Symbolic Data Analysis: A New Tool in Data Mining

N.Carlo Lauro
Department of Mathematics and Statistics, University Federico II of Naples
Monte Sant' Angelo, Via Cinthia
I-80126 Napoli, Italy
clauro@unina.it

Rosanna Verde
Dep. of Business Management and Quantitative Methodologies – Second University of Naples
Piazza Umberto I
I-81043 Capua, Italy
rosanna.verde@unina2.it

1. Introduction

In the present paper we intend to underline some advantages of a conceptual interpretation of the data extracted from a database and of their modelling through Symbolic Objects (SO). Many (SDA) techniques have been recently developed (mainly, inside two European research projects: SODAS, ISO3D) to visualize, summarize and analyse complex data (Bock & Diday, 2000). SO are characterized by multi-valued variables, as well as by logical relations and taxonomical structures defined on their descriptors. They can be obtained by expert descriptions of natural concepts (e.g. families, enterprises, species of animals, etc.) or by typologies on classical data (e.g. group of customers, group of products or services, behaviours, etc.), or by queries on databases. The last two cases put in evidence some links between SDA and Knowledge Discovery (KD) process based on Data Mining (DM). In fact both of them aim to identify meaningful patterns, understandable relations and rules by extracting information from huge databases. This suggests their integrated use to offer new and powerful tools not only from a descriptive point of view but also from a decisional/ confirmatory one. It must be noticed that Symbolic Data Warehouses allow keeping the knowledge domain for the logical dependency and hierarchical rules, usually lost in a classical Data Warehouse. At the meantime SDA methods seem particularly useful to solve some typical problems of pre-processing in DM.

2. Symbolic objects for modelling *conceptual* information

In the field of knowledge discovery a relevant role can be played by the Symbolic Data approach, as clearly pointed out in the introduction to the recent treatise edited by Bock & E. Diday (2000): *"In many domains of human activities it is now quite common to record huge sets of data in large databases. It becomes a task of first importance to summarize [and analyse] these data in terms of their underlying concepts in order to extract new knowledge from them".*

Symbolic objects are suitable modelling of concepts. They where introduced in order to overcome the inadequacy of the classical statistical data to represent the complexity of the real world better portrayed by concepts and expressed through suitable mapping of *human language and expert knowledge*. The definition of a SO is based on the dualism *intent-extent*. However, according to Diday (2000 - in Bock & Diday, 2000 pp.1-23), a SO is defined in intent, and, meantime, it contains the way to compute its extent that it is not inside in the conceptual description. Formally, a SO σ is a triple (a, R, d), where d is a description, R is a relation between descriptions, and a is a *mapping function* which allows to compute the extent of σ. The set d of description of σ is constituted by multi-valued variables (multi-categorical, intervals, modals), which are usually in relation by logical dependencies and taxonomies. In the framework of KD is interesting the way to compute the extent of σ through a function a. It is given by the set of elements w belonging to a set of observations Ω which respect the properties of σ by means of the relation R: $a(w)=R(y(w),d)\in L$; where $L=\{true, false\}$ or $L=[0,1]$ according to σ is a Boolean or modal SO. Thus, w takes part of the extent of σ if $a(w)=true$ or $a(w)>\alpha$, with α is a threshold equal to the minimum relative amount of the required properties of σ.

In knowledge discovery, a SO represent a first view on the mined data. In particular, the symbolic representation of the classes extracted from databases allows keeping the *knowledge domain* in a suitable Data Warehouse, as well as the conceptual description of the classes together with the structure of relational databases. According to the conceptual classification approach, the classes assume "*a more complex meaning than the simple union of their single element*". Thus, looking for their structure corresponds to the capability of representing data in the KD process.

Moreover, the strong relationship between the symbolic data description and the databases structure is even more evident when the information is extracted from object-oriented databases. In fact, the constraints of hereditariness are respected in the symbolic interpretation of the classes and in the process of *generalization* and *specialization* of the classes, which is usually performed in order to obtain SO of different order. The main advantages of introducing a symbolic structure in *Data Mining* (*Symbolic Mining*), can be shown in all the phases of the KD process: pre-processing, mining step, refining and post-processing. Many advantages are definitely in the pre-processing step. In fact, if the classes are described at a conceptual level, many problems, typical of the pre-treatment are automatically solved i.e.: (i) *missing values* by imputing the incomplete record to the most similar conceptual class; (ii) o*utliers* are captured in the conceptual data definition; (*iii) mixed variables* coexist in the SO description and the most part of the SDA techniques are able to deal with such descriptors; (iv) *data coding*, is part of the *symbolic-numerical-symbolic* approach to SDA;(v) *selection of the descriptors*, based on expert knowledge, as well as on the optimisation of an internal consistency criterion. Related to this phase is the problematic choice of the number of the SO to extract and the selection of the variables to consider in their descriptions. Briefly, we can say that the amount of conceptual data stored in a *Symbolic Data Warehouse* (SDW) should not be, *a-priori*, a small one in order to avoid an over generalization of concepts that can cause a loss of precious information, (e.g. rare cases or special patterns). A more effective refining of SDW can be later on obtained, in the mining step, by suitable SDA techniques (e.g. pyramidal clustering).

