure. We will, however, not deal with this question.

The use and analysis of PATDEX has lead to the conclusion that its performance concerning the classification problem showed some weaknesses. Ultimately this was a problem of the similarity measure in two respects as already indicated. First, the type of the measure (as an abstraction of the Tversky measure) was too simple in order to reflect information of the objects which are necessary for the classification. Secondly, even if the type of the measure would have been optimal one would still face the problem of chosing the parameters of the measure. To overcome this problem a learning process will be introduced.

We will first describe the structural improvements of the measure. They get their motivation from the actual use of the system for diagnostic purposes rather than from purely mathematical considerations. The information reflected by the improvements is usually available in the intended applications. The improvements are contained in the system PATDEX/2 (cf. Wess 91).

The underlying pattern of the new features in PATDEX/2 is that not all attributes are equally important for determining the class of an object description. This leads to the notion of relevance. The relevances are numbers $w_{ij} \in [0,1]$; where the index i points to an attribute A_i resp. its value and the index j refers to a class C_j. The w_{ij} should indicate the degree with which a_i points at C_j. The relevances give rise to the *relevance* matrix $R[w_{ij}]$. The main problem is now to determine the entries (also called weights) of the relevance matrix. These weights are exactly the elements which will be learned later on.

It is convenient to normalize the matrix such that

(i) For all i and j $0 \leq w_{ij} \leq 1$ holds;

(ii) For all j we have $\sum_{i=1}^{n} w_{ij} = 1$.

We will now discuss the possibilities for the weights. This leads to some changes in the computation of sim_{PAT}.

Local and global weights: Global weights satisfy $w_{ij} = w_{ik}$ for all j and k; otherwise weights are called local. Global weights are less precise but easier to determine.

Conflicting attribute values: If two objects have different values for an attribute A with domain D then the form of the difference should play a role. This can be achieved by introducing a function $\omega : D^2 \rightarrow [0,1]$ which has to represent the similarities of the attribute values. If one of the values is unknown then the similarity ω_i evaluates to zero.

Redundant attribute values: Redundant attribute values for the actual object description count negative in the measure. This has the undesired effect of decreasing the similarity by the acquisition of more and e.g. completely uninteresting attribute values. This leads to the notion of *classifying* and *not classifying* attributes for redundant attributes, depending on their values. This division of the attributes has to be made by the user; in applications to diagnosis this is usually not so difficult because classifying attributes there correspond to attributes with an abnormal value. The impact on the measure is that only classifying attributes enter the computation of R in sim_{PAT}.

Unknown attribute values: Unknown values for the actual object description also count negatively in the computation of the measure. This may not be justified because the known values may determine, at least with some probability, the missing ones. Hence for such unknown values a value should be substituted which has a probability above some (user defined) threshold θ. The probability can be estimated by the frequencies in the base of object descriptions.

These remarks lead to a redefinition of the similarity measure. For the similarity between values the user chooses a threshold λ. We put:

$x_{act} = (w_{i_1}, ..., w_{i_k}), x = (v_{r_1}, ..., v_{r_j})$; here we list only the coordinates with a known value or where the value in x_act can be predicted with probability $\geq \theta$.

$\mathrm{H} = \{i_1, ..., i_k\}$,

$\mathrm{K} = \{r_1, ..., r_j\}$;

$\mathrm{E}' = \{i | \omega(w_i, v_i) \geq \lambda\}$,
the set of attributes with sufficiently similar values;

$\mathrm{C}' = \{i | \omega(w_i, v_i) < \lambda\}$,
the set of attributes with not sufficiently similar values;

$\mathrm{U}' = \mathrm{H} \setminus \mathrm{K}$,
the set of attributes with a known or estimated value for x but unknown value for x_{act}.

$\mathrm{R}' = \mathrm{K} \setminus \mathrm{H}$,
the set of attributes with a redundant and classifying value for x_{act}.

