Revista de Matemática: Teoría y Aplicaciones 2000 **7**(1-2) : 1–22
cimpa – ucr issn: 1409-2433

AN INTRODUCTION TO SYMBOLIC DATA ANALYSIS AND ITS APPLICATION TO THE SODAS PROJECT

Edwin Diday*

Received: 21 September 1999

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).

Keywords: Symbolic Data Analysis, SODAS software, Symbolic objects, Relational Data Bases, Boolean symbolic objects, Modal symbolic objects, Extent, Complete symbolic object, Robustness.

Resumen

Las descripciones de los datos de las unidades se llaman "simbólicas" cuando son más complejas que las estándar debido al hecho que contienen variación interna y están estructuradas. Los datos simbólicos aparecen a través de diversas fuentes, por ejemplo para resummir grandes Bases de Datos Relacionales por sus conceptos fundamentales. "Extracción del conocimiento" significa la obtención de resultados explicativos, por lo que se introducen los "objetos simbólicos" y se estudian en este artículo. Ellos modelan

*CEREMADE, Université de Paris IX – Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 15, France; E-Mail: diday@ceremade.dauphine.fr

conceptos y constituyen una salida explicativa para el análisis de datos. Es más, pueden ser usados para definir consultas a una Base de Datos Relacional y propagar conceptos entre Bases de Datos. Definimos el "Análisis de Datos Simbólico" (SDA) como una extensión del Análisis de Datos estándar a tablas de datos simbólicos como entrada, con el fin de encontrar objetos simbólicos como salida. En este artículo damos un panorama de desarrollos recientes en SDA. Presentamos herramientas y métodos de SDA, e introducimos el prototipo de software SODAS (resultado del trabajo conjunto de 17 equipos de nueve países que participan en un proyecto europeo de EUROSTAT).

Palabras clave: Análisis Simbólico de Datos, software SODAS, Objetos Simbólicos, Bases de Datos Relacionales, Objetos Simbólicos Booleanos, Objetos Simbólicos Modales, Extensión, Objeto Simbólico Completo, Robustez.

Mathematics Subject Classification: 62-07, 62H25, 68T35

Introduction

Knowledge extraction from large data bases is our main aim as in *Data Mining*. 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 sets of data. They 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. For instance, a cell can contain, a distribution (Schweitzer (1984) says that "distributions are the number of the future"!), or several values linked by a taxonomy, or intervals with 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. We define *Symbolic Data Analysis* (SDA) as the extension of standard Data Analysis to such tables. *Extracting knowledge* means getting explanatory results, that why, *symbolic objects* are introduced. They constitute an explanatory output of a SDA and moreover they can be used in order to define queries of a Data Base.

Now, we try to look for the historical and practical origin of the Symbolic Data Analysis field. The Aristotle Organon (IV BC) 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 interquartile 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* and by a way of finding its *extent*. For instance, the description of a 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 at the same time:

i) the intent of a class of individuals by 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, one of them, is there 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). 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: "*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, tree 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 superiors to them. Thus the idea of triangle in general extends to all different kinds of triangle*".

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 modelled 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 recognising that a given entity of the real world is a car. A concept of the real world can be modelled by a symbolic object and its extent. Whereas, *concepts* or *entities* of the real world are mathematically modelled by *symbolic objects*, their computing model provided by the so called *objects* used in the *object oriented language* and for instance, in computer languages as C++ or JAVA.

In the Aristotelian tradition, concepts are characterised 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 characterised by a set of similar individuals. In contrast , with the aritotelician tradition, were all the members of the extent of a concept are equivalent, a third tendency derived from psychology and cognitive science 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 Aristotelian tradition as they can have the explanatory power of a logical description.
- The Adansonian tradition as the member of the extent of a symbolic object are similar in the meaning that they must satisfy at the best the same properties.
- The Rosch point of view, as their membership function is able to provide prototypical instances characterised 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 (see for instance, Diday (1998)).

Symbolic Data Analysis is born from the simultaneous influence of several fields , from:

- standard exploratory data analysis (Tukey (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 $\mathbb{R}^p$ 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.

