Available online at www.sciencedirect.com

SCIENCE DIRECT®

Environmental Modelling & Software XX (2003) XXX–XXX

Environmental Modelling & Software

www.elsevier.com/locate/envsoft

A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations

Héctor Núñez [a,*], Miquel Sànchez-Marrè [a], Ulises Cortés [a], Joaquim Comas [b], Montse Martínez [b], Ignasi Rodríguez-Roda [b], Manel Poch [b]

[a] *Artificial Intelligence Section, Knowledge Engineering and Machine Learning Group (KEMLG), Universitat Politècnica de Catalunya, Campus Nord-Edifici C5, Jordi Girona 1-3, 08034 Barcelona, Catalonia, Spain*

[b] *Laboratori d'Enginyeria Química i Ambiental, Universitat de Girona, Campus de Montilivi, 17071 Girona, Spain*

Received 22 October 2002; received in revised form 6 February 2003; accepted 5 March 2003

Abstract

The step of identifying to which class of operational situation belongs the current environmental system (ES) situation is a key element to build successful environmental decision support systems (EDSS). This diagnosis phase is especially difficult due to multiple features involved in most environmental systems. It is not an easy task for environmental managers to acquire, to integrate and to understand all the increasing amount of data obtained from an environmental process and to get meaningful knowledge from it. Thus, a deeper classification task in a EDSS needs a full integration of gathered data, including the use of statistics, pattern recognition, clustering techniques, similarity-based reasoning and other advanced information technology techniques. Consequently, it is necessary to use automatic knowledge acquisition and management methods to build consistent and robust decision support systems. Additionally, some environmental problems can only be solved by experts who use their own experience in the resolution of similar situations. This is the reason why many artificial intelligence (AI) techniques have been used in recent past years trying to solve these classification tasks. Integration of AI techniques in EDSS has led to more accurate and reliable EDSS.

Case-based reasoning (CBR) is a good technique to solve new problems based on previous experience. Main assumption in CBR relies on the hypothesis that similar problems should have similar solutions. When working with labelled cases, the retrieval step in CBR cycle can be seen as a classification task. The new cases will be labelled (classified) with the label (class) of the most similar case retrieved from the case base. In environmental systems, these classes are operational situations. Thus, similarity measures are key elements in obtaining a reliable classification of new situations. This paper describes a comparative analysis of several commonly used similarity measures, and a study on its performance for classification tasks. In addition, it introduces *L'Eixample* distance, a new similarity measure for case retrieval. This measure has been tested with good accuracy results, which improve the performance of the classification task. The testing has been done using two environmental data sets and other data sets from the UCI Machine Learning Database Repository.

Keywords: Environmental situation classification; Case retrieval; Case-based reasoning; Similarity metric

1. Introduction

The management of environmental system (ES) is a very complex and dangerous task. The step of identifying to which class of operational situation belongs the current ES situation is a key element to build successful environmental decision support systems (EDSS). If EDSS are able to make reliable diagnostics, then the pro-

[*] Corresponding author. Tel.: +34-93-4017994; fax: +34-93-401-7014.

E-mail addresses: hnunez@lsi.upc.es (H. Núñez); miquel@lsi.upc.es (M. Sànchez-Marrè); ia@lsi.upc.es (U. Cortés); quim@lequia.udg.es (J. Comas); montse@lequia.udg.es (M. Martínez); ignasi@lequia.udg.es (I. Rodríguez-Roda); manel@lequia.udg.es (M. Poch).

1364-8152/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2003.03.003

posed plan to deal with the situation will be accurate and optimal enough to lead the environmental system to a normal operation state.

1.1. Identification of current situations in classical EDSS

This diagnosis phase is especially difficult due to multiple features involved in most environmental systems, including chemical, biological, physical, inflow-variability, microbiological, subjectivity and temporal effects, which implies the analysis of different kind of data (numerical and qualitative) and uncertainty to identify a particular situation.

Progress in instrumentation, in computer technology and in process sensors has enabled data gathering, which implies more available information. However, it is not an easy task for experts to acquire, to integrate and to understand all the increasing amount of data obtained from an environmental process and to get meaningful knowledge from it.

Usually, the first step in the diagnosis of specific situations in environmental systems is to determine which are the key parameters that identify an operational state and that need to be checked to deal with that situation. However, it is important to remark that what defines a situation is not a single parameter, but a group of variables and their interrelations. In this sense, there exist several physical, biological and chemical indicators that are commonly used for a fast monitoring of the state of environmental processes. The main objective of indicators, which can be numerical and/or symbolic, is to transform data and statistics into synthetic information easy to understand by several groups of people involved in the domain such as scientists, politicians, administration and citizens. For example, the number and type of filamentous organisms and protozoa in the activated sludge process, or the presence of *Cladophora* in the streams are biological indicators, while the suspended solids concentration or the biological oxygen demand (a global measure of the biodegradable organic matter) are chemical indicators in wastewater treatment plant processes. Another example of indicator, in this case for atmospheric pollution, is the National Ambient Air Quality Standards in the USA or the Catalan Index of air quality, which is a single unit-free figure denoting the effect of the different pollutants measured on overall air quality including particles, such as SO_2, NO_2, O_3, or CO.

Commonly, a first diagnosis assessing the performance and behaviour of environmental systems compares the results of indicators with respect to the environmental goals, normally based on the legal requirements. Thus, the goal of this first diagnosis is to determine if the process is in a normal or abnormal state. However, scientists and environmental managers are more interested in predicting or making an early diagnosis of any abnormal process situation and unfamiliar situations. They want to infer the causes to these problems to generate a sophisticated action plan, while applying the recommended list of tasks to return the process to a “normal” situation. For this reason, a second and deeper classification task of the environmental system status should be covered by an EDSS.

