STATISTICA, anno LXIII, n. 2, 2003

PROXIMITY MEASURES IN SYMBOLIC DATA ANALYSIS

L. Nieddu, A. Rizzi

1. INTRODUCTION

The analysis of symbolic data has led to a new branch of Data Analysis called Symbolic Data Analysis (SDA) where the objects considered are new entities for which the representation as points in $S = Y_1 \times Y_2 \times \cdots \times Y_n$ (which usually reduces to $\Re^n$) no longer holds. In SDA features characterizing symbolic objects may take more than one value, or may be in the form of interval data or be qualitative data or subsets of qualitative data sets. The object or symbolic data analysis are elements of the type: "the weight is between 60 and 70 kg, the color is black or white and if age is lower then 18 then the occupation is 'student'".

In real life problems, very often researchers and data analysts come across this type of features which are not only difficult to handle with classical data analysis techniques but are also quite difficult to put in a standard data matrix where units and variables are represented by columns and rows and where it is not possible to represent logical relations such as "if age is lower then 18 then occupation is 'student'". Moreover, in particular occasions, it is also possible that the data to be processed are collected in the form of classes or groups of individuals instead of single individuals, e.g. census data which, to guarantee privacy, are made available aggregating data for subsets of units of the population.

Another example of data available in aggregated form is the output of a query in a relational database. These data are often characteristic in Data mining. We are concerned with different types of data set coming from different sources as accountancy, marketing researches, Delphi methods, open opinions expressed by customers, clients and so on. To handle in an appropriate manner this type of data the extension of standard statistical techniques to symbolic data has then become necessary.

Symbolic data can be generated in different ways. First of all one must consider that each category of a categorical variable or any logical combination of variables is a concept (Diday, Lechevallier, 2000). Each concept of this variable can be obtained from a *query* to a database.

The paper is common job of both the A. Paragraphs 1, 2.1, and 3, are basically referred to L. Nieddu, paragraphs 2.2, and 4 to A. Rizzi.

Symbolic data can also be used after clustering in order to summarize a huge set of data to describe in an exploratory way the obtained clusters and their internal variation.

A crucial issue in the adaptation of standard statistical techniques to symbolic data lays in the specification of resemblance measures (similarity and dissimilarity) between objects.

For example, in cluster analysis, where the aim is to determine a partition of *n* units in *k* homogeneous clusters, it is usually assumed that a similarity measure is set on the data in order to obtain clusters which are composed of units for whom the similarity is greater then the dissimilarity and it would be fair to assume that the dissimilarity between units belonging to different clusters be greater then the dissimilarity between units belonging to the same cluster.

Various techniques that have been developed for exploratory data analysis and multidimensional classification manage to handle almost only numerical variables. In the last decade there has been a flurry of activity aimed at extending these techniques to symbolic data (Gowda and Diday, 1994; Nagabhushan *et al.*, 1995; Lauro and Palumbo, 2000; Chouakria *et al.*, 2000; Lauro *et al.*, 2000; Périnel and Lechevallier, 2000; Chavent and Lechevallier, 2002; Mali and Mitra, 2003.

This paper presents a series of well-established similarity and dissimilarity indexes for binary symbolic objects. Some distance measures will be suggested for binary symbolic objects which could also be useful for probabilistic symbolic objects.

In section 2.1 the definition of symbolic data will be recalled, in section 2.2 we recall some algebraic structure that is involved in SDA, while in section 3 several resemblance measures will be considered. In section 4 the due conclusions will be drawn.

2. SYMBOLIC DATA ANALYSIS

A symbolic object is defined as a description that is expressed as a conjunction of statements regarding the values assumed by the variables. Let Ω be the set of observed objects, each one characterized by *p* variables $y_i, i = 1,\ldots,p$. Formally a variable $y_i : \Omega \to O_i$ can be considered as a function, where O_i is the observation set of y_i. The observation set of a variable is composed only by those values that the observed objects can actually assume. The variable y_i may be measured on a nominal, ordinal, interval, ratio or absolute scale.

An **elementary event** is an event of the type $e = [y_i \in V_i]$ indicating that variable y_i takes values in $V_i \subseteq O_i$. The elementary event *e* can be true or false, therefore. a mapping of the type $\mathrm{e} = \Omega \to \{true, false\}$ can be associated to the elementary event *e*.

An **assertion object** is composed by the logical conjunction of elementary events:

$a = \wedge_i [y_i \in V_i]$.

Again a mapping of the type $\text{a} = \Omega \rightarrow \{true, false\}$ will be associated to the assertion object *a*.

Boolean assertion objects (De Carvalho, 1995) can be used to provide, via the logical conjunction of elementary events, a precise description of a concept and allow to take into account, when describing concepts, the range of values of the variables considered.

For instance, if **Ω** is a set of patients and the variables considered to describe them are "temperature", "blood pressure" and "status", an elementary event could be $[Status \in \{Conscious,\ Partially\ Consciuous\}]$ or $[Temperature \in [37°,39°]]$ and an assertion object could be used to describe very severe conditions such as

$$[Blood\ \ Pressure \in [60,90]] \wedge [Status \in \{Unconsciuous\}] \wedge [Temperature \in [30°,35°]]$$

Each assertion object could be used to describe one class of patients. The *extension* of each assertion object *a* on the set **Ω** is defined as the subset of **Ω** for which the assertion *a* is true.

