

Some proposals for symbolic descriptors selection

Alcune proposte per la selezione di descrittori di oggetti simbolici

Simona Balzano

Rosanna Verde

*Seconda Università di Napoli,
P.zza Umberto I – 81043, Capua (CE)*

Antonio Irpino

*Università degli studi di Napoli “Federico II”
via Cintia – Monte S. Angelo,
80126 Napoli*

Riassunto: In questo lavoro si propongono alcuni criteri per la selezione di descrittori oggetti simbolici, nel contesto di riferimento dell'estrazione di conoscenza da basi di dati. I criteri differiscono a seconda della natura dei descrittori: la strategia proposta procede, in una prima fase numerico-geometrica, a selezionare tra i descrittori ad intervalli quelli maggiormente ridondanti; mentre, in una seconda fase logico-numerica, calcola, sia per i descrittori multinominali che per quelli ad intervalli che hanno superato la prima selezione, una misura del potere discriminante. In base a tale misura sarà selezionato un insieme di descrittori tale da fornire delle descrizioni di oggetti quanto più possibile separati tra loro. Per la misura del potere discriminante si propongono due indici di dissimilarità per descrittori multinominali e ad intervalli.

Keywords: symbolic data, knowledge extraction, dissimilarity measures.

1. Introduction

In this paper we propose a methodology for the selection of symbolic descriptors, aiming at knowledge extraction from complex data structures, and its representation in terms of well discriminated symbolic objects. This matter has been mostly pursued in the context of the extraction of information from large databases. Methods for the extraction of symbolic data by RDBMS are usually based on queries (Stéphan, V., 1998). A generalization step is even performed on extracted data in order to arrange them into a symbolic data table. The quality of extracted symbolic data can be evaluated in terms of *completeness, simplicity, covering, discrimination, added information*.

In order to overcome the arbitrary choice of variables and, then, to define a more robust knowledge extraction method, we propose a procedure for variables selection. The definition of the query must be based on the known values that attributes assume in the relation(s). Without *a priori* knowledge, we need some other criteria for identifying attributes to be considered. The proposed methodology introduces numerical indices for selecting the most discriminant descriptors of symbolic objects, in a context mostly close to the conceptual clustering field. In fact, according to the definition of symbolic objects as modelling of concepts, they are described by means of the characteristics of conceptual clusters. The extension of the conceptual-class is retrieved and quantified, getting a numerical criterion for finding the most discriminant attributes. The measures of discriminant power of the hereafter introduced descriptors are different according to nature of the symbolic descriptors, (i.e. multi-nominal, interval or a modal variables).

2. A measure of the discriminant power of the symbolic multi-nominal variables

The proposed strategy is based on a ranking of descriptors with respect to a measure of their global discriminant power. Each attribute defines subclasses in each relation and contributes to their description with the sets of values taken by the elements belonging to the subclasses. In order to select the most discriminant descriptors (attributes) we perform a generalization by a dropping procedure.

The discriminant power of the several attributes (i.e. multi-nominal and interval variables) can be measured through the degree of generalization induced by the exclusion of each of them (one by one) from the description of the whole structure.

Let $Y_1, \dots, Y_p$ be the multi-nominal variables of a relation and n_j the number of categories of the attribute Y_j ($j=1, \dots, p$). The maximum specialized classes are described by the different ($C = n_1 \times n_2 \times \dots \times n_p$) combinations of the categories of the Y_j (for $j=1, \dots, p$). Such classes are even described in terms of symbolic assertions.

The discriminant power of the Y_j is evaluated by the decreasing of a suitable global measure of the classes *within variability*, due to the shift from C to C/n_j classes after the exclusion of Y_j by the description of the symbolic object assertions. That is equivalent to comparing the variability *between* the defined classes when one descriptor at a time is kept constant. In this way the real effect of the excluded descriptor is evaluated.

In particular, we consider a mutual variability index $v_i(a_h, a_k)$, based on a dissimilarity measure between assertions a_h and a_k . This measure is additive with respect to the p symbolic descriptors. The descriptors can be so sorted according to the decreasing dissimilarity measure induced by the descriptor dropping. Attributes with the highest *global* dissimilarity value can be considered to have the highest discriminant power.

Let $V_{i(h)}$ and $V_{i(k)}$ be the sets of categories or the intervals, respectively for the i^{th} multi-nominal or continuous attribute which describe the h^{th} and k^{th} objects. The dissimilarity between a_h and a_k can be expressed by a real value in $[0,1]$, computed by two different formulas for the multi-nominal and continuous case. In the multi-nominal case we have:

$$v_i^\circ(a_h, a_k) = \begin{cases} 1 & \text{se } V_{i(h)} \cap V_{i(k)} = \emptyset \\ 1 - \frac{\text{card}[V_{i(h)} \cap V_{i(k)}]}{\text{card}[V_{i(h)} \cup V_{i(k)}]} & \text{se } V_{i(h)} \cap V_{i(k)} \neq \emptyset \end{cases} \quad (1)$$

The more the values of attribute Y_j in the several classes are different, the more the attribute Y_j will be discriminant. Extreme values, 0 and 1, mean a null and maximum discriminant power respectively. Internal values in $[0,1]$ can be interpreted like higher or lower degree of dissimilarity between a_h and a_k , and higher or lower discriminant power, according to whether they are closer to 1 or 0 respectively.

In order to mitigate the effect of the different number of modalities of descriptors, $\text{card}[V_{i(h)} \cap V_{i(k)}]$ can be normalized to the dimension of the (finite) domain of Y_j , rather than to $\text{card}[V_{i(h)} \cup V_{i(k)}]$.