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 Aristotelician tradition and consists in defining top down the classes by a good choice of the properties which characterise them and from the most general to the most specific. In that way we obtain a decision tree where each node is characterised 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 Agglomerate 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 Sokal (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 generalise them in terms of disjunction of conjunction of properties, that why these classes are called *polytheistic* in opposition with classes generalised by a conjunction of properties and called *monotheistic*. Whereas, the first tendency yields monotheistic classes by a top- down process, the second produces polytheistic 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 generalisation of hierarchical trees, allowing overlapping clusters) of polytheistic classes is described. In Brito and Diday (1991), Brito (1994) an ascending pyramid produces monotheistic 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*.

Since the first papers announcing the main principles of Symbolic Data Analysis ((Diday (1987) a, (1987) b, (1989)) many work have been done in the same direction. In

factorial analysis, P. Cazes, A. Chouakria, E. Diday, Y. Schektman (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 dependence 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. Perinel has a chapter in this book on *symbolic discrimination rules*, M.C. Bravo, J.M. Garcia- Santesmases (1998) on *segmentation trees for stratified data* and J.P. Rasson and S. Lissoir(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 distiguish 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), 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 in chapter 5 and Gettler-Summa (1997) in section 9.4 of 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).

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). Stéphan (1998) for extracting symbolic objects from a data base (see chapter 5). 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). More generally, the most recent algorithms in Symbolic Data Analysis are in this book.

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 "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) Multivalued: 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) 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 age(w) is less than 2 months then height(w) is less than 10".

Many examples of such symbolic data are given in the chapter 3 of this book. Table 1 gives some examples of such data:

WAGES	TOWN	SOCIO-ECONOMIC GROUP
3.5	London	Personal of service
$[3, 8]$	{Paris, London }	
$\{3.1, 4.6, 7.2\}$		{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".

2 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 meaning 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.

Example: We have two relations of a Relational Data Base defined as follows. The first one called "delivery" is given in table 2. 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 2: Relation "Delivery".

The supplying are described by the relation "Supplying" defined in the following table 3.

Supplying	Supplier	Town
FT1	F1	Paris
FT2	F2	Toulouse
FT3	F1	Lannion
FT4	F3	Clamart
FT5	F3	Clamart

Table 3: Relation "Supplying".

From these two relations we can deduce the following data table 4, 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	FT4, FT5	Clamart

Table 4: 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 to the two 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 4.

ii) Multivalued with weights: this is the case for the towns of the supplier F1. The weights 1/2 mean that "the town of the supplier F1 is Paris or Lannion with a frequency equal to 1/2".

iii) Rules: some rules have to be given as input in addition to the data table 4. 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".

3 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 "generalisation" process which give also a way, by starting with this description, to find at least, the individuals of this class.

More formally, let Ω be a set of individuals, D a set containing descriptions of individuals or of class of individuals, y a mapping defined from Ω 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 $h_R : D \times D \to \{0, 1\}$ such that $h_R(x, y) = 1$ iff xRy. We generalise the mapping h_R by the mapping $H_R : D \times D \to L$ 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 Ω 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 $[y(w)Rd] \in L = \{\text{true,false}\}$. In this case, if $y(w) = (y_1, \ldots, y_p)$, 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 $y(w) = (\text{colour}(w), \text{height}(w))$, $d = (\{\text{red,blue,yellow}\}, [10, 15]) = (d_1, d_2)$,$\text{colour}(u) = \{\text{red,yellow}\}$,$\text{height}(u) = \{21\}$, then $a(u) = [\text{colour}(u) \subseteq \{\text{red,blue,yellow}\}] \vee [\text{height}(u) \subseteq [10, 15]] = \text{true} \vee \text{false} = \text{true}$.

- "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: $rR_iq = \sum_{j=1,k} r_j q_j e^{(r_j - \min(r_j, q_j))}$. 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 $d = (\{(0.2)12, (0.8)[20, 28]\}, \{(0.4)\text{employee}, (0.6)\text{worker}\})$ by:

  $$a(u) = [\text{age}(u) R_1 \{(0.2)12, (0.8)[20, 28]\}] \vee^* [SPC(u) R_2 \{(0.4)\text{employee}, (0.6)\text{worker}\}]$$

Syntax of symbolic objects in the case of "assertions"