To actually represent data, the description of concepts by Boolean assertions must take into account various types of logical dependencies between variables, such as:

– *Hierarchical dependences* (mother-daughter): this type of dependence establishes conditions for which a variable could not be measured ("NA", i.e. "not-applicable") when another variable takes values in a particular subset. For instance, the variables "Pregnant" or "number of pregnancies" are not applicable when the variable "gender" is equal to "male". In these cases the domain of the variable is extended considering the code "NA". It is worth noticing that the numeric value "zero" could not be used instead of "NA", because having zero pregnancies has different implications for a woman then for a man.

– *Logical dependences*: which restrict the set of possible values of a variable according to the values taken by another variable. For instance, if the variable "age" is greater then 70 then the variable "occupation" will probably be "retired".

In this paragraph we recall some important structure involved in SDA.

A non empty set M with a relation ≤ is said to be an ordered set if the following conditions are satisfied (Schaefer, 1974):

1) $x \leq x$ for every $x \in M$
2) $x \leq y$ and $y \leq x$ implies $x = y$
3) $x \leq y$ and $y \leq z$ implies $x \leq z$

Let *A* be a subset of an ordered set *M*. The element $x \in M$ (resp. $z \in M$) is called an *upper bound* (resp. *lower bound*) of *A* if $y \leq x$ for all $y \in A$ (resp. $z \leq y$ for all $y \in A$). Moreover, if there is an upper bound (resp. lower bound) of *A*, then A is said to be *bounded from above* (*bounded from below* resp). If A is bounded from above and below, then A is called *ordered bound*. Let $x, y \in M$ such that $x \leq y$. We denote by:

$$[x, y] := \{z \in M \mid x \leq z \leq y\}$$

the *order interval* between x and y. It is obvious that a subset A is order bounded if and only if it is contained in some order interval.

Definition 1. A real vector space E which is ordered by some order relation $\leq$ is called a *vector lattice* if any two elements $x, y \in E$ have at least a upper bound denoted as sup(x,y) and a greatest lower bound denoted by inf(x,y) and the following properties are satisfied:

1) $x \leq y$ implies $x+z \leq y+z$ for all $x,y,z \in$ E
2) $0 \leq x$ implies $0 \leq tx$ for all $x \in E$ and $t \in \Re^+$.

Let E be a vector lattice. We denote by $E_+ := \{x \in E \mid 0 \leq x\}$ the positive cone of E. For $x \in E$ let $x^+ :=$ sup(x, 0), $x^- :=$ inf(-x, 0), $|x|:=$sup(x, -x) be the *positive, negative,* and the *absolute value* of x respectively. Two elements $x, y \in E$ are called orthogonal (or *lattice disjoint*), denoted by $x \perp y$) if sup(x,y)=0.

For a vector lattice E we have the following properties:

Proposition. For all $x,y,z \in E$ and $a \in I\!R$ the following assertions are satisfied:

1) $x+y$ = sup(x,y) + inf(x,y);
   sup(x,y) = -inf(-x,-y)
   sup(x,y) + z = sup{(x+z,y+z)} and
   inf(x,y) + z = inf{(x+z,y+z)}
2) $x = x^+ - x^-$
3) $|x| = x^+ + x^-$; $|ax| = |a||x|$ and $|x+y| = |x| + \lfloor y|$
4) $x^+ \perp x^-$ and the decomposition of x into the difference of two orthogonal positive elements is unique.
5) $x \leq y$ is equivalent to $x^+ \leq y^-$ and $y \leq x^-$
6) $x \perp y$ is equivalent to sup($|x| + \lfloor y|$) = $|x| + \lfloor y|$. In this case we have $|x+y| = |x| + \lfloor y|$.

A norm on a vector lattice E is called a *lattice norm.*

$|x| + \lfloor y|$ implies $\|x\| \leq \|y\|$ for $x,y \in E$

Definition 2. A Banach lattice is a real Banach space E endowed with an ordering $\leq$ such that (E, $\leq$) is a vector lattice and the norm on E is a lattice norm.

Symbolic data tables, constitute the main input of SDA (Bock and Diday, 2000). The column of the input data table are associated to symbolic variables wich are used in order to describe a set of units or objects. Rows are called *symbolic descrip-*

tions of these objects because they are not only vectors of single quantitative or categorical values. Each cell of this symbolic data tables contains data of different types:

1) single quantitative value
2) single categorical value
3) interval
4) multivalued with associated weights

The algebraic structure of vector lattice is involved in the first three variables. It is always very important to understand the data structure: *order* for Boolean variables, *vector lattices* for intervals and so on.

3. SIMILARITY AND DISSIMILARITY MEASURES

A dissimilarity measure D on a set of elements E is a real valued function $D: E \times E \rightarrow \Re$ such that:

1) $D(a,b) = D(b,a) \qquad \forall a,b \in E$
2) $D(a,b) \geq D(a,a) \qquad \forall\, b \in E$
3) $D(a,b) \leq +\infty \qquad \forall a,b \in E$

Usually $D(a,a) = 0$ and sometimes it is also required for the dissimilarity measure to take values in $[0,1]$. This measures and the others considered in this paragraph are defined in the algebraic structure of the vector lattice.