In presence of interval variables the index is computed as follows:

$$v_i^*(a_h, a_k) = \begin{cases} 1 & \text{se } V_{i(h)} \cap V_{i(k)} = \emptyset \\ 1 - \frac{\mu(V_{i(h)} \cap V_{i(k)})}{\mu(V_{i(h)} \cup V_{i(k)})} & \text{se } V_{i(h)} \cap V_{i(k)} \neq \emptyset \end{cases} \quad (2)$$

where $V_{i(h)} \cup V_{i(k)}$ and $V_{i(h)} \cap V_{i(k)}$ are, respectively, the union and the intersection of the intervals for a_h and a_k and m indicates its wideness (i.e. the description potential). Both these measures are closely similar to a dissimilarity index between pairs of observations, proposed by J.C. Gower in 1971, for both nominal and numerical variables.

3. A selection criterion of symbolic interval variables

The dissimilarity index under (2), when the symbolic object are described by intervals, is insensible to the different kinds of overlapping. In particular, the dissimilarity value $n_i^*(a_h, a_k)$ does not distinguish partial intersection of intervals from total inclusion cases. It is possible to demonstrate it can assume the same value in both cases.

We propose a new graphical-numerical criterion based on the comparison of the intervals in the descriptions of two symbolic objects (a_h and a_k) in order to evaluate the discriminant power of each descriptor. Such criterion allows to distinguish the two situations: the total inclusion of an interval in the other one, no-empty intersection.

Let $a_i = [Y_1 = [\underline{y}_{1i}; \overline{y}_{1i}]] \wedge [Y_2 = [\underline{y}_{2i}; \overline{y}_{2i}]] \wedge \dots \wedge [Y_p = [\underline{y}_{pi}; \overline{y}_{pi}]]$, ($i = 1, \dots, n$) be an assertion described by p interval variable, $\underline{y}_{ji}$ and $\overline{y}_{ji}$ representing Y_j interval lower and upper bounds respectively. For an object a_i , all intervals can be represented according to $a(\underline{y}_{ji}; \overline{y}_{ji})$ representation on a Cartesian plane, where intervals lower bounds are taken as coordinates on the x axis and upper bounds as coordinates on the y axis. With respect to each *interval-point*, the plane can be shared in *sections*, identifying all the possible relationships of intersection and/or inclusion relations between a_i and the other objects.

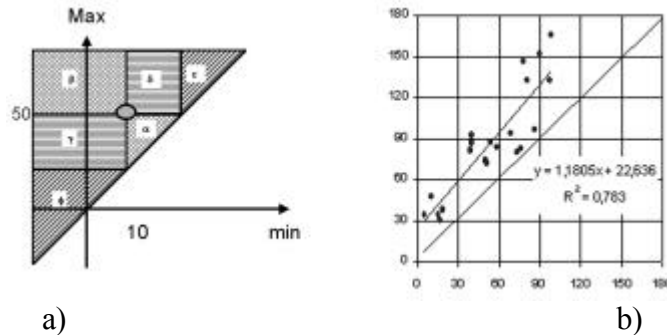


Figure 1: a) Interval [10, 50] representation; b) Intervals interpolation

In figure 1, a) α is the locus of all intervals including [10,50]; β is the locus of all intervals included in [10,50]; γ is the locus of all intervals partially overlapping [10,50] on the right; δ is the locus of all intervals partially overlapping [10,50] on the left; ϵ is the locus of all intervals with lower bound higher than 50; ϕ is the locus of all intervals with upper bound less than 10; line r is the locus of all intervals with null size.

This representation can be useful to detect redundant descriptors: for example, the proportion of intervals laying in γ and δ , or in ε and ϕ can give a measure of the discriminant power of the corresponding descriptor. Moreover, the descriptors selection is carried out through the p regressions between lower and upper bounds. In such a context, the regression coefficients b_j and the R^2_j represent numerical indexes of how much the descriptor characterizes the objects. The values of the b_j allow to get information about intervals wideness and location. In particular, for example, $b_j > 0$ means that the intervals tend to be separated; the more b_j is close to 1, the more intervals have the same size; $b_j < 0$ means that intervals tend to be overlapped (all included each other for $b_j = -1$). Of course, the meaningfulness of R^2_j is essential for the strength of the considerations above. A score can be calculated according to the sign of b_j and its deviation from 1.

4. Conclusions

It is worth noting that the proposed tools for variables selection, being based on a discriminant criteria, require no more arbitrary choices or a priori knowledge, and can be referred to an exploratory approach. In this sense, our strategy provides a robust tool for data mining and knowledge extraction.

The proposed procedure performs a selection for interval and multi-nominal descriptors. In prospective, the problem of the selection of modal variables in the symbolic objects description can be tackled by proposing a suitable measure of discrimination power based some homogeneity measures defined on the distributions of the classes. Such measure uses the concept of *capacity*, in the Choquet sense, and of *credibility*, in the Shafer sense (Irpino, 2000).

References

- Bock H-H., Diday E., (2001), *Analysis of Symbolic Data*, Springer-Verlag, Heidelberg
- Gower J.C. (1971), A general coefficient of similarity and some of its properties, *Biometrics*, **27**, p. 857-874
- Irpino A., Verde R., (2000), Consensus of Galois lattices of modal symbolic objects, Acts of Conference *OSDA 2000*, Namur, Belgio.
- Moore, R.E.(1966), *Interval Analysis*, Prentice-Hall, Inc.
- Stéphan V. (1998), *Construction d'objets symboliques par synthèse des résultats de requêtes SQL*, Doctorat Thesis, Université Paris IX, Dauphine