

Analysis of Symbolic Data under Constrains⁽¹⁾

Analisi di Dati Simbolici in presenza di Vincoli

Rosanna Verde, Antonio Irpino

Dipartimento di Strategie Aziendali e metodologie quantitative,
Seconda Università degli Studi di Napoli,
e-mail: rosanna.verde@unina2.it , irpino@unina.it

Riassunto: Nel presente lavoro è trattato il problema delle regole e tassonomie nella descrizione di dati complessi e strutturati, espressi in forma di oggetti simbolici. La struttura dell'informazione contenuta in questo tipo di dati è fornita proprio da relazioni espressa da implicazioni di dipendenza logica tra i descrittori, così come da possibili gerarchie sulle modalità degli stessi descrittori. Il primo tipo di vincoli, incide sulla riduzione dello spazio di descrizione, in base ad un principio di coerenza; il secondo, invece, permette di tener in considerazione il diverso ruolo giocato dalle categorie di una variabile nella descrizione dei dati e consente, inoltre, una riduzione dello spazio di descrizione per fusione di modalità, in modalità a un livello gerarchico più elevato, secondo un prefissato o generato schema gerarchico.

Keywords: symbolic data, logical rules, hierarchical rules, taxonomy

1. Introduction

The present paper deals with coherency problems in symbolic data representation whenever some relationships are defined among the data descriptors.

As known, symbolic objects (SO's) are complex data described by multi-valued variables, called symbolic descriptors, which can be of different nature: intervals, multi-categorical, modal. Symbolic data (SD) correspond to the set of descriptions of the SO's. The main advantage to analyse SO's rather than classical statistical units consists in the capability of these complex units to interpret, in a more suitable way, the whole information describing typology, homogeneous groups of individuals, collections of units and so on. SO is defined (Bock and Diday, 2000) as a modelling of a concept, and then the homogeneous set that a SO describes can be considered as the extent of a concept.

SO's are not only characterized by the set of multi-values taken by their descriptors, but the most relevant information which is contained in their definition, is given by the relations among the descriptive components. Such relations can be expressed as logical dependency rules between symbolic descriptors and taxonomies on the categories of multi-categorical variables. The structure of the data is properly defined by these constrains; some authors have treated this topic from some different points of view (Csernel and De Carvalho, 1998; Verde, De Carvalho, 1999).

⁽¹⁾ The present paper is financially supported by COFIN 2001 grant "Metodi computazionali per la validazione di regole di decisione basati su descrizioni e rappresentazioni ottimali di dati complessi" (coord. Nazionale Prof.ssa R. Siciliano).

In the framework of the Symbolic Data Analysis, the presence of constraints on the symbolic description induces some effects in the representation of symbolic data and in their analysis (comparison, classification and SO's visualization in reduced subspaces).

2. Symbolic objects representation and logical rules

Classical SO definition is given in Bock and Diday (2000) as a triplet (a, R, d) , where d is the descriptor space expressed by the multi-values of the descriptors $y=(y_1, \dots, y_p) \in D$ (D is the set of all the admissible descriptions) and by the system of constraints defined on such descriptors; R is the set of relations of comparison between the descriptions of the SO and of an element w of a set Ω , described on the same set of descriptors y of the SO, and a is a mapping function allowing to compare the description of the element w to the description of the SO.

Logical relations between SO's descriptors can be expressed in terms of hierarchical (HD) and logical dependence (LD) rules:

- A *hierarchical dependence rule* (HD) restricts the domains of a descriptor y_i if another descriptor y_j assumes a sub-set of values S_j of its domain D_j :

iff $[y_j \in S_j]$ *then* $[y_i = NA]$.

It means that if y_j assumes values in $S_j \subseteq D_j$ then y_i becomes *Not Applicable* (NA), while y_i is inapplicable if and only if y_j assumes value in S_j . HD can be expressed as a *logical equivalence* $r_1: (p \leftrightarrow q) \leftrightarrow (\neg(p \wedge \neg q) \wedge \neg(q \wedge \neg p))$. According to the first De Morgan's law, it is equal to: $((\neg p \vee q) \wedge (\neg q \vee p)) \leftrightarrow ((p \rightarrow q) \wedge (q \rightarrow p))$, where p is the set of values S_j that the variables y_j assumes in the rules and q is the set of values that can be assumed by the variables y_i , while the symbols " $\neg \wedge \vee \rightarrow \leftrightarrow$ " correspond to the *five connectives*: "not", "and", "or", "implication", "equivalence".