A dissimilarity measure for which $D(a,a) = 0$ and that fulfils the triangle inequality is called a metric or distance. Sometimes it is named a semi-metric or semi-distance, and the terms "metric" and "distance" are left for those dissimilarities which also fulfill the definiteness condition (see for instance, Rizzi, 1985 or Esposito *et al.*, 2000). It is also called an ultrametric if it fulfils the ultrametric condition:

$$D(a,b) \leq \max\{D(a,c), D(c,b)\} \qquad \forall a,b,c \in E$$

Obviously if a dissimilarity is an ultrametric it is also a metric.

Analogously a similarity measure S on a set of elements E is a real valued function $S: E \times E \rightarrow \Re$ such that:

1) $S(a,b) = S(b,a) \qquad \forall a,b \in E$
2) $S(a,b) \leq S(a,a) \qquad \forall\, b \in E$
3) $S(a,b) \geq 0 \qquad \forall a,b \in E$

More specifically S is usually required to be a function having domain in $E \times E$ and taking values in $[0,1]$.

If a resemblance measure fulfils an inequality dual to the ultrametric condition, it is named an ultraminima, i.e.:

$$S(a,b) \geq \min\{S(a,c), S(c,b)\} \qquad \forall a,b,c \in E$$

Given a similarity measure S on $E \times E$, and a strictly decreasing function ξ in $[0,1]$ then the mapping $D(a,b) = \xi[S(a,b)]$ is a dissimilarity index. Conversely, if ξ is also non-negative in $[0,1]$, then the quantity $S(a,b) = \xi[D(a,b)]$ is a similarity index. Usual transformations are: $\xi(x) = \max(x) - x$, $\xi(x) = \sqrt{\max(x) - x}$ or $\xi(x) = \cos(90x)$, but any strictly decreasing function in $[0,1]$ would do, depending on the particular purpose of the resemblance index considered.

Given two symbolic objects, $a = \wedge_i [y_i \in a_i]$ and $b = \wedge_i [y_i \in b_i]$ the dissimilarity between these two objects can be computed aggregating, with an appropriate aggregation function, the comparison functions, which are dissimilarities measures computed independently for each variable.

The usually applied aggregation function is the generalized Minkowski metric,

$$D(a,b) = \left(\sum_{k=1}^{p} D^r(a_k, b_k) \right)^{\frac{1}{r}}$$

where $D(a_k,b_k)$ is a dissimilarity measure for feature k.

First then a comparison function must be chosen to compute similarity or dissimilarity between variable, and then the resemblance between symbolic objects will be computed aggregating those similarity or dissimilarity indexes variable-wise.

To compute comparison functions for each variable, agreement-disagreement indices can be used (De Carvalho, 1994) according to the following table:

TABLE 1

Agreement-disagreement table

	Agreement	Disagreement
Agreement	$\alpha = \pi(a_k \cap b_k)$	$\beta = \pi(a_k \cap \dot{b}_k)$
Disagreement	$\gamma = \pi(\dot{a}_k \cap b_k)$	$\delta = \pi(\dot{a}_k \cap \dot{b}_k)$

Where $\dot{b}_k$ is the complementary set of b_k in the domain O_k and $\pi(a_k)$ is a function that accounts for the description potential of a_k and that can be defined as:

$$\pi(a_k) = \begin{cases} |a_k| & \text{if the variable is integer, nominal or ordinal.} \\ \\ |\overline{a}_k - \underline{a}_k| & \text{if the variable is a continuous interval} \end{cases}$$

where with the symbols $\overline{a}$ and $\underline{a}$ the upper and the lower bounds of an interval of the real line have been represented.

According to the previous definitions, classical similarity and dissimilarity indexes have been extended for symbolic data. Namely, some similarity measures are:

Sokal-Michener (simple matching)	$S=\frac{\alpha+\delta}{\alpha+\beta+\gamma+\delta}$	Sokal-Sneath	$S=\frac{\alpha}{\alpha+2(\beta+\gamma)}$
Jaccard	$S=\frac{\alpha}{\alpha+\beta+\gamma}$	Dice-Czekanowski-Sorenson	$S=\frac{2\alpha}{2\alpha+\beta+\gamma}$
Roger-Tanimoto	$S=\frac{\alpha+\delta}{\alpha+\delta+2(\beta+\gamma)}$	Russel-Rao	$S=\frac{\alpha}{\alpha+\beta+\gamma+\delta}$

The Sokal-Michener similarity index is invariant to inversion of “agreement” and “disagreement”. This property does not hold for Jaccard similarity.

The Russel-Rao similarity index is peculiar, since in all the other indexes when the description potential of $(\dot{a}_k \cap \dot{b}_k)$ (i.e. the description potential of what is not in a_k or in b_k) is considered in the index then it is present both at the numerator and at the denominator of the fraction.

Roger-Tanimoto and Soak-Sneath indexes double weight mismatches (i.e. $a_k \cap \dot{b}_k$ and $\dot{a}_k \cap b_k$) and the former ignores conjoint absence (i.e. $\dot{a}_k \cap \dot{b}_k$). On the other hand Dice-Czekanowski-Sorenson index double weights conjoint presence without considering conjoint absence.

It is worth noticing that Jaccard, Sokal-Sneath and Dice-Czekanowski-Sorenson indexes are all indeterminate if $\alpha=\beta=\gamma=0$, which could happen even if in very special cases, such as, for instance, when a_k and b_k are two degenerate intervals.