1.2. Identification of current situations in advanced EDSS

This deeper classification task in a EDSS needs a full integration of gathered data, including the use of statistics, pattern recognition, clustering techniques, and other advanced information technology techniques. Also, the use of modelling techniques (both mechanistic and black box models) must be considered to simulate and predict the conditions of the process. The experts of the domain, who occasionally carry out the study and interpretation of the database to increase the knowledge about the process, usually do this mathematical analysis. Clearly, it is necessary to use automatic knowledge acquisition and management methods to build consistent and robust decision support systems. In addition, some environmental problems can only be solved by experts using their own experience in the resolution of similar situations. These experts are not always accessible when dealing with risk situations, and it is crucial to record each new experience to learn about the process, while reusing this specific knowledge in the future. This is the reason why many artificial intelligence (AI) techniques have been used in recent past years trying to solve these classification tasks. Integration of AI techniques in EDSS has led to obtain more accurate and reliable EDSS. Furthermore, case-based reasoning (CBR) can be a good technique to make diagnosis based on previous experience.

Main assumption in CBR relies in retrieving the most similar cases or experiences among those stored in the case base. Then, previous solutions given to these most similar past-solved cases can be adapted to fit new solutions for new cases or problems in a concrete domain, instead of deriving them from scratch. When working with labelled cases, the retrieval step in CBR cycle can be seen as a classification task. The new cases will be labelled (classified) with the label (class) of the most similar case retrieved from the case base. In environmental systems, these classes are operational situations. Thus, *similarity measures* are key elements in obtaining a reliable classification of new situations.

1.3. Related work

Theoretical frameworks for the systematic construction of similarity measures have been described in

Osborne and Bridge (1996, 1997) and Bridge (1998). Other research work introduced new measures for a practical use in CBR systems, such as Bayesian distance measures in Kontkanen et al. (2000) and some heterogeneous difference metrics in Wilson and Martínez (1997). Also, a review of some used similarity measures was done in Liao and Zhang (1998).

This paper aims to analyse and study the performance of several commonly used measures in practical use, for a better classification of environmental situations. In addition, *L'Eixample* distance, a new similarity measure for case retrieval, is introduced. This measure tries to improve the competence of a CBR system, providing flexibility and adaptation to environmental domains where some attributes have a substantial higher importance than others do. This similarity measure has been tested against some other related and well-known similarity measures with good results. Measures are evaluated in terms of classification accuracy on unseen cases, measured by a 10-fold cross-validation process. In this comparative analysis, we have selected two basic similarity measures (Euclidean and Manhattan), two unweighted similarity measures (Clark and Canberra) and two heterogeneous similarity measures (heterogeneous value difference metric and interpolated values difference metric). Although all these are dissimilarity measures, we can refer to similarity measures normalized in interval [0,1], by means of the relation:

$$SIM(x,y) = 1-DISS(x,y)$$

where $DISS(x,y)$ means the dissimilarity measure between cases x and y, which is commonly computed as a sum of attribute differences within the interval [0,1]. For this reason throughout the paper, these measures will be equally referred as similarity or dissimilarity measures.

1.4. Overview

The paper is organized in the following way. Section 2 outlines main features about case-based reasoning. In Section 3, background information on selected distance measures is provided. Section 4 introduces *L'Eixample* distance measure. Section 5 presents the results comparing the performance of all measures for classification tasks tested on two environmental databases and 14 databases from the UCI Machine Learning Repository. Finally, in Section 6, conclusions and future research directions are outlined.

2. Case-based reasoning

CBR systems have been used in a broad range of domains to capture and organize past experience and to learn how to solve new situations from previous past solutions.

The basic reasoning cycle of a CBR agent can be summarized by a schematic cycle (see Fig. 1) and detailed in the following steps (Kolodner, 1993):

- *Retrieve* the most similar case(s) to the new case. Similarity measures are involved in this step.
- *Adapt* or *reuse* the information and knowledge in that case to solve the new case. The selected best case has to be *adapted* when it does not match perfectly the new case.
- *Evaluate* or *revise* the proposed solution. A CBR-agent usually requires some feedback to know what is going right and what is going wrong. Usually, it is performed by simulation or by asking a human.
- *Learn* or *retain* the parts of this experience likely to be useful for future problem solving. The agent can learn both from successful solutions and from failed ones (repair).

Case-based reasoning in continuous situations has been applied in CIDA (Joh, 1997), an assistant for conceptual internetworking design, and NETTRAC (Brandau et al., 1991) as a case-based system for planning and execution monitoring in traffic management in public telephone networks. In environmental sciences, CBR has been applied in different areas with different goals, because of its general applicability. It has been used in information retrieval from large historical meteorological databases (Jones and Roydhouse, 1995), in optimization of sequence operations for the design of wastewater treatment systems (Krovvidy and Wee, 1993), in supervisory systems for supervising and controlling WWTP management (Rodríguez-Roda et al., 1999; Sànchez-Marrè et al., 1997), in decision support systems for planning forest fire fighting (Avesani et al., 2000), in case-based prediction for rangeland pest management advisories by Branting et al. (1997), or in case-based design for process engineering (Surma and Brauschweig, 1996).

