

# KNOWLEDGE DISCOVERY FROM SYMBOLIC DATA AND THE SODAS SOFTWARE

Edwin Diday<sup>1</sup>

University Paris 9 Dauphine,  
Ceremade. Pl. Du Ml de L. de Tassigny. 75016  
diday@ceremade.dauphine.fr

**Abstract.** The data descriptions of the units are called “symbolic” when they are more complex than the standard ones due to the fact that they contain internal variation and are structured. Symbolic data happen from many sources, for instance in order to summarise 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. In this paper we give an overview on recent development on 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 values. In a symbolic data table, a cell can contain, a distribution (Schweitzer (1984) says that “distributions are the number of the futur”!), or intervals, or several values linked by a taxonomy and logical rules, etc.. 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 summarise extensive data sets contained in Data Bases.

### 1.1 Historical and practical origin of the Symbolic Data Analysis field.

The key idea of SDA has been given by Aristotle, four century B. C. . The Aristotle Organon (IV B. C.) clearly distinguishes “first order individuals” (as

"a horse" or "a man") considered as a unit associated to an individual of the world, from "second order individuals" (as "the horse" or "the man") also taken as a unit associated to a class of individuals.

Our first aim is to extend standard data analysis to second order individuals. For instance, in a census of a country, each individual of each region is described by a set of numerical or categorical variables given in several relations of a Data Base.

Such individual is considered as a "first order individual". In order to study the regions considered as "second order individuals", we can describe each of them in summarising the values taken by its inhabitants, by inter-quartile intervals, or subsets of categorical values, or histograms or probability distributions, etc. depending on the concerned variable. In such a way, we obtain a "symbolic data table" where each row defines the "description" of a region and each column is associated to a symbolic variable. An extension of standard Data Analysis to such data table is the first aim of what we have called "Symbolic Data Analysis".

Another important aim is to obtain (or "mine") explanatory results (i.e. knowledge) by extracted, the so called "symbolic objects" which modelize a "concept" or a "physical entity" of the real world. A "symbolic object" is defined by its "intent" which contains a way of finding its "extent". For instance, the description of the inhabitant of a region and the way of allocating an individual to this region, is called "intent", the set of individuals which satisfy this intent is called "extent". The syntax of symbolic objects must have an explanatory power. For instance, the symbolic object defined by the following expression (see section 4, for a formal definition):  $a(w) = [\text{age}(w) \in [30, 35]] \wedge [\text{Number of children}(w) \leq 2]$ , gives the intent of a class of individuals by at the same time:

- i) the description  $d = ([30, 35], 2)$ , where  $[30, 35]$  is the inter-quartile interval of the random variable associated to the region for the variable age,
- ii) a way of calculating the extent by the mapping "a" defined with the help of the relation  $R = (\in, \leq)$ .

It means that an individual " $w$ " satisfies this intent (i.e. belongs to the "extent") if his age is between 30 and 35 years old and he has less than 2 children.

This very simple kind of symbolic object can be extended at least in the following way: the individuals are of second order (as towns or regions) and represent classes of individuals of first order; therefore the descriptions of the individuals are defined by distributions (the histogram of the age in a town, for instance). In this case we have to define a different kind of relation " $R$ " and a threshold in order to calculate the extent.

There are several advantages in the use of symbolic objects to model "concepts", one of them, is their ability to be translated in a query of a Data Base and therefore, to propagate the "concepts" that they describe from one data base to another database (i.e. from a country to another country).

## 1.2 Symbolic objects model "concepts" but what do we call a "concept"?

There are two kinds of "concepts":

i) The “concepts of the real world” as a town, a region, a scenario of road accident, a kind of unemployment,... That kind of concept is defined by an ”intent” and an “extent” notions brightly defined by Arnault and Nicole(1662) in the framework of Port–Royal school<sup>1</sup>.

ii) The ”concepts of our mind” (among the so called ”mental objects” by J.P. Changeux (1983)) which represents in our mind concepts of the real world by their intent and a ”way of computing their extent” and not the extent itself as (for sure!) there is no room for all the possible extents. A concept of our mind can be mathematically modeled by a symbolic object which is defined by a description ”d” (i.e. its intent) and a mapping ”a” able to compute its extent , for instance, the description of what we call a ”car” and a way of recognizing that a given entity of the real world is a car. A concept of the real world can be modeled by a symbolic object and its extent. Whereas, ”concepts” or ”entities” of the real world are mathematically modeled by ”symbolic objects”, their computing modelization is provided by the so called ”objects” used in the ”object oriented language” and for instance, in computer languages as C++ or JAVA.

### 1.3 Concepts: four tendencies

In the Aristotelian tradition, concepts are characterized by logical conjunction of properties. In the Adansonian tradition (Adanson (1727-1806) was a french naturalist very much ahead of his time), a concept is characterized by a set of ”similar” individuals. In contrast , with the aristotelician tradition, were all the members of the extent of a concept are equivalent, a third tendency derived from psychology and cognitive science (see Rosch (1978)), is to consider that concepts must be represented by classes which ”tend to become defined in terms of prototyped or prototypical instances that contain the attributes most representative of items inside the class” .Wille (1981), following Wagner (1973) says as ”in traditional philosophy things for which their intent describes all the properties valid for the individual of their extent are called ”concept”.

Symbolic objects combine the advantages of these four tendencies:

- The Aristotelician tradition as they can have the explanatory power of a logical description of the concepts that they represent.

- The Adansonian tradition as the member of the extent of a symbolic object are similar in the sens that they must satisfy at the best the same properties (not necessarily boolean). In that sens the concepts that they represent are polythetic.

---

<sup>1</sup> “Now, in these universal ideas there are two things which is important to keep quite distinct: comprehension and extension. I call the comprehension of an idea the attributes which it contains and which cannot be taken away from it without destroying it; thus the comprehension of the idea of a triangle includes, to a superficial extent, figure, three lines, three angles, the equality of these three angles to two right angles to two right angles, etc. I call the extension of an idea the subjects to which it applies, which are also called the inferiors of a universal term, that being called superioir to them. Thus the idea of triangle in general extends to all different kinds of triangle”.

- The Rosch point of view, as their membership function is able to provide prototypical instances characterized by the most representative attributes.

