An Introduction to Symbolic Data Analysis and the Sodas Software

Edwin Diday

University Paris 9 Dauphine
Ceremade. Pl. Du Mle de L. de Tassigny. 75016 Paris, FRANCE

Abstract. The data descriptions of the units are called "symbolic" when they are more complex than standard ones due to the fact that they contain internal variation and are structured. Symbolic data arise from many sources, for instance in order to summarize huge Relational Data Bases by their underlying concepts. "Extracting knowledge" means getting explanatory results, that why, "symbolic objects" are introduced and studied in this paper. They model concepts and constitute an explanatory output for data analysis. Moreover they can be used in order to define queries of a Relational Data Base and propagate concepts between Data Bases. We define "Symbolic Data Analysis" (SDA) as the extension of standard Data Analysis to symbolic data tables as input in order to find symbolic objects as output. Any SDA is based on four spaces: the space of individuals , the space of concepts, the space of descriptions modelling individuals or classes of individuals, the space of symbolic objects modelling concepts. Based on these four spaces, new problems appear such as the quality, robustness and reliability of the approximation of a concept by a symbolic object, the symbolic description of a class, the consensus between symbolic descriptions, and so on. In this paper we give an overview on recent development in SDA. We present some tools and methods of SDA and introduce the SODAS software prototype (issued from the work of 17 teams of nine countries involved in an European project of EUROSTAT) .

1 Introduction

As input, when large data sets are aggregated into smaller more manageable data sizes we need more complex data tables called "symbolic data tables" because a cell of such data table does not necessarily contain as usual, a single quantitative or categorical value.
In a symbolic data table, a cell can contain a distribution (Schweitzer (1985) says that "distributions are the number of the future!"), or intervals, or several values linked by a taxonomy and logical rules. The need to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination,...) to symbolic data table is

increasing in order to get more accurate information and summarize extensive data sets contained in Data Bases.

Since the first papers announcing the main principles of Symbolic Data Analysis (Diday 1987a, 1987b, 1989) much work has been done up to the most recent book published by Bock and Diday (2000) and the proceedings of IFCS'2000 (Kiers et al. 2000) which contains a large chapter devoted to this field. In factorial analysis, Cazes, Chouakria, Diday and Schechtman (1997) have defined a principal component analysis of individuals described by a vector of numerical intervals. In the same direction is the work by Verde and De Carvalho (1998) that takes care on given dependence rules (see also Lauro and Palumbo 1998). In the case where the individuals are described by symbolic data, Conruyt (1994) in the case of structured data, Ciampi et al. (1995), Périnel (1996), have developed an extension of standard decision trees. In the same direction is the work by Bravo and Garcia-Santesmases (1998) on "segmentation trees for stratified data" and Rasson and Lissoir (1998). See also (Auriol 1995) for a link with the domain of "Case Based Reasoning". In order to select the symbolic variables which distinguish at the best individuals or classes of individuals, several works have been done such as Vignes (1991) and more recently Ziani (1996). It is often useful to calculate dissimilarities between symbolic objects; in that direction mention should be made of Gowda and Diday (1992), De Carvalho (1994, 1998a). A complete review is reported in the work by Esposito et al. (2000). If each cell of the data table is a random variable represented by a histogram (for instance, the histogram of the inhabitant age of a town), a histogram of histogram can be calculated for instance, by taking care of rules between the variable values in De Carvalho (1998b), or by using the capacity theory (Diday & Emilion 1995, 1997, Diday et al. 1996). Noirhomme and Rouard (1998, 2000) give a way of representing multidimensional symbolic data (see also Gigout 1998).

Starting from standard data has been proposed a way for extracting symbolic objects from a factorial analysis (Gettler-Summa 1992), and a way for extracting symbolic objects from a partition (Stephan et al. 2000). Starting from time-series, Ferraris, Gettler-Summa, Pardoux, Tong (1995), have defined a way for providing symbolic objects (see also Gettler-Summa & Pardoux 2000).

More recently, several dissertations have been presented in the Paris 9 - Dauphine University. Mfoumoune (1998) for the sequential building of a pyramid where each node is associated to a symbolic object. Chavent (1997), in order to build a partition of a set of symbolic objects by a top-down algorithm which provides also a symbolic object associated to each obtained class (see chapter 11 in Bock , Diday (2000)). Stéphan (1998) for extracting symbolic objects from a data base (see also Stéphan et al. 2000). Hillali (1998) for describing classes of individuals described by a vector of probability distributions. Pollaillon (1998), for extending Galois lattices and extracted pyramid to symbolic data at input and "complete" symbolic objects at output (Pollaillon 2000). Tang (1998) for extending Factorial Correspondence Analysis and O. Rodriguez (2000) for extending regression and Multidimensional Scaling to interval data.

1.1 The Input of a Symbolic Data Analysis: a "Symbolic Data Table"

"Symbolic data tables" constitute the main input of a Symbolic Data Analysis. They are defined in the following way: columns of the input data table are «symbolic variables» which are used in order to describe a set of units called "individuals". Rows are called «symbolic descriptions» of these individuals because they are not as usual, only vectors of single quantitative or categorical values. Each cell of this «symbolic data table» contains data of different types:

(a) Single quantitative value: for instance, if «height» is a variable and w is an individual: height(w) = 3.5.
(b) Single categorical value: for instance, town(w) = London.
(c) Multi-valued: for instance, in the quantitative case height(w) = {3.5, 2.1, 5} means that the height of w can be either 3.5 or 2.1 or 5. Notice that (a) and (b) are special cases of (c).
(d) Interval: for instance height(w) = [3, 5], which means that the height of w varies in the interval [3, 5].
(e) Multi-valued with weights: for instance a histogram or a membership function (notice that (a) and (b) are special cases of (e) when the weights are equal to 1 or 0).