All these coefficients can be generalized according to the following formulas:

$$S=\frac{\alpha+t\delta}{\alpha+w\delta+\vartheta(\beta+\gamma)} \qquad t=\{0,1\};\ w=\{0,1\};\vartheta>0$$

Analogous class of dissimilarity measures can be obtained from the previous one simply considering the dissimilarity index $D=1-S$:

Other similarity measures that do not fit in the previous class are:

Kulczynski	$S=\frac{1}{2}\left(\frac{\alpha}{\alpha+\beta}+\frac{\alpha}{\alpha+\gamma}\right)$	Occhiai-Driver-Kroeber	$S=\frac{\alpha}{\sqrt{(\alpha+\beta)(\alpha+\gamma)}}$

which can be considered, respectively, as the arithmetic and geometric mean of the quantities $\alpha/(\alpha+\beta)$ and $\alpha/(\alpha+\gamma)$ which represent the proportion of agreements on the marginal distributions.

Another class of resemblance measures for symbolic objects is based on the notion of description potential of a symbolic object *a*. This type of measure does not require a variable-wise function and an aggregation function to obtain an aggregate similarity or dissimilarity measure.

Gowda and Diday have proposed various types of similarity and dissimilarity measures (see Gowda and Diday, 1992, 1991a, 1991b). To overcome some disadvantages of the previous measure, Gowda and Ravi have proposed (see Gowda and Ravi, 1995) modified similarity and dissimilarity measures defined on the basis of position, span and content of symbolic objects which can be used on symbolic data composed of qualitative and quantitative values (mixed feature type).

Given two symbolic objects, *a* and *b*, dissimilarity between these two objects can be written as:

$$D(a,b)=\sum_{k=1}^{p} D(a_k,b_k)$$

the dissimilarity between the *k*-th feature $D(a_k,b_k)$ is computed considering the contribution of three different components which incorporate different types of dissimilarities (Gowda and Diday, 1991a; Gowda and Ravi, 1995; Ravi and Gowda, 1999):

a) $D_p(a_k,b_k)$ dissimilarity due to position, defined for quantitative data; represents the relative positions of the two features values on the real line.

b) $D_s(a_k,b_k)$ dissimilarity due to span, defined for qualitative and quantitative data, is due to the relative dimensions of the feature values without taking into account their intersection.

c) $D_c(a_k,b_k)$ dissimilarity due to content, takes into account the common part of the two features.

where $D_p(a_k,b_k)$ is the dissimilarity due to position and is computed, for quantitative interval data, as:

$$D_p(a_k,b_k)=\cos\left[90\left(1-\frac{|\underline{a}_k-\underline{b}_k|}{u_k}\right)\right]$$

where $\underline{a}_k$ and $\underline{b}_k$ are the lower limits of the two intervals a_k and b_k for the *k*-th feature and u_k is the length of the maximum interval for that feature.

$D_s(a_k,b_k)$ represents the part of the dissimilarity due to span and is calculated, for interval data, as:

$$D_s(a_k,b_k)=\cos\left(45\frac{\pi(a_k)+\pi(b_k)}{|\max(\overline{a}_k,\overline{b}_k)-\min(\underline{a}_k,\underline{b}_k)|}\right)$$

while for qualitative data the component due to span is given by:

$$D_s(a_k, b_k) = \cos\left(45 \frac{\pi(a_k) + \pi(b_k)}{\pi(a_k) + \pi(b_k) - \pi(a_k \cap b_k)}\right)$$

The dissimilarity component due to content is defined as:

$$D_c(a_k, b_k) = \cos\left(90 \frac{\pi(a_k \cap b_k)}{\pi(a_k) + \pi(b_k) - \pi(a_k \cap b_k)}\right)$$

Dissimilarity is then computed, for quantitative interval data, as $D(a_k, b_k) = D_p(a_k, b_k) + D_s(a_k, b_k)$ while for qualitative data as $D(a_k, b_k) = D_s(a_k, b_k) + D_c(a_k, b_k)$ (Gowda and Ravi, 1995; Ravi and Gowda, 1999).

De Baets *et al.* (2001) have examined twenty-eight measures of similarity between crisp subsets of a finite universe. They propose a class of rational similarity measures based solely on the cardinality of the sets involved:

$$S(a,b) = \frac{r \min\{\#(a \setminus b), \#(b \setminus a)\} + s \max\{\#(a \setminus b), \#(b \setminus a)\} + t[\#(a \cap b)] + u[\#(\dot{a} \cap \dot{b})]}{r' \min\{\#(a \setminus b), \#(b \setminus a)\} + s' \max\{\#(a \setminus b), \#(b \setminus a)\} + t'[\#(a \cap b)] + u'[\#(\dot{a} \cap \dot{b})]}$$
$$r, r', s, s', t, s', u, u' \in \{0,1\}$$

where the symbol “#” denotes the cardinality of a set and “\” is the set difference operator.

