Analysing Similarity Assessment in Feature-Vector Case Representations

Héctor Núñez[1], Miquel Sànchez-Marrè[1], Ulises Cortés[1], Quim Comas[2], Ignasi Rodríguez-Roda[2] and Manel Poch[2]

[1]KnowledgeEngineering & Machine Learning Group, Technical Univ. of Catalonia
Campus Nord-Edifici C5, 08034 Barcelona, Catalonia, EU
{hnunez,miquel,ia}@lsi.upc.es
[2]Chemical and Environmental Engineering Laboratory (LEQUIA),
University of Girona, Campus de Montilivi, 17071 Girona, Catalonia, EU
{quim,ignasi,manel}@lequia.udg.es

Abstract. 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. CBR systems retrieve 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 particular domain, instead of derive them from scratch. Thus, similarity measures are key elements in obtaining reliable similar cases, which will be used to derive solutions for new cases. This paper describes a comparative analysis of several commonly used similarity measures, including a measure previously developed by the authors, and a study on its performance in the CBR retrieval step for feature-vector case representations. The testing has been done using sixteen data sets from the UCI Machine Learning Database Repository, plus two complex environmental databases.

1 Introduction

In CBR, similarity is used to decide which instance is closest to a new current case, and similarity measures have attracted the attention of many researchers in the field. Theoretical frameworks for the systematic construction of similarity measures have been described in [6], [5] and [1]. Other research work introduced new measures for a practical use in CBR systems, such as Bayesian distance measures in [2] and some heterogeneous difference metrics in [9]. Also, a review of some used similarity measures was done in [4]. This paper aims at analysing and studying the performance of several commonly used measures in practical use. In addition, *L'Eixample* distance, a similarity measure previously developed by the authors is introduced. This measure tries to improve the competence of a CBR system, providing flexibility and adaptation to real application domains where some attributes have a substantial higher impor-

tance 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 predictive accuracy on unseen cases, by means of a ten-fold cross-validation process. In this comparative analysis, 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) were selected. Although all these measures are really distance measures, you can refer to similarity measures by means of the following relation, where both similarity values and distance values are normalised:

$$Sim(x, y) = 1 - dist(x, y)$$

The paper is organised in the following way. Section 2 outlines main features about case retrieval and similarity assessment step in CBR systems. In Section 3, background information on selected distance measures is provided. Section 4 describes *L'Eixample* measure. Section 5 presents the results comparing the performance of all measures tested on 16 databases from the UCI Machine Learning Repository plus 2 complex environmental databases. Finally, in Section 6 conclusions and future research directions are outlined.

2 Case Retrieval and Similarity Assessment

A retrieval method should try to maximize the similarity between the actual case and the retrieved one(s). And this task usually implies the use of general domain knowledge. *Selecting* the best similar case(s), it is usually performed in most feature-vector CBR systems by means of some evaluation heuristic functions or distances, possibly domain dependent. Commonly, each attribute or dimension of a case has a determined importance value (weight), which is incorporated in the evaluation function. This weight could be static or dynamic depending on the CBR system purposes. Also, the evaluation function computes an absolute match score (a numeric value), although a relative match score between the set of retrieved cases and the new case can also be computed.

A large amount of CBR systems represent cases as a plain structure composed by a vector of feature-value pairs. In such a situation, these systems use a generalised weighted distance function, which can be described as:

$$dist(x, y) = \frac{\sum_{k=1}^{K} w_k * atr\_dist(x_k, y_k)}{\sum_{k=1}^{K} w_k}$$

where k is the number of attributes, x and y are whatever pair of cases, x_k is the value of the case x for the attribute k, and w_k is the weight or importance of the attribute k.

3 Similarity Measures

Currently, there are several similarity measures that have been used in feature-vector CBR systems, and some comparison studies exist among these similarity measures (see [9] and [4]). 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 lineal or quantitative data (continuous), with entire or qualitative (discrete) or nominal (discrete not ordered). To give a greater distance contribution to an attribute than other less important attributes is necessary, too. In this study, several similarity measures were tested.