There are several case representation formalisms ranging from flat structures to hierarchical structures such as a graph representation. One of the most used by its simplicity and applicability is the feature vector approach, where the cases are represented by means of a set of attribute-value pairs. For this reason, the paper will focus on feature vector CBR systems. Similarity measures are intrinsically related to the used case representation formalism in the CBR system. Within the feature vector approach, the similarity measures are normally computed as an aggregation of attribute differences between two cases. If similarity measures do not capture the actual differences between cases, the retrieval step, and the whole CBR performance will be

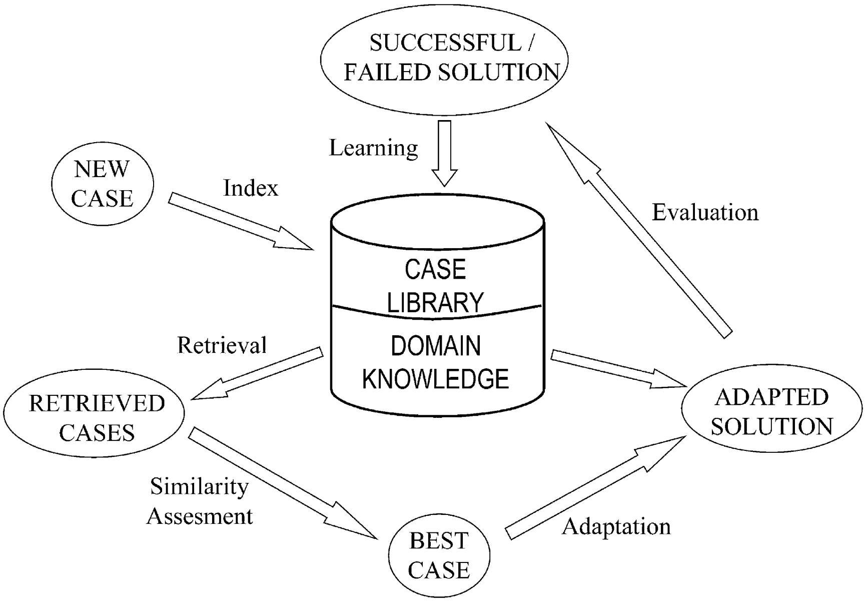

Fig. 1. The general case-based reasoning paradigm.

bad. Thus, the selection of an appropriate similarity measure is a key point in CBR systems.

3. Similarity measures

Most case-based reasoners use a generalized weighted dissimilarity measure such as,

$$diss(C_i,C_j) = \frac{\sum_{k=1}^{n} w_k * atr\text{—}diss(C_{ik},C_{jk})}{\sum_{k=1}^{n} w_k}$$

where C_i and C_j are two cases; w_k is the weight or importance assigned to attribute k; and $atr\text{—}diss(C_{ik},C_{jk})$ is the dissimilarity degree between the value of attribute k in cases i and j.

Currently, there are several similarity measures that have been used in CBR systems, and some comparison studies exist among these similarity measures (see Wilson and Martínez, 1997; Liao and Zhang, 1998). The results obtained in these studies show that the different similarity measures have a performance strongly related to the type of attributes representing the case and to the importance of each attribute. Thus, it is very different to deal with only continuous data, with ordered discrete data or non-ordered discrete data. Also, it is necessary to give a greater distance contribution to an important attribute than to other less important ones. In this study, our new proposed similarity measure, *L'Eixample*, is compared against some other measures that had been used before, with a very good performance in tests done in prior studies carried out. These selected similarity measures are described in the following subsections.

3.1. Measures derived from Minkowski's metric

$$d(C_i,C_j) = \left(\sum_{k=1}^{n} |C_{ik}-C_{jk}|^r\right)^{1/r} \qquad r\geq 1$$

where n is the number of input attributes. When $r = 1$, *Manhattan* or *City-Block* distance function is obtained. If $r = 2$, *Euclidean* distance is obtained. When including weights for all the attributes, the general formula becomes the following:

$$d(C_i,C_j) = \left(\frac{\sum_{k=1}^{n} w_k^r * |d(C_{ik},C_{jk})|^r}{\sum_{k=1}^{n} w_k^r}\right)^{1/r}$$

where for non-ordered attributes, their contribution to the distance is,

$$d(C_{ik},C_{jk}) = 1-\delta_{qlv(C_{ik}),qlv(C_{jk})}$$

and δ is the Kronecker δ.

3.2. Unweighted similarity measures

We include in this study two similarity measures that ignore attribute weight:

These similarity metrics, defined in Lance and Williams (1966), are very sensitive to small changes close to $x_{ik} = 0 = x_{jk}$, and can be less reliable if the (x_{ik}) are sample estimates of some quantities. An advantage of these metrics is that they do not need a previous normalization.

Clark:

$$d(C_i,C_j) = \sum_{k=1}^{n} \frac{|C_{ik}-C_{jk}|^2}{|C_{ik}+C_{jk}|^2}$$

and *Canberra*:

$$d(C_i,C_j) = \sum_{k=1}^{n} \frac{|C_{ik}-C_{jk}|}{|C_{ik}+C_{jk}|}$$

3.3. Heterogeneous similarity measures

To obtain a broader study and results, two other distance measures that show very high values of efficiency have been included. These functions were proposed in Wilson and Martínez (1997).

Heterogeneous value difference metric (HVDM):

$$HVDM(C_i,C_j) = \sqrt{\sum_{k=1}^{n} d_k^2(C_{ik},C_{jk})}$$

where m is the number of attributes. The function $d_k(C_{ik},C_{jk})$ returns a distance between the two values C_{ik} and C_{jk} for attribute k, and is defined as:

$$d_k^2(C_{ik},C_{jk})$$

$$= \begin{cases} 1, & \text{if } C_{ik} \text{ or } C_{jk} \text{ is unknown; otherwise} \\ normalized\text{—}vdm_k(C_{ik},C_{jk}), & \text{if } k \text{ is nominal} \\ normalized\text{—}diff_k(C_{ik},C_{jk}), & \text{if } k \text{ is linear} \end{cases}$$

where $normalized\text{—}vdm_k(C_{ik},C_{jk})$, is defined as follows:

$$normalized\text{—}vdm_k(C_{ik},C_{jk}) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{k,C_{ik},c}}{N_{k,C_{ik}}} - \frac{N_{k,C_{jk},c}}{N_{k,C_{jk}}} \right|^2}$$

where $N_{k,x}$ is the number of instances that have value C_{ik} for attribute k; $N_{k,C_{ik},c}$ is the number of instances that have value C_{ik} for attribute k and output class c; C is the number of output classes in the problem domain.