- The Wille property is satisfied by the so called "complete symbolic objects" which can be proved that they constitute a Galois lattice on symbolic data (see for instance, Diday (1998)). Symbolic Data Analysis is born from which influence? There was a simultaneous influence of several fields:

- standard exploratory data analysis (Tuckey (1958), Benzécri (1973), Diday et al (1984) , Saporta (1990), Lebart et al (1998)) where more importance is given to individuals than in standard statistics and where the symbolic approach extend the methods to more complex descriptions of the units and give more explanatory results .

- Artificial Intelligence (AI) where much efforts has been devoted in finding good languages in order to represent complex knowledge instead of the simple IRp vectors of the standard statistical units. Notice that the simple language used in order to represent symbolic objects is more inspired from languages based on first order logic ((Michalski (1973), Hayes Roth and McDermot (1977)) then from graph representation (Winston (1979), Sowa (1984)). Notice also, that in symbolic data analysis we are not much interested in the computer language (SQL, C++, JAVA, ...) used in order to represent symbolic objects but much more by their mathematical model, the way of inducing them from the data, their graphical representation, etc.

- Numerical Taxonomy in biology, Learning Machine in AI, Classification in Data Analysis.

#### **1.4 Symbolic Data Analysis is born from which influence?**

There was a simultaneous influence of several fields:

- standard exploratory data analysis (Tuckey (1958), Benzécri (1973), Diday et al (1984) , Saporta (1990), Lebart et al (1998)) where more importance is given to individuals than in standard statistics and where the symbolic approach extend the methods to more complex descriptions of the units and give more explanatory results .

- Artificial Intelligence (AI) where much efforts has been devoted in finding good languages in order to represent complex knowledge instead of the simple IRp vectors of the standard statistical units. Notice that the simple language used in order to represent symbolic objects is more inspired from languages based on first order logic ((Michalski (1973), Hayes Roth and McDermot (1977)) then from graph representation (Winston (1979), Sowa (1984)). Notice also, that in symbolic data analysis we are not much interested in the computer language (SQL, C++, JAVA, ...) used in order to represent symbolic objects but much more by their mathematical model, the way of inducing them from the data, their graphical representation, etc.

- Numerical Taxonomy in biology, Learning Machine in AI, Classification in Data Analysis.

### 1.5 In all these fields a natural question arose: how does one obtain classes and their description?

Historically, we may say briefly that there are three tendencies:

The first proposed by A. de Jussieu (1748) is in the Aristotelian tradition and consists in defining top down the classes by a good choice of the properties which characterize them from the most general to the most specific. In that way we obtain a decision tree where each node is characterized by a conjunction of properties. Many others have continued this tendency. By starting from individuals of first order: Belson (1959), Morgan and Sonquist (A.I.D. program (1963)), Lance and Williams (1967), Breiman and al. (1984), Quinlan (1986). By starting from individuals of second order: Pankurst (1978), Payne (1975), Gower (1975), J. Lebbe, R. Vigne (1991), H. Ralambondrainy (1991), Ganascia (1991).

The second tendency, put forward by Adanson (1757) who gave the first "Sequential Agglomerative Hierarchical Clustering" (SAHC) algorithm. This well known "bottom up" algorithm, starting by classes reduced to individuals, merges at each stage the most "similar" classes. This tendency is well represented by Ward (1963), Lerman (1970), Jardine and Sibson (1971), Sneath and Sokhal (1973), Jambu (1978), Roux (1985), Bock (1974), Celeux, Diday, Govaert, Lechevallier and Ralambondrainy (1989), etc. The classes obtained in this way contain similar objects. It is then possible to generalize them in terms of disjunction of conjunction of properties, that why these classes are called "polythetic" in opposition with classes generalized by a conjunction of properties and called "monothetic". Whereas, the first tendency yields monothetic classes by a top-down process, the second produces polythetic classes by a "bottom up" process. In this framework, a family of methods called "Conceptual Clustering" has been developed in the eighties such as Langley and Sages (1984), Lebowitz (1983), Fisher D.H. (1987), Fisher and Langley (1986) for a review. Instead of producing trees, in Diday (1984), Bertrand (1986) for instance, an ascending process building a pyramid (a generalization of hierarchical trees, allowing overlapping clusters) of polythetic classes is described. In Brito and Diday (1991), Brito (1994) an ascending pyramid produces monothetic classes.

The third tendency consists in looking directly for classes and their representation. For instance, the "Dynamic Clustering Method" (Diday (1971), Diday and al (1979)), Diday and Simon (1976)), defines a general framework and algorithms which aim to discover simultaneously classes and their representation in such a way that they "fit" together as well as possible. This approach has been used with several kinds of inter-class structure (partitions, hierarchies, ...) and representation modes for each class (seeds, probability laws, factorial axis, regressions,...). In Diday (1976), a logical representation of clusters is proposed. With regards to the "Conceptual Clustering" algorithm based on the Dynamic Clustering Method or inspired by it, mention should be made of Diday, Govaert, Lechevallier, Sidi (1980), Michalski, Diday, Stepp (1982), Michalski, Stepp (1983) among other pioneers papers in "Conceptual Clustering".

## 1.6 The Symbolic Data Analysis field:

Since the first papers announcing the main principles of Symbolic Data Analysis ((Diday (1987) a, (1987) b, (1989)) many works have been done until the most recent book published by Bock, Diday (2000) and the proceedings of IFCS'2000 , published by Kiers and al (2000) which contains a large chapter devoted to this field.. In factorial analysis, P. Cazes, A. Chouakria, E. Diday, Y. Schecktmann (1997)) have defined a principal component analysis of individuals described by a vector of numerical intervals and in the same direction R. Verde, F.A.T. De Carvalho (1998) by taking care on given dependance rules, see also Lauro, Palumbo (1998) and the section 9.3 in this book. In the case where the individuals are described by symbolic data, Conruyt (1993) in the case of structured data, Ciampi, E. Diday, J. Lebbe, E. Périnel, R. Vigne (1995), Périnel (1996), have developed an extension of standard decision trees. In the same direction E. Périnel has (see chapter 10 in Bock , Diday (2000)) on "symbolic discrimination rules" , M.C. Bravo, J.M. Garcia-Santesmases (1998) on "segmentation trees for stratified data" and J.P. Rasson and S. Lissouir(1998) starting from a dissimilarity between symbolic descriptions. See also E. Auriol (1995) for a link with the domain of "Case Based Reasonning". In order to select the symbolic variables which distinguish at the best the individuals or classes of individuals, several works have been done as R. 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 C. Gowda and E. Diday (1992), De Carvalho (1994, 1998 a). 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 variables values in De Carvalho (1998) b, or by using the capacity theory in Diday, Emilion ((1995, 1997), Diday, Emilion, Hillali (1996). Noirhomme and Rouard (1998) give a way of representing multidimensional symbolic data (see chapter 7 in Bock , Diday (2000)), see also E. Gigout (1998) .