If the initial data table contains p variables we denote $y(w) = (y_1(w), \ldots, y_p(w))$, $D = (D_1, \ldots, D_p)$, $d \in D : d = (d_1, \ldots, d_p)$ and $R' = (R_1, \ldots, R_p)$ 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 $[d\prime Rd] = \wedge_{i=1,p}[d\prime_i R_i d_i]$ 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 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:
$a(w) = [\text{age}(w) \subseteq \{12, 20, 28\}] \wedge [SPC(w) \subseteq \{\text{employee,worker}\}]$. If the individual u is described in the original symbolic data table by $age(u) = \{12, 20\}$ and $SPC(u) = \{\text{employee}\}$ then: $a(u) = [\{12, 20\} \subseteq \{12, 20, 28\}] \wedge [\{\text{employee}\} \subseteq \{\text{employee,worker}\}]$ =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 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 $[d'Rd] = \wedge *_{i=1,2}$ $p_i[d\prime_i R_i d_i]$ 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:

$$a(w) = [\text{age}(w)R1\{(0.2)12, (0.8)[20, 28]\}] \wedge^* [SPC(w)R2\{(0.4)\text{employee}, (0.6)\text{worker}\}]$$

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: $Extent(a) = \{w \in \Omega / a(w) = \text{true}\}$. In the modal case, given a threshold α, it is defined by $Ext_\alpha(s) = Extent_\alpha(a) = \{w \in \Omega / a(w) \geq \alpha\}$.

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_e 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 this book in chapter 3. 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 drd', then $Ext(s_1) \subseteq Ext(s_2) if s_1 r_s s_2$. If R is such that $[dRd']$ =true implies $d'rd$ then $Ext(s_2) \subseteq Ext(s_1)$ if $s_1 r_s s_2$.

3.1 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 Ω) 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 $T_{c\in C}y(c) = \cup_{c\in C}y(c)$, $R \equiv \ll \subseteq \gg$, $y(u) = \{\text{pink,blue}\}$, $C = \{c, c'\}$, $y(c) = \{\text{pink,red}\}$, $y(c) = \{\text{blue,red}\}$, then $a(u) = [y(w)RT_{c\in C}y(c)] = [\{pink, blue\} \subseteq \{pink, red\} \cup \{blue, red\}) = \{pink, red, blue\}] = true$ and $u \in Ext(s)$.
- G: from S in $P(\Omega)$ 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 to a factorial axis, from a decision tree, etc. Finally we can summarise the mathematical framework of a symbolic data analysis in the following way (figure 1):

4 Some advantages in the use of symbolic objects

We can observe at least five 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, in order to characterise a concept, they are able to join easily several properties based on different variables coming from different arrays and different underlying populations.

5 Some symbolic data analysis methods

Symbolic Data Analysis methods are mainly characterised 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

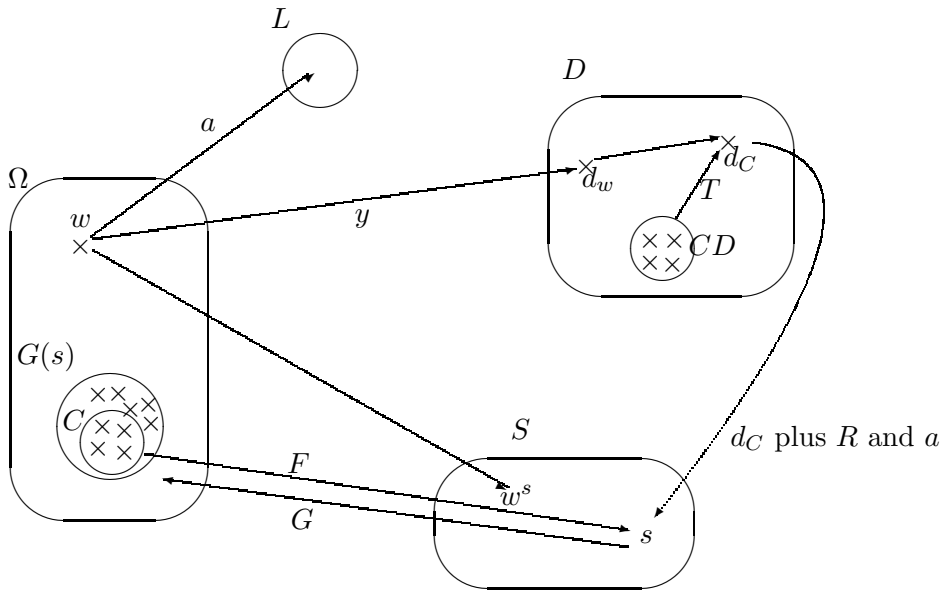

Figure 1: Ω: set of individuals. D: description set. $L = \{\text{true,false}\}$ or $L = [0,1]$. S: set of symbolic objects. y: description function. a: membership function from Ω in $L = \{\text{true,false}\}$ or $L = [0,1]$. R: comparison relation. T: generalisation mapping. F: intension mapping, G: 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 $F(C) = (a, R, d_C)$. $G(s)$ is the extension of s.

close from the one of the user and moreover have all the advantages mentioned in 5).

ii) They use efficient generalisation 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 this book 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 which means that the individuals are not represented in the factorial mapping 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.

