

# Classifying Environmental System Situations by means of Case-Based Reasoning: a Comparative Study

Héctor Núñez<sup>a</sup>, Miquel Sànchez-Marrè<sup>a</sup>, Ulises Cortés<sup>a</sup>, Quim Comas<sup>b</sup>,  
Montse Martínez<sup>b</sup> and Manel Poch<sup>b</sup>

<sup>a</sup>*Knowledge Engineering & Machine Learning Group. Universitat Politècnica de Catalunya. Campus Nord, Edifici C5. Jordi Girona 1-3. 08034 Barcelona. {hnunez, miquel, ia}@lsi.upc.es*

<sup>b</sup>*Laboratori d'Enginyeria Química i Ambiental. Universitat de Girona. Campus de Montilivi. 17071 Girona. {quim, montse, manel}@lequia.udg.es*

**Abstract:** The step of identifying to which class of operational situation belongs the current Environmental System situation is a key element to build successful Environmental Decision Support Systems (EDSS). Case-Based Reasoning (CBR) is a good technique to solve new problems based in previous experience. Main assumption in CBR relies in 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:** similarity assessment, environmental situation classification, case retrieval, case-based reasoning.

## 1. INTRODUCTION

The management of Environmental Systems is a very complex and dangerous task. The step of identifying to which class of operational situation belongs the current Environmental Systems situation is a key element to build successful Environmental Decision Support Systems (EDSS). If EDSS are able to make reliable diagnostics, then the proposed solutions by EDSS will be accurate and optimal enough to lead the Environmental System to a normal operation state. This diagnosis phase is especially difficult due to multiple features involved in most Environmental Systems, such as chemical, biological, physical, inflow-variability, microbiological and temporal. This is the reason why many Artificial Intelligence techniques have been used in recent past years, to try to solve these classification tasks. Integration of AI techniques in EDSS has led to obtain more accurate and reliable EDSS. Case-Based Reasoning (CBR) can be a good technique to make

diagnosis based in 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 derive 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. Theoretical frameworks for the systematic construction of similarity measures have been described in Osborne and Bridge [1996], Osborne and Bridge [1997], 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 to study the performance of several commonly used measures in practical use, for a better classification of environmental situations. In addition, *L'Exemple* 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. 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 ten-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 distance measures, we can refer to similarity measures by means of the relation:

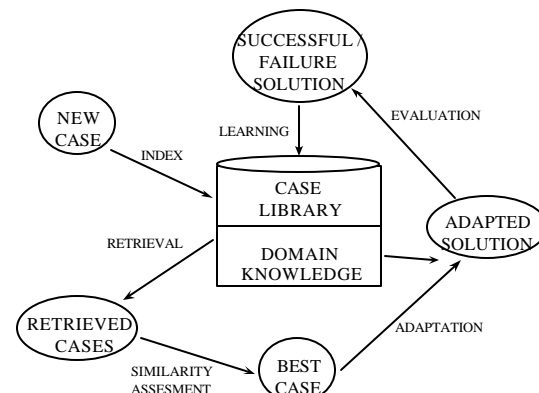
$$SIM(x, y) = 1 - DIST(x, y)$$

The paper is organised 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'Exemple* distance measure. Section 5 presents the results comparing the performance of all measures for classification tasks tested on two environmental databases and six 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 organise past experience and to learn how to solve new situations from previous past solutions. Case-based Reasoning in continuous situations has been applied in CIDA Joh [1997], an assistant for conceptual internetwork 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 optimisation of sequence operations for the design of wastewater treatment systems Krovvidy and Wee [1993], in supervisory systems for supervising and controlling WWTP management R-Roda *et al.* [1999], Sánchez-Marrè *et al.* [1997], in decision support systems for planning forest fire fighting Avesani *et al.* [1995], 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].



**Figure 1.** The general case-based reasoning paradigm

The basic reasoning cycle of a CBR agent can be summarised by a schematic cycle (see figure 1). In Aamodt and Plaza [1994] they adopt the four REs schema:

- *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* of 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 to a human oracle.
- *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).

## 3. SIMILARITY MEASURES

Most Case-based reasoners use a generalized weighted distance function such as,

$$dist(C_i, C_j) = \frac{\sum_{k=1}^n w_k * atr\_dist(C_{i,k}, C_{j,k})}{\sum_{k=1}^n w_k}$$

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] and 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, is very different to deal with only continuous data, with ordered discrete data or non-ordered discrete data. To give a greater distance contribution to an attribute than others less important attributes is necessary, too. In this study, our new proposed similarity measure, *L'Exemple*, is compared against some others measures that had been used before, with a very good performance in tests done in prior studies carried out. These selected similarity measures are:

### 3.1 Measures derived from Minkowski's metric

$$d(x_i, x_j) = \left( \sum_{k=1}^K |x_{ik} - x_{jk}|^r \right)^{1/r} \quad r \geq 1$$

