Chapter IX
Principles on Symbolic Data Analysis

Héctor Oscar Nigro
Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina

Sandra Elizabeth González Císaro
Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina

INTRODUCTION

Today's technology allows storing vast quantities of information from different sources in nature. This information has missing values, nulls, internal variation, taxonomies, and rules. We need a new type of data analysis that allows us represent the complexity of reality, maintaining the internal variation and structure (Diday, 2003).

In Data Analysis Process or Data Mining, it is necessary to know the nature of null values - the cases are by absence value, null value or default value-, being also possible and valid to have some imprecision, due to differential semantic in a concept, diverse sources, linguistic imprecision, element resumed in Database, human errors, etc (Chavent, 1997). So, we need a conceptual support to manipulate these types of situations. As we are going to see below, Symbolic Data Analysis (SDA) is a new issue based on a strong conceptual model called Symbolic Object (SO).

A "SO" is defined by its "intent" which contains a way to find its "extent". For instance, the description of habitants in a region and the way of allocating an individual to this region is called "intent", the set of individuals, which satisfies this intent, is called "extent" (Diday 2003). For this type of analysis, different experts are needed, each one giving their concepts.

Basically, Diday (Diday, 2002) distinguishes between two types of concept:

1. The *concepts of the real world*: That kind of concept is defined by an "intent" and an "extent" which exist, have existed or will exist in the real world.
2. The *concepts of our mind* (among the so called "mental objects" by J.P. Changeux (1983)) which frame in our mind concepts of our imagination or of the real world by their properties and a "way of finding their extent" (by using the senses), and not the

> extent itself as (undoubtedly!), there is no room in our mind for all the possible extents (Diday, 2003).

A "SO" models a concept, in the same way our mind does, by using a description "d" (representing its properties) 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 in the real world is a car. Hence, whereas a concept is defined by intent and extent, it is modeled by intent and a way of finding its extent is by "SOs" like those in our mind. It should be noticed that it is quite impossible to obtain all the characteristic properties of a concept and its complete extent. Therefore, a SO is just an approximation of a concept and the problems of quality, robustness and reliability of this approximation arise (Diday, 2003).

The topic is presented as follows: First, in the background section, the History and Fields of Influence and Sources of Symbolic Data. Second, in the focus section Formal definitions of SO and SDA, Semantics applied to the SO Concept and Principles of SDA. Third: Future Trends. Then Conclusions, References, Terms and Definitions.

BACKGROUND

Diday presented the first article on 1988, in the Proceedings of the First Conference of the International Federation of Classification Societies (IFCS) (Bock & Diday 2000). Then, much work has been done up to the publication of Bock, Diday (2000) and the Proceedings of IFCS'2000 (Bock & Diday 2000). Diday has directed an important quantity of PhD Thesis, with relevant theoretical aspects for SO. Some of the most representatives works are: Brito P. (1991), De Carvalho F. (1992), Auriol E. (1995), Périnel E. (1996), Stéphan V. (1996), Ziani D. (1996), Chavent M. (1997), Polaillon G. (1998), Hillali Y. (1998), Mfoummoune E. (1998), Vautrain F. (2000), Rodriguez Rojas O. (2000), De Reynies M. (2001), Vrac M. (2002), Mehdi M. (2003) and Pak K. (2003).

Now, we are going to explain the fundamentals that the SDA holds from their fields of influence and the most representative authors:

- ***Statistics:*** From Statistics the SO counts. It *knows* the distributions.
- ***Exploratory analysis***: The capacity of showing *new relations* between the descriptors {Tukey, Benzecri}(Bock & Diday 2000).
- ***Cognitive sciences and psychology***: The membership function of the SO is to provide *prototypical instances* characterized by the most representative attributes and individuals {Rosch} (Diday, 2003).
- ***Artificial intelligence***: The representation of *complex knowledge*, and the form of manipulation. This is more inspired from languages based on first order logic {Aristotle, Michalski, Auriol} (Diday, 2003).
- ***Biology:*** They use taxonomies and solutions widely investigated in this area {Adamson}(Bock & Diday 2000).
- ***Formal concept analysis***: The Complete Object Symbolic is a Galois Lattice {Wille R.} (Polaillon, 1998).

In some of these sciences, the problem is how to obtain classes and their descriptions (Bock & Diday, 2000)

The "*Symbolic data tables*" constitute the main input of a SDA, helped by the *Background Knowledge* (Diday, 2003). In the chapter 1 of Bock H and Diday E's book are mentioned as follow: each cell of this symbolic data table contains data of different types:

(a) *Single quantitative value*: if « height » is a variable and w is an individual: height (w) = 3.5.

(b) *Single categorical value*: Town (w) = Tandil.