Using this we define:

$$E_0 \quad = \quad \sum_{i \in E'} w_{ij} \cdot \omega(w_i, v_i);$$

$$
\begin{aligned}
C_0 &= \sum_{i \in C'} w_{ij} \cdot (1 - \omega(w_i, v_i)); \\
R_0 &= |R'|; \\
U_0 &= \sum_{i \in U'} w_{ij} \quad .
\end{aligned}
$$

This leads finally to the measure of PATDEX/2:

$$\mathrm{sim}(\mathrm{x_{act}}, \mathrm{x}) = \frac{\alpha \cdot |\mathrm{E_0}|}{\alpha \cdot |\mathrm{E_0}| + \beta \cdot |\mathrm{C_0}| + \gamma \cdot |\mathrm{U_0}| + \eta \cdot |\mathrm{R_0}|}$$

α, β, γ and η can be chosen as before. The partially user defined parameters are a step towards the idea of the Tversky measure. The approach takes into account that the precise form of the measure is a priori (i.e. when the problem is given) not available; the user can fill in as much knowledge as he has about the problem. Given a base of correctly classified object descriptions experiments with PATDEX/2 showed that the similarity measure did not even classify the cases from the base correctly. This was expected and here a learning process starts. What is learned are the weights, i.e. the entries of the relevance matrix. This process has an initial phase and a learning phase; the training set is the case base.

Initial phase: The initial weights w_{ij} are determined according to the observed frequencies in the base.

Learning phase: The cases (x_{act},C) are taken from the case base. The system selects the most similar case (x,D) from the case base (similarity of cases means similarity of their object descriptions). If C = D, then nothing will be changed.

For C $\neq$ D we distinguish two possibilities:

(1) x contains less known attribute values than x_{act}. Here the class D was obviously only correct by accident and the case (x,D) is eliminated from the case base.

(2) In all other situations (x,D) remains in the case base but the weights are updated.

The numerical form of the learning rule is not of interest; the leading principles are:

- $sim_{PAT/2}(x_{act}, x) < \delta$ should be achieved, they are not anymore sufficiently similar;
- weights for attributes in C$'$ and U$'$ are increased;
- weights for attributes in E$'$ are decreased;
- weights for attributes in R$'$ remain invariant;
- the weights w_{ij} are still normalized according to $\sum_{i=1}^{n} w_{ij} = 1$ for each j.

Rules of this type are known in unsupervised neural networks; an example is the Grossberg rule resp. the rule in competive learning, cf. Rumelhart, Zipser (1985).

After each erroneous diagnosis the weights of the relevance matrix are changed. In summary, the measure *sim* (and therefore the relation $S(x, y, z)$ has been built up in two steps:

a) The first approximation is done by modifying the measure sim_{PAT} using knowledge about the classification task.

b) The result of a) is the starting point for an adaptive learning process where only the success in the classification task plays a role.

7 Acknowledgement

The author thanks Stefan Wess for helpful discussions and the referee for useful remarks.

8 References

ALTHOFF, K.-D. (1992), Lernen aus Fallbeispielen zur Diagnose technischer Systeme, in: Doctoral Dissertation, University of Kaiserslautern.

ARROW, K.J. (1963), Social Choice and Individual Values, New York.

RICHTER, M.M., WESS, S. (1991), Similarity, Uncertainty and Case-Based Reasoning in PATDEX, in: *Automated Reasoning, Essays in Honor of Woody Bledsoe* (ed. R.S. Boyer), Kluwer Acad. Publ..

RUMELHART, D.E., ZIPSER, D. (1985), Feature Discovery by Competitive Learning, *Cognitive Science* 9, 75-112.

TVERSKY, A. (1977), Features of Similarity, *Psychological Review* 84,p.327-352.

WAGENER, M. (1983), Kardinalität in der Nutzentheorie, *Mathematical Systems in Economy* 81.

WESS, S. (1991), PATDEX/2 - Ein System zum fallfokusierenden Lernen in Technischen Diagnose Situationen, Seki-Report SWP-01-91

WESS, S., JANETZKO, D., MELIS, E. (1992), Goal-Driven-Similarity Assessment, Seki-Report, Universität Kaiserslautern.