The function $normalized\text{—}diff_k(C_{ik},C_{jk})$, is defined as shown below:

$$normalized\text{—}diff_k(C_{ik},C_{jk}) = \frac{|C_{ik}-C_{jk}|}{4\sigma_a}$$

where σ_k is the standard deviation of the numeric values of attribute k.

Interpolated value difference metric (IVDM):

$$IVDM(C_i,C_j) = \sum_{k=1}^{n} ivdm_k(C_{ik},C_{jk})^2$$

where $ivdm_k$ is defined as:

$$ivdmk(C_{ik},C_{jk})$$

$$= \begin{cases} vdm_k(C_{ik},C_{jk}) & \text{if k is discrete} \\ \sum_{c=1}^{C} |p_{k,c}(C_{ik})-p_{k,c}(C_{jk})|^2 & \text{otherwise} \end{cases}$$

where $vdm_k(C_{ik},C_{jk})$ is defined as follows:

$$vdm_k(C_{ik},C_{jk}) = \sum_{c=1}^{C} |P_{k,C_{ik},c}-P_{k,C_{jk},c}|^2$$

C is the number of classes in the database. $P_{k,C_{ik},c}$ is the conditional probability that the output class is c given that attribute k has the value C_{ik}. And:

$$P_{k,C_{ik},c} = \frac{N_{k,C_{ik},c}}{N_{k,C_{ik}}}$$

where $N_{k,C_{ik}}$ is the number of instances that have value C_{ik} for attribute k; $N_{k,C_{ik},c}$ is the number of instances that have value C_{ik} for attribute k and output class c.

$P_{k,c}(x)$ is the interpolated probability value of a continuous value C_{ik} for attribute k and class c, and is defined:

$$P_{k,c}(x) = P_{k,u,c} + \left(\frac{x-mid_{k,u}}{mid_{k,u+1}-mid_{k,u}}\right)*(P_{k,u+1,c}-P_{k,u,c})$$

In this equation, $mid_{k,u}$ and $mid_{k,u+1}$ are midpoints of two consecutive discretized ranges such that $mid_{k,u} \leq C_{ik} < mid_{k,u+1}$. $P_{k,u,c}$ is the probability value of the discretized range u, which is taken to be the probability value of the midpoint of range u. The value of u is found by first setting $u = discretize_k(C_{ik})$, and then subtracting 1 from u if $C_{ik} < mid_{k,u}$. The value of $mid_{k,u}$ can be found as follows:

$$mid_{k,u} = min_k + width_k*(u+0.5)$$

4. *L'Eixample* heterogeneous weight-sensitive measure

After a theoretical and experimental analysis of some measures in real domains, it was assumed that an exponential weighting transformation would lead to a better attribute relevance characterization when the number of attributes, n, is very high. This exponential transformation allows amplifying the differences among attributes, when n becomes a large number. It has been experimentally tested that experts do not assign very extreme weights to attributes, as they do not want to be considered as very rigid experts in the field. After a preliminary competence study, a normalized weight-sensitive similarity measure was developed, and named as *L'Eixample* distance (Sànchez-Marrè et al., 1998). It takes into account the different nature of the quantitative or qualitative values of the continuous attributes depending on its relevance.

Main feature of *L'Eixample* measure is the sensitivity to weights for continuous attributes. For the most important continuous attributes, that is weight $> \alpha$, the distance is computed based on their qualitative values. This implies that relevant attributes having the same qualitative value are equal, and having different qualitative values are very different, even when a continuous measure would be very small. And for those less relevant ones, that is weight$\leq\alpha$, the distance is computed based on their quantitative values. This implies that non-relevant attributes having the same qualitative value are not equal, and those having different qualitative values are more similar (see Fig. 2).

For example, think of a continuous attribute like pH of water that is a common feature for many environmental system tasks such as wastewater treatment plants supervision, river water quality management or lake water quality management. The pH value is computed as follows:

$$\mathrm{pH} = -\log[\mathrm{H_3O^+}]$$

The pH values range from 0 to 14, with a common agreed discretization in three categories named as low (acidic), normal (neutral) and high (alkaline). Low values range from 0 to 6, normal values range from 6 to 8, and high values range from 8 to 14. In this situation, current values for pH measures of 5.8 and 6.3 could be very similar in a simple quantitative scale measurement, but in fact, both values are extremely different because they represent an acidic and a neutral characteristic of water. If this qualitative difference is really important (i.e. the attribute is relevant), then it should be measured in a better way. Thus, it would be better to compute the degree of similarity between both values within a qualitative scale. On the other hand, if the attribute relevance were not very important, then a quantitative scale could be used. This is what *L'Eixample* measure performs.

