Comparing Dissimilarity Measures for Symbolic Data Analysis

Donato MALERBA, Floriana ESPOSITO,
Vincenzo GIOVIALE and Valentina TAMMA[†]
Dipartimento di Informatica, University of Bari
Via Orabona 4 – 70126 Bari, Italy
e-mail: {malerba, esposito}@di.uniba.it,
vincenzo_gio@yahoo.it, valli@csc.liv.ac.uk

Abstract: Symbolic data analysis aims at extending classical data analysis techniques to manage symbolic data, which are a form of aggregated data. This extension passes through the definition of a new set of dissimilarity measures. Many of them have been reported in the literature, but no comparative study has been performed. This paper presents an empirical evaluation of dissimilarity measures proposed for a restricted class of symbolic data, namely Boolean symbolic objects. To define a ground truth for the empirical evaluation, a data set with a fully understandable and explainable property has been selected. Empirical results show a variety of, sometimes unexpected, behaviours of the compared measures.

Keywords: Symbolic data analysis, Boolean symbolic objects, dissimilarity measure.

1. Introduction

Most of statistical techniques for data analysis have been designed for a relatively simple situation: the unit for statistical analysis is an individual (e.g., a person or an object) described by a well defined set of random variables (either qualitative or quantitative), each of which result in just one single value.

Nowadays, data analysts are confronted with new challenges: they are asked to process data that go beyond the classical framework, as in the case of data concerning more or less homogeneous classes or groups of individuals (*second-order objects*) instead of single individuals (*first-order objects*). A typical situation is that of census data, which raise privacy issues in all governmental agencies that distribute them. To guarantee that data analysts cannot identify an individual or a single business establishment, data are made available in aggregate form. Data aggregations by census tracts or by enumeration districts are examples of second-order objects.

[†] Presently at the Department of Computer Science, The University of Liverpool, Chadwick Building, Peach St., Liverpool L69 7ZF, UK.

Aggregated data describe a group of individuals by *set-valued* or *modal* variables. A variable Y defined for all elements k of a set E is termed set-valued with the domain $\mathcal{Y}$ if it takes its values in $\boldsymbol{P}(\mathcal{Y})=\{U \mid U \subseteq \mathcal{Y}\}$, that is the power set of $\mathcal{Y}$. When $Y(k)$ is finite for each k, than Y is called *multi-valued*. A single-valued variable is a special case of set-valued variable for which $|Y(k)|=1$ for each k. When an order relation $\prec$ is defined on $\mathcal{Y}$ then the value returned by a set-valued variable can be expressed by an interval $[\alpha,\beta]$, and Y is termed an *interval* variable. More generally, a *modal* variable is a set-valued variable with a measure or a (frequency, probability or weight) distribution associated to $Y(k)$.

A class or group of individuals described by a number of set-valued or modal variables is termed *symbolic data.* Symbolic data lead to more complex data tables called *symbolic data tables*. The extension of classical data analysis techniques to such tables is termed *symbolic data analysis* [1]. Implementing an integrated software environment for both the construction of symbolic data tables from records of individuals and the analysis of symbolic data has been the general aim of the three-years ESPRIT project SODAS[1] (Symbolic Official Data Analysis System), concluded in November 1999. The recently started three-years IST project ASSO (Analysis System of Symbolic Official Data) is intended to improve the SODAS prototype with respect to several aspects (management of symbolic data, new or improved data analysis methods, new visualization techniques).

An important module of the SODAS software is that concerning the computation of some dissimilarity measures. Indeed, the extension of statistical techniques to symbolic data requires the specification of some dissimilarity (or conversely, similarity) measures. Many formulations of dissimilarity measures for symbolic data have been reported in literature [5], but no comparative study on their suitability to real-world problems has been performed. This paper presents an empirical evaluation of dissimilarity measures proposed for a restricted class of symbolic data, namely Boolean symbolic objects (BSOs). To define a ground truth for the empirical evaluation, a data set with a fully understandable and explainable property has been selected. Empirical results show a variety of, sometimes unexpected, behaviours of the compared measures. The reported experimentation is the first step towards the fulfillment of one of the objectives of the ASSO project, namely comparing various dissimilarity measures to support users in selecting the best measure for their data analysis problem.