Starting from standard data, Gettler-Summa (1992), Smadhi (1995) have proposed a way for extracting symbolic objects from a factorial analysis ; in order to extract symbolic objects from a partition, see Stephan, Hébrail, Lechevallier (see chapter 5 in Bock , Diday (2000)) and Gettler-Summa in this book. Starting from time-series, Ferraris, Gettler-Summa, C. Pardoux, H. Tong (1995), have defined a way for providing symbolic objects (see chapter 12 in Bock , Diday (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 (1998), in order to build a partition of a set of symbolic objects by a top-down algorithm which provide also a symbolic object associated to each obtained class (see chapter 11 in Bock , Diday (2000)). Stéphane (1998) for extracting symbolic objects from a data base(see chapter 5 in Bock , Diday (2000)). Hillali (1998) for describing classes of individuals described by a vector of probability distributions. Pollaillon (1998), for extending Galois lattices to symbolic data at input and "complete"

symbolic objects at output (see section 11.4 in Bock , Diday E. (2000)). More generally, the most recent algorithms in Symbolic Data Analysis are in this book.

## 2 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 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 :  $\text{height}(w) = 3.5$ .

(b) Single categorical value: for instance,  $\text{Town}(w) = \text{London}$ .

(c) Multivalued: for instance, in the quantitative case  $\text{height}(w) = 3.5, 2.1, 5$  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  $\text{height}(w) = [3, 5]$ , which means that the height of  $w$  varies in the interval  $[3, 5]$ .

(e) Multivalued with weights: for instance a histogram or a membership function (notice that (a) and (b) and (c) are special cases of (e) when the weights are equal to 1). Variables can be: (g) Taxonomic: for instance, the colour is considered to be “light” if it is “yellow”, “white” or “pink” . (h) 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.

(i) With logical dependencies, for instance: if  $\text{age}(w)$  is less than 2 months then  $\text{height}(w)$  is less than 10 . Many example of such symbolic data are given in the chapter 3 in Bock , Diday (2000). Table 1 gives some examples of such data:

WAGES	TOWN	SOCIO-ECONOMIC GROUP
3.5	London	Personal of service
[3, 8]	Paris,London	{0.1 Manager, 0.6 Manual, ...}
[(0.4)[2, 3[, (0.6)[3, 8]]		

Table 1. A “symbolic data table”: each cell contains an example of “symbolic data”.

## 3 Sources of Symbolic Data:

Symbolic data are generated when we summarise huge sets of data. The need of such summary can appear by 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  $|SPC| \times |A| \times |R|$  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 sens 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 serie 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.

**Example:** We have two relations of a Relational Data Base defined as follows. The first one called "delivery" is given in table 1. It describes five types of deliveries characterised by the name of the supplier, its company and the town from where the supplying is coming.

Delivery	Supplier	Company	Town
Liv1	F1	CNET	Paris
Liv2	F2	MATRA	Toulouse
Liv3	F3	EDF	Clamart
Liv4	F1	CNET	Lannion
Liv5	F3	EDF	Clamart

Table 1 Relation "Delivery"

The supplying are described by the relation "Supplying" defined in the following table 2.

Supplying	Supplier	Town
FT1	F1	Paris
FT2	F2	Toulouse
FT3	F1	Lannion
FT3	F3	Clamart
FT5	F3	Clamart

Table 2: Relation "Supplying"

From these two relations we can deduce the following data table 3, which describes each supplier by his company, his supplying and his towns:

Supplier	Company	Supplying	Town
F1	CNET	FT1, FT3	$\frac{1}{2}$ Paris, $\frac{1}{2}$ Lannion
F2	MATRA	FT2	Toulouse
F3	EDF	T4, FT5	Clamart

Table 3: Relation “Supplier” obtained by merging the relations “Delivery” and “Supplying”.

Hence, we can see that in order to study a set of suppliers described by the variables associated with the two first relations we are naturally required to take in account the four following conditions which characterise symbolic data:

i) Multivalued: this happens when the variables ”Supplying” and ”Town” have several values as shown in the table 3.

ii) Multivalued with weights: this is the case for the towns of the supplier F1. The weights means that the town of the supplier F1 is Paris or Lannion with a frequency equal to .

iii) Rules: some rules have to be given as input in addition to the data table 3. For instance, ”if the town is Paris and the supplier is CNET, then the supplying is FT1.

iv) Taxonomy: by using regions we can replace for instance Paris, Clamart by ” Parisian Region ”.

#### 4 Main output of Symbolic Data Analysis algorithms:

Most of these algorithms give in their output the description “ $d$ ” of a class of individuals by using a “generalization” process which give also a way, by starting with this description, to find at least, the individuals of this class.

More formally, let  $\Omega$  be a set of individuals,  $D$  a set containing descriptions of individuals or of class of individuals,  $y$  a mapping defined from  $\Omega$  into  $D$  which associates to each  $w \in \Omega$  a description  $d \in D$  from a given symbolic data table. We denote by  $R$ , a relation defined on  $D$ . It is defined by a subset  $E$  of  $D \times D$ . If  $(x, y) \in E$  we say that  $x$  and  $y$  are connected by  $R$  and this is denoted by  $xRy$ . The characteristic mapping of  $R$  is  $h_R : D \times D \rightarrow \{0, 1\}$  such that  $h_R(x, y) = 1$  iff  $xRy$ . We generalise the mapping  $h_R$  by the mapping  $h_R : D \times D \rightarrow L$  and we denote  $[d'Rd] = H_R(d', d)$  the result of the “comparison” of  $d'$  and  $d$  by  $H_R$ . We can have  $L = \{true, false\}$ , in this case  $[d'Rd] = true$  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,  $[d'Rd]$  can be interpreted as the “true value” of  $xRy$  or ” the degree to which  $d'$  is in relation  $R$  with  $d$  (see in Bandemer and Nather (1992), the section 5.2 on fuzzy relations).