L'Eixample measure is defined as:

$$d(C_i,C_j) = \frac{\sum_{k=1}^{n} e^{w_k} \times d(C_{ik},C_{jk})}{\sum_{k=1}^{n} e^{w_k}}$$

where

$$d(C_{ik},C_{jk}) = \begin{cases} \dfrac{|qtv(C_{ik})-qtv(C_{jk})|}{upperval(k)-lowerval(k)} & \text{if } k \text{ is continuous and } w_k\leq\alpha \\ \dfrac{|qlv(C_{ik})-qlv(C_{jk})|}{\#mod(k)-1} & \text{if } k \text{ is continuous and } w_k > \alpha \\ & \text{or } k \text{ is ordered discrete} \\ 1-\delta_{qlv(C_{ik}),qlv(C_{jk})} & \text{if } k \text{ is non } - \text{ ordered discrete} \end{cases}$$

and C_i and C_j are two different cases. w_k is the weight of attribute k; C_{ik} is the value of the attribute k in the case i; C_{jk} is the value of the attribute k in the case C_j; $qtv(C_{ik})$ is the quantitative value of C_{ik}; $qtv(C_{jk})$ is the quantitative value of C_{jk}; $upperval(k)$ is the upper quantitative value of k; $lowerval(k)$ is the lower quantitative value of k; α is a cut point on the weight of the attributes; $qlv(C_{ik})$ is the qualitative value of C_{ik}; $qlv(C_{jk})$ is the qualitative value of C_{jk}; $\#mod(k)$ is the number of modalities (categories) of k; $\delta_{qlv(C_{ik}),qlv(C_{jk})}$ is the δ of Kronecker.

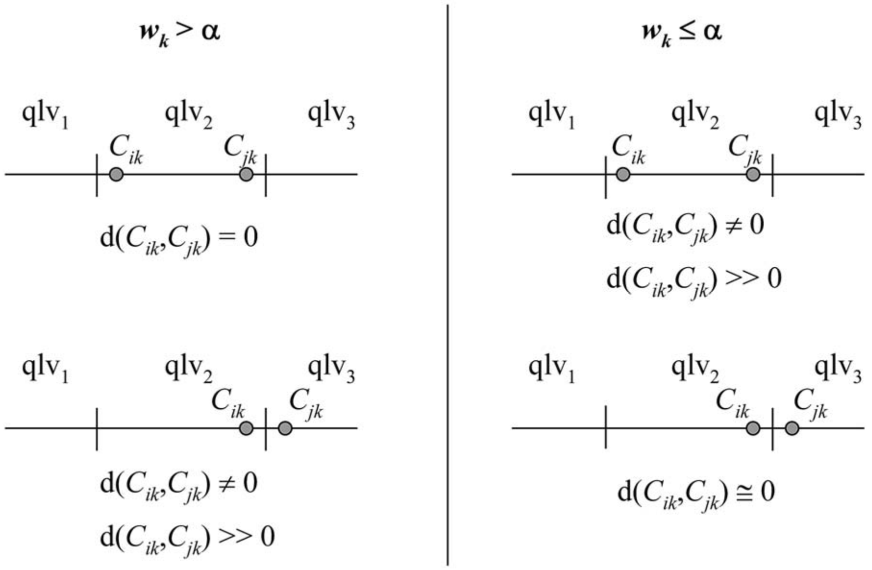

Fig. 2. Continuous attribute scenarios depending on weight w_k and values of C_{ik} and C_{jk}.

5. Experimental test

To test the efficiency of all similarity measures tested, a nearest neighbour classifier was implemented using each one of the seven similarity measures: HVDM, IVDM, Euclidean, Manhattan, Clark, Canberra and *L'Eixample*. Each measure was tested in two environmental databases as well as in 14 databases from the UCI database repository. Two real environmental databases were selected and tested: air pollution database and wastewater treatment plant database (WWTP). These databases were selected for several reasons. One is that they are the most easily available environmental databases for the study. Another one is that they represent extreme difficulty cases. The air pollution database has no missing values, while the WWTP database has an average of 35.8% of missing values and an imprecise labelling of cases due to multiple label setting by the experts. Finally, in both environmental domains, there were human experts available to help in the validation and interpretation of results.

The air pollution database contains information about the contamination level of the air in the central area of Mexico City. There are five continuous attributes indicating the presence of substances affecting the air quality: ozone, sulphur dioxide, nitrogen dioxide, carbon monoxide and total suspended particles. According to these values, a pollution-degree state is assigned to each case, which can be: Normal, No—satisfactory, Bad, and Too—bad. This database is available at http://www.sma.df.gob.mx/imecaweb/base—datos.htm.

The WWTP database describes the daily operation of a WWTP located in Catalonia. There are 15 attributes used for its characterization measured at different location points in the WWTP: the influent flow rate, the concentration of organic matter measured as chemical oxygen demand at the influent, the concentration of suspended solids at the influent, the concentration of total kjeldhal nitrogen (TKN) at the influent, the concentration of organic matter measured as chemical oxygen demand at the primary effluent, the concentration of suspended solids at the primary effluent, the concentration of biomass in the biological reactor, the settleability index of activated sludge (SVI), the sludge residence time, the food to microorganism ratio (F/M) in the biological reactor, the predominant filamentous organism in the biological reactor, the dissolved oxygen concentration of the biological reactor, the concentration of organic matter measured as chemical oxygen demand at the effluent, the concentration of suspended solids at the effluent, and the concentration of total nitrogen (TN) at the effluent (see Rodríguez-Roda et al., 2002 for a detailed description). Taking into account these features, an operational state label is assigned as the environmental situation. Twenty-four classes are used. Some of them have very few examples, making the classification process still more difficult.

5.1. Missing values

In Euclidean, Manhattan, Clark, Canberra and *L'Eixample* measures, a pre-processing task was carried out to substitute the missing input values by the average value of the instances with valid values. This was done for all the attributes. In the case of HVDM, a distance of 1 is given when one of the values compared or both are unknown. IVDM treats the unknown values as any another value. Thus, if the two values compared are both missing, the distance between them is 0 (Table 1).