2. Dissimilarity measures for Boolean Symbolic Objects

Henceforth, the term *dissimilarity measure d* on a set of objects E refers to a real valued function on $E\times E$ such that: $d_a^* = d(a,a) \leq d(a,b) = d(b,a) < \infty$ for all $a,b\in$E. Conversely, a *similarity measure s* on a set of objects E is a real valued function on $E\times E$ such that: $s_a^* = s(a,a) \geq s(a,b) = s(b,a) \geq 0$ for all $a,b\in$E. Generally, $d_a^* = d^*$ and $s_a^* = s^*$ for each object a in E, and more specifically, $d^* = 1$ while $s^* = 0$. Studies on their properties can be limited to dissimilarity measures alone, since it is always possible to transform a similarity

[1] The SODAS software is freely distributed by CISIA at the following URL address: http://www.cisia.com/download.htm.

measure into a dissimilarity one with the same properties. Many methods have been reported in the literature to derive dissimilarity measures from a matrix of observed data [4], or, more generally, for a set of symbolic objects [5]. In the following, only some measures proposed for BSOs are briefly reported.
Let a and b be two BSOs:

$a = [Y_1=A_1] \wedge [Y_2=A_2] \wedge ... \wedge [Y_p=A_p]$

$b = [Y_1=B_1] \wedge [Y_2=B_2] \wedge ... \wedge [Y_p=B_p]$

where each variable Y_j takes values in a domain γ_j and A_j and B_j are subsets of γ_j. It is possible to define a dissimilarity measure between two BSOs a and b by aggregating dissimilarity values computed independently at the level of single variables Y_j (*componentwise dissimilarities*). A classical aggregation function is the generalized Minkowski metric. However, another class of measures defined for BSOs is based on the the notion of *description potential,* $\pi(a)$, which is defined as the *volume* of the Cartesian product $A_1 \times A_2 \times ... \times A_p$. For this class of measures, no componentwise decomposition is necessary, so that no function is required to aggregate dissimilarities computed independently for each variable.

The list of dissimilarity measures considered in this study are reported in Table 1, where they are denoted as in the SODAS software, namely:
U_1: Gowda and Diday's dissimilarity measure [6];
U_2: Ichino and Yaguchi's first formulation of a dissimilarity measure [7];
U_3: Ichino and Yaguchi's dissimilarity measure normalized w.r.t. domain length [7];
U_4: Ichino and Yaguchi's normalized and weighted dissimilarity measure [7];
SO_1: De Carvalho's dissimilarity measure [2];
SO_2: De Carvalho's extension of Ichino and Yaguchi's dissimilarity [2];
SO_3: De Carvalho's first dissimilarity measure based on description potential [3];
SO_4: De Carvalho's second dissimilarity measure based on description potential [3];
SO_5: De Carvalho's normalized dissimilarity measure based on description potential [3];
C_1: De Carvalho's normalized dissimilarity measure for constrained BSOs [3].

The term constrained BSO refers to the fact that some dependences between variables are defined, namely either *hierarchical dependences* (mother-daughter) which establish conditions for some variables being not measurable (not-applicable values), or *logical dependences* which establish the set of possible values for a variable Y_j conditioned by the set of values taken by another variable Y_k. An investigation of the effect of constraints on the computation of dissimilarity measures is out of the scope of this paper, nevertheless it is always possible to apply the measures defined for constrained BSOs to unconstrained BSOs. This explains why C_1 has been considered in the empirical comparison reported in the next section.

3. Experimental evaluation

Many dissimilarity measures have been proposed for symbolic data analysis, nevertheless they have never been compared in order to understand both their common and their peculiar properties. In this section, an empirical evaluation is reported with reference to a data set for which a desirable behaviour of a dissimilarity measure can be defined. The data set is called "Abalone data" and is available at the UCI Machine Learning Repository

Table 1. Dissimilarity measures available in the DI method for BSO's, and related parameters.