For instance,  $R \in \{=, \equiv, \leq, \subseteq\}$  or is an implication, a kind of matching, etc.  $R$  can also use a set of such operators.

The description of an individual, is called “individual description”. The description of a class of individuals is an ”intensional description”. For instance, the description of a scenario of accidents, of a class of failures, etc. is an intensional description. A symbolic object is defined both by a description “ $d$ ” (generally, intensional) and a way of comparing it to individual descriptions defined by a mapping “ $a$ ” called “membership function”. More formally:

##### Definition of a symbolic object

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  $\Omega$  in  $L$  depending on  $R$  and  $d$ .

Symbolic Data Analysis in SODAS concerns usually classes of symbolic objects where  $R$  is fixed, “ $d$ ” varies among a finite set of coherent descriptions and  $a(w) = [y(w)Rd]$ . 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  $[y(w)Rd] \in L = \{true, false\}$ . In this case, if  $y(w) = (y_1, \dots, y_p)$ , the  $y_i$  are of type (a) to (d), defined in section 1.

**Example:**

Let be  $a(w) = [y(w)Rd]$  with  $R : [d'Rd] = \vee_{i=1,2}[d'_i R_i d_i]$  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  $[y(w)Rd] \in L = [0, 1]$ .

**Example:**

Let be  $a(u) = [y(u)Rd]$  where for instance  $R : [d'Rd] = \max_{i=1,2}[d'_i R_i d_i]$  with  $\sum_{i=1,2} p_i = 1$  and where the “matching” of two probability distributions is defined for two discrete probability distributions  $d'_i = r$  and  $d_i = q$  of  $k$  values by:  $rR_i q = \sum_{j=1,k} r_j q_j e^{(r_j - \min(r_j, q_j))}$ . By analogy with the boolean case we denote  $[d'Rd] = \vee_{i=1,2}^*[d'_i R_i d_i]$  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. (see Diday (1995), for instance).

**Syntax of symbolic objects in the case of “assertions”:**

If the initial data table contains  $p$  variables we denote  $y(w) = (y_1(w), \dots, y_p(w))$ ,  $D = (D_1, \dots, D_p)$ ,  $d \in D: d = (d_1, \dots, d_p)$  and  $R' = (R_1, \dots, R_p)$  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  $[d'Rd] = \wedge_{i=1,p}[d'_i R_i d_i]$  where “ $\wedge$ ” has the standard logical meaning and “ $a$ ” is defined by:  $a(w) = [y(w)Rd]$  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 objects  $s = (a, R, d)$ . 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 multivalued and

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

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

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

**Other possible classes of symbolic objects:** if for instance the mapping “ $a$ ” is of the following kind:  $a(w) = [h_e(y(w))h_J(R)h_i(d)]$ , different classes of symbolic objects may be defined depending on the choice of  $h_e$ ,  $h_J$  and  $h_i$ . In practice, these mappings may be used for instance, in the following way:  $h_e$  is a filter of the extension of the symbolic object,  $h_J$  is a filter of the descriptive variables and  $h_i$  is a filter on the descriptions. More details may be found in Diday (1998) and in chapter 3 of Bock, Diday (2000). The following example illustrate a kind of filter.

**Example of filter on the extension:**

We associate to each town a symbolic object defined by  $a(w) = [h_e(y(w))Rd]$  where “ $d$ ” is the description of its inhabitant by using for instance, the histogram associated to each variable (as the histogram of the age). In order that the extension of such symbolic object contains only members of its associated town, the mapping  $h_e$  is defined in the following way:  $h_e(y(w)) = y(w)$  if  $w$  is member of the town and if not  $h_e(y(w)) = HS$  where  $HS$  is a dummy value such that  $[h_e(y(w))Rd] = 0$  for any description  $d$ .

**Order between symbolic objects:** if  $r$  is a given order on  $D$ , then the induced order on the set of symbolic objects denoted by  $r_s$  is defined by:  $s_1 r_s s_2$  iff  $d_1 r d_2$ . If  $R$  is such that  $[dRd'] = true$  implies  $d r d'$ , then  $Ext(s_1) \subseteq Ext(s_2)$  if  $s_1 r_s s_2$ . If  $R$  is such that  $[dRd'] = true$  implies  $d' r d$  then  $Ext(s_2) \subseteq Ext(s_1)$  if  $s_1 r_s s_2$ .

**Tools for symbolic objects:** Tools between symbolic objects (Diday (1995)) are needed such as similarities (F. de Carvalho (1998), Esposito et al (1998)), matching, merging by generalisation where a t-norm or a t-conorm (Schweizer, Sklar (1983) and Diday, Emilion (1995), (1997)) denoted  $T$  can be used, splitting by specialisation (Ciampi et al. (1995)). Under some assumption on the choice of  $R$  and  $T$  it can be shown that the underlying structure of a set of symbolic objects is a Galois lattice (Brito(1994), Polaillon, Diday (1997), Polaillon (1998)), where the vertices are closed sets defined by complete symbolic objects. More precisely, the associated Galois correspondence is defined by two mappings  $F$  and  $G$ :

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

–  $G$ : from  $S$  in  $P(\Omega)$  such that:  $G(s) = \text{Ext}(s)$ .

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 to a factorial axis, from a decision tree, etc. Finally we can summarize the mathematical framework of a symbolic data analysis in the following way (figure 1):

**Fig. 1.**  $\Omega$ : set of individuals.  $D$ : description set.  $L = \{\text{true, false}\}$  or  $L = [0, 1]$ .  $S$ : set of symbolic objects.  $y$ : description function.  $a$ : membership function from  $\Omega$  in  $L = \{\text{true, false}\}$  or  $L = [0, 1]$ .  $R$ : comparison relation.  $T$ : generalization mapping.  $F$ : intension mapping,  $G$ : extension mapping.  $d_w : y(w) = d_w$  is an individual description.  $w^s : w^s = F(w) = (a, R, y(w))$  is an individual symbolic object.  $d_C$ : description of class  $C$ .  $s$ : intensional symbolic object given by  $F(C) = (a, R, d_C)$  where  $a = [y(w)Rd_C]$ .  $G(s)$  is the extension of  $s$ .