Where k 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:

$$d(C_i, C_j) = \left( \frac{\sum_{k=1}^K weight^k * |d(A_{ki}, A_{kj})|^r}{\sum_{k=1}^K weight^k} \right)^{1/r}$$

Where for not ordered attributes, their contribution to the distance is,

$$d(A_{ki}, A_{kj}) = 1 - \delta_{kl} v(A_{ki}), v(A_{kj})$$

and  $\delta$  is the  $\delta$  of Kronecker

### 3.2 Unweighted similarity measures

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

*Clark*:

$$d(x_i, x_j) = \sum_{k=1}^K \frac{|x_{i,k} - x_{j,k}|^2}{|x_{i,k} + x_{j,k}|^2}$$

and *Canberra*:

$$d(x_i, x_j) = \sum_{k=1}^K \frac{|x_{i,k} - x_{j,k}|}{|x_{i,k} + x_{j,k}|}$$

### 3.3 Heterogeneous similarity measures

To obtain a broader study and results, other two 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)*:

$$HVDM(i, j) = \sqrt{\sum_{a=1}^m d_a^2(x_a, y_a)}$$

Where m is the number of attributes. The function  $d_a(x_a, y_a)$  returns a distance between the two values x and y for attribute a, and is defined as:

$$d_a^2(x_a, y_a) = \begin{cases} 1, & \text{if } x \text{ or } y \text{ is unknown} \\ \text{normalized\_vdm}_a(x, y), & \text{if } a \text{ is nominal} \\ \text{normalized\_diff}_a(x, y), & \text{if } a \text{ is linear} \end{cases}$$

Where *normalized\_vdm<sub>a</sub>(x,y)*, is defined as follows:

$$\text{Normalized\_vdm}_a(x, y) = \sqrt{\sum_{c=1}^C \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2}$$

Where:

- $N_{a,x}$  is the number of instances that have value x for attribute a;
- $N_{a,x,c}$  is the number of instances that have value x for attribute a and output class c;
- C is the number of output classes in the problem domain

The function *normalized\_diff<sub>a</sub>(x,y)*, is defined as showed below:

$$\text{normalized\_diff}_a(x, y) = \frac{|x - y|}{4\sigma_a}$$

where  $\sigma_a$  is the standard deviation of the numeric values of attribute a.

*Interpolated Value Difference Metric (IVDM)*:

$$IVDM(x, y) = \sum_{a=1}^m ivdm_a(x_a, y_a)^2$$

Where *ivdm<sub>a</sub>* is defined as:

$$ivdm_a(x, y) = \begin{cases} vdm_a(x, y) & \text{if } a \text{ is discrete} \\ \sum_{c=1}^C |p_{a,c}(x) - p_{a,c}(y)|^2 & \text{otherwise} \end{cases}$$

where  $vdm_a(x,y)$  is defined as follows:

$$vdm_a(x,y) = \sum_{c=1}^C |P_{a,x,c} - P_{a,y,c}|^2$$

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

$$P_{a,x,c} = \frac{N_{a,x,c}}{N_{a,x}}$$

Where  $N_{a,x}$  is the number of instances that have value x for attribute a;  $N_{a,x,c}$  is the number of instances that have value x for attribute a and output class c.

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

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

In this equation,  $mid_{a,u}$  and  $mid_{a,u+1}$  are midpoint of two consecutive discretized ranges such that  $mid_{a,u} \leq x < mid_{a,u+1}$ .  $P_{a,u,c}$  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  $u = discretize_a(x)$ , and then subtracting 1 from u if  $x < mid_{a,u}$ . The value of  $mid_{a,u}$  can be found as follows:

$$mid_{a,u} = min_a + width_a * (u+.5)$$

#### 4. L'EIXAMPLE DISTANCE MEASURE

We assumed that an *exponential* weighting transformation would be required for a better attribute relevance characterisation, when the number of attributes is very high. After a competence study, we developed a *normalised weight-sensitive distance function*, which was named as *L'Example* distance. It takes into account the different nature of the quantitative or qualitative values of the continuous attributes depending on its relevance.