Name	***Componentwise dissimilarity measure***	***Objectwise dissimilarity measure***
U_1	$D^{(j)}(A_j, B_j) = D_\pi(A_j, B_j) + D_s(A_j, B_j) + D_c(A_j, B_j)$ where $D_\pi(A_j, B_j)$ *is* due to *position*, $D_s(A_j, B_j)$ is due to *spanning* and $D_c(A_j,B_j)$ is due to *content*.	$d(a, b) = \sum_{j=1}^{p} \mathrm{D}^{(\mathrm{j})}(\mathrm{Aj}, \mathrm{Bj})$
U_2	$\phi(A_j, B_j) = \lvert A_j \oplus B_j \rvert - \lvert A_j \otimes B_j \rvert + \gamma(2\lvert A_j \otimes B_j \rvert - \lvert A_j \rvert - \lvert B_j \rvert)$ where *meet* ($\otimes$) and *join* ($\oplus$) are two Cartesian operators.	$d_q(a, b) = \sqrt[q]{\sum_{j=1}^{p} \left[\phi(A_j, B_j)\right]^q}$
U_3	$\psi(A_j, B_j) = \frac{\phi(A_j, B_j)}{\lvert Y_j \rvert}$	$d_q(a, b) = \sqrt[q]{\sum_{j=1}^{p} \left[\psi(A_j, B_j)\right]^q}$
U_4	$\psi(A_j, B_j) = \frac{\phi(A_j, B_j)}{\lvert Y_j \rvert}$	$d_q(a, b) = \sqrt[q]{\sum_{j=1}^{p} w_j \left[\psi(A_j, B_j)\right]^q}$
SO_1	$d_i(A_j,B_j)$ $i=1, \ldots,5$	$\mathrm{d}^{\mathrm{i}}_q(a, b) = \sqrt[q]{\sum_{j=1}^{p} \left[w_j\, d_i(A_j, B_j)\right]^q}$
SO_2	$\psi'(A_j, B_j) = \frac{\phi(A_j, B_j)}{\mu(A_j \oplus B_j)}$	$d'_q(a, b) = \sqrt[q]{\sum_{j=1}^{p} \frac{1}{p} \left[\psi'(A_j, B_j)\right]^q}$
SO_3	none	$d'_1(a, b) = \pi(a \oplus b) - \pi(a \otimes b) + \gamma(2\pi(a \otimes b) - \pi(a) - \pi(b))$
SO_4	none	$d'_2(a,b) = \frac{\pi(a \oplus b) - \pi(a \otimes b) + \gamma(2\pi(a \otimes b) - \pi(a) - \pi(b))}{\pi(a^E)}$ where $\pi(a^E) = [Y_1 \in \mathsf{Y}_1] \wedge \ldots \wedge [Y_p \in \mathsf{Y}_p]$
SO_5	none	$d'_3(a,b) = \frac{\pi(a \oplus b) - \pi(a \otimes b) + \gamma(2\pi(a \otimes b) - \pi(a) - \pi(b))}{\pi(a \oplus b)}$
C_1	$d_i(A_j,B_j)$ $i=1, \ldots,5$	$\mathrm{d}^{\mathrm{i}}_q(a, b) = \sqrt[q]{\frac{\sum_{j=1}^{p} \left[w_j\, d_i(A_j, B_j)\right]^q}{\sum_{J=1}^{p} \delta(j)}}$, where $\delta(j)$ is the indicator function

Table 2. Attributes of the Abalone data set

Attribute Name	***Data Type***	***Unit***	***Description***
Sex	Nominal		M, F. I (infant)
Length	Continuous	mm	Longest shell measurement
Diameter	Continuous	mm	Perpendicular to length
Height	Continuous	mm	Measured with meat in shell
Whole weight	Continuous	grams	Weight of the whole abalone
Shucked weight	Continuous	grams	Weight of the meat
Viscera weight	Continuous	grams	Gut weight after bleeding
Shell weight	Continuous	grams	Weigh of the dried shell
Rings	Integer		Number of rings

(URL: http://www.ics.uci.edu/~mlearn/MLRepository.html). It contains 4177 cases of marine crustaceans, which are described by means of the nine attributes listed in Table 2. There are no missing values in the data.