– **Quality, robustness and reliability of a symbolic object modeling a concept known by its extent in a sample**

In figure 2 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 have a “concept”  $C$  (insurance companies, for instance) which extent is known in a sample  $\Omega$  of individuals (for instance, 300 insurance companies among a sample of 10000 companies). Each individual of the extent of  $C$

is described by using the mapping  $Y$  and then this set of individual descriptions is generalised with the operator  $T$  in order to produce the description  $d_C$ . The comparison relation  $R$  is chosen in relation with the  $T$  choice. The membership function is then defined by  $a_C(w) = [y(w)Rd_C]$  and then the symbolic object modelling the concept  $C$  is the triples  $= (a_C, R, d_C)$ . By definition of  $T$ , in the Boolean case the extent of  $s$  contains the extent of  $C$  but can contain also individuals which are not in the extent of  $C$ . In the modal case we have the same result if the threshold  $\alpha$  is well chosen . We consider the case where  $s$  is a modal symbolic object, for a given  $\alpha$   $Ext_\alpha(s) = \{w/a(w)\} \geq \alpha$  can contain two kinds of errors: i) individuals who satisfies the concept and are not in the extent of  $s$  and ii) individuals who don't satisfy the concept but are in the extent of  $s$ .

The “quality” and “robustness” of the symbolic object  $s$  can be defined in several ways. We suggest the following bootstrapping method:

Step1. Resample  $n$  times with replacement 50% of the initial sample  $\Omega'$ .

Step2. Calculate the symbolic object  $s$  modelling  $C$  by following the scheme of figure 2.

Step3. Calculate the extent of  $s$  in  $\Omega'$ .

Step4. Build the two histograms of the frequency of errors of kind i) and ii)

The quality and robustness of the symbolic object  $s$  is the higher when the mean and the mean square of the two histograms is the lowest. In other words, let  $X1$  (resp.  $X2$ ) be the random variable which associates to each resample the frequency of error of type i (resp. ii). Then, the lowest is the mean and the mean square of these two random variables, the higher is the quality and robustness of the symbolic object  $s$ . If the operator  $T$  is the Min  $t$ -norm or the Max  $t$ -conorm it is easy to show under a natural assumption that the smallest (resp. largest) symbolic object obtained by resampling converge in probability towards the symbolic object obtained by the Min  $t$ -norm (resp. Max  $t$ -conorm) on the initial sample. For sure the process can become more accurate if can obtain other samples  $\Omega'$  and other samples in  $\Omega'$  but it has already the following advantages:

- It reduces the importance of the outliers.
- It gives an idea on the errors distribution  $s$  without any model assumption.
- It allows the comparison of different choices of the operator  $T$ .

The “reliability” of the membership of an individual  $w$  to the extent of  $s$  is measured by the mean and mean square of the histogram of  $a_C(w)$ . If the  $i$ th resample gives the value  $a_i(w)$  then the reliability of  $s$  is defined by:  $W(s) = (\sum_{w \in Ext(C) \cap \Omega'} \sum_{i=1, n} a_i(w))/n \cdot |Ext(C) \cap \Omega'|$ .

The highest is  $W(s)$  (i.e. the closest from 1) the better is the reliability of  $s$ .

– **Quality, robustness , reliability and characteristic of symbolic objects modeling concepts induced by a classification**

In figure 3 the “real world” and the “modeled world” are the same but the steps are different. Instead of starting from a given concept, here we start from a given classification of the sample  $\Omega'$ . Then, from each class  $A$  of this classification we induce a concept  $C$  in the set of concepts and a set of description associated to each individual  $A$ . By applying an operator  $T$  we obtain a description  $d_C$  of the class  $A$ , in the set of description. The membership function is then defined

by  $a_C(w) = [y(w)Rd_C]$  and then the symbolic object modelling the concept  $C$  is the triple  $s = (a_C, R, d_C)$ .

The quality, robustness and reliability of each class  $A$  of the given classification can be obtained as in the preceding section if  $A$  play the role of  $\Omega'$ .

It is possible to build an operator  $T$  which produce instead a unique symbolic object, a set of symbolic objects which extent covers  $A$ , the extent of each one covering partially the class  $A$ . In that way, see for instance *M. Chavent (1997)* by a top-down clustering tree or *Brito and Diday (1991)* in a bottom-up clustering pyramid in the unsupervised case and *E. Perinel (1996)* or *Gettler-Summa (1995)* in the supervised case. As  $A$  is a cluster it is also interesting to describe it by a set of symbolic objects which satisfies simultaneously an unsupervised and a supervised criteria. For instance, in a top-down clustering tree where at each step a splitting variable is choosen, the criterion to optimise can express (in its unsupervised part), the sum of the two by two distances of the individuals of the class associated to a node and simultaneously (for its supervised part), the Gini impurity criterion of this class. We prune the terminal nodes which doesn't improve the criterion and in the final tree, we associate easily to each terminal node a symbolic object by the conjunction of of the values of the splitting variables used in the branches of the path which defines this node. The extent of each obtained symbolic object covers the individuals of each node and only those and together they cover  $A$ .

In order to see how much a cluster modeled by a set of symbolic objects modeling  $A$  is characteristic, an hypergeometric distribution can be used. Let  $N$  be the size of  $\Omega'$ ,  $n$  the size of  $A$ ,  $p$  the proportion in  $\Omega'$  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:

$$Pr(X = x) = \frac{C_x^{Np} C_{N-Np}^{n-x}}{C_N^n},$$

where  $C_N^n = N!/n!(N-n)!$  is the number of possible samples of size  $n$  in  $N$ ,  $C_{Np}^x = Np!/(Np-x)!x!$  is the number of groups of  $x$  individuals belonging in the extent of  $s$  in  $\Omega'$  and  $C_{N-Np}^{n-x} = (N-Np)!/(n-x)!(N-Np-n+x)!$  is the number of groups of  $(n-x)$  individuals which are not belonging in the extent of  $s$  in  $\Omega'$ . If the operator  $T$  produces  $k$  symbolic objects of extent with size  $x_1, \dots, x_k$  then the more  $Y = \sum_{i=1,k} Pr(X = x_i)/k$  is small, the more the class  $A$  is characteristic. This happen for instance, when  $p$  is small and  $x/n$  large. Notice that in the case where  $A$  is a complete symbolic object the size of the extent is  $n$  and  $p = n/N$ , so  $Pr(X = n) = C_n^n C_{N-n}^0 / C_N^n = 1.1/C_N^n = ((N-n)!n!)/N!$  which is the probability of a complete symbolic object of size  $n$  in a population of size  $N$ . When resampling, if the mean of the random variable  $Y$  is out of the chosen confident interval then the more its standard deviation is low the more the characterisation 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 (see for instance Bock, Diday (2000), chapter 8). The closest are considered to be the same. Instead of studying robustness, reliability and characteristic of symbolic objects, by using their extent, another way consists in using their description part. In that case, we need to extend the notion of “mean” and “standard deviation” to symbolic data, in order to use, for instance, a Fisher and a Student test. In that way, a first effort can be found in Bertrand, Goupil (2000) and Billard, Diday (2000).