- A *logical dependence rule* (LD) provides a restriction of the domain of a descriptor y_i to a sub-set $S_i \subseteq D_i$ of values if another descriptor assumes a sub-set of values $S_j \subseteq D_j$:

if $[y_j \in S_j]$ *then* $[y_i \in S_i]$.

This rule is based on a unidirectional logical implication, so that if y_j assumes some values in subset S_j then y_i takes values in subset $S_i \subseteq D_i$, but if y_i assumes values in S_i it does not mean that y_j should take values in S_j . LD can be expressed as a *logical implication* $r_1: (p \rightarrow q) \leftrightarrow (\neg p \vee q)$.

We deal with a redefinition of the symbolic objects in the presence of a sequence of inductions between their descriptors. SO's are geometrically represented by hypercubes, where each dimension corresponds to the values taken by a descriptor. Of course, the original description space of SO's is reduced whenever dependency rules are introduced. Consequently, the real dimension of the hypercubes associated to the SO's goes down. In fact, an *induction* allows to identify a subset of the space of the description that is not admissible $(p \wedge \neg q)$. This means that the presence of inductions can reduce the total variability of the set of objects.

Therefore, we aim at looking for a subset of disjoint symbolic sub-objects consistently with restrictions given by the rules. A possible solution was proposed by Csernel and De Carvalho (1998) using NSF *Normal Symbolic Form* algorithm when the dependence rules are structured in tree of dependence graphs. Alternatively, we can redefine the logical rules in terms of *implication* (LD) and *equivalence* (HD) relations between

variables, typical of first-order logic and consider their properties in order to search a coherent decomposition of the SO's, independently from the order of the multiple rules. Let $r_1, r_2, \dots, r_k$ be a sequence of *multiple* dependence rules between variables. For simplicity we consider a sequence of two rules r_1 : LD and r_2 : HD. Indicating with p, q, s and t the set of values assumed by the variables describing symbolic objects, the two logical rules are $r_1: (p \rightarrow q)$ and $r_2: (s \leftrightarrow t) \leftrightarrow ((s \rightarrow t) \wedge (t \rightarrow s))$. The multiple HD and LD rules, expressed respectively by equivalence and implication relations between the variables y_i, y_j and y_k , satisfy associative and distributive properties:

$$(r_1 \wedge r_2) \leftrightarrow (p \rightarrow q) \wedge ((s \rightarrow t) \wedge (t \rightarrow s)) \leftrightarrow \\ ((\neg p \vee q) \wedge \neg s \wedge \neg t) \vee ((\neg p \vee q) \wedge \neg s \wedge s) \vee ((\neg p \vee q) \wedge t \wedge \neg t) \vee ((\neg p \vee q) \wedge t \wedge s)$$

By removing the logical incoherencies: $(\neg s \wedge s) = \emptyset$, $(t \wedge \neg t) = \emptyset$ and $t(q \wedge t) = \emptyset$, the decomposition of the constrained SO's is realised as a disjunction of the different set of values assumed by the variables according to the rules. Each symbolic object results then decomposed into three sub-objects consistent with the following values assumed by the variables involved in the logical relations: $(\neg p \wedge \neg s \wedge \neg t) \vee (q \wedge \neg s \wedge \neg t) \vee (\neg p \wedge t \wedge s)$.

Finally, the problem of eventual overlapping in the sub-objects description space can be avoided by the application of a sequence of logical relations, that for a single LD rule:

$(p \rightarrow q) \leftrightarrow (\neg p \vee q)$ are given by the following equivalences:

$$((p \vee \neg p) \wedge (p \rightarrow q)) \leftrightarrow ((p \vee \neg p) \wedge (\neg p \vee q)) \leftrightarrow ((p \wedge q) \vee \neg p)$$

having introduced the *third excluded law* $(p \vee \neg p)$ (Verde, De Carvalho, 1999).

The last result: $\neg p \vee (p \wedge q)$ allows to decompose the original object into a subset of disjoint symbolic sub-objects consistently with restrictions given by the rules.