5.2. Discretization

Some similarity measures have a good performance when the attributes are all continuous or all discrete. Others incorporate mechanisms to deal appropriately all types of attributes. Our proposal is to perform a discretization pre-process on the continuous attributes in such a way that the general accuracy can be improved (Dougherty et al., 1995). Discretization may serve to mark differences that are important in the problem domain. There exist many discretization algorithms in the literature, and had been compared among them to prove their general accuracy (Dougherty et al., 1995; Ventura and Martínez, 1995). To improve the retrieval accuracy, a global and supervised method to discretize all the continuous attributes, the CAIM algorithm proposed by Kurgan and Cios (2001), was selected. This algorithm tries to maximize the dependency relationship between the class label and the continuous-values attribute, and at the same time, to minimize the number of discrete intervals. In our approach, all the continuous attributes were divided in a number of intervals equal to the number of present classes in the database, or in five intervals when the number of present classes was less than 5. The class-attribute interdependency maximization (CAIM) criterion which measures the dependency between the class variable C and the discretization variable D for attribute F is defined as:

$$CAIM(C,D|F) = \frac{\sum_{r=1}^{n} \frac{max_r^2}{M_{+r}}}{n}$$

where n is the number of intervals; r iterates through all intervals, i.e. $r = 1,2,\ldots,n$; max_i is the maximum value among all q_{ir} values (maximum value within the rth column of the quanta matrix), $i = 1,2,\ldots,S$; M_{+r} is the total number of continuous values of attribute F that are within the interval $(d_{r-1},d_r]$ (Table 2).

The CAIM criterion is a heuristic measure that quantifies the interdependence between classes and the num-

Table 1
Major properties of databases considered in the experimentation

Database	Database characteristics					
	#Inst.	Cont.	Disc Ord.	Disc NOrd.	#Class.	%Mis.
Air pollution	365	5	0	0	4	0
WWTP	793	14	0	1	24	35.8
Auto	205	15	0	8	7	0.004
Bridges	108	3	0	8	3	0.06
Cleveland	303	5	2	6	2	0
Glass	214	9	0	0	7	0
Hepatitis	155	6	0	13	2	5.7
Horse-colic	301	7	0	16	2	30
Ionosphere	351	34	0	0	2	0
Iris	150	4	0	0	3	0
Liver disorders	345	6	0	0	2	0
Pima Indians diabetes	768	8	0	0	2	0
Soyabean (large)	307	0	6	29	19	21.7
Votes	435	0	0	16	2	7.3
Wine	178	13	0	0	3	0
Zoo	90	0	0	16	7	0

Table 2
Quanta matrix. Frequency matrix for attribute F and discretization scheme D

Class	Interval					Class total
	$[d_0,d_1]$	$\dots$	$[d_{r-1},d_r]$	$\dots$	$[d_{n-1},d_n]$	
C_1	q_{11}	$\dots$	q_{1r}	$\dots$	Q_{1n}	M_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_i	q_{i1}	$\dots$	q_{ir}	$\dots$	q_{in}	M_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_S	q_{S1}	$\dots$	q_{Sr}	$\dots$	q_{Sn}	M_{S+}
Interval total	M_{+1}	$\dots$	M_{+r}	$\dots$	M_{+n}	M

ber of unique values of the continuous attribute. For complete details about the CAIM algorithm, see Kurgan and Cios (2001).

5.3. *Weight assignment*

Although there are some global weighting schemes in the literature (Wettschereck et al., 1997; Jarmulak et al., 2000; Mohri and Tanaka, 1994), we use a new approach named as the class-value distribution (CVD), which was proposed in Núñez et al. (2002). This approach is based on estimated probabilities and correlation. The calculated values are in the range from 0 to 10 in ascending order of relevance. In this algorithm, a *correlation matrix is filled for each attribute*, representing the correlation between attribute values and class value as shown in Table 3. In this table, V_i is the i value of a discrete attribute. When the attribute is continuous, V_i represents one interval after the discretization process. C_j is the class j. q_{ij} is the number of instances that have value i and belong to class j. q_{+j} is the number of instances belonging to class j. q_{i+} is the number of instances that have value i. q_{++} is the number of instances in the training set.

Two main issues must be taken into account to set appropriates weights: the distribution of the values of the attribute across the classes, and the values associated to a class across the attribute values, which happens when all row values are 0 except one.

The first one shows how a single attribute value can determine a class. In the correlation matrix, this fact can be easily noticed by observing a single row. By observing a column, it is possible to determine the different attribute values that predict a class. In both cases, it will be ideal to find only one value different to 0 in each row and in the column where that value is. This indicates that one attribute value can predict a single class, and at the same time, one class is determined only by a single attribute value. The perfect attribute can be seen like a near-diagonal matrix. To take into account both class and value distribution, a score for each attribute must be calculated:

Table 3
Correlation matrix of an attribute

	C_1	C_2	…	C_n	Value total
V_1	q_{11}	q_{12}	…	q_{1n}	q_{1+}
V_2	q_{21}	q_{22}	…	q_{2n}	q_{2+}
⋮	⋮	⋮	…	⋮	⋮
V_m	q_{m1}	q_{m2}	…	q_{mn}	q_{m+}
Class total	q_{+1}	q_{+2}	…	q_{+n}	q_{++}

$$H_a = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_{max,i}}{q_{+,i}} * \frac{q_{max,i}}{q_{max,+}}\right)$$

where $q_{max,i}$ is the maximum value among all $q_{x,i}$ (maximum value within the ith column of the correlation matrix), q_{max+} is the number of instances that have the value $q_{max,i}$ (total of the row where the maximum values are).

The lowest limit will be $1/(|a| * n)$, where $|a|$ is the number of different feature values and n is the number of classes. The weight of the attribute is finally obtained by means of a scaling process:

$$W_a = \mathrm{int}\left(\frac{H_a - \frac{1}{|a| * n}}{1 - \frac{1}{|a| * n}} * 10\right)$$

In this approach, small addends are necessary to prevent possible zero division in very special conditions.