**Fig. 2.** Modelisation by a symbolic object of a concept known by its extent

## 5 Some advantages in the use of symbolic objects

We can observe at least fifth kinds of advantages in the use of symbolic objects. First, 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 factorial axes). Second, they can be easily transformed in term of query of a Data base. Third, by being independent of the initial data table they are able to identify any matching individual described in any data table. Fourth, 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. Fifth ,

**Fig. 3.** Symbolic objects modélisation of concepts induced by a classification

in order to characterize a concept, they are able to join easily several properties based on different variables coming from different arrays and different underlying populations.

## 6 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 from 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 on the internal variation of the symbolic objects.

The following methods are developed in Bock, Diday (2000) and in the SO-DAS 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 sens 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 statistic (central object, histograms, dispersion, co-dispersion, etc. from a symbolic data table) to symbolic data.
- mining 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 to a complete symbolic object.
- dissimilarities between boolean or probabilistic symbolic objects,
- extension of decision trees on probabilistic symbolic objects, extension of a Parzen discrimination method to classes of symbolic objects,
- generalisation 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.

## 7 Symbolic Data Analysis in the SODAS software

### 7.1 The general aim

The general aim of SODAS can be stated in the following way: building symbolic data in order to summarise huge data sets and then, analyse them by Symbolic Data Analysis. For instance, if a set of households is characterized by its region, its socio-economic group, the number of bedrooms and of dining-living, we obtain a data table of the kind of table 4:

Household number	Region	Bedroom	Dining Living	Socio Econ
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	3	3

Table 4 : Standard Data Table of Households

In census data there is a huge set of households, 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 5:

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

Table 5: The first column of table 4 concerning the household number has been deleted.

We can now describe each town by the histogram of the categories of each variable. This is done in table 6 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, 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").

Region	Bedroom	Dining-Liv	Socio-Ec gr
Northern-Metro	(2/3) 2, (1/3) 3	(2/3) 2, (1/3) 3	(1/3) 1, (2/3) 3
East-anglia	(2/3) 1, (1/3) 2	(2/3) 2, (1/3) 3	(2/3) 1, (1/3) 2
Greater London			

Table 6 A symbolic data table where each cell contains a histogram

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, ...). Then define a context by giving: the units (individuals, households,...), 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 visualisation of symbolic descriptions,...).

## 7.2 Examples of applications strategy in SODAS:

We start from data provided by the three Statistical institute involved in SODAS (ONS (England), INE (Portugal), EUSTAT (Span)), as household consuming,

census, labour force survey or road transportation. Units are for instance, defined as "regions" or as "unemployment type" defined by each category of a new variable obtained by the cartesian product: "unemployment people categories x age categories x countries" given by a query to the relational data base. Then, DB2SO associates to each unit a symbolic description . Hence, we get a symbolic data table on which symbolic data analysis methods can be applied. In order to summarize and to get an overview on this symbolic data table, we can for instance, apply the following steps: we apply DIV (see chapter 11 in Bock , Diday (2000) ) which provides classes of units. It is then possible to apply again DB2SO on the same units but with the classes given by DIV. Therefore, each class represents a set of regions or a set of unemployment type. Hence, we obtain a new symbolic data table where each unit represents one of these classes. Several symbolic data analysis methods can then be applied: for instance, a principal component analysis (PCA, see chapter 9 in Bock , Diday (2000)) in order to get a graphical overview on these classes, a graphical visualisation of each class by "stars" (see chapter 7 in Bock , Diday (2000)), a description of each class by a disjunction of assertions (DSD, see section 9.4 in Bock , Diday (2000) ), etc.

### 7.3 SODAS software overview

In figure 4 an overview on the SODAS software is given. The input of DB2SO (see chapter 5 in Bock , Diday (2000)) is a query to a data base. Its output is a symbolic data table. Having obtained this data table any symbolic data analysis method can be applied.

### 7.4 SODAS future

The next steps in the future of SODAS will mainly consists first to extract symbolic objects from the clustering, factorial analysis, decision tree or discrimination (standard or symbolic) methods. Second, to induce from these symbolic objects, a new symbolic data table in order to study them, by a symbolic data analysis of higher level. Third, to select the "best" symbolic objects and prototypes, by using good criteria . Fourth, to propagate the obtained symbolic objects (the concepts that they represent). This propagation can be done towards the same Data Base for instance, at different times (in order to study the time evolution of the retained concepts) or towards other data bases associated to different countries. In any case, we have to compare sets of concepts and their associated symbolic objects obtained from different data bases. This may be done in several ways. For instance, by looking for a consensus tree or pyramid, between the concepts obtained in two different countries. Among many other ways, we can also calculate the extent of the symbolic objects obtained from a country in another country and then comparing the concepts associated to the symbolic objects of the first country to the concepts of the second country induced by the "complete symbolic objects" obtained from these extension. An overview on the next steps for the research and development of SODAS project are given in figure 5.

**Fig. 4.** A SODAS software overview

## 8 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. 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

**Fig. 5.** The future for the research and software development of the SODAS project

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. 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.