Variables can be:

(f) Taxonomic: for instance, «the colour» is considered to be "light" if it is "yellow", "white" or "pink".
(g) Hierarchically dependent: for instance, we can describe the kind of computer of a company only if it has a computer, hence the variable "does the company has computers? " and the variable " kind of computer " are hierarchically linked.
(h) With logical dependencies, for instance: «if age(w) is less than 2 months then height(w) is less than 10».

Many example of such symbolic data are given in the chapter 3 in (Bock & Diday 2000).

Sources of Symbolic Data. Symbolic data are generated when we summarize huge sets of data. The need of such summary can appear in different ways, for instance, from any query to a data base which induces categories and descriptive variables. These categories can be, for instance, simply the towns or in a more complex way, the socio-professional categories (SPC) crossed with categories of age (A) and regions (R). Hence, in this last case, we obtain a new categorical variable of cardinality $|\text{SPC}| \times |A| \times |R|$ where $|X|$ is the cardinality of X. The descriptive variables of the households can then be used in order to describe these categories by symbolic data.
Symbolic Data can also appear after a clustering in order to describe in an explanatory way (by using the initial variables) the obtained clusters.
Symbolic data may also be "native" in the sense that they result from expert knowledge (scenario of traffic accidents, type of emigration, species of insects, ...), from the probability distribution, the percentiles or the range of any random variable associated to each cell of a stochastic data table, from time series (in representing each time series by the histogram of its values or in describing intervals of time), from confidential data (in order to hide the initial data by less accuracy), etc. They result

also from Relational Data Bases, in order to study a set of units whose description needs the merging of several relations as is shown in the following example.

1.2 Output of Symbolic Data Analysis

Most of the symbolic data analysis algorithms give in their output the symbolic description "d" of a class of individuals (which are the partial or complete extent of a given concept), by using a "generalization" process. By starting with this description, symbolic objects model the underlying concept and give a way, to find at least, the individuals of this class.

Example: The age of two individuals $w1$, $w2$ which satisfy a given concept (for instance they leave in the same town), are age($w1$) = 30, age($w2$) = 35, the description of the class C = {$w1$, $w2$} obtained by a generalization process can be [30, 35]. The extent of this description contains at least $w1$ and $w2$ but may contain other individuals. In this simple case the symbolic object "s" is defined by a triple: $s = (a, R, d)$ where d = [30, 35], R = "$\in$" and "a" is the mapping: $W \rightarrow$ {true, false} such that $a(w)$ = the true value of "age(w) R d" denoted with [age(w) R d]. An individual w is in the extent of s if $a(w)$ = true.

More formally (see figure 1), let W be a set of individuals, D a set containing descriptions of individuals d_w or of a class of individuals d_C, "y" a mapping defined from W into D which associates to each $w \in W$ a description $d_w \in D$ from a given symbolic data table. We denote by R, a relation defined on D. It is defined by a subset W of $D \times D$. If $(x, y) \in W$ we say that x and y are connected by R and this is denoted by $x\ R\ y$. More generally we say that $x\ R\ y$ take its value in a set L. We can have L={true,false}, in this case [$d'\ R\ d$] = true means that there is a connection between d and d'. We can also have L= [0, 1] if d is more or less connected to d'. In this case, [$d'\ R\ d$] can be interpreted as the "true value" of $x\ R\ y$ or " the degree to which d' is in relation R with d". For instance, $R \in \{=, \equiv, \leq, \subseteq\}$ or R is an implication, a kind of matching taking care of missing values, etc. R can also use a logical combination of such operators.

2 Symbolic Objects

A «symbolic object» is defined by a description "d", a relation "R" for comparing d to the description d_w of an individual and a mapping "a" called "membership function". More formally: "a symbolic object is a triple $s = (a, R, d)$ where R is a relation between descriptions, d is a description and a is a mapping defined from W in L depending on R and d".
Symbolic Data Analysis concerns usually classes of symbolic objects where R is fixed, "d" varies among a finite set of coherent descriptions and "a" is such that: $a(w)$=[$y(w)\ R\ d$] which is by definition the result of the comparison of the description of the individual w to d. More generally, many other cases can be considered. If, for instance, the mapping "a" is of the following kind: $a(w) = [h_e(y(w))\ h_j(R)\ h_i(d)]$

where the mappings h_e , h_j and h_i are "filters" which will be discussed hereunder. There are two kinds of symbolic objects:

- «Boolean symbolic objects» if $[y(w)\ R\ d] \in L = \{\text{true, false}\}$. In this case, if $y(w) = (y_1,...,y_p)$, the y_i are of type (a) to (d), defined in section 1.
  Example: Let be $a(w) = [y(w)\ R\ d]$ with R: $[\ d'\ R\ d\ ] = \vee_{\ i=1,\ 2}\ [d'_i\ R_i\ d_i\ ]$ where $\vee$ has the standard logical meaning and $R_i = \subseteq$. If $y(w)$ = (colour(w), height(w)), d=({red, blue, yellow}, [10,15])=(d_1, d_2), colour(u)={red, yellow}, height(u)={21}, then $a(u)$=[colour(u) $\subseteq$ {red, blue, yellow}]$\vee$ [height(u) $\subseteq$ [10,15]]= true $\vee$ false = true.
- «Modal symbolic objects» if $[y(w)\ R\ d] \in L = [0,1]$.
  Example: Let be $a(u) = [y(u)\ R\ d]$ where for instance R: $[d'\ R\ d\ ]$ = Max $_{\ i=1,\ 2}$ $[d'_i\ R_i\ d_i\ ]$. The choice of the Max is among many other possible choices related to copulas theory (Diday 2000). The "matching" of two probability distributions is defined for two discrete probability distributions $d'_I = r$ and $d_i = q$ of k values by: $r\ R_i\ q = \sum_{j=1,k}\ r_j\ q_j\ e^{(r_j\ -\ min\ (r_j,\ q_j))}$. By analogy with the Boolean case we denote $[\ d'R\ d\ ]= \vee^*_{\ i=1,\ 2}[\ d'_i\ R_i\ d_i\ ]$ where $\vee^*$ = Max . With these definitions it is possible to calculate the mapping "a" of a symbolic object $s = (a, R, d)$ where SPC means «socio-professional-category» and d = ({(0.2)12, (0.8)[20 ,28]}, {(0.4)employee, (0.6)worker}) by: $a(u)$ = [age(u) R_1{(0.2)12, (0.8) [20 ,28]}] $\vee^*$ [SPC(u) R_2{(0.4)employee, (0.6)worker}].
  Notice that in this example the weights (0.2), (0.8), (0.4), (0.6) represent frequencies but more generally other kinds of weights may be used as "possibilities", "necessities", "capacities", etc. Notice that the R_i depends on this choice, (Diday 1995).

2.1 Syntax of Symbolic Objects in the Case of "Assertions"

If the initial data table contains p variables we denote $y(w) = (y_1(w),..., y_p\ (w))$, $D = (D_1,...,D_p)$, $d \in D$: $d = (d_1,..., d_p)$ and $R' = (R_1,...,R_p)$ where R_i is a relation defined on D_i. We call «assertion» a special case of a symbolic object defined by $s = (a, R, d)$ where R is defined by $[d'\ R\ d] = \wedge_{\ i=1,\ p}\ [\ d'_i\ R_i\ d_i]$ where "$\wedge$" has the standard logical meaning and "a" is defined by: $a(w) = [\ y(w)\ R\ d\ ]$ in the Boolean case. Notice that considering the expression $a(w) = \wedge_{\ i=1,\ p}\ [y_i\ (w)\ R_i\ d_i]$ we are able to define the symbolic object $s = (a, R, d)$. Hence, we can say that this explanatory expression defines a symbolic object called "assertion".

For example, a Boolean assertion is: $a(w)$ = [age(w) $\subseteq$ {12, 20 ,28}] $\wedge$ [SPC(w) $\subseteq$ {employee, worker}]. If the individual u is described in the original symbolic data table by age(u)={12, 20} and SPC (u) = {employee } then: $a(u)$ = [{12, 20 }$\subseteq$ {12, 20 ,28}] $\wedge$ [{employee}$\subseteq${employee, worker}]= true.

In the modal case, the variables are multi-valued and weighted, an example is given by $a(u) = [y(u)\ R\ d\ ]$ with $[d'\ R\ d\ ] = f(\{[y_i(w)\ R_i\ d_i]\}_{i=1,...,p})$ where for instance, $f(\{[y_i(w)\ R_i\ d_i]\}_{i=1,...,p}) = \Pi_{\ i=1,\ 2}\ [\ d'_i\ R_i\ d_i\ ]$ where in case of probability distributions, the "matching" is defined for two discrete density distributions $d'_i = r = (r_1, ...,r_k)$ and $d_i = q = (q_1, ... , q_k)$ of k values by: $r\ R_i\ q = \sum_{j=1,k}\ r_j\ q_j\ e^{(r_j\ -\ min(r_j,q_j))}$.

By analogy with the Boolean case we denote $[\, d' R\, d\,] = \wedge^*_{i=1,2} p_i [\, d'_i\, R_i\, d_i\,]$ where the meaning of "$\wedge^*$" is given by the definition of the mapping "*f*". For instance, with these choices, a modal assertion $I = (a, R, d)$ is completely defined by the equality: $a(w)$ = [age(w) R_1 {(0.2)12, (0.8) [20 ,28]}] $\wedge^*$ [SPC(w) R_2 {(0.4)employee, (0.6)worker}].

Extent of a symbolic object s. In the Boolean case, the extent of a symbolic object is denoted *Ext*(I) and defined by the extent of a, which is: $Extent(a) = \{w \in W / a(w) = \text{true}\}$. In the modal case, given a threshold α, it is defined by $Ext_a(s) = Extent_a(a) = \{w \in W / a(w) \geq \alpha\}$.