6 Symbolic Data Analysis in the SODAS software

6.1 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 characterised 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 5.

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

Table 5: Standard Data Table of Households.

In census data there is a huge set of households, we can summarise 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 6.

We can now describe each town by the histogram of the categories of each variable. This is done in table 7 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 (each cell of the associated data table contains a frequency) and if the variable are symbolic (in this case each cell contains a histogram). If in the first case each branch of the decision

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			

Table 6: The first column of table 5 concerning the household number has been deleted.

tree represents an interval of frequency (for instance, the frequency of the category $[20, 30]$ years old), whereas in the second case it represents an interval of values (for instance the interval $[0, 30]$ years old).

Region	Bedroom	Dining-Liv	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	(2/3) 1, (1/3) 2
Greater London			

Table 7: 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, ACESS, ...). 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 7 or by a generalisation process applied to its members. 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,...).

6.2 Examples of applications strategy in SODAS

We start from data provided by the three Statistical institute involved in SODAS (ONS (England), INE (Portugal), EUSTAT (Spain)), 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 as "unemployment people categories x age categories x country" gives 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 summarise and to get an overview on this symbolic data table, we can for instance, apply the following

steps: we apply DIV (see chapter 11) 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 order to get a graphical overview on these classes, a graphical visualisation of each class by "stars" (see chapter 7), a description of each class by a disjunction of assertions (DSD, see section 9.4), etc.

6.3 SODAS software overview

In figure 2 an overview on the SODAS software is given. The input of DB2SO (see chapter 5) 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.

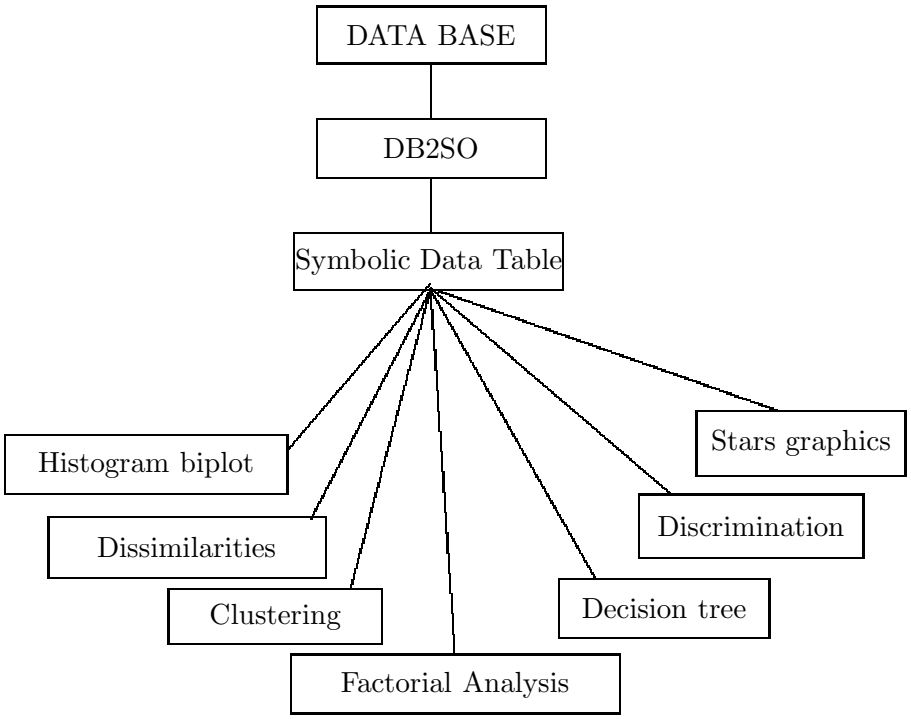

Figure 2: A SODAS software overview.

6.4 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 base, for instance, at different times (in order to study the time evolutions 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 3.