3. Symbolic Data Analysis methods in Data Mining

SDA methods play an important role in Data Mining, whereas the knowledge is expressed by means of an underlying conceptual meaning. They can be distinguished in relation to the different kind of knowledge that they are able to mine. A suitable visualization of conceptual data can be realized through three-dimensional reflecting techniques (i.e. Zoom-Stars; Noirhomme, Rouard, 2000 – in Bock & Diday, 2000 pp.125-138) as well as generalized factorial methods, which furnish an immediate perception of the conceptual data complexity also in terms of objects/concepts size and shape. The Zoom-star is a descriptive simple method. It allows visualizing concepts through stars by representing the variability of each descriptor on a different axis. This graphical tool furnishes an easy way to compare different conceptual data by overlapping the star-configurations. The factorial techniques extended to the SO (Lauro et *al*., 2000) allow visualizing them on a Cartesian plan. Furthermore, the factorial axes, in this context, are interpretable as high order SO (Summa, 1993). On factorial plans it is possible to visualise the internal variability of each symbolic object so to analyse their shape and size (Lauro & Palumbo, 2000). The measure of the geometrical proximity between objects offer a view of their similarity with respect to a suitable synthesis of the original symbolic descriptors.

Moreover, a re-fining of the discovered concepts can be achieved by clustering methods, in order to associate to each partition or to a nested symbolic class a proper conceptual meaning, or represent it by a suitable prototype. The clustering strategy used presents the advantage to considerably reduce the computational complexity of the partitioning algorithm by optimising a numerical criterion as in the classical clustering algorithms. The conceptual meaning of the classes is kept, while the optimisation process guarantees (Verde *et al*., 2000) a numerical convergence of the algorithm to the best representation of the classes in terms of minimum internal variability (*partition algorithm;* Chavent, 2000 – in Bock & Diday, 2000 pp.299-311) and minimum generalization (*hierarchical algorithm;* Brito, 2000 – in Bock & Diday, 2000 pp.312-322) of the concepts.

Some SDA techniques, as discrimination and classification ones, can support Data Mining for decisional or confirmatory scopes. The symbolic segmentation tree (Perinel, Lechevallier, 2000; Bravo et al. 2000 – in Bock & Diday, 2000 pp.244-265; 266-293) allows to define structure of concepts (symbolic decision trees) on the basis of series of rules defined on the SO descriptors. Each node of the decision tree corresponds to a symbolic description and the results of this technique are decisional paths based on logical sequence of concepts, more and more specialised. According to the symbolic approach the disadvantage of the classical segmentation techniques, which do not take into account logical (and not only numerical) links among the characters in the partition process, is overcome. Generally, the symbolic partition techniques, both supervised and unsupervised, allow keeping the coherence among the concepts ever more detailed.

In the first ones come out the confirmative aspect with respect to the a priori classes (or high order concepts), while in the second ones, the explorative way aims at defining and representing concepts, which require, in every case, to be verified at a logical-conceptual level.

Similarly, new symbolic approaches to neural networks and to Kohonen maps have been proposed with the goal of representing neurons as symbolic and the connections as relations among concepts at several specialization and generalization degrees. Concerning the discrimination techniques they present as main advantage to distinguish a priori concepts on the basis of the characteristics and rules, which define the objects belonging to them. Furthermore, Factorial Discriminant Analysis (Lauro et al. 2000 – in Bock & Diday, 2000 pp.212-233), extended to SO and classes allow to take the description of new elements in the conceptual one, on the basis of geometrical rules.

Data Mining has usually an explorative strategy of analysis in KD process and the advantage obtained by the symbolic approach can be seen as aimed at the same goal, but a confirmative role of the Data Mining could be promoted when SDA techniques aim at re-fining (or post re-fining) hypothesis on the bases of a specialization or generalization of the conceptual meaning of the extracted data.

REFERENCE

Bock, H.H., Diday, E. (2000) *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data.* Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.

Gettler-Summa, M. (1993) Factorial Axis Interpretation by Symbolic Objects. In: *Journées Symbolique - Numerique*. Diday, Kodratoff (Eds), Pinson. Paris.

Lauro, N.C, Palumbo, F. (2000), Principal Component Analysis of Interval Data: a Symbolic Data Analysis Approach. Computational Statistics Vol. 15 n.1 pp. 73-87.

Lauro, N.C., Verde, R. & Palumbo, F. (2000) Factorial Methods with Cohesion Constraints on Symbolic Objects. In H. Kiers, J.P. Rasson, P. Groenen & M. Shader (Eds), *Data analysis, classification and related methods*: Springer-Verlag, Heidelberg, pp. 381-386.

Verde, R., de Carvalho, F.A.T, Lechevallier, Y. (2000) A Dynamic clustering Algorithm for Multi-nominal Data. In: Kiers, H.A.L. *et al.* (Eds.) *Data Analysis, Classification, and Related Methods*, Springer-Verlag, 387-394.

RESUME

Dans ce papier nous entendons mettre en évidence les avantages d' une modélisation, par des Objets Symboliques, de données extraites d' une base, ainsi que l' utilisation de techniques pour l' Analyse des données symboliques (développée dans le cadre de deux projets européens: SODAS et ISO3D) à support du procès d' extraction de connaissance. Nous proposons une utilisation intégrée des techniques d' ADS dans ce procès afin d' identifier de patterns significatifs, relations et règles, non pas seulement avec un objectif descriptif mais aussi avec des objectifs confirmatifs et décisionnels.