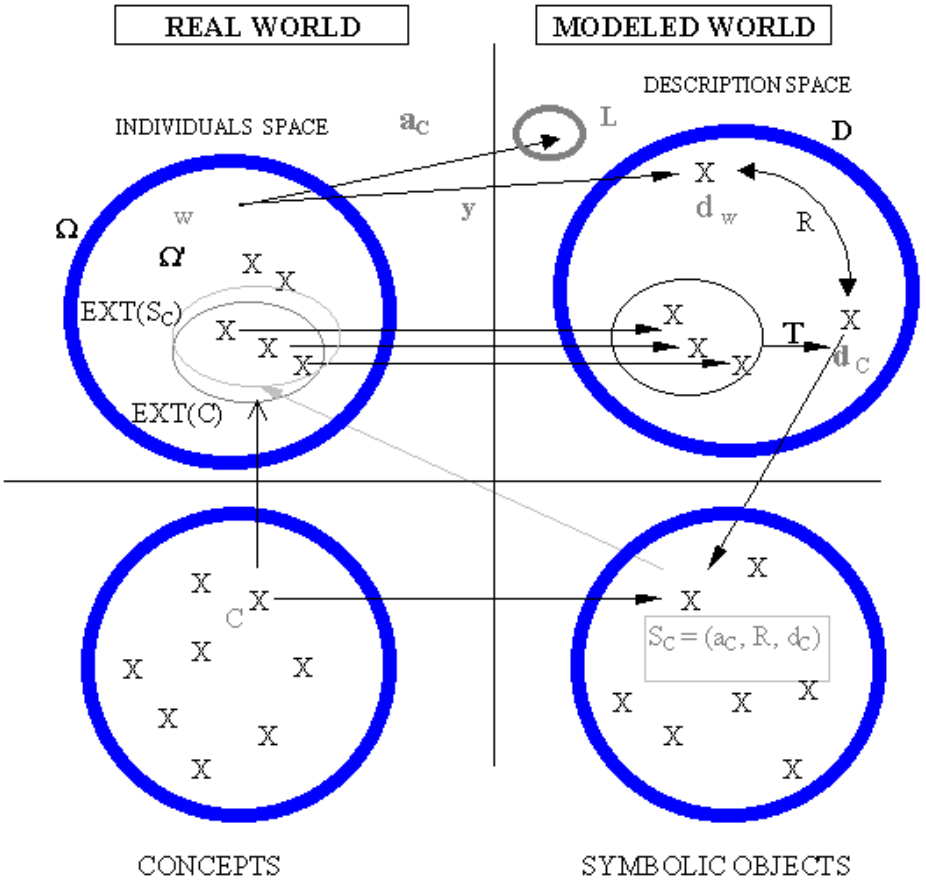

Fig. 1. Modeling by a symbolic object of a concept known by its extent

2.2 Underlying Structures of Symbolic Objects: a Generalized Conceptual Lattice

Under some assumptions on the choice of R and T (for instance $T \equiv$ Max if $R \equiv \leq$ and $T \equiv$ Min if $R \equiv \geq$) it can be shown that the underlying structure of a set of symbolic objects is a Galois lattice (Diday 1991, Brito 1994, Diday & Emilion 1995, 1997), Polaillon & Diday (1997), Polaillon (1998), Bock & Diday (2000)), where the vertices are closed sets defined thereunder by «complete symbolic objects». More precisely, the associated Galois correspondence is defined by two mappings F and G:

- F: from $P(W)$ (the power set of W) into S (the set of symbolic objects) such that $F(C)= s$ where $s = (a, R, d)$ is defined by $d = T_{c \hat{I} C}\, y(C)$ and so $a(w) = [y(w)\ R\ T_{c \hat{I} C}\ y(C)]$, for a given R. For example, if $T_{c \hat{I} C}\, y(C) = \cup_{c \in C}\, y(C)$, $R \equiv$ «$\subseteq$», $y(u) = \{$pink, blue$\}$, $C = \{$c, c'$\}$, $y(C) = \{$pink, red$\}$, $y(c') = \{$blue, red$\}$, then $a(u) =[y(w)\ R\ T_{c \hat{I} C}\ y(C)] = [\{$pink, blue$\} \subseteq \{$pink, red$\} \cup \{$blue, red$\}) = \{$pink, red, blue$\}] =$ true and $u \in Ext\ (s)$.
- G: from S in $P(W)$ such that: $G(s) = Ext\ (s)$.

A «complete symbolic object» s is such that $F(G(s)) = s$. Such objects can be selected from the Galois lattice but also, from a partitioning, a hierarchical or a pyramidal clustering, from the most influential individuals in a factorial axis, from a decision tree, etc.

In order to see how much a given symbolic object is characteristic of a class A, an hypergeometric distribution can be used. Let N be the size of W', n the size of A, $p = Ext(s/W')/N$ the proportion in W' of individuals belonging in the extent of s, X a random variable whose value at each resample is the proportion in A of individuals belonging in the extent of s. Then, the hypergeometric law gives the probability of $X = x$ by: $Pr(X=x) = C^{x}_{Np}\, C^{n-x}_{N-Np} / C^{n}_{N}$ where $C^{n}_{N} = N! / n!(N-n)!$ is the number of possible samples of size n in N, $C^{x}_{Np} =$ Np!/(Np-x)!x! is the number of groups of x individuals belonging in the extent of s in W' and $C^{n-x}_{N-Np} = (N-Np)!/ (n-x)!(N-Np-n+x)!$ is the number of groups of $(n-x)$ individuals which are not belonging in the extent of s in W'. If the operator T produces k symbolic objects of extent in A with size $x_1,\ldots, x_k$ then the more $Y = \mathring{a}_{i\,=\,1,k}\, Pr(X = x_i)/k$ is small, the more these symbolic objects are characteristic of the class A. This happen for instance, when p is small and x/n large or p large and x/n small. Notice that in the case where s is a complete symbolic object the size of the extent is n and $p = n/N$, so $Pr\ (X = n) = \boldsymbol{C^{n}_{n}}\ x\ \boldsymbol{C^{0}_{N-n}} / \boldsymbol{C^{n}_{N}} = 1x1/\ \boldsymbol{C^{n}_{N}} = ((N-n)!\ n!)/\ N!$ which is the probability of a complete symbolic object of size n in a population of size N. When bootstrapping W', if the mean of the random variable Y is out of the chosen confidence interval, then the more its standard deviation is low the more the characterization is reliable. If we are interested by the variation of the characteristic of a specific symbolic objet, notice that at each resample we have to recognize each symbolic object. This can be done by the use of a dissimilarity measure between symbolic objects from one resample to the next (Esposito et al. 2000). The closest are considered to be the same.