7 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 European Community project called SODAS for a " Symbolic Official Data Analysis System" in which 15 institutions of 9 European countries are concerned is to produce a first software of Symbolic Data Analysis. Three Official Statistical Institutions are involved in this project: EUSTAT (Spain), 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. Moreover, the next steps will be to improve the actual methods explained in this book (robustness, validity of the results etc.) and extend the symbolic data analysis methodology to regression, multidimensional scaling, neural networks, etc.

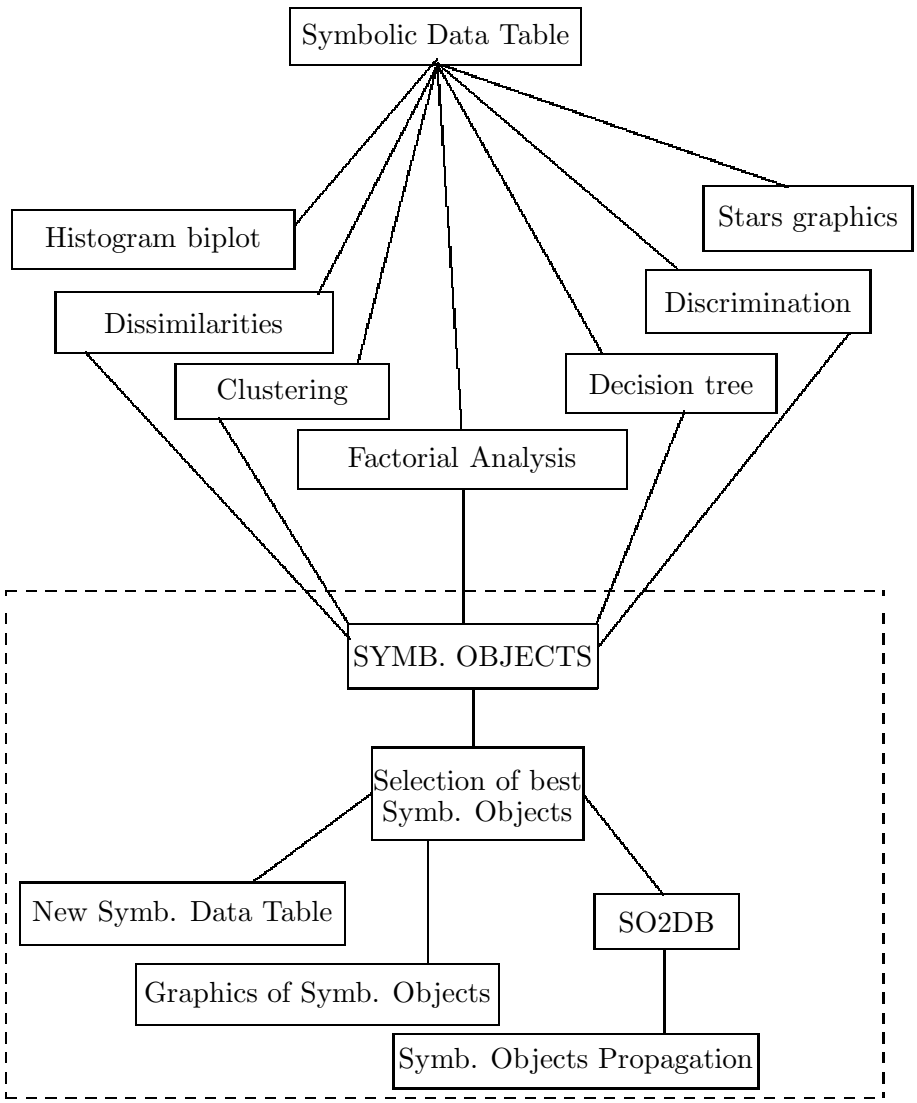

Figure 3: The future for the research and software development of the SODAS project.

References

Adanson, M. (1757) *Hipster Natural du Sénégal-Coquilles.* Gauche, Paris.

Aristotle. (IV BC) *Organon, Vol. I Catégories, Vol. II De l'Interprétation.* J. Vrin Edit., Paris (1994).

Arnault, A.; Nicole, P. (1662) *La Logique ou l'Art de Penser.* Froman, Stuttgart (1965).

Auriol, E. (1995) *Intégration d'Approches Symboliques pour le Raisonnement à Partir d'Exemples.* Thèse de Doctorat, Université Paris 9 Dauphine.