Or, according to the notation used in table 1:

$$S(a,b) = \frac{r \min\{\beta, \gamma\} + s \max\{\beta, \gamma\} + t\alpha + u\delta}{r' \min\{\beta, \gamma\} + s' \max\{\beta, \gamma\} + t'\alpha + u'\delta}.$$

To obtain a reflexive similarity index the conditions $t = t'$ and $u = u'$ must hold. Indeterminacy cases such as 0/0 are handled setting the index to 1. Some of the measures that can be obtained for particular choices of the coefficients are well known in the literature. For instance, the choice $r = s = u = u' = 0, r' = s' = t = t' = 1$ gives the Jaccard’s index, while the choice $r = s = u = 0, r' = s' = t = t' = u' = 1$ yields the Russel-Rao measure that, for binary vectors can be considered the normalized inner product. Besides the usual properties that should be verified by a similarity index, De Baets *et al.* propose three boundary conditions that similarity indexes should verify, regarding similarity to the empty set, similarity to the universe and similarity between complementary sets. The only two indexes that verify all the three boundary conditions are obtained for $r = s = r' = 0, s' = t = t' = u = u' = 1$ and $r = s = 0, r' = s' = t = t' = u = u' = 1$ and are respectively:

$$S(a,b)=\frac{\alpha+\delta}{\max\{\beta,\gamma\}+\alpha+\delta}; \qquad S(a,b)=\frac{\alpha+\delta}{\alpha+\beta+\gamma+\delta}$$

the second one being the well known Sokal-Michener simple matching coefficient. Besides boundary conditions, some monotonicity properties have been studied regarding three and four sets which have been extensively considered for all the indexes allowing for a complete characterization and classification of the rational similarity measures that have been considered.

An effort to empirically compare dissimilarity measures has been carried out by Malerba *et al.* (2001) where an evaluation of dissimilarity measures for Boolean symbolic objects has been proposed. The data set considered for testing is the well-known Abalone Fish dataset available form the University of California at Irvine Machine Learning Repository. This dataset contains 4177 records of Abalone fishes described by nine mixed (qualitative and quantitative) attributes. This dataset is usually used to predict the age of an abalone fish only considering such attributes like sex, weight, shell weight etc[1] (see Malerba *et al.*, 2001). The argument made in the paper of Malerba *et al.*, is that, considered that the performance of techniques such regression-tree on the abalone fish dataset is quite high, the eight attributes are sufficient enough to predict the age of an abalone. They then "expect that the degree of dissimilarity between crustacean computed on the independent attributes do actually be proportional to the dissimilarity in the dependent attribute" (Malerba *et al.*, 2001). Abalone data have been aggregated into nine symbolic objects using SODAS software (Symbolic Official Data Analysis System http://www.cisia.com/download.htm) and ten dissimilarity indexes (including De Carvalho's, Ichino and Yaguchi's and Gowda and Diday's indexes) have been computed, comparing their performance. It is not clear, however, how the proportionality of the degree of dissimilarity stated above should still hold when the 4177 abalone fishes have been grouped into symbolic objects.

Vladutu *et al.* (Vladutu *et al.*, 2001) have proposed a distance for symbolic data in the context of Generalized Radial Basis Function networks. The proposed distance is tailored only for discrimination purposes, i.e. it is assumed that there is a training set available where data have previously been assigned to one of N classes. The distance for a specific feature is defined in the following fashion:

$$d(v_a,v_b)=\sum_{i=1}^{N}\left|\frac{c_{a_i}}{c_a}-\frac{c_{b_i}}{c_b}\right|^s$$

where v_a and v_b are two possible values for the feature under consideration, c_{a_i} and c_{b_i} are the number of element for which values v_a and v_b have been classified in class i and c_a and c_b denote the total number of element which present, respec-

[1] All the information regarding the dataset are available at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/abalone/abalone.names

tively, value *a* and *b* for the feature under study. The overall distance between two elements is then calculated by a weighted sum of the distances between features, i.e. if k is the total number of features and w_i is the weight assigned to feature *I* then:

$$D(X,Y) = \sum_{i=1}^{k} w_i d(v_{X_i}, v_{Y_i})^r$$

The use of this type of distance even if proved useful on a number of test-sample (Vladutu *et al.*, 2001) is restricted to supervised learning frameworks where it reduces to a distance between row profiles in a matrix where the rows are the possible values of the character and the columns are the classes.

Let *a* be a symbolic object $a = \wedge_i [y_i \in V_i]$: the definition of description potential varies according to the type of symbolic object consider (constrained or unconstrained). For an unconstrained symbolic object the description potential is given by $\pi(a) = \prod_{j=1}^{p} \pi(a_j)$, where $\pi(a_k)$ has been previously defined. For a constrained Boolean symbolic object the definition of description potential needs to be slightly modified in order to take into consideration hierarchical and logical dependences. For logical dependence, i.e. dependences of the type *if* $(y_j \in s_j \subseteq O_j)$ *then* $(y_i \in s_i \subseteq O_i)$ where s_j and s_i are subsets of the domains of, respectively, variables y_j and y_i, the description potential becomes $\pi(a) = \prod_{j=1}^{p} \pi(a_j) - \pi(a')$ where $\pi(a')$ is the description potential of the incoherent restriction of a which includes all the description vectors fulfilling a but that are incoherent. For a hierarchical dependence of the type *if* $(y_j \in s_j \subseteq O_j)$ *then* $(y_i \in \{NA\})$, the description potential becomes $\pi(a) = \pi(a_j \cup \{NA\}) \prod_{j=1, j \neq i}^{p} \pi(a_j) - \pi(a') - \pi(a'')$ where variable y_i takes values in an enlarged domain which contains, as one of its categories, the label "NA", $\pi(a')$ is the description potential including all description vectors where $y_i \notin NA$ even if the assumption of the relation is true, and $\pi(a'')$ is the description potential including all vectors where $y_i \in NA$ even if the "if" part of the relation is false. This extended definition of description potential can be applied to the determination of dissimilarity measures which are a trivial extension of Ichino & Yaghuchi's (Ichino and Yaguchi, 1994) distances such as:

$$D(a,b) = \frac{\pi(a \oplus b) - \pi(a \cap b) + \gamma[2\pi(a \cap b) - \pi(a) - \pi(b)]}{R} \qquad \gamma \in [0, 0.5]$$

where R can be equal to 1, or be the potential of the entire domain of the p variables or be $\pi(a \oplus b)$ where $\oplus$ is the Cartesian join operator. For the first two choices of R the dissimilarity measures are equivalent and they are not metric functions because the triangular inequality does not hold. The third choice for R ends up in a dissimilarity measures which is also a metric.

It is worth noticing that the previous dissimilarity indexes are closely related to the concept of symmetric difference between two sets. Indeed an interesting class of distances based principally on the idea of symmetric difference between sets can be applied to the computation of dissimilarity for symbolic data. Given two sets, **a** and **b**, the symmetric difference is $\mathbf{a}-\mathbf{b}=(\mathbf{a}\setminus\mathbf{b})\cup(\mathbf{b}\setminus\mathbf{a})=(\mathbf{a}\cup\mathbf{b})\setminus(\mathbf{a}\cap\mathbf{b})$. Let μ be a measure for a set, a possible distance between two sets a_k and b_k is given by the quantity

$$D(a_k,b_k)=\mu(a_k-b_k)$$

and if μ coincides with the description potential π then the previous quantity, for qualitative datasets, reduces to $D(a_k,b_k)=\pi(a_k \oplus b_k)-\pi(a_k \cap b_k)$ which is also a liable option for a dissimilarity measure for interval data, for it is equivalent to Ichino and Jaguchi's distance choosing $\gamma=0$. This distance is easily extended to compare two functions f_{a_k} and f_{b_k} (which could be, for instance, two density functions) defined over an interval O_k

$$D(a_k,b_k)=\int_{0_k}\left|f_{a_k}-f_{b_k}\right|d\mu$$

A distance assuming values in [0,1] is:

$$D(a_k,b_k)=\begin{cases}\dfrac{\mu(a_k-b_k)}{\mu(a_k\cup b_k)} & \mathit{if}\ \ \mu(a_k\cup b_k)>0\\[2ex] 0 & \mathit{if}\ \ \mu(a_k\cup b_k)=0\end{cases}$$

that reduces to:

$$D(a_k,b_k)=\frac{\int_{0_k}\left|f_{a_k}-f_{b_k}\right|d\mu}{\int_{0_k}\max(f_{a_k}-f_{b_k})d\mu}$$

when applied to functions defined over the same set.

The previous distance can be slightly modified, taking into account the measure of the domain $\boldsymbol{O}$ where the sets are embedded

$$D(a_k,b_k)=\begin{cases}\dfrac{\mu(a_k-b_k)}{\mu(\boldsymbol{O})-\mu(a_k\cap b_k)} & \mathit{if}\ \ \mu(\boldsymbol{O})-\mu(a_k\cap b_k)>0\\ \\ 0 & \mathit{if}\ \ \mu(\boldsymbol{O})-\mu(a_k\cap b_k)=0\end{cases}$$

This quantity depends on the part of the domain that is not common to the two sets, compared to the part of the union of the two sets which is not in common.

All the quantities proposed above can be used on single features of symbolic data or on the whole symbolic data considering the notion of description potential, which, for a Boolean symbolic object $a=[y_1=a_1]\wedge[y_2=a_2]\wedge\ldots\wedge[y_p=a_p]$ can be considered a measure of the volume of the Cartesian product $\underset{i=1}{\overset{p}{\times}}a_i$

Another way to compute dissimilarity between symbolic objects as dissimilarity between sets is to use the Hausdorff distance, which was initially defined to compare two sets. Given the function $h(\mathbf{a},\mathbf{b})=\sup_{a\in\mathbf{a}}\inf_{b\in\mathbf{b}}\|b-a\|$, the Hausdorff distance between two sets **a** and **b** both in $\Re^p$ is defined as

$$D(a,b)=\max\{h(a,b),h(b,a)\}$$

In the particular case of vectors of intervals the Hausdorff distance (Chavent and Lechevallier, 2002) can be computed as

$$D(\mathbf{a},\mathbf{b})=\sum_{i=1}^{p}\max\{|\underline{b}_i-\underline{a}_i|,|\overline{b}_i-\overline{a}_i|\}$$

that, reduces to the city-block distance for degenerate intervals corresponding to points in $\Re^p$.

4. CONCLUSIONS

Symbolic data analysis has been introduced by E. Diday in the late 80's. In the last decade we have had so many papers, national and international research groups, specific international research for implementing adequate software. The well known European Community Project, called SODAS, for a *Symbolic Official Data Analysis System*, implemented by 17 institutions of 9 European countries, has produced a prototype software for SDA. All these researches have a common denominator: a relation measure between two or more symbolic objects. The measure of similarity or dissimilarity is different for Booleans objects and for different kinds of variables (defined for intervals, categorical variables with multiple values) or probabilistic objects.