While the HD rules $(p \leftrightarrow q) \leftrightarrow ((p \rightarrow q) \wedge (q \rightarrow p))$ do not induce overlapping among symbolic sub-objects, considering reciprocity implication between the two variables.

3. Symbolic objects variability and taxonomies

Association rules on the categories of SO's descriptors have defined in Bock & Diday (2000) as *taxonomic* or tree-structured variables.

A taxonomic variable y_j assumes nominal values in the domain $D_j = \{a, b, \dots\}$ and a hierarchy $\mathcal{H}_j$: *i.* each category $b \in D_j$ is called *node* of the hierarchy tree $\mathcal{H}_j$; *ii.* a category c is a *descendant* of a (or, equivalently, a is an *ancestor* of c) if $c \prec a$; *iii.* the hierarchy $\mathcal{H}_j$ contains a *unique* root, that is an element $r \in D_j$ which has no ancestor at all; *vi.* a category $f \in D_j$ which has no successor is a leaf or terminal node of the tree, while the other nodes are internal nodes; *v.* partial tree below a internal node b is denoted $L(b)$; *vi.* a proximity $\delta(a, b)$ between two categories $a, b \in D_j$ (i.e. two nodes a, b of $\mathcal{H}_j$) is defined according to the hierarchical structure *which* connects the two nodes; such proximity may be provided by the *path distance* in tree descending $\mathcal{H}_j$.

Taxonomies are typically related to hierarchical relations between the categories of nominal and, and as their generalization, of multi-categorical variables. The presence of hierarchical structures in the SO's description, provides effects in analysis of symbolic data. Therefore, in order to take into account such relations among categories we suggest to relate a suitable coding of symbolic data to each level of the hierarchical structure. A sequence of levels on a hierarchy can be defined as follows: starting to the root node r (level 0), the first level is constituted by the direct descendant nodes (both terminals or internals) of r , e.g. $L_1(r)$, the second level is formed by the direct descendant nodes (both terminals or internals) of the nodes in $L_1(r)$, e.g. $L_2(a)$, $L_2(b)$, ..., with $a, b, \dots \in L_1(r)$; the h -th level contains as elements all the nodes descending from all the nodes f (terminals or internals) belonging to the level $h-1$; the last level contains all the terminal nodes of the tree structure.

The system of coding weights we here propose is based on a simple measure associated to the nodes at each level. Let be $a, b \in L_1(r)$, at the level 1, we associate weights proportional to the path distance: $\theta_a = \theta_b = 1/\#L_1(r) = 0.5$ (where $\#L_1(r)$ is the number of nodes at first level). The weights taken by the elements $x, y, \dots \in L_h(g)$, $s, t, \dots \in L_h(k)$, ... (with $g, k, \dots \in L_{h-1}(f)$, $\forall f$ at level $h-1$) at the level h are: $\theta_x = \theta_g \cdot 1/\#L_h(g)$ and so on. In such away, the system of coding results normalised to 1 at each level.

In SDA techniques, taxonomies on the SO's description induce a reduction of the categories in the description space. However, the proposed system of coding, based on *fuzzy* transformation, allows to keep the structure information on the SD's.

For sake of brevity, we can just make mention to the factorial techniques on multi-categorical variables (e.g. Generalized Canonical Analysis on Symbolic Objects) where going up in the hierarchical structure, the configuration of the categories on the factorial plane move through the barycentre of the nodes configuration at the below level. Similarly, in the dual representation space, the images of the hypercubes associated to the SO's result influenced by the different variable configuration and change their orientation according to the most weighted categories in the analysis.

References

- Bock H-H., Diday E., (2000), *Analysis of Symbolic Data*, Springer-Verlag, Heidelberg.
- Casari E. (1959) *Lineamenti di logica matematica*, Feltrinelli Ed., Milano.
- Csernel M., De Carvalho F.A.T., (1998) Towards a Normal Symbolic Form, *Data Science, Classification and Related Methods*, Springer, 379-386.
- De Carvalho F.A.T. (1998) Extension based proximities between constrained Boolean symbolic objects, *Data Science, Classification and Related Methods*, Eds: Hayshi C. *et al.*, Springer-Verlag, 386-393.
- Verde R., De Carvalho F.A.T. (1999) Dependence rule influence on factorial representation of Boolean Symbolic Objects, KESDA'98, published by Eurostat – Studies & Research, Luxemburg, pp. 287-300.