## 9 References

1. Adanson M. (1757) "Histoire Naturelle du Sénégal- Coquillages". Bauche Paris.
2. Aristotele (IV BC) "Organon" Vol. I Catégories, II De l'interprétation. J. Vrin edit. (Paris) (1994).
3. Arnault A., Nicole P. (1662), "La logique ou l'art de penser", Froman, Stuttgart (1965).
4. Auriol E. (1995) "Intégration d'approches symboliques pour le raisonnement à partir d'exemples" Thèse de doctorat, Université Paris 9 Dauphine.
5. Barbut M., Monjardet B. (1971, "Ordre et classification", T.2 Hachette, Paris.
6. Belson (1959), "Matching and prediction on the principle of biological classification", Applied Statistics, vol. VIII.
7. Benzecri J.P. et al. (1973) "L'Analyse de Données", Dunod, Paris.
8. Bertrand P. , Goupil F. (2000) "Descriptive statistics for Symbolic Data". In "Analysis of Symbolic Data". Bock, Diday edit. Study in Classification, Data Analysis and Knowledge Organisation. Springer Verlag.
9. Bertrand P. (1986) "Etude de la représentation pyramidale", Thèse de 3 cycle, Université Paris IX-Dauphine.
10. Billard L., Diday E. (2000) "Regression Analysis for Interval-Valued Data" . In Proc. Of IFCS-2000. "Data Analysis, Classification and related methods". Kiers and all editors. Springer Verlag.
11. Bock H.H. (1974) "Automatische Klassifikation". Vandenhoeck and Ruprecht, Gottingen.
12. Bock H.H., Diday E. (2000) "Analysis of Symbolic Data". Study in Classification, Data Analysis and Knowledge Organisation. Springer Verlag.
13. Breiman L., Friedman J.H., Olshen R.A., Stone C.S. (1984) "Classification and regression trees", Belmont, Wadsworth.
14. 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.
15. Brito P. (1994) "Order structure of symbolic assertion objects". IEEE TR. on Knowledge and Data Engineering Vol.6, n 5, October.
16. Bandemer H., Nather W. (1992) "Fuzzy Data Analysis". Kluwer Academic Publisher.
17. 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, n3, 1990,pp 35-51.
18. Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H. (1989), "Classification Automatique: environnement Statistique et Informatique". Dunod.
19. Changeux J.P. (1983) "L'homme neuronal". Fayard, Collection Pluriel.
20. Chavent M. (1997) "Analyse des Données symboliques. Une méthode divisive de classification". Thèse de doctorat, Université Paris 9 Dauphine.

21. 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).
22. 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.
23. De Carvalho F.A.T. (1998) a a "New metrics for constrained boolean symbolic objects" Proc. KESDA'98, Eurostat. Luxembourg.
24. De Carvalho F.A.T. (1998) b "Statistical proximity functions of boolean symbolic objects based on histograms" IFCS, Roma, Springer-Verlag.
25. Diday E.(1971) "La méthode des nuées dynamiques" ; Revue de Statist. Appliquée. Vol XIX, n 2, pp. 19-34.
26. Diday E.(1976) "Sélection typologique de variables". Rapport INRIA. Rocquencourt 78150, France.
27. Diday E.(1976) "Cluster analysis" in K.S. Fu (ed.). Digital Pattern Recognition. Springer Verlag. PP. 47-94.
28. Diday E. et al. (1979) "Optimisation en classification automatique". INRIA edit. Rocquencourt 78150, France.
29. Diday E., Govaert G., Lechevallier Y., Sidi J. (1980) "Clustering in Pattern Recognition". Proceed. NATO Adv. Study Institute on Digital Processing and Analysis, Bonas, J.C. Simon edit.
30. Diday E. (1984) "Une représentation visuelle des classes empiétantes". Rapport INRIA n 291. Rocquencourt 78150, France.
31. Diday E., Lemaire J., Pouget J., Testu F. (1984) "Eléments d'Analyse des données". Dunod , Paris.
32. Diday E. (1986) "Orders and overlapping clusters by pyramids". Proceed. Multidimensional Data Analysis. Edits. J.D. De Loeuw et al, DSWO Press, Leiden, The Netherlands.
33. Diday E. (1987 a) "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.
34. Diday E. (1987 b) "Introduction à l'approche symbolique en Analyse des Données ". Première Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987.
35. Diday E. (1989) "Introduction à l'approche symbolique en analyse des données". RAIRO (Revue, d'Automatique, d'informatique et de Recherche Opérationnelle), vol. 23, n2.
36. Diday E. (1995) " Probabilist, possibilist and belief objects for knowledge analysis " .Annals of Operations Research . 55, 227-276.
37. 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).
38. 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.

39. Diday E., Emilion R., Hillali Y. (1996) "Symbolic data analysis of probabilist objects by capacities and credibilities. XXXVIII Societa Italiana Di Statistica. Rimini, Italy.
40. Diday E.(1998) "L'Analyse des Données Symboliques: un cadre théorique et des outils" . Cahiers du CEREMADE.
41. Esposito F., Malerba D., Lisi F. (1998) "Flexible matching of boolean symbolic objects" Proc. NTTS'98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).
42. Ferraris, Gettler-Summa, C. Pardoux, H. Tong (1995) "Knowlege extraction using stochastic matrices: Application to elaborate a fishing strategy" Proc. Ordinal and Symbolic Data Analysis. Paris ;
43. Diday, Lechevallier, Opitz edit.Springer Studies in Classification.
44. Fisher D.H., Langley P. (1986) "Conceptual clustering and its relation to Numerical Taxonomy". Workshop on Artificial Intelligence and Statistics" Addison-Wesley, W. Gale édit.
45. Fisher D.H.,(1987) a, "Conceptual clustering learning from examples and inference". Proceed. 4th Workshop on Machine Learning. Irvine, California.
46. Ganascia J.G. (1991) "Charade: apprentissage de bases de connaissances". Cepadues, Kodratoff, Diday edit.
47. Gettler-Summa M. (1992) "Factorial axis interpretation by symbolic objects". Journées - Symbolique - Numérique. Université Paris IX- Dauphine. Lise-Ceremade.
48. Gettler-Summa M. (1997) "Symbolic marking: application on car accidents scenari" Proc. AMSDA, Capri, Italy.
49. Gigout E. (1998) " Graphical interpretation of symbolic objects resulting from data mining". Proc. KESDA'98, Eurostat. Luxembourg.
50. Gowda K.C., Diday E. (1992) "Symbolic clustering using a new similarity measure". IEEE Trans. Syst. Man and Cybernet. 22 (2), 368-378.
51. Gower J.C. (1974) "Maximal predictive classification". Biomet. Vol. 30, p. 643-644.
52. Hayes-Roth F., McDermott J. (1978) "An interference matching technique for inducing abstractions"Comm. ACM. Artificial Intelligence, Language processing.
53. Hebrail G. (1996) " SODAS (Symbolic Official Data Analysis System) ". Proceedings of IFCS'96, Kobe , Japan. Springer Verlag.
54. Jambu M. (1978) "Classification Automatique pour l'Analyse des Données". Dunod, Paris. J
55. ardine N., Sibson R. (1971) "Mathematical Taxonomy". John-Wiley and Sons. New-York.
56. Jussieu A.L. (1748) "Taxonomy. Coup d'oeil sur l'histoire et les principes des classifications botaniques". Dictionnaire d'Histoire Universelle.
57. Lance G.N. , Williams W.T. (1967) "A general theory of Classification sorting strategies: hierarchical systems". Comp. Jorn. Vol. 9 n4.
58. Langley P., Sage S. (1984) "Conceptual clustering as discrimination learning". Proceed. Fifth Biennial Conf. the Canadian Soc. for Comp. Studies of Intelligence. Labowitz M. (1983) " Generalization from natural language text" Cognit. Science 7, 1.