The methods for the synthesis of different measure of relation are another crucial point in SDA.

The problem of data codification is open particularly with regard to the stability of the conclusions that can be deduced from the dataset.

The representation of the data is in complex algebraic structures. These structures are fundamental for a scientific approach to SDA. Although they are not generally known to the scholars of social sciences.

Often the objects are characterised by different kinds of variables. Many of these variables have been studied for the first time in Statistics just with reference to this type of analysis. We refer for instance to algebra of intervals.

Reality is very simple but usually our simple models cannot manage to explain this reality! Linear models can give information on the complexity of data but cannot show all the relations between the objects we are concerned with. Measures of similarity and dissimilarity are the basis for every kind of data process.

We believe that SDA can improve the approach to explain data. We need to process these data to reduce our information and to gain some understanding of the phenomenon we are concerned with.

SDA has specific applications in Data mining and, particularly, in the elaboration of large data sets. Knowledge extraction from large databases is the crucial issue in Data mining.

In these researches the stability of the conclusions is very important when we do new revisions of the input data or we change slightly the data that we are processing. SDA has had many kinds of applications but it is not yet very well known by scholars, particularly in the Anglo-Saxon academic world. Nonetheless the applications also are limited and generally done in academic circles. Applications referee to classical data, as Fisher's Iris. It is very complicated to obtain data from firms because of the privacy issues. It is also very complicated to codify large data sets in the logic of SDA.

It is our specific opinion that SDA can find very important applications in different sectors of the economy, social sciences, technology and in many other important branches of research. The Software is not yet well known to different people in firms.

The heavy formalization of SDA can limit the utilization by scholars not specifically expert in mathematics and particularly in abstract algebra.

Dipartimento di Statistica, Probabilità e
Statistiche Applicate
Università degli Studi di Roma "La Sapienza"

ALFREDO RIZZI
LUCIANO NIEDDU

BIBLIOGRAPHY

L. BOCCI, A. RIZZI (2000), *Misure di prossimità nell'analisi dei dati simbolici*, in Società Italiana di Statistica, "Atti della XL riunione scientifica", Firenze, 2000.

H.H. BOCK, E. DIDAY (eds.) (2000), *Analysis of symbolic data. Exploratory Methods for extracting statistical information from complex data*, Studies in classification, data analysis and knowledge organization, vol. 15, Springer-Verlag, Berlin.

P. BRITO (1995), *Symbolic objects: order structure and pyramidal clustering*, "Annals of Operations Research", 55, pp. 277-297.

M. CHAVENT, Y. LECHEVALLIER (2002), *Dynamical Clustering of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance*, in K. JAJUGA, A. SOKOLOWSKI and H.H. BOCK (eds.), "Classification, Clustering, and Data Analysis, Recent Advances and Applications", Studies in classification, data analysis and Knowledge Organization, Springer-Verlag, Berlin.

A. CHOUAKRIA, P. CAZES, E. DIDAY (2000), *Symbolic Principal Component Analysis*, in H.H. BOCK and E. DIDAY (eds.), "Analysis of Symbolic Data. Exploratory Methods for extracting statistical information from complex data", Studies in classification, data analysis and knowledge organization, vol. 15, Springer-Verlag, Berlin.

B. DE BAETS, H. DE MEYER, H. NAESSENS (2001), *A class of rational cardinality-based similarity measures*, "Journal of Computational and Applied Mathematics", 132, 51-69.

F.A.T. DE CARVALHO (1994), *Proximity Coefficients between Boolean Symbolic Objects*, in E. DIDAY & Y. LECHEVALLIER & M. SCHADER and P. BERTRAND and B. BURTSCHY (eds.), "New Approaches in classification and Data Analysis", Springer-Verlag, Berlin, pp. 387-394.

F.A.T. DE CARVALHO (1995), *Histograms in symbolic data analysis*, "Annals of Operations Research", 55, pp. 299-322.

F.A.T. DE CARVALHO (1998), *Extension based proximity coefficients between constrained boolean symbolic objects*, in C. HAYASHI *et al.* (eds.), "Proceedings of IFCS'96", Springer-Verlag, Berlin, pp. 370-378.

E. DIDAY (1988), *The symbolic approach in clustering and related methods of data analysis: The basic choices*, in H.H. BOCH (ed.), IFCS-87.

E. DIDAY (1995), *Probabilistic, possibilist and belief objects for knowledge analysis*, "Annals of Operations Research" 55, pp. 227-276.

E. DIDAY, Y. LECHEVALLIER (2000), *From Data Mining to Knoledge Mining:An Introduction to Symbolic Data Analysis*, in GAULL, OPITZ, SCHADER (eds), "Data Analysis", Springer-Verlag, Berlin.

F. ESPOSITO, D. MALERBA, V. TAMMA, H. H. BOCK (2000), *Classical Resemblance Measures*, in H.H. BOCK and E. DIDAY (eds.), "Analysis of Symbolic Data. Exploratory Methods for extracting statistical information from complex data", Sstudies in classification, data analysis and Knowledge Organization, vol. 15, Springer-Verlag, Berlin.