A «complete symbolic object» s is such that $F(G(s)) = s$. Such objects can be selected from the Galois lattice but also, from a partitioning, a hierarchical or a pyramidal clustering, from the most influential individuals in a factorial axis, from a decision tree, etc.

2.3 Modeling Individuals, Classes of Individuals and Concepts

In figure 1 the "set of individuals" and the "set of concepts" is considered to be in the "real world", the "modeled world" is the "set of descriptions" which models individuals (or classes of individuals) and the "set of symbolic objects" which models concepts. We start with a "concept" C whose extent denoted $Ext(C/W')$ is known in a sample W' of individuals. For instance, if the concept is "insurance companies", for instance, 30 insurance companies among a sample W' of 1000 companies. Each individual w of the extent of C in W' is described by using the mapping Y such that $Y(w)$ describe the individual w. We generalize the set of descriptions of the individuals of $Ext(C/W')$ with the operator T in order to produce the description d_C (which can be a set of Cartesian products of intervals and (or) distributions).

i. The comparison relation R is chosen in relation with the T choice. For instance, if $T = \cup$ then $R =$ "$\subseteq$", it $T = \cap$, then $R =$ "$\supseteq$".
ii. The membership function is then defined by $a_C(w) = [y(w)\ R_C\ d_C]$ and then the symbolic object modelling the concept C is the triple $s = (a_C, R, d_C)$.

When we don't have concepts as input, we get them in the following way:

i. A clustering of W' by using the description of the individuals produces a set of classes.
ii. To each interesting class denoted A, we associate a concept C and a symbolic object $s_A = (a_A, R_A, d_A)$ with $a_A = [Y(w)\ R_A\ d_A]$ where d_A is obtained by using an operator T on the set of the descriptions of the individuals of A, as in the preceding case.
iii. The concept C is considered to be modeled by s_A.

2.4 Some Advantages in the Use of Symbolic Objects

We can observe at least five kinds of advantages in the use of symbolic objects.

1. They give a summary of the original symbolic data table in an explanatory way, (i.e. close to the initial language of the user) by expressing descriptions based on properties concerning the initial variables or meaningful variables (such as indicators obtained by regression or factorial axes).
2. They can be easily transformed in terms of a query of a Data base and so they can be used in order to propagate concepts between data bases (for instance, from one country to another country).
3. By being independent of the initial data table they are able to identify any matching individual described in any data table.
4. In the use of their descriptive part, they are able to give a new symbolic data table of higher level on which a symbolic data analysis of second level can be applied.
5. In order to characterize a concept, they are able to join easily several properties based on different variables coming from different relations in a Data Base and different samples of a population.
6. In order to apply exploratory data analysis to several data bases, instead of merging them in a huge data base, an alternative is to summarize each Data Base by symbolic objects and then to apply Symbolic Data Analysis to the whole set of obtained symbolic objects.

3 Some Symbolic Data Analysis Methods

Symbolic Data Analysis methods are mainly characterized by the following principle:

i. they start as input with a symbolic data table and they give as output a set of symbolic objects. These symbolic objects give explanation of the results in a language close to the one of the user and moreover have all the advantages mentioned in 5).
ii. They use efficient generalization processes during the algorithms in order to select the best variables and individuals.
iii. They give graphical descriptions taking account of the internal variation of the symbolic objects.

The following methods are developed in Bock & Diday (2000) and in the SODAS software:

- Principal Component and Discriminate Factorial Analysis of a symbolic data table. The output of these methods preserves the internal variation of the input data in the sense that the individuals are not represented in the factorial plane by a point as usual but by a rectangle which allows the definition of a symbolic object with explanatory factorial axes as variables;
- extension of elementary descriptive statistics to symbolic data (central object, histograms, dispersion, co-dispersion, etc. from a symbolic data table);
- extracting symbolic objects from the answers to queries of a relational data base;
- partitioning, hierarchical or pyramidal clustering of a set of individuals described by a symbolic data table such that each class be associated with a complete symbolic object;
- dissimilarities between Boolean or probabilistic symbolic objects;
- extension of decision trees on probabilistic symbolic objects;
- generalization by a disjunction of symbolic objects of a class of individuals described in a standard way;
- inter-active and ergonomic graphical representation of symbolic objects.