L'Eixample measure showed a better performance than some other measures, as it was found in a preliminary, but restricted comparison in a unique domain and only against Minkowski's metrics [7]. Here, it is analysed and compared against some others measures that had been used before in the CBR community. These selected similarity measures for the study were:

3.1 Measures derived from Minkowski's metric

$$d(x,y) = \left(\sum_{k=1}^{K} |x_k - y_k|^r \right)^{1/r} \quad r \geq 1$$

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

$$d(x,y) = \left(\frac{\sum_{k=1}^{K} w_k^r * |atr\_dist(x_k, y_k)|^r}{\sum_{k=1}^{K} w_k^r} \right)^{1/r}$$

Where $atr\_dist(x_k,y_k)$ is:

$$atr\_dist(x_k, y_k) = \begin{cases} |x_k - y_k| & \text{if } k \text{ is continuous} \\ 0 & \text{if } x_k = y_k \text{ and } k \text{ is discrete} \\ 1 & \text{if } x_k \neq y_k \text{ and } k \text{ is discrete} \end{cases}$$

3.2 Unweighted similarity measures

In this study, two similarity measures ignoring attribute's weight were included:

Clark:

$$d(x,y)=\sum_{k=1}^{K}\frac{|x_k-y_k|^2}{|x_k+y_k|^2}$$

and *Canberra:*

$$d(x,y)=\sum_{k=1}^{K}\frac{|x_k-y_k|}{|x_k+y_k|}$$

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 [9]:

Heterogeneous Value Difference Metric (HVDM):

$$HVDM\ (x,y)=\sqrt{\sum_{k=1}^{K}d_k^2(x_k,y_k)}$$

Where K is the number of attributes. The function $d_k(x_k,y_k)$ returns a distance between the two values x_k, y_k for attribute k, and is defined as:

$$d_k^2(x_k,y_k)=\begin{cases}1, & \text{if } x_k \text{ or } y_k \text{ is missing, otherwise}\\ normalized\_vdm_k(x_k,y_k), & \text{if } k \text{ is discrete}\\ normalized\_diff_k(x_k,y_k), & \text{if } k \text{ is continuous}\end{cases}$$

Where $normalized\_vdm_k(x_k,y_k)$, is defined as follows:

$$normalized\_vdm_k(x_k,y_k)=\sqrt{\sum_{c=1}^{C}\left|\frac{N_{k,x,c}}{N_k}-\frac{N_{k,y,c}}{N_{k,y}}\right|^2}$$

Where:

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

The function $normalized\_diff_k(x_k,y_k)$, is defined as showed below:

$$normalized\_diff_k(x_k,y_k)=\frac{|x_k-y_k|}{4\sigma_k}$$

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

Interpolated Value Difference Metric (IVDM):

$$IVDM(x, y) = \sum_{k=1}^{K} ivdm_k(x_k, y_k)^2$$

Where $ivdm_k$ is defined as:

$$ivdm_k(x_k, y_k) = \begin{cases} vdm_k(x_k, y_k) & \text{if } k \text{ is discrete} \\ \sum_{c=1}^{C} \left| p_{k,c}(x_k) - p_{k,c}(y_k) \right|^2 & \text{otherwise} \end{cases}$$

where $vdm_k(x_k,y_k)$ is defined as follows:

$$vdm_k(x_k, y_k) = \sum_{c=1}^{C} \left| P_{k,x_k,c} - P_{k,y_k,c} \right|^2$$

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

$$P_{k,x_k,c} = \frac{N_{k,x_k,c}}{N_{k,x_k}}$$

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

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

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

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

$$mid_{k,u} = min_k + width_k * (u+.5)$$

4 *L'Eixample* 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 characterisation, when the number of attributes, k, is very high. This exponential transformation allows amplifying the differences among attributes, when k becomes a large number. It has been experimentally tested that experts don't assign very extreme weights to attributes, as they don't want to be considered as very rigid experts in the field. After a preliminary competence study, a normalised weight-sensitive distance function was developed, and named as *L'Eixample* distance [7]. It takes into account the different nature of the quantitative or qualitative values of the continuous attributes depending on its relevance.

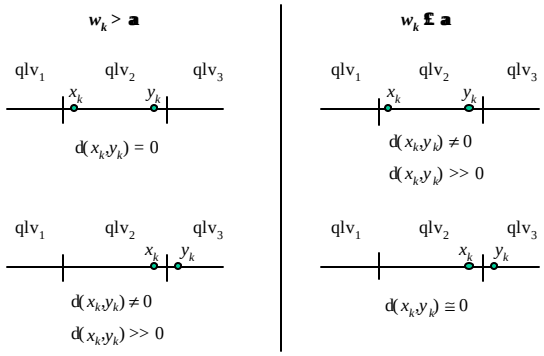

Fig. 1. Continuous attribute scenarios depending on weight w_k and values of x_k and y_k.