*L'Example distance* is sensitive to weights. 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 equals, 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 equals, and having different qualitative values, are more similar. *L'Example distance* used to rank the best cases is:

$$d(C_i, C_j) = \frac{\sum_{k=1}^n e^{w_k} \times d(A_{ki}, A_{kj})}{\sum_{k=1}^n e^{w_k}}$$

where

$$d(A_{ki}, A_{kj}) = \begin{cases} \frac{|quantval(A_{ki}) - quantval(A_{kj})|}{upperval(A_k) - lowerval(A_k)} & \text{if } A_k \text{ is a continuous attribute and } w_k \leq \alpha \\ \frac{|qualval(A_{ki}) - qualval(A_{kj})|}{\# \text{ mod}(A_k) - 1} & \text{if } A_k \text{ is a continuous attribute and } w_k > \alpha \\ & \text{or } A_k \text{ is an ordered discrete attribute} \\ 1 - \delta_{qualval(A_{ki}), qualval(A_{kj})} & \text{if } A_k \text{ is a non ordered discrete attribute} \end{cases}$$

and,

$C_i$  is the case i;  $C_j$  is the case j;  $W_k$  is the weight of attribute k;  $A_{ki}$  is the value of the attribute k in the case i;  $A_{kj}$  is the value of the attribute k in the case j;  $qtv(A_{ki})$  is the quantitative value of  $A_{ki}$ ;  $qtv(A_{kj})$  is the quantitative value of  $A_{kj}$ ;  $A_k$  is the attribute k;  $upperval(A_k)$  is the upper quantitative value of  $A_k$ ;  $lowerval(A_k)$  is the lower quantitative value of  $A_k$ ;  $\alpha$  is a cut point on the weight of the attributes;  $qlv(A_{ki})$  is the qualitative value of  $A_{ki}$ ;  $qlv(A_{kj})$  is the qualitative value of  $A_{kj}$ ;  $\# \text{ mod}(A_k)$  is the number of modalities (categories) of  $A_k$ ;  $\delta_{qlv(A_{ki}), qlv(A_{kj})}$  is the  $\delta$  of Kronecker.

#### 5. EXPERIMENTAL TEST

To test the efficiency of all similarity measures tested, a nearest neighbour classifier was implemented using each one of the seven distance measures: HVDM, IVDM, Euclidean, Manhattan, Clark, Canberra and *L'Example*. Each distance measure was tested in two environmental databases as well as in six databases from the UCI database repository. Two real environmental data bases were selected and tested: Air Pollution database and Wastewater Treatment Plant database (WWTP). These databases were selected for several reasons. One is that they were the most easily available environmental databases for the study. Another one is that they represent extreme difficulty cases. The Air Pollution databases has no missing values, while the WWTP database has an average of 35.8% of missing values. 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 5 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, Too\_bad. This database is available at [www.sma.df.gob.mx/imecaweb/base\\_datos.htm](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. 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 very difficult.

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 over all 10 trials is reported for each data test, and the highest accuracy achieved for each data set is shown in boldface in table 1. 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, 7 points were given to the best measure, until 1 point to the worst measure. The table 1 also shows the number of instances in each database (#Inst.), the number of continuous attributes (Cont), ordered discrete attributes (Disc Ord), not ordered discrete attributes (Disc NOrd.), number of classes (#Class) and missing values percent (%Mis.).

## 5.1 Missing values

In Euclidean, Manhattan, Clark, Canberra and *L'Example* distance measures, a pre-processing task was carried out to substitute the missing input values by the average value obtained 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.** Generalization Accuracy

Database	Similarity Measures							Database Characteristics					
	HVDM	IVDM	Euclid	Manh	Clark	Canberra	<i>L'Example</i>	# Inst	Con	Disc Ord.	Disc NOrd	# Class	% Mis
WWTP	44.65	29.12	45.29	45.16	43.64	43.26	<b>45.42</b>	793	14	0	1	24	35.8
Air Pollution	91.93	92.74	97.23	96.14	90.98	89.90	<b>100</b>	365	5	0	0	4	0
Breast Cancer	94.99	95.57	95.68	<b>96.55</b>	96.35	96.54	<b>96.55</b>	699	0	9	0	2	0
Hepatitis	76.67	82.58	81.45	79.87	81.69	80.21	<b>83.45</b>	155	6	0	13	2	5.7
Horse-Colic	60.53	76.78	<b>78.72</b>	76.82	73.07	72.86	77.61	301	7	0	16	2	30
Iris	94.67	94.67	96	95.33	96	94.66	<b>97.33</b>	150	4	0	0	3	0
Pima Indians Diabetes	<b>71.09</b>	69.28	67.93	67.67	66.84	67.88	68.23	768	8	0	0	2	0
Soybean (large)	90.88	<b>92.18</b>	90.91	91.06	91.65	90.76	91.06	307	0	6	29	19	21.7
Average Accuracy:	78.17	79.11	81.65	81.07	80.02	79.51	<b>82.45</b>						
Accuracy ordering:	22	34	38	37	30	18	<b>50</b>						

## 5.2 Discretization and Weight Assignment

Some of the similarity measures have a good performance when the attributes are all continuous or all discrete. Others incorporate mechanisms to deal appropriately all the types of attributes. Our proposal is to make a discretization on the continuous attributes. Discretization serves to mark differences that are important in the problem domain. The continuous attributes were divided in a number of intervals equal to the number of present classes in the database. This division was made through a statistical analysis of the distribution of the classes and the values for each attribute.