4 Symbolic Data Analysis in the SODAS Software

The general aim of SODAS can be stated in the following way: building symbolic data in order to summarize huge data sets and then, analyze them by Symbolic Data Analysis. For instance, if a set of households is characterized by its region, the number of bedrooms and of dining-living, its socio-economic group, we obtain a data table of the kind of table 1:

Table 1. Standard Data Table where the units are Households

Household number	Region	Bedroom	Dining-Living	Socio-Econ group
11404	Northern-Metropolitan	2	1	1
11405	Northern-Metropolitan	2	1	3
11406	Northern-Metropolitan	1	3	3
12111	Northern-Metropolitan			
12112	East anglia	1	3	3
12112	East anglia	2	2	1
12112	Greater London N-E	1	2	3

In census data there is a huge set of households. In order to compare the regions, we can summarize them by describing each region by the households of their inhabitants. In order to do so, we delete the first column of this table and we obtain the table 2:

Table 2. The first column of table 4 concerning the household number has been deleted

Region	Bedroom	Dining-Liv	Socio-Ec gr
Northern-Metropolitan	2	1	1
Northern-Metropolitan	2	1	3
Northern-Metropolitan	1	3	3
Northern-Metropolitan			
East-anglia	1	3	3
East-anglia	2	2	1
East-anglia	1	2	3
Greater London North-East			

We can now describe each town by the histogram of the categories of each variable. This is done in table 3 which is a symbolic data table as each cell contains a histogram and not a quantitative or categorical number as in the standard data tables. It is easy to see that standard data analysis methods will not apply in the same way with these kind of symbolic data. For instance that a decision tree will not be the same if the variables are categories and each cell of the associated data table contains a frequency and if the variable are symbolic and each cell contains a histogram. In the first case each branch of the decision tree represents an interval of frequency (for instance, “the frequency of the category [20, 30] years old is less then 0.3”), whereas in the second case it represents an interval of values (for instance, “the age is less then 50 years old”). For more details see in Bock & Diday (2000) the chapter 11.

Table 3. A symbolic data table where the units are now regions

Region	Bedroom	Dining-Living	Socio-Ec gr
Northern Metropolitan	(2\3) 2, (1\3) 3	(2\3) 1, (1\3) 3	(1\3) 1, (2\3) 3
East-anglia	(2\3) 1, (1\3) 2	(2\3) 2, (1\3) 3	(1\3) 1, (2\3) 3
Greater London			

The main steps for a symbolic data analysis in SODAS can then be defined as following:
If there is more than one data table, put the data in a relational data base (ORACLE, ACCESS, and so on). Then, define a context by giving: the units (individuals, households, and so on), the classes (regions, socio-economics groups,...), the descriptive variables of the units. Then, build a symbolic data table where the units are the preceding classes, the descriptions of each class is obtained by a histogram as in table 6 or by a generalization process applied to its members. This is done by a computer program of SODAS called "DB2SO" (from Data Bases Two Symbolic Objects). Finally, apply to this symbolic data table, symbolic data analysis methods (histogram of each symbolic variable, dissimilarities between symbolic descriptions, clustering, factorial analysis, discrimination of a symbolic data table, graphical visualization of symbolic descriptions, and so on).

5 Conclusion

The need to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination,...) to symbolic data tables in order to extract new knowledge, is increasing due to the expansion of information technology, now able to store an increasing amount of huge data sets. This need, has led to a new methodology called "Symbolic Data Analysis" whose aim is to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination, decision trees,...) to new kind of data table called "symbolic data table" and to give more explanatory results expressed by real world concepts mathematically represented by easy readable "symbolic objects". The aim of the EUROSTAT European Community project called SODAS for a «Symbolic Official Data Analysis System» in which 17 institutions of 9 European countries are concerned was to produce a first software of Symbolic Data Analysis (fig. 2). Three Official Statistical Institutions was involved in this project: EUSTAT (Span), INE (Portugal) and ONS (England). An example of future application proposed on their Census data consists in finding clusters of unemployed people and their associated mined symbolic objects in a country, calculating its extent in the census of another country and describing this extent by new symbolic objects in order to compare the behaviour of the two countries. In that way, several new theoretical development are needed as the selection and the stochastic convergence of symbolic objects. Also, as the consensus between set of symbolic objects and their associated concepts extracted from different data bases. New software development are also needed as a tool in order to be able to transform a symbolic object extracted from a data base in a query of this data base or of another data base. This new tool may be called SO2DB as it is complementary to the actual DB2SO (Malerba et al,

2002). Moreover, the next steps will be to improve the actual SDA methods (robustness, validity of the results, extending standard tests to symbolic data, etc.) and extend the symbolic data analysis methodology to regression, multidimensional scaling, neural network etc. The SODAS software is free and available at http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm.