But main feature of *L'Eixample* distance 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 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. See Fig. 1.

L'Eixample distance used to rank the best cases is:

$$d(x,y) = \frac{\sum_{k=1}^{K} e^{w_k} \times d(x_k, y_k)}{\sum_{k=1}^{K} e^{w_k}}$$

where

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

and,

x and y are two different cases. W_k is the weight of attribute k; x_k is the value of the attribute k in the case x; y_k is the value of the attribute k in the case y; $qtv(x_k)$ is the quantitative value of x_k; $qtv(y_k)$ is the quantitative value of y_k; $upperval(k)$ 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; $qlv(x_k)$ is the qualitative value of x_k; $qlv(y_k)$ is the qualitative value of y_k; $\#mod(Ak)$ is the number of modalities (categories) of k; $\delta_{qlv(xk),qlv(yk)}$ is the δ of Kronecker.

5 Experimental Set-up and Evaluation

To test the efficiency of all similarity measures analysed, a nearest neighbour classifier was implemented using each one of the 7 distance measures: HVDM, IVDM, Euclidean, Manhattan, Clark, Canberra and *L'Eixample*. Each distance measure was tested in the 16 selected databases from the UCI database repository, and 2 complex environmental databases. 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 retrieval in a CBR system, a test by means of a 10-fold cross-validation process was implemented. The table 1 shows the number of instances in each database (#Inst.), the number of continuous attributes (Cont), ordered discrete attributes (OrdDisc), no ordered discrete attributes (NoOrdDisc), number of classes (#Class) and missing values percentage (%Mis.).

5.1 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 that are very important in the classification of the cases. Discretization serves to mark differences that are important in the problem domain. However, all the continuous attributes are not discretized. A value alpha that is set by the expert is incorporated, in

such a way that the continuous attributes having a weight higher than alpha (very important attribute), are discretized to make their similarity more determining. The continuous attributes were divided in a number of intervals equal to the number of present classes in the database.

As there is not any information about the relevance of attributes in the UCI databases, weights for all databases were set for each attribute using the global weighting assignment method *CV* described in [8] using the correlation level between the attribute and the class label. The assigned weights are in a rank of 0 to10. We fixed a value of 8.0 for alpha in the *L'Eixample* similarity measure.

Table 1. Major properties of databases considered in the experimentation

	Database Characteristics					
Database	#Inst	Cont	OrdDisc	NoOrdDisc	#Class	%Mis
Air pollution	365	5	0	0	4	0
Auto	205	15	0	8	7	0.004
Bridges	108	3	0	8	3	0.06
Breast Cancer	699	0	9	0	2	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
Monks-1	432	0	0	6	2	0
Monks-2	432	0	0	6	2	0
Monks-3	432	0	0	6	2	0
Pima Indians Diabetes	768	8	0	0	2	0
Soybean (large)	307	0	6	29	19	21.7
Votes	435	0	0	16	2	7.3
Wine	178	13	0	0	3	0
WWTP	793	14	0	1	24	35.8
Zoo	90	0	0	16	7	0

5.2 Evaluation

The average accuracy and standard deviation of 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 2. 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.

From the experiments, it can be argued that *L'Eixample* measure accuracy mean seems to be better than the other measures in most tested domains. Also, its standard deviation is the lowest one. To ensure the experimental results, statistical significance tests were done to decide whether the differences between each of the measures and *L'Eixample* measure were really significant or not. Results have shown that at 95% level of confidence, the differences between mean accuracy are statistically significant.

Thus, *L'Eixample* measure is significatively better than the other ones in the context of the experimental work done.

Table 2. Generalization accuracy results.