Generally this data set is used for prediction tasks. The number of rings (last attribute) is the value to be predicted from which it is possible to know the age in years of the crustacean by adding 1.5 to the number of rings. Since the *dependent* attribute is integer-valued, this database has been extensively investigated in empirical studies concerning regression-tree induction [8,10]. The number of rings varies between 1 and 29, with sample mean equal to 9.934 and sample standard deviation equal to 3.224 (there are few cases of crustaceans with less than 3 or more than 25 rings). The performance of regression-tree induction systems reported in the literature is generally high, meaning that the eight *independent* attributes are actually sufficiently informative for the intended prediction task. In other words, two abalones with the same number of rings should also present similar values for the attributes sex, length, diameter, height, and so on. Basing upon this consideration we expect to observe that *the degree of dissimilarity between crustaceans computed on the independent attributes do actually be proportional to the dissimilarity in the dependent attribute* (i.e., difference in the number of rings). We will call this property as *monotonic increasing dissimilarity* (shortly, *MID property*).

Abalone data can be aggregated into symbolic objects, each of which correspond to a range of values for the number of rings. In particular, nine BSOs have been generated by applying the DB2SO facility [9] available in the SODAS software (see Table 3).

The dissimilarity measures briefly presented in Section 2 have been applied to the BSOs previously illustrated. In this example the value of the parameter γ is set to 0.5, the order of power q is 2 and the weights are uniformly distributed. Results are depicted in Figure 1. Dissimilarities are reported along the vertical axis, while BSOs are listed along the horizontal axis, in ascending order with respect to the number of rings. Each line represents the dissimilarity between a given BSO and the subsequent BSOs in the list. The number of lines in each graph is eight, since there are nine BSOs. For the sake of clarity, the lower triangular matrix of the dissimilarities depicted in the graph labeled 'Abalone- U_1' is reported in Table 4.

Table 3. Boolean symbolic objects generated by partitioning the *Rings* attribute into nine intervals of equal length

BSO	Rings	Sex	Length	Diameter	Height	Whole	Shucked	Viscera	Shell
1	**1-3**	I,M	[0.08:0.24]	[0.05:0.17]	[0.01:0.06]	[0.00:0.07]	[0.00:0.03]	[0.00:0.01]	[0.00:0.02]
2	**4-6**	I,M,F	[0.13:0.66]	[0.09:0.47]	[0.00:0.18]	[0.01:1.37]	[0.00:0.64]	[0.00:0.29]	[0.00:0.35]
3	**7-9**	I,M,F	[0.20:0.75]	[0.16:0.58]	[0.00:1.13]	[0.04:2.33]	[0.02:1.25]	[0.01:0.54]	[0.02:0.56]
4	**10-12**	I,M,F	[0.29:0.78]	[0.22:0.63]	[0.06:0.51]	[0.12:2.78]	[0.04:1.49]	[0.02:0.76]	[0.04:0.73]
5	**13-15**	I,M,F	[0.32:0.81]	[0.25:0.65]	[0.08:0.25]	[0.16:2.55]	[0.06:1.35]	[0.03:0.57]	[0.05:0.80]
6	**16-18**	I,M,F	[0.40:0.77]	[0.31:0.60]	[0.10:0.24]	[0.35:2.83]	[0.11:1.15]	[0.06:0.48]	[0.12:1.00]
7	**19-21**	I,M,F	[0.45:0.74]	[0.35:0.59]	[0.12:0.23]	[0.41:2.13]	[0.11:0.87]	[0.07:0.49]	[0.16:0.85]
8	**22-24**	M,F	[0.45:0.80]	[0.38:0.63]	[0.14:0.22]	[0.64:2.53]	[0.16:0.93]	[0.11:0.59]	[0.24:0.71]
9	**25-29**	M,F	[0.55:0.70]	[0.47:0.58]	[0.18:0.22]	[1.06:2.18]	[0.32:0.75]	[0.19:0.39]	[0.38:0.88]