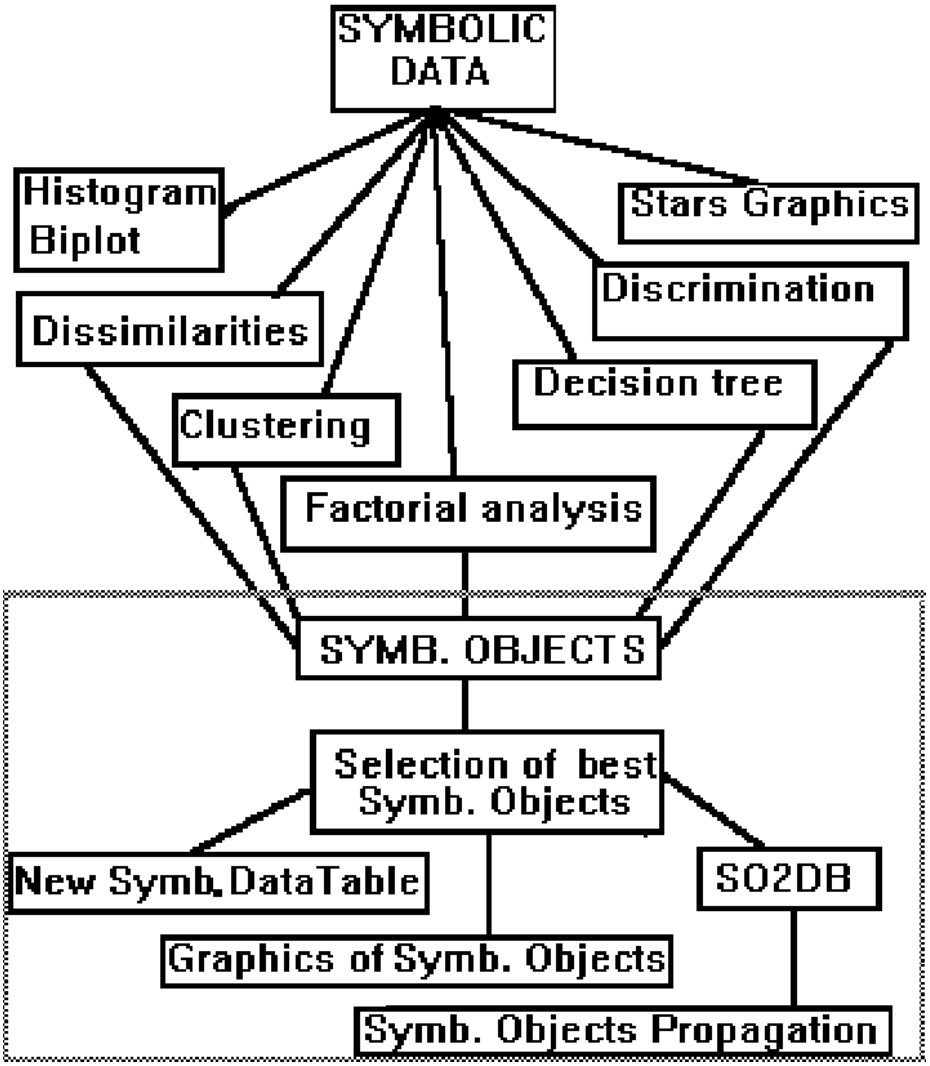

Fig. 2. Software development of the SODAS project

References

Auriol E. (1995) “Intégration d’approches symboliques pour le raisonnement à partir d’exemples” Thèse de doctorat, Université Paris 9 Dauphine.

Bock H.H., Diday E. (2000) “Analysis of Symbolic Data”. Study in Classification, Data Analysis and Knowledge Organization. Springer Verlag.

Bravo C., Garcia-Santesmas J. (1998) "Symbolic objects description of strata by segmentation trees". Proc. NTTS. Ph. Nanopoulos, Garonna, Lauro edit, Eurostat, Sorrento (Italy).

Brito P., Diday E. (1991) “Pyramidal representation of symbolic objects” NATO ASI Series, Vol. F 61. Proc. Knowledge Data and computer-assisted Decisions. Schader and Gaul edit. Springer-Verlag.

Brito P. (1994) “Order structure of symbolic assertion objects”. IEEE TR. on Knowledge and Data Engineering Vol.6, n° 5, October.

Cazes P., Chouakria A., Diday E., Schechtman Y.(1997) “Extension de l’Analyse en Composantes Principales à des données intervalles”. Revue de Statistiques Appliquée, vol. XXXVIII, n°3, 1990,pp 35-51.

Chavent M. (1997) “Analyse des Données symboliques. Une méthode divisive de classification”. Thèse de doctorat, Université Paris 9 Dauphine.

Ciampi A., Diday E., Lebbe J., Périnel E., Vigne (1995) R. “Recursive partition with probabilistically imprecise data”. OSDA’95. Editors: Diday, Lechevallier, Opitz Springer Verlag (1996).

Conruyt N. (1994) “Amélioration de la robustesse des systèmes d’aide à la description, à la classification et à la détermination des objets biologiques”. Thèse de doctorat, Université Paris 9 Dauphine.

De Carvalho, F.A.T. (1994) “Proximity coefficients between Boolean symbolic objects”. In: Diday, E., Lechevallier, Y., Schader, M., Bertrand, P. and Burtschy, B. (Eds.): *New Approaches in Classification and Data Analysis*, Springer-Verlag, Heidelberg, Germany, 387-394.

De Carvalho F.A.T. (1998a) “New metrics for constrained boolean symbolic objects” Proc. KESDA’98, Eurostat. Luxembourg.

De Carvalho F.A.T. (1998b) “Statistical proximity functions of boolean symbolic objects based on histograms” IFCS, Roma, Springer-Verlag.

Diday E. (1987a) “The symbolic approach in clustering and related methods of Data Analysis” in “Classification and Related Methods of Data Analysis”, Proc. IFCS, Aachen, Germany. H. Bock ed.North-Holland.

Diday E. (1987b) “Introduction à l’approche symbolique en Analyse des Données ”. Première Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987.

Diday E. (1989) “Introduction à l’approche symbolique en analyse des données”. RAIRO (Revue, d’Automatique, d’informatique et de Recherche Opérationnelle), vol. 23, n°2.

Diday E. (1991) “Des objets de l’Analyse des Données à ceux de l’Analyse des Connaissances” in “Induction symbolique et numérique”. Y. Kodratoff and E. Diday edit. CEPADUES-EDITIONS, Toulouse, France.

Diday E. (1995) “Probabilist, possibilist and belief objects for knowledge analysis”. Annals of Operations Research . 55, 227-276.