K. C. GOWDA, E. DIDAY (1991a), *Symbolic Clustering using a new dissimilarity measure*, "Pattern Recognition" 24 (6), 657-578.

K. C. GOWDA, E. DIDAY (1991b), *Unsupervised learning through symbolic clustering*, "Pattern Recognition Letters" 12, 259-264.

K. C. GOWDA, E. DIDAY (1992), *Symbolic Clustering using a new similarity measure*, "IEEE Transaction on Systems, Man and Cybernetics" 22 (2), 368-378.

K. C. GOWDA, E. DIDAY (1994), *Symbolic clustering algorithms using similarity and dissimilarity measures*, in E. DIDAY, Y. LECHEVALLIER, M. SCHADER *et al.* (eds.), "IFCS-93", 1994, 412-422.

K. C. GOWDA, T.V. RAVI (1995), *Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity*, "Pattern Recognition Letters" 16, 647-652.

M. ICHINO, H. YAGUCHI (1994), *General Minkowski metrics for mixed type data analysis*, "IEEE Transaction on System, Man and Cybernetics", 24, 4, pp. 698-708.

N.C. LAURO, F. PALUMBO (2000), *Principal Component Analysis of Interval Data: a Symbolic Data Analysis Approach*, "Computational Statistics", 15, 1, pp. 73-87.

N.C. LAURO, R. VERDE, F. PALOMBO (2000), *Factorial Discriminant Analysis on Symbolic Objects*, in H.H. BOCK and E. DIDAY (eds.), "Analysis of Symbolic Data. Exploratory Methods for extracting statistical information from complex data", Studies in classification, data analysis and Knowledge Organization, Vol. 15, Springer-Verlag, Berlin.

D. MALERBA, F. ESPOSITO, V. GIOVIALE & V. TAMMA (2001), *Comparing dissimilarity measures in Symbolic Data Analysis*, Proceedings of the Joint Conferences on "New Techniques and Technologies for Statistics" and "Exchange of Technology and Know-how" (ETK-NTTS'01), 473-481.

K. MALI, S. MITRA (2003), *Clustering and its validation in a symbolic framework*, "Pattern Recognition Letters" 24, 2367-2376.

P. NAGABHUSHAN, K.C. GOWDA, E. DIDAY (1995), *Dimensionality reduction of symbolic data*, "Pattern Recognition Letters", 16, 219-223.

E. PÉRINEL, Y. LECHEVALLIER (2000), *Symbolic Discrimination Rules*, in H.H. BOCK and E. DIDAY (eds.), "Analysis of Symbolic Data. Exploratory Methods for extracting statistical information from complex data", Studies in classification, data analysis and Knowledge Organization, vol. 15, Springer-Verlag, Berlin.

T.V. RAVI, K. GOWDA (1999), *An ISODATA clustering procedure for symbolic objects using a distributed genetic algorithm*, "Pattern Recognition Letters" 20, 659-666.

A. RIZZI (1985), *Analisi dei dati*, Nuova Italia Scientifica.

A. RIZZI (1998), *Metriche nell'analisi dei dati simbolici*, "Statistica", 4, 577-588.

H.H. SCHAEFER (1974), *Banach Lattices and Positive Operators*, Springer-Verlagm Berlin.

L. VLADUTU, S. PAPADIMITRIOU, S. MAVROUDI, A. BEZERIANOS (2001), *Generalised RBF Networks Trained Using and IBL Algorithm for Mining Symbolic Data*, in D. CHEUNG, G. J. WILLIAMS and Q. LI (eds.), "Advances in Knowledge Discovery and Data Mining", 5th Pacific-Asia Conference, 2001, Hong Kong, China, Series, Lecture Notes in Artificial Intelligence Volume 2035, pp. 587-593, Springer-Verlag Heidelberg.

RIASSUNTO

Misure di prossimità nell'analisi di dati simbolici

Gli autori considerano il problema della determinazione di misure di prossimilità tra dati simbolici. Inizialmente vengono richiamate le definizioni di evento elementare, asserzione e dipendenze logiche e gerarchiche. Quindi alcune bene note misure di prossimità tra due oggetti vengono considerate (Sokal-Michener, Roger-Tanimoto, Sokal-Sneath, Dice-Czekanowski-Sorenson, Russel-Rao). Come misure di prossimità basate su funzioni di aggregazione vengono prese in considerazione le proposte di Gowda-Diday, De Baets *et al.*, Vladutu *et al.*, e Ichino-Iyaghuchi. Le strutture algebriche sono prese in considerazione, con particolare riferimento a *reticoli vettoriali e interni* nello spazio di Banach.

SUMMARY

Proximity measures in symbolic data analysis

The Authors consider the general problem of similarity and dissimilarity measures in Symbolic Data Analysis. First of all they examine the classical definitions of elementary

event, assertion object, hierarchical dependences and logical dependences. Then they consider some well-known measures of similarity and dissimilarity between two objects (Sokal-Michener, Roger-Tanimoto, Sokal-Sneath, Dice-Czekanowski-Sorenson, Russel-Rao). For resemblance measures based on aggregation functions, the authors consider the proposals of Gowda-Diday, De Baets *et al.*, Malerba *et al.*, Vladutu *et al.*, and Ichino-Iyaghuchi. A paragraph is dedicated to the general algebraic structure; particularly to *intervals and vector lattices* in Banach space.