It is noteworthy that the MID property does not hold when the dissimilarity among BSOs is computed by means of Gowda and Diday's measure (U_1). Surprisingly, old crustaceans with a high number of rings (25-29) are considered more similar to very young crustaceans with low number of rings (1-3) than to middle aged abalones with 16÷18 rings. Actually, for all numeric variables the dissimilarity components due to position (D_π) and content (D_c) increase along the horizontal axis, while the component due to spanning (D_s) first increases and then decreases. Spanning measures the difference between two interval widths and indeed BSO1 and BSO9 seem pretty similar due to the fact that continuous variables have intervals with the same (small) width despite the fact that the intervals are quite distant. Incidentally, such intervals are relatively small since there are few cases of abalones aggregated into BSO1 and BSO9. Thus, U_1 can lead to unexpected results in those cases in which BSOs are generated from unequally distributed cases with respect to a given class variable, such as *rings*. Also for the dissimilarity measure U_2 the MID property does not hold and the first BSO is the most atypical. In particular, BSO1 and BSO7 are more similar than BSO1 and BSO4. This can be explained by noting that the dissimilarity component due to meet ($\otimes$) has no effect when γ is set to 0.5, while the join ($\oplus$) of intervals in BSO1 and BSO7 is lower than the join of intervals in BSO1 and BSO4, since all numerical intervals of BSO7 are included in the corresponding intervals for BSO4. Generalizing, we note that it is advisable not to nullify the effect of the meet operator by setting $\gamma = 0.5$, otherwise anomalous similarities can be found. Similar considerations apply to the dissimilarity measures U_3 and U_4, although their normalization factor (i.e., domain

Table 4. Dissimilarity values computed by means of the dissimilarity U_1.

Rings	**1-3**	**4-6**	**7-9**	**10-12**	**13-15**	**16-18**	**19-21**	**22-24**	**25-29**
1-3	0								
4-6	12.7126	0							
7-9	13.8107	6.1026	0						
10-12	13.6988	7.4184	3.8644	0					
13-15	13.2341	6.6741	4.0135	2.7576	0				
16-18	12.4039	7.7761	6.1147	5.2025	3.3863	0			
19-21	11.6926	7.8082	7.263	6.8322	5.1771	3.0946	0		
22-24	11.3287	9.1946	8.0059	7.6176	6.1752	4.8742	3.5518	0	
25-29	9.2101	11.7497	12.1851	12.0065	11.1622	9.8778	8.1261	7.4043	0