Barbut, M.; Monjardet,B. (1971) *Ordre et Classification*, T.2. Hachette, Paris.

Belson. (1959) "Matching and prediction on the principle of biological classification", *Applied Statistics* **VIII**.

Benzécri, J.P. et al. (1973) *L'Analyse de Données.* Dunod, Paris.

Bertrand, P. (1986) *Etude de la Représentation Pyramidale.* Thèse de 3-ème cycle, Université Paris IX-Dauphine, Paris.

Bock, H.H. (1974) *Automatische Klassifikation.* Vandenhoeck and Ruprecht, Göttingen.

Breiman, L.; Friedman, J.H.; Olsken, R.A.; Stone, C.S. (1984) *Classification and Regression Trees.* Belmont, Wadsworth.

Brito, P.; Diday, E. (1991) "Pyramidal representation of symbolic objects", *NATO ASI Series* **F 61**. M. Schader and W. Gaul (Eds.), *Knowledge Data and Computer-Assisted Decisions.* Springer-Verlag, Berlin.

Brito, P. (1994) "Order structure of symbolic assertion objects", *IEEE TR. on Knowledge and Data Engineering* **6**(5).

Bandemer, H..; Nather, W. (1992) *Fuzzy Data Analysis.* Kluwer Academic Publisher, Dordrecht.

Cazes, P.; Chouakria, A.; Diday, E.; Schektman, Y.(1997)) "Extension de l'analyse en composantes principales des données intervalles", *Revue de Statistiques Appliquée* **38**(3): 35–51.

Celeux, G.; Diday, E.; Govaert, G.; Lechevallier, Y.; Ralambondrainy, H. (1989) *Classification Automatique: Environnement Statistique et Informatique.* Dunod, Paris.

Changeux, J.P. (1983) *L'Homme Neuronal.* Collection Pluriel, Fayard, Paris.

Chavent, M. (1997) *Analyse des Données Symboliques. Une Méthode Divisive de Classification.* Thèse de doctorat, Université Paris 9 Dauphine, Paris.

Ciampi, A.; Diday, E.; Lebbe, J.; Périnel, E.; Vigne, R. (1995) "Recursive partition with probabilistically imprecise data", *OSDA'95,* E. Diday, Y. Lechevallier & O. Opitz (Eds.) Springer Verlag, Berlin.

Conruyt, N. (1994) *Amlioration de la Robustesse des Systèmes d'Aide à la Description, la Classification et la Détermination des Objets Biologiques.* Thése de doctorat, Université Paris 9 Dauphine, Paris.

De Carvalho, F.A.T. (1998) "New metrics for constrained boolean symbolic objects" *KESDA'98*, Eurostat, Luxembourg.

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

Diday, E. (1971) "La méthode des nuées dynamiques", *Revue de Statistique Appliquée* **19**(2): 19–34.

Diday, E.(1976) "Sélection typologique de variables", Rapport INRIA, Rocquencourt 78150, France.

Diday, E.(1976) "Cluster analysis", *Digital Pattern Recognition*, K.S. Fu (Ed.), Springer Verlag, Berlin: 47-94.

Diday, E. et al. (1979) *Optimisation en Classification Automatique.* INRIA, Rocquencourt.

Diday, E.; Govaert, G.; Lechevallier, Y.; Sidi, J. (1980) "Clustering in pattern recognition", *NATO Adv. Study Institute on Digital Processing and Analysis*, Bonas, J.C. Simon (ed.)

Diday, E. (1984) "Une représentation visuelle des classes empiétantes", Rapport INRIA n. 291, Rocquencourt.

Diday, E.; Lemaire, J.; Pouget, J.; Testu, F. (1984) *Eléments d'Analyse des Données.* Dunod, Paris.

Diday, E. (1986) "Orders and overlapping clusters by pyramids", *Multidimensional Data Analysis,* J.D. De Leeuw et al. (Eds.), DSWO Press, Leiden.

Diday E. (1987a) "The symbolic aproach in clustering and related methods of data analysis", *Classification and Related Methods of Data Analysis*, H. Bock (Ed.), North-Holland, Amsterdam.

Diday, E. (1987b) "Introduction à l'approche symbolique en analyse des donn'ées", *Première Journées Symbolique-Numérique,* Université Paris IX Dauphine, Paris.

Diday, E. (1989) "Introduction à l'approche symbolique en analyse des données", *RAIRO (Revue, d'Automatique, d'Informatique et de Recherche Opérationnelle)* **23**(2).