59. Lauro C., Palumbo F. (1998) "New approaches to Principal Component Analysis of Interval Data". Proc. NTTS'98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).
60. Lebart L., Morineau A., Piron M. (1995) "Statistique Exploratoire Multidimensionnelle" . Dunod, Paris.
61. Lebbe J. and Vignes R. (1991) "Génération de graphes d'identification à partir de descriptions de concepts", in *Induction Symbolique-Numérique*. Kodratoff, Diday edit. Cepadues (Toulouse).
62. Lerman I.C. (1970) "Les bases de la classification automatique" Gautier-Villars Paris.
63. 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).
64. Mfoumoune E. (1998) "Les aspects algorithmiques de la classification ascendante pyramidale et incrémentale" . Thèse de doctorat, Université Paris 9 Dauphine.
65. Michalski, R. (1973), Aqual/1 -Computer Implementation of a variable-valued logic system VL1 and examples in Pattern Recognition". Proc. Int. Joint Conf. on Pattern Recognition, Washington D.C., pp 3-17.
66. Michalski R., Step R.E. (1983) "Automated construction of classifications Conceptual Clustering versus Numerical Taxonomy", *IEEE Trans. on Pattern Analysis and Machine Intelligence*. Vol. 5, n4.
67. 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.
68. Morgan J.N., Sonquist J.A. (1963) "Problems in the analysis of survey data : a proposal". *J.A.S.A.* 58, p. 417-434.
69. Pankhurst R.J. (1978) "Biological identification. The principle and practice of identificatin methods in biology". London,
70. Edward Arnold. Payne R.W. (1975) "Genkey: a program for construction diagnostic keys". *Biological Identification with Computer* .Pankhurst edit. P. 65-72. Acad. Press. London
71. Périnel E. (1996) "Segmentation et Analyse de Données Symboliques: Application à des données Probabilistes Imprécises". Thèse de doctorat, Université Paris 9 Dauphine.
72. Pollaillon G., Diday E. (1997) " Galois lattices of symbolic objects " Rapport du Ceremade University Paris9- Dauphine (February).
73. 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.
74. Rasson J.P., Lissoir S. (1998) "Symbolic Kernel discriminante analysis" Proc. NTTS'98 Sorrento, Italy. Nanopoulos, Garonna, Lauro edit. Eurostat (Luxembourg).
75. Quinlan J.R. (1986) "Induction of decision trees". *Machine Learning* 1, pp 81-106. Kluwer Acad. Publishers, Boston.

76. Ralambondrainy H. (1991) "Apprentissage dans le contexte d'un schéma de base de Données" Kodratoff, Diday edit. CEPADUES, Toulouse.
77. Rosch E. (1978) "Principle of categorization" . E. Rosch , B. Lloyd edits. Cognition and Categorization , pp 27-48 . Hillsdale, N.J.: Erlbaum.
78. Roux M. (1985) "Algorithmes de classification", Masson..
79. Saporta G. (1990) "Probabilités, Analyse des Données et Statistiques". Edit. Technip Paris.
80. Schweizer B. (1985) "Distributions are the numbers of the futur" . Proc. sec. Napoli Meeting on "The mathematics of fuzzy systems". Instituto di Mathematica delle Faculta di Mathematica delle Faculta di Achitectura, Universita degli studi di Napoli. p. 137-149. Schweizer B. ,
81. Sklar A. (1983) " Probabilist metric spaces ". Elsever North-Holland, New-York. Sneath P.H.A.,
82. Sokal R.R. (1973) "Numerical Taxonomy" Freeman and Comp. Publishers. San Francisco.
83. Sowa J. (1984) Conceptual Structures: Information processing in mind and machine. Addison Wesley.
84. Stéphan (1998) "Construction d'objets symboliques par synthèse des résultats de requêtes SQL. Thèse de doctorat, Université Paris 9 Dauphine.
85. Tukey J. W. (1958) "Exploratory Data Analysis". Addison Wesley, Reading, Mass.
86. Vignes (1991) "Caractérisation automatique de groupes biologiques" . Thèse de doctorat, Université Paris 9 Dauphine.
87. Verde R., F.A.T. De Carvalho (1998) "Dependance rules influence on factorial representation of boolean symbolic objects". Proc. KESDA '98, Eurostat. Luxembourg.
88. Wagner H. (1973) "Begriff", Hanbuck Philosophischer Grundbegriffe, eds H. Krungs, H.M. Baumgartner and C. Wild, Kosel, Munchen ; PP. 191- 209.
89. Ward J.H. (1963) "3hierarchical groupings to optimize an objective function". J. Amer. Stat. Assoc. 58, pp. 236-244.
90. Wille R. (1982) "Restructuring lattice theory: an approach based on hierarchies of concepts." Proceed. Symp. Ordered Sets (I. Rival ed.), Reidel, Dordrecht-Boston.
91. Wille R. (1989) "Knowledge Acquisition by methods of formal concepts analysis, in Data Analysis, Learning symbolic and Numeric Knowledge. Diday edit. Nova Sciences Publishers.
92. Winston P. (1979) "Artificial Intelligence". Addison Wesley
93. . Ziani D. (1996 ) "Sélection de variables sur un ensemble d'objets symboliques" Thèse de doctorat, Université Paris 9 Dauphine.