5.4. Evaluation

To verify the accuracy of the environmental situation classification in both environmental databases, and class prediction in the other databases, a test was implemented by means of a 10-fold cross-validation process. The average accuracy and standard deviation over all 10 trials are reported for each data test, and the highest accuracy achieved for each data set is shown in boldface in Table 4. Another feature was taken into account: the accuracy ordering among the measures, in order to show the accuracy quality of all measures, and not only the best one. For each data test, seven points were given to the best measure, until 1 point to the worst measure. Table 1 shows the number of instances in each database (#Inst.), the number of continuous attributes (Cont.), ordered discrete attributes (Disc Ord.), non-ordered discrete attributes (Disc NOrd.), number of classes (#Class.) and percentage of missing values (%Mis.).

From the experiments, it can be argued that *L'Eixample* measure accuracy mean seems to be better than the other measures in several tested domains. To ensure the experimental results, statistical significance tests were done to decide whether the differences between each one of the measures and *L'Eixample* measure were really significant or not. Results have shown that at 90% level of confidence, the differences between mean accuracy are statistically significant in most cases. Thus, *L'Eixample* measure is significatively better than the other ones, except HVDM measure, in the context of the experimental work done. At an 80% level of confidence, *L'Eixample* measure is significatively better than all other measures. And finally, at a 95% level of confidence, *L'Eixample* measure is significatively better than IVDM, Clark and Canberra measures.

From the accuracy results and from the definition of *L'Eixample* measure, it can be stated that in general, this measure is very well-suited for databases with a high number of continuous attributes where it exploits the domain knowledge about feature relevance to improve its performance.

6. Conclusions and future work

The main result of this paper is to show a comparison of several similarity measures to improve the classification of environmental situations. From Table 4, it can be argued that *L'Eixample* measure seems to outperform the others in a general case improving the performance of a CBR system. Thus, using *L'Eixample* similarity measure, the classification of environmental system situations can be improved. The average accuracy on seven of 16 databases is the highest, and also, the accuracy ordering punctuation is the best. This improvement is due to the fact that the domain knowledge of the experts has been taken into account in the measure, as it has been recognized by some researchers (Leake et al., 1997). For example, the weights assigned to the attributes have actually split them between important and irrelevant. Another feature is the proposal of an exponential weight transformation that gives more importance to separate important from irrelevant attributes. On the other hand, the most important contribution is the proposal of a weight-sensitive and heterogeneous function, in the sense of discretizing the most important continuous attributes to improve the retrieval process and to apply a different criterion of distance for continuous attributes. Some previous measures were presented as heterogeneous only by the fact of applying

Table 4
Generalization accuracy

Database	Similarity measures						
	HVDM	IVDM	Euclid	Manh	Clark	Canberra	*L'Eixample*
Air pollution	90.72	82.87	93.19	91.31	90.20	89.68	**99.44**
WWTP	49.64	32.73	**52.14**	50.47	46.42	47.26	50.47
Auto	81.98	**82.04**	74.87	77.87	71.29	77.41	79.37
Bridges	87.75	83.29	83.45	85.29	84.45	83.45	**92.37**
Cleveland	74.85	72.58	77.21	77.54	74.88	75.59	73.88
Glass	71.20	72.31	68.03	70.72	64.94	68.63	**73.18**
Hepatitis	78.47	79.51	81.16	79.83	**82.28**	80.98	75.83
Horse-colic	78.37	**79.60**	72.46	72.46	72.55	73.54	76.56
Ionosphere	90.01	83.76	85.46	89.73	88.59	90.02	**92.86**
Iris	92.66	92.66	**96.00**	94.00	95.33	94.00	**96.00**
Liver disorders	62.09	62.08	60.89	61.75	60.87	59.08	**65.48**
Pima Indians diabetes	69.65	65.63	71.08	69.89	64.30	65.60	67.35
Soyabean (large)	90.33	90.76	**91.95**	**91.95**	91.80	91.50	**91.95**
Votes	97.06	**97.58**	93.48	93.48	93.48	93.48	94.95
Wine	**98.50**	82.91	95.64	96.82	96.14	97.32	96.14
Zoo	**97.00**	95.00	96.00	96.00	96.00	96.00	95.00
Average accuracy	81.89	78.46	80.81	81.19	79.60	80.22	**82.55**
Standard deviation of accuracy	13.83	15.69	13.52	13.54	15.05	14.57	14.27
Accuracy ordering	75.00	49.00	74.00	76.00	54.00	60.00	83.00

different functions of distance to the different attribute types (Wilson and Martínez, 1997). A final remark in the analysis result must be made; a very poor accuracy is obtained for the WWTP database with all measures. This is principally due to the large amount of missing values present in all the attributes (35.8%). Moreover, there are six attributes, of a total of 15 attributes, which have more than 50% of missing values, even reaching an 88.9% in one feature, and also the labelling of cases is very imprecise due to multiple label setting by the experts, as mentioned before.

Although only the experimental results with a unique discretization method, i.e. the CAIM method, and with a unique feature weighting technique, i.e. the CVD method, have been presented in this paper, the performance of *L'Eixample* measure holds independently of the used discretization method or weighting scheme as shown in Núñez et al. (2003).

The direction of future investigations will be focussed mainly on studying the sensitivity of similarity assessment in the process of automatic discretization and in the automatic assignment of weights, and additionally, in assigning different weights for each interval found in the discretization step (local weighting schemes). Some preliminary work was reported in Núñez et al. (2002). Also, new environmental databases will be searched and tested. Only two were tested in this study, as they were the only two available to the authors.

Acknowledgements

This work has been supported by the Spanish CICyT projects TIC2000-1011 and REN2000-1755-C03, and also by EU project A-TEAM (IST 1999-10176). The authors wish to thank the useful comments of the anonymous reviewers to the preliminary version of the paper.