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", *OSDA'95 (Ordinal and Symbolic Data Analysis),* Springer Verlag, Berlin.

Diday, E.; Emilion, R. (1997) "Treillis de Galois maximaux et capacités de Choquet", *Compte Rendus à l'Académie des Sciences. Analyse Mathématique*, série 1, **t.324**.

Diday, E.; Emilion, R.; Hillali, Y. (1996) "Symbolic data analysis of probabilist objects by capacities and credibilities", *XXXVIII Società Italiana Di Statistica,* Rimini.

Diday, E.(1998) "L'analyse des données symboliques: un cadre théorique et des outils", *Cahiers du CEREMADE*, Université Paris IX Dauphine, Paris.

Esposito, F.; Malerba, D.; Lisi, F. (1998) "Flexible matching of boolean symbolic objects", *NTTS'98 Sorrento*, Nanopoulos, Garonna, Lauro (Eds.), Eurostat, Luxembourg.

Ferraris; Gettler-Summa, M.; Pardoux, C.; Tong, H. (1995) "Knowlege extraction using stochastic matrices: Application to elaborate a fishing strategy" , *Ordinal and Symbolic Data Analysis*, E. Diday, Y. Lechevallier, & O. Opitz (Eds.), Springer Studies in Classification, Paris.

Fisher, D.H.; Langley, P. (1986) "Conceptual clustering and its relation to numerical taxonomy", *Workshop on Artificial Intelligence and Statistics,* W. Gale (Ed.), Addisson-Wesley.

Fisher, D.H.(1987a) "Conceptual clustering learning from examples and inference", *4th Workshop on Machine Learning,* Irvine, California.

Ganascia, J.G. (1991) "Charade: apprentissage de bases de connaissances", *Cepadues*, Diday (Ed.), Kodratoff.

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

Gettler-Summa, M. (1997) "Symbolic marking: application on car accidents scenari", *ASMDA*, Capri.

Gigout, E. (1998) " Graphical interpretation of symbolic objects resulting from data mining", *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.

Gower, J.C. (1974) "Maximal predictive classification", *Biomet* **30**: 643–644.

Hayes-Roth, F.; McDermott, J. (1978) "An interference matching technique for inducing abstractions", *Comm. ACM. Artificial Intelligence, Language processing.*

Hebrail, G. (1996) "SODAS (Symbolic Official Data Analysis System)", *IFCS96*, Springer Verlag, Japan.

Jambu, M. (1978) *Classification Automatique pour l'Analyse des Données.* Dunod, Paris.

Jardine, N.; Sibson, R. (1971) *Mathematical Taxonomy.* John-Wiley and Sons, New-York.

Jussieu, A.L. (1748) *Taxonomy. Coup d'oeil sur l'histoire et les principes des classifications botaniques.* Dictionnaire d'Histoire Universelle.

Lance, G.N.; Williams, W.T. (1967) "A general theory of classification sorting strategies: hierarchical systems", *Comp. Jorn* **9** (4).

Langley, P.; Sage, S. (1984) "Conceptual clustering as discrimination learning", *Fifth Biennial Conf. the Canadian Soc. for Comp. Studies of Intelligence.*

Labowitz, M. (1983) "Generalization from natural language text",*Cognit. Science* **7**(1).

Lauro, C.; Palumbo, F. (1998) "New approaches to principal component analysis of interval data", *NTTS'98*, Nanopoulos, Garonna, Lauro (Eds.), Eurostat, Sorrento.

Lebart, L.; Morineau, A.; Piron, M. (1995) *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

Lebbe, J.; Vignes, R. (1991) "Génération de graphes d'identification partir de descriptions de concepts", *Induction Symbolique-Numérique*, Kodratoff, Diday (Eds.), Cepadues, Toulouse.

Lerman, I.C. (1970) *Les Bases de la Classification Automatique*. Gautier-Villars, Paris.

Noirhomme-Fraiture; Rouard, M. (1998) "Representation of sub-populations and correlation with zoom star", *NTTS'98*, Eurostat, Nanopoulos, Garonna, Lauro (Eds.), Sorrento.

Mfoumoune, E. (1998) *Les Aspects Algorithmiques de la Classification Ascendante Pyramidale et Incrémentale* . Thèse de Doctorat, Université Paris 9 Dauphine, Paris.