	Similarity Measures						
Database	HVDM	IVDM	Euclid	Manh	Clark	Canberra	*L'Eixample*
Air pollution	95.88	88.30	97.25	97.25	91.05	90.50	**98.64**
Auto	78.76	68.26	74.34	81.65	72.86	75.86	**82.67**
Bridges	81.36	88.27	87.59	85.18	79.36	81.23	**89.18**
Breast Cancer	94.99	95.57	95.68	**96.55**	96.35	96.54	**96.55**
Hepatitis	76.67	82.58	81.45	79.87	81.69	80.21	**83.45**
Horse-Colic	60.53	76.78	**78.72**	76.82	73.07	72.86	77.61
Ionosphere	86.32	91.17	84.05	91.19	83.18	88.88	**91.47**
Iris	94.67	94.67	96	95.33	96	94.66	**97.33**
Liver Disorders	62.92	58.23	60.73	60.25	64.16	60.75	**65.72**
Monks-1	68.09	68.09	70.20	68.20	61.08	61.08	**73.92**
Monks-2	65.72	66.55	66.23	66.21	79.85	79.85	**82.68**
Monks-3	95.50	93.45	93.15	93.15	92.40	92.40	**97.29**
Pima Indians Diabetes	**71.09**	69.28	67.93	67.67	66.84	67.88	68.23
Soybean (large)	90.88	**92.18**	90.91	91.06	91.65	90.76	91.06
Votes	95.17	95.17	94.89	93.68	93.84	93.84	**95.97**
Wine	99.41	96.42	**99.58**	**99.58**	96.64	98.23	98.40
WWTP	**44.65**	29.12	40.50	43.70	36.31	37.19	43.50
Zoo	97.78	**98.89**	97	97	96	96	98
Average Accuracy	81.13	80.72	82.01	82.46	80.69	81.04	**85.09**
St. Dev. of Accuracy	16.00	18.07	16.06	15.73	16.19	16.24	**14.86**
Accuracy ordering	71	71	79	78	56	55	**116**

6 Conclusions and Future Work

The main result of this paper is to show a comparison of several similarity measures. From the table 1 and from the statistical test carried out, it can be argued that *L'Eixample* measure outperforms the other ones in a general case improving the performance of a CBR system. The average accuracy on all the databases is the highest, the standard deviation is the lowest, 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 recognised by some researchers [3]. For example, the weights assigned to the attributes have actually split them between important and irrelevant. Another important contribution is the proposal of an exponential weight transformation that helps 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 different criteria of distance for different attribute types. Some previous measures were presented as heterogeneous only by the fact of applying different functions of distance to the different attribute types [9].

Main drawback of the approach is that *L'Eixample* measure is very sensitive to the discretization process and to the weight assignment. This fact was found out in a sensitivity analysis done with the databases. For this reason, the direction of future investigations is being mainly focused on working in the process of automatic discretization and in the automatic assignment of weights, and additionally, in assigning different weights in each interval found in the discretization step (local weighting schemes). Some preliminary work was reported in [8].

Acknowledgements

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

References

1. D. Bridge. Defing and combining symmetric and asymmetric similarity measures. Procc. of *4th Eur. Work.. on Case-based Reasoning (EWCBR'98)*. LNAI-1488, pp. 52-63, 1998.
2. P. Kontkanen, J. Lathinen, P. Myllymäki and H. Tirri.. An unsupervised Bayesian distance measure. Procc. of *5th E. W. on Case-based Reasoning (EWCBR'2000)*. LNAI-1898, pp. 148-160, 2000.
3. D.B. Leake, A. Kinley and D. Wilson. Case-based similarity assessment: estimating adaptability from experience. Procc. of *National Conference on Artificial Intelligence (AAAI'97)*. pp. 674-679, 1997.
4. T.W. Liao, and Z. Zhang. Similarity measures for retrieval in case-based reasoning systems, *Applied Artificial Intelligence*,12,267-288,1998.
5. H.R. Osborne and D. Bridge. Similarity metrics: a formal unification of cardinal and non-cardinal similarity measures. Procc. of *2nd Int. Conf. On Case-based Reasoning (ICCBR'97)*. LNAI-1266, pp. 235-244, 1997.
6. H.R. Osborne and D. Bridge. A case-based similarity framework. Procc. of *3rd Eur. Work.. on Case-based Reasoning (EWCBR'96)*. LNAI-1168, pp. 309-323, 1996.
7. M. Sànchez-Marrè, U. Cortés, I. R-Roda & M. Poch. 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.
8. H. Núñez, M. Sànchez-Marrè, U. Cortés, J. Comas, Ignasi R-Roda and M. Poch. Feature Weighting Techniques for Prediction Tasks in Environmental Processes. *ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI'2002)*. Lyon, France, July 2002
9. D.R. Wilson and T.R.Martínez. Improved Heterogeneous Distance Functions, *Journal of Artificial Intelligence Research*, 6, 1-34, 1997.