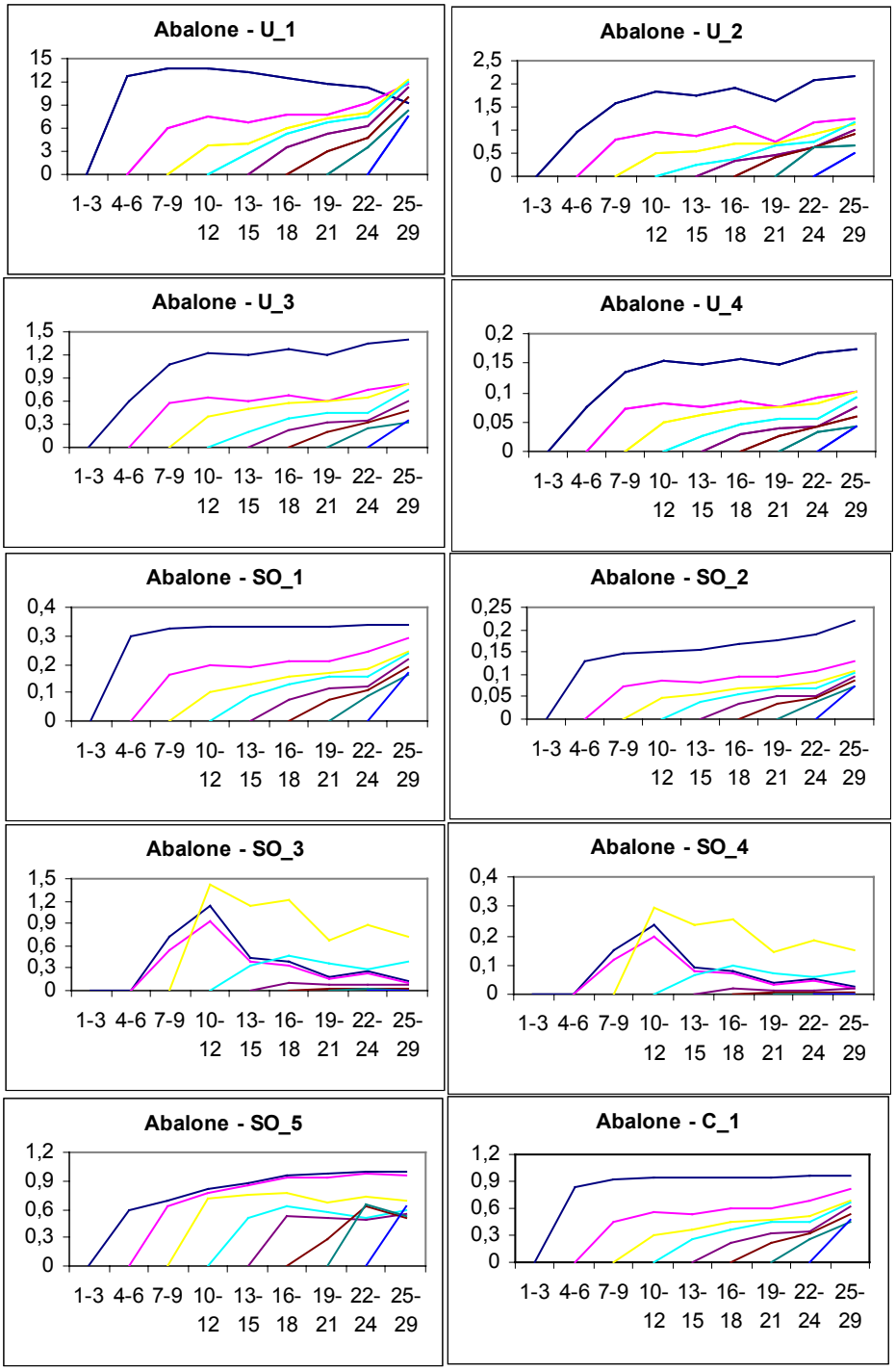

Figure 1. Graphs of ten dissimilarity measures for the abalone data.

cardinality) tend to reduce the effect of the missing "meet" component. On the contrary, the normalization by the join volume $\mu(A_j \oplus B_j)$, proposed by De Carvalho in SO_2, totally removes the problem even when $\gamma = 0.5$, and the expected MID property is satisfied.

The MID property is generally valid for SO_1 as well, and the atypical symbolic object is still BSO1. However, in this case, the numerical variables have almost no effect on the computation of the dissimilarity, since the agreement index at the numerator of each d_i, namely $\alpha=\mu(A_j \cap B_j)$, is very small. As a matter of fact, the dissimilarity between two symbolic objects is computed on the basis of the only nominal variable "sex."

In the case of SO_3, the atypical object is the third, while BSO1 is considered similar to all other symbolic objects, including BSO9. Once again, the contribution of the meet is nullified by setting $\gamma = 0.5$, nevertheless the situation is quite different from that presented in the case of U_2. The effect of the join is multiplicative in the computation of the description potential defined by De Carvalho, while it is additive in Ichino and Yaguchi's dissimilarity measures. In SO_3 small intervals can zero the dissimilarity, while in U_2 they have practically no effect. On the contrary, a single large interval can have a strong impact in U_2, while it may have no effect in SO_3. In the Abalone data set, small intervals are obtained by applying the join operator to continuous variables of symbolic objects BSO_i, $i \neq 3$. This explains the strange behaviour of lines depicted in the graph "Abalone – SO_3". Also in SO_5, which is a normalized version of SO_3, the MID property is not satisfied even though it has a more regular behaviour than SO_3 since the contribution of the join operator on numerical variables is reduced.

Finally, in the case of C_1, the MID property is satisfied. According to this measure all symbolic objects are quite dissimilar (the maximum distance is 1.0), and the most atypical is that representing young abalones (BSO1).