As there is not any information about the relevance of attributes in the UCI databases, weights were set for each attribute to a value depending on the correlation level between the attribute and the class label. The assigned weights are in a rank of 0..10.

## 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 the table 1, can be argued that *L'Example* measure seems to outperform the other ones in a general case improving the performance of

a CBR system. Thus, the classification of Environmental System situations will be improved. The average accuracy on all the databases is the highest, and also the accuracy ordering punctuation is also 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 recognised by some researchers Leake *et al.* [1997]. For example, the weights assigned to the attributes have actually split them between important and irrelevant. Another important contribution is the proposal of a novel exponential weight transformation that gives more importance to separate important from irrelevant attributes. On the other hand, a heterogeneous function is proposed 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 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. This is principally due to the large amount of missing values present in all the attributes (35.8%). Moreover, there are 6 attributes, of a total of 15 attributes, which have more than 50% of missing values, even reaching an 88.9% in one feature. The direction of future investigations will be focused mainly on working 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.

## 7. ACKNOWLEDGEMENTS

This work has been supported by the Spanish CICYT project TIC2000-1011, and EU project A-TEAM (IST 1999-10176).

## 8. REFERENCES

- Aamodt A. and Plaza E. Case-based reasoning: fundamental issues, methodological variations and system approaches. *AI Communications* 7(1):39-59, 1994.
- Avesani, P., Perini, A. and Ricci, F. Interactive case-based planning for forest fire management. *Applied Intelligence* 13(1):189-206, 2000.
- Brandau R., Lemmon A. And Lafond C. Experience with extended episodes: cases with complex temporal structure. In *Proc. of Workshop on case-based reasoning (DARPA)*. Washington D.C., 1991.
- Branting. L.K., Hastings, J.D., and Lockwood, J.A. Integrating cases and models for prediction in biological systems. *AI Applications* 11(1):29-48, 1997.
- Bridge D. Defining and combining symmetric and asymmetric similarity measures. *Proc. of 4<sup>th</sup> Eur. Work. on Case-based Reasoning (EWCBR'98)*. LNAI-1488, pp. 52-63, 1998.
- Joh D.Y. CBR in a changing environment. *Proc. of 2<sup>nd</sup> Int. Conf. On Case-based Reasoning (ICCBR'97)*. LNAI-1266, pp. 53-62, 1997.
- Jones, E. and Roydhouse, A. Retrieving structured spatial information from large databases: a progress report. *Proc. of IJCAI Workshop on Artificial Intelligence and the Environment*, pp. 49-57, Montréal, 1995.
- Kontkanen P., Lathinen J., Myllymäki P. and Tirri H. An unsupervised Bayesian distance measure. *Proc. of 5<sup>th</sup> Eur. Work. on Case-based Reasoning (EWCBR'2000)*. LNAI-1898, pp. 148-160, 2000.
- Krovvidy S. and Wee W.G. Wastewater Treatment Systems from Case-Based Reasoning. *Machine Learning* 10, pp. 341-363, 1993.
- Leake D.B., Kinley A. and Wilson D. Case-based similarity assessment: estimating adaptability from experience. *Proc. of National Conference on Artificial Intelligence (AAAI'97)*. pp. 674-679, 1997.
- Liao T.W., and Zhang Z. Similarity measures for retrieval in case-based reasoning systems, *Applied Artificial Intelligence*, 12, 267-288, 1998.
- Osborne H.R. and Bridge D. Similarity metrics: a formal unification of cardinal and non-cardinal similarity measures. *Proc. of 2<sup>nd</sup> Int. Conf. On Case-based Reasoning (ICCBR'97)*. LNAI-1266, pp. 235-244, 1997.
- Osborne H.R. and Bridge D. A case-based similarity framework. *Proc. of 3<sup>rd</sup> Eur. Work. on Case-based Reasoning (EWCBR'96)*. LNAI-1168, pp. 309-323, 1996.
- R-Roda I., Poch M., Sánchez-Marrè M., Cortés U. and Lafuente J. Consider a Case-Based System for Control of Complex Processes. *Chemical Engineering Progress* 95(6):39-48, June 1999.
- Sánchez-Marrè M., Cortés U., R-Roda I., Poch M. and Lafuente J. Learning and Adaptation in WWTP through Case-Based Reasoning. Special issue on Machine Learning of *Microcomputers in Civil Engineering* 12(4):251-266. July, 1997.
- Surma, J. and Brauschweig, B. Case-Based Retrieval in Process Engineering: Supporting Design by Reusing Flowsheets. *Engineering Applications of Artificial Intelligence* 9(4): 385-391, 1996.
- Wilson D.R. and Martínez T.R. Improved Heterogeneous Distance Functions, *Journal of Artificial Intelligence Research*, 6, 1-34, 1997.