Diday E., Emilion R. (1995) “Lattices and Capacities in Analysis of Probabilist Objects”. Proceed. of OSDA’95 (Ordinal and Symbolic Data Analysis). Springer Verlag Editor (1996).

Diday E., Emilion R. (1997) “Treillis de Galois maximaux et Capacités de Choquet” Compte rendu à l’Académie des Sciences. Analyse Mathématique, t. 324, série 1.

Diday E., Emilion R., Hillali Y. (1996) “Symbolic data analysis of probabilist objects by capacities and credibilities. XXXVIII Societa Italiana Di Statistica. Rimini, Italy.

Diday E.(1998) “L’Analyse des Données Symboliques: un cadre théorique et des outils”. Cahiers du CEREMADE n° 9821.

Diday E. (2000) “Partitioning concepts described by distributions with copulas modelling” OSDA ‘2000. Bruxelles.

Esposito, F., Malerba, D., & Tamma, V.(2000) “Dissimilarity Measures for Symbolic Objects” (Section 8.3), 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, pp. 165-185.

Ferraris, Gettler-Summa, C. Pardoux, H. Tong (1995) “Knowledge extraction using stochastic matrices: Application to elaborate a fishing strategy” Proc. Ordinal and Symbolic Data Analysis. Paris ; Diday, Lechevallier, Opitz edit.Springer Studies in Classification.

Gettler-Summa M. (1992) “Factorial axis interpretation by symbolic objects”. Journées - Symbolique - Numérique. Université Paris IX- Dauphine. Lise-Ceremade.

Gettler-Summa, M., Pardoux, C. (2000) Noirhomme-Fraiture, Rouard M. (2000) “Symbolic Approaches for Three-way Data” (Chapter 12), 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, pp. 342-354.

Gigout E. (1998) “Graphical interpretation of symbolic objects resulting from data mining”. Proc. KESDA’98, Eurostat. Luxembourg.

Gowda K.C., Diday E. (1992) “Symbolic clustering using a new similarity measure”. IEEE Trans. Syst. Man and Cybernet. 22 (2), 368-378.

Hillali, Y. (1998) “Analyse et modélisation des données probabilistes: capacités et lois multidimensionnelles”, Thèse de doctorat, University Paris 9 Dauphine.

Kiers, H., Rasson, J.P., Groenen, P.J.F., Schader, M. (eds.) (2000) Data Analysis, Classification, and Related Methods. Series: Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, Berlin.

Lauro C., Palumbo F. (1998) “New approaches to Principal Component Analysis of Interval Data”. Nanopoulos, Ph., Garonna, Lauro, C. (eds.) Proc. NTTS’98 Sorrento, Italy. Eurostat, Luxembourg.

Malerba, D., Esposito, F., Monopoli M. (2002). Estrazione e matching di oggetti simbolici da database relazionali. Atti del Decimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati SEBD’2002, 265-272.

Mfoumoune E.-M. (1998) “Analyse de données symbolique incrémentale et apprentissage”, Thèse de doctorat, University Paris 9 Dauphine.

Michalski R., Diday E., Step R.E. (1982) “A recent advances in Data Analysis: clustering objects into classes characterized by conjonctive concepts”. Progress in Pattern Recognition, vol 1. L; Kanal and A. Rosenfeld Eds.

Noirhomme-Fraiture, Rouard M. (1998) “Representation of Sub-Populations and Correlation with Zoom Star”. Proc. NTTS’98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).

Noirhomme-Fraiture, Rouard M. (2000) “Visualizing and Editing Symbolic Objects” (Chapter 7), 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, pp. 125-138.

Périnel E. (1996) “Segmentation et Analyse de Données Symboliques: Application à des données Probabilistes Imprécises”. Thèse de doctorat, Université Paris 9 Dauphine.

Pollaillon G., Diday E. (1997) “Galois lattices of symbolic objects” Rapport du Ceremade University Paris9- Dauphine (February).

Pollaillon G. (1998) “Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme”. Thèse de doctorat, Université Paris 9 Dauphine.

Pollaillon G. (2000) “Pyramidal Classification for Interval Data Using Galois Lattice Reduction” (Section 11.4), 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, pp. 324-341.

Rasson J.P., Lissoir S. (1998) “Symbolic Kernel discriminante analysis” Proc. NTTS’98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).

Rodriguez O. (2000) “Classification et modèles linéaires en Analyse des Données Symboliques”. Thèse de doctorat, University Paris 9 Dauphine.

Stéphan (1998) “Construction d’objets symboliques par synthèse des résultats de requêtes SQL”. Thèse de doctorat, Université Paris 9 Dauphine.

Stéphan, V., Hébrail, G., Lechavallier, Y. (2000) “Generation of Symbolic Objects from Relational Database” (Chapter 5), 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, pp. 78-105.

Tang Ahanda, B. (1998) “Extensions des méthodes d’analyse factorielle sur des données symboliques. Thèse de doctorat, Université Paris 9 Dauphine.

Vignes (1991) “Caractérisation automatique de groupes biologiques”. Thèse de doctorat, Université Paris 9 Dauphine.

Verde R., F.A.T. De Carvalho (1998) “Dependance rules influence on factorial representation of boolean symbolic objects”. Proc. KESDA’98, Eurostat. Luxembourg.

Ziani D. (1996) “Sélection de variables sur un ensemble d’objets symboliques” Thèse de doctorat, Université Paris 9 Dauphine.