Michalski, R. (1973) " Aqual/1 -computer implementation of a variable-valued logic system VL1 and examples in pattern recognition", *Int. Joint Conf. on Pattern Recognition*, Washington D.C., 3-17.

Michalski, R.; Stepp, R.E. (1983) "Automated construction of classifications conceptual clustering versus numerical taxonomy", *IEEE Trans. on Pattern Analysis and Machine Intelligence* **5**(4).

Michalski, R.; Diday, E.; Stepp, 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, A. Rosenfeld (Eds.).

Morgan, J.N.; Sonquist, J.A. (1963) "Problems in the analysis of survey data: a proposal",*J.A.S.A.* **58**: 417–434.

Pankhurst, R.J. (1978) *Biological Identification. The principle and Practice of Identificatin Methods in Biology*. Edward Arnold, London.

Payne, R.W. (1975) "Genkey: a program for construction diagnostic keys", *Biological Identification with Computer*, Pankhurst (Ed.), Acad.Press, London.

Prinel (1996) *Segmentation et Analyse de Données Symboliques: Application des Données Probabilistes Imprécises*. Thèse de Doctorat, Université Paris 9 Dauphine, Paris.

Pollaillon, G.; Diday, E. (1997) "Galois lattices of symbolic objects", Rapport du Ceremade, Université Paris 9-Dauphine, Paris.

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, Paris.

Rasson, J.P.; Lissoir, S. (1998) "Symbolic kernel discriminante analysis" *NTTS'98*, Nanopoulos, Garonna, Lauro (Eds.), Eurostat, Sorrento.

Quinlan, J.R. (1986) "Induction of decision trees",*Machine Learning* **1**: 81–106.

Ralambondrainy, H. (1991) "Apprentissage dans le contexte d'un schéma de base de données", *Induction Symbolique-Numérique*, Kodratoff, Diday (Eds.), CEPADUES, Toulouse.

Rosch, E. (1978) "Principle of categorization", *Cognition and Categorization*, E. Rosch, B. Lloyd (Eds.), Erlbaum, Hillsdale: 27–48 .

Roux, M. (1985) *Algorithmes de Classification*. Masson, Paris.

Saporta, G. (1990) *Probabilités, Analyse des Données et Statistiques*. Technip, Paris.

Schweizer, B. (1985) "Distributions are the numbers of the futur", *Napoli Meeting on The Mathematics of Fuzzy Systems*, Instituto di Mathematica delle Facultà di Mathematica delle Facultà di Achitectura, Universita degli studi di Napoli, Napoli: 137–149.

Schweizer, B.; Sklar, A. (1983) *Probabilistic Metric Spaces*. Elsevier North-Holland, New-York.

Sneath, P.H.A.; Sokal, R.R. (1973) *Numerical Taxonomy*. Freeman and Comp. Publishers, San Francisco.

Sowa, J. (1984) *Conceptual Structures: Information Processing in Mind and Machine*. Addison Wesley, Reading

Stphan. (1998) *Construction d'Objets Symboliques par Synthèse des Résultats de requêtes SQL*. Thèse de doctorat, Université Paris 9 Dauphine, Paris.

Tukey, J. W. (1958) *Exploratory Data Analysis*. Addisson Wesley, Reading.

Vignes, R. (1991) *Caractérisation Automatique de Groupes Biologiques*, Thèse de doctorat, Université Paris 9 Dauphine, Paris.

Verde, R.; De Carvalho, F. A. T. (1998) "Dependance rules influence on factorial representation of boolean symbolic objects", *KESDA'98*, Eurostat, Luxembourg.

Wagner, H. (1973) "Begriff", *Hanbuck Philosophischer Grundbegriffe*, H. Krungs, H.M. Baumgartner , C. Wild (Eds.), Kosel, München: 191–209.

Ward, J.H. (1963) "Hierarchical groupings to optimize an objective function", *J. Amer. Stat. Assoc* **58**: 236–244.

Wille, R. (1982) "Restructuring lattice theory: an approach based on hierarchies of concepts", *Symp. Ordered Sets*, I. Rival (Ed.), Reidel, Dordrecht-Boston.

Wille, R. (1989) "Knowledge acquisition by methods of formal concepts analysis", *Data Analysis, Learning Symbolic and Numeric Knowledge*, Diday (Ed.), Nova Sciences Publishers.

Winston, P. (1979) *Artificial Intelligence*. Addison Wesley, Reading.

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