References

Avesani, P., Perini, A., Ricci, F., 2000. Interactive case-based planning for forest fire management. Applied Intelligence 13 (1), 189–206.

Brandau, R., Lemmon, A., Lafond, C., 1991. Experience with extended episodes: cases with complex temporal structure. In: Proceedings of the Workshop on Case-Based Reasoning (DARPA), Washington, DC,

Branting, L.K., Hastings, J.D., Lockwood, J.A., 1997. Integrating cases and models for prediction in biological systems. AI Applications 11 (1), 29–48.

Bridge, D., Defining and combining symmetric and asymmetric similarity measures. Proceedings. of the Fourth European. Workshop on Case-based Reasoning (EWCBR'98). LNAI-1488, pp. 52-63, 1998.

Dougherty, J., Kohavi, R., and Sahami, M., Supervised and Un^{su}pervised Discretization of continuous Features. Procc. Of the 12th International Conference on Machine Learning, pp. 194-202, 1995.

Jarmulak, J., Craw, S., and Rowe, R., Self-Optimising CBR Retrieval. Proceedings of the 12th IEEE International Conference on Tools with Arti^{f}icial Intelligence. pp. 376-383. 2000.

Joh, D.Y., CBR in a changing environment. Proceedings of the Second International Conference on Case-based Reasoning (ICCBR'97). LNAI-1266, pp. 53-62, 1997.

Jones, E., Roydhouse, A., 1995. Retrieving structured spatial information from large databases: a progress report. In: Proceedings of the IJCAI Workshop on Artificial Intelligence and the Environment, Montréal,, pp. 49–57.

Kolodner, J., 1993. Case-Based Reasoning. Morgan Kaufmann, Palo Alto, CA.

Kontkanen, P., Lathinen, J., Myllymäki, P., and Tirri, H., An unsupervised Bayesian distance measure. Proceedings of the Fifth European Workshop on Case-based Reasoning (EWCBR'2000). LNAI-1898, pp. 148-160, 2000.

Krovvidy, S., Wee, W.G., 1993. Wastewater treatment systems from case-based reasoning. Machine Learning 10, 341–363.

Kurgan, L., and Cios, K.J, Discretization Algorithm that Uses Class-Attribute Interdependence Maximisation, Proceedings of the 2001 International Conference on Artificial Intelligence (IC-AI 2001), pp.980-987, Las Vegas, Nevada.

Lance, G.N., Williams, W.T., 1966. Computer programs for hierarchical polythetic classification ("similarity analyses"). Computer Journal 9, 60–64.

Leake, D.B., Kinley, A., and Wilson, D., Case-based similarity assessment: estimating adaptability from experience. Procc. of National Conference on Artificial Intelligence (AAAI'97). pp. 674-679, 1997.

Liao, T.W., Zhang, Z., 1998. Similarity measures for retrieval in case-based reasoning systems. Applied Artificial Intelligence 12, 267–288.

Mohri, T., Tanaka, H., 1994. An optimal weighting criterion of case indexing for both numeric and symbolic attributes. In: Aha, D.W. (Ed.), Case-Based Reasoning papers from the 1994 Workshop. AAAI Press, Menlo Park, CA.

Núñez, H., Sànchez-Marrè, M., Cortés, U., Comas, J., Rodríguez-Roda, I., Poch, M., 2002. Feature weighting techniques for prediction tasks in environmental processes. In: ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI'2002), Lyon, France, July,

Núñez, H., Sànchez-Marrè, M., Cortés, U., 2003. Similarity measures in instance-based reasoning. Artificial Intelligence (submitted).

Osborne, H.R., and Bridge, D., A case-based similarity framework. Procc. of 3rd European. Workshop on Case-based Reasoning (EWCBR'96). LNAI-1168, pp. 309-323, 1996.

Osborne, H.R., and Bridge, D., Similarity metrics: a formal unification of cardinal and non-cardinal similarity measures. Procc. of 2nd International Conference on Case-based Reasoning (ICCBR'97). LNAI-1266, pp. 235-244, 1997.

Rodríguez-Roda, I., Poch, M., Sànchez-Marrè, M., Cortés, U., Lafuente, J., 1999. Consider a case-based system for control of complex processes. Chemical Engineering Progress 95 (6), 39–48.

Rodríguez-Roda, I., Comas, J., Colprim, J., Poch, M., Sànchez-Marrè, M., Cortés, U., Baeza, J., Lafuente, J., 2002. A hybrid supervisory system to support wastewater treatment plant operation: implementation and validation. Water Science & Technology 45 (4-5), 289–297.

Sànchez-Marrè, M., Cortés, U., Rodríguez-Roda, I., Poch, M., Lafuente, J., 1997. Learning and adaptation in WWTP through case-based reasoning. Microcomputers in Civil Engineering 12 (4), 251–266 (special issue on machine learning).

Sànchez-Marrè, M., Cortés, U., R-Roda, I., and Poch, M., L'Eixample Distance: a New Similarity Measure for Case retrieval. 1st Catalan Conference on Artificial Intelligence (CCIA'98). ACIA Bulletin 14-15:246-253. Tarragona, Catalonia. October, 1998.

Surma, J., Brauschweig, B., 1996. Case-based retrieval in process engineering: supporting design by reusing flowsheets. Engineering Applications of Artificial Intelligence 9 (4), 385–391.

Ventura, D., and Martínez, T.R., And Empirical Comparison of Discretization Methods. Procc. Of the 10th International Symposium on Computer and Information Sciences, pp. 443-450, 1995.

Wettschereck, D., Aha, D.W., Mohri, T., 1997. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review (special issue on lazy learning algorithms) 11, 273–314.

Wilson, D.R., Martínez, T.R., 1997. Improved heterogeneous distance functions. Journal of Artificial Intelligence Research 6, 1–34.