Summarizing, the following conclusions can be drawn from this empirical evaluation:

1. Only three dissimilarity measures proposed by de Carvalho, namely SO_1, SO_2 and C_1, satisfy the MID property. For all these measures the less typical object is that representing very young abalones (BSO1).
2. When BSOs are generated from unequally distributed cases with respect to a given class variable, the actual size of variable value sets A_i might simply reflect the original distribution of cases. In particular, small value sets may be due to the scantiness of cases used in the BSO generation process, while large value sets may occur because of the natural variability in a large population of cases used to synthesize BSOs. When this happens, distance measures based on the spanning factor (e.g., U_1) may lead to unexpected results.
3. In Ichino and Yaguchi's measures (i.e., U_2, U_3 and U_4), the contribution of the meet operator should not be nullified by setting the parameter γ to 0.5. An intermediate value between 0 and 0.5 is generally recommended. Moreover, the normalization proposed by de Carvalho (see SO_2) show better results.
4. In the case of continuous variables, the width of value intervals is critical, since dissimilarities measures based on additive aggregation tend to return high values when only one componentwise dissimilarity is quite large, while measures based on description potential return small values when only one componentwise dissimilarity is quite small.

4. Conclusions

In symbolic data analysis a key role is played by the computation of dissimilarity measures. Many measures have been proposed in the literature, although a comparison that investigates their applicability to real data has never been reported. The main difficulty was due to the lack of a standard in the representation of symbolic objects and the necessity of implementing many dissimilarity measures. The software produced by the ESPRIT Project SODAS has partially solved this problem by defining a suite of modules that enable the generation, visualization and manipulation of symbolic objects. In this work, a comparative study of the dissimilarity measures for BSOs is reported with reference to a particular data set for which an expected property could be defined. Interestingly enough, such a property has been observed only for some dissimilarity measures, which actually show very different behaviours. There are a number of possible directions for future research. One is to experiment whether other data sets with fully understandable and explainable properties related to the proximity concept. Another direction is to extend the empirical evaluation to dissimilarity measure defined on probabilistic symbolic objects. A third direction is to develop new dissimilarity measures for symbolic data that remove the two basic assumptions, namely variable independence and equal attribute relevance.

Acknowledgements

This work was supported partly by the European IST project no. 25161 (ASSO).

References

[1] Bock, H.H., Diday, E. (eds.): Analysis of Symbolic Data.Exploratory Methods for Extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag, Berlin, (2000), ISBN 3-540-66619-2

[2] de Carvalho, F.A.T.: Proximity coefficients between Boolean symbolic objects. In: Diday, E. et al. (eds.): New Approaches in Classification and Data Analysis, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 5, Springer-Verlag, Berlin, (1994) 387-394.

[3] de Carvalho, F.A.T.: Extension based proximity coefficients between constrained Boolean symbolic objects. In: Hayashi, C. et al. (eds.): Proc. of IFCS'96, Springer, Berlin, (1998) 370-378.

[4] Esposito, F., Malerba, D., Tamma, V., and Bock, H.H.: Classical resemblance measures. In: Bock, H.H, Diday, E. (eds.): Analysis of Symbolic Data. Exploratory Methods for extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag, Berlin, (2000) 139-152

[5] Esposito, F., Malerba, D., Tamma, V.: Dissimilarity Measures for Symbolic Objects. In: Bock, H.H., Diday, E. (eds.): Analysis of Symbolic Data. Exploratory Methods for extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag, Berlin, (2000) 165-185

[6] Gowda, K. C., Diday, E.: Symbolic clustering using a new dissimilarity measure. In Pattern Recognition, Vol. 24, No. 6, (1991) 567-578.

[7] Ichino, M., Yaguchi, H.: Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 24, No. 4, (1994) 698-707

[8] Robnik-Šikonja, M., Kononenko, I.: Pruning regression trees with MDL. In: Prade, H. (ed.): Proc. of the 13th European Conference on Artificial Intelligence, John Wiley & Sons, Chichester, England, (1998) 455-459.

[9] Stéphan, V., Hébrail, G., and Lechevallier, Y.: Generation of Symbolic Objects from Relational Databases. In: Bock, H.H., Diday, E. (eds.): Analysis of Symbolic Data. Exploratory Methods for extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 15, Springer-Verlag, Berlin, (2000) 78-105.

[10] Torgo, L.: Functional Models for Regression Tree Leaves. In: Fisher, D. (ed.): Proc. of the International Machine Learning, Morgan Kaufmann, San Francisco, CA, (1997) 385-393.