

Post-Analysis of Association Rules in a Symbolic Framework¹

Confronto tra Regole di Associazione in un Contesto Simbolico

D. Bruzzese, C. Davino

Dipartimento di Matematica e Statistica

Università Federico II, Via Cintia, I-80126 Napoli

email: dbruzzes/cdavino@unina.it

Riassunto: La necessità di definire una misura di distanza tra Regole di Associazione, al fine di cogliere l'evoluzione temporale del fenomeno indagato, è di primaria importanza quando si dispone di basi di dati riferiti alle stesse variabili ma osservate in momenti di tempo successivi. L'obiettivo è quello di separare la naturale variabilità di un fenomeno nel tempo, da variazioni sistematiche delle relazioni che lo descrivono, allo scopo di predisporre tempestivamente opportuni interventi. In questa lavoro viene introdotta una misura di distanza tra regole che si ispira alla teoria degli oggetti simbolici e che rappresenta uno strumento di analisi e di controllo per monitorare le trasformazioni avvenute nel tempo.

Keywords: Symbolic Data, Association Rules, Distance Measures.

1. Introduction

Association Rules (Agrawal et al., 1993) represent a consolidated tool to analyze relationships in social and economical frameworks. Since their introduction, they revealed some drawbacks: the number of discovered rules is often so huge that the analysis and interpretation is almost impossible (Bruzzese and Davino, 2001) and user-defined thresholds are imposed to reduce their number. At this regard, clustering techniques (Toivonen et al., 1995; Gago and Bentos 1998) are also used to obtain homogeneous groups of rules describing similar behaviours. These approaches are based on the computation of distances between the rules that refer only to the items belonging to both the rules. Moreover, in case of rules extraction in different time occasions, the proposed measures cannot be used if continuous variables are divided into a different number of categories in each time occasion.

The aim of this paper is to introduce a distance measure between rules that undertakes to overcome the main drawbacks of association rules and of the related distance measures. At this regard, an interpretation of the rules in a symbolic data analysis (Bock and Diday, 2000) framework is proposed so that both the antecedent and the consequent parts of the rules are treated as symbolic objects.

¹ This paper is financially supported by "Data Mining e Analisi Simbolica" COFIN2000 grant (Prof. C. Lauro).

2. Basic Notations

Let D be a set of databases in T different occasions (where $t = 1, \dots, T$) and let X be a set of observed variables (categorical, discrete, continuous) common to the T occasions. Each continuous variable is transformed in an interval discrete variable where the intervals are data-driven and depend from the distribution of the variable in each t . In this paper the term item will be referred to a class of a categorized variable or to a modality of a categorical variable or to a single value of a discrete variable.

From each database D_t , a set of rules R_t is mined; each rule is an implication of the form $P \rightarrow C$ where P is a set of n different items linked by the logical conjunction “ $\wedge$ ” and C is the consequent item.

Let $R_i^{t_1} : P_i^{t_1} \rightarrow C_i^{t_1}$ and $R_j^{t_2} : P_j^{t_2} \rightarrow C_j^{t_2}$ be two rules extracted into two different time occasions (t_1 and t_2) where $P_i^{t_1}$ and $P_j^{t_2}$ are the item sets in the antecedent part and $C_i^{t_1}$ and $C_j^{t_2}$ those in the consequence. The two rules can share one or more variables both in the antecedence and in the consequence while one or more variables can characterize only $R_i^{t_1}$ or $R_j^{t_2}$.

It is possible to define with $NP_i^{t_1}$ ($NP_j^{t_2}$) the set of variables present in $P_j^{t_2}$ ($P_i^{t_1}$) but not in $P_i^{t_1}$ ($P_j^{t_2}$); in this way the item sets $P_i^{t_1} \cup NP_i^{t_1}$ and $P_j^{t_2} \cup NP_j^{t_2}$ will refer to the same set of variables. In the symbolic objects notation, each variable in the set $P_i^{t_1} \cup NP_i^{t_1}$ ($P_j^{t_2} \cup NP_j^{t_2}$) together with the related frequency distribution, defines a probabilistic elementary event of the modal symbolic object s_i^p (s_j^p).

In the definition of a generic symbolic object s :

$$s = \bigwedge_{m=1}^M a_m = \bigwedge_{m=1}^M \left[Y_m = \{a_{mn}, p_{mn}\}_{n=1, \dots, N} \right] \quad (1)$$

we consider that each a_m is one of the variables in the set $P_i^{t_1} \cup NP_i^{t_1}$ (or in $P_j^{t_2} \cup NP_j^{t_2}$), a_{mn} is one of the N modalities or of the N classes in which a_m is categorized and p_{mn} is the corresponding support of a_{mn} in the considered rule.

The same notation can be used in case of the consequences, with the only exception that the modal symbolic object s_i^c (s_j^c) derives from the logical conjunction of the variables in the set $P_i^{t_1} \cup C_i^{t_1} \cup NC_i^{t_1}$ ($P_j^{t_2} \cup C_j^{t_2} \cup NC_j^{t_2}$) because in an association rule the consequence cannot leave out of consideration its premise.

3. A symbolic distance measure

In order to measure the distance between two rules, both the antecedent and the consequent parts are compared. The distance between the premises of two rules, $d(P_i^{t_1}, P_j^{t_2})$, is based on the comparison between the two modal symbolic objects s_i^p and s_j^p while the distance between the consequences of two rules, $d(C_i^{t_1}, C_j^{t_2})$, is based on the comparison between the two modal symbolic objects s_i^c and s_j^c .

Even if many distance measures among modal symbolic objects have been proposed in literature (Bock and Diday, 2000 – Bocci and Rizzi, 2000), in the framework of Association Rules it is important to provide a distance measure that allows both to take into account symbolic objects containing different categorizations of the same continuous variable and to obtain a maximum value when the two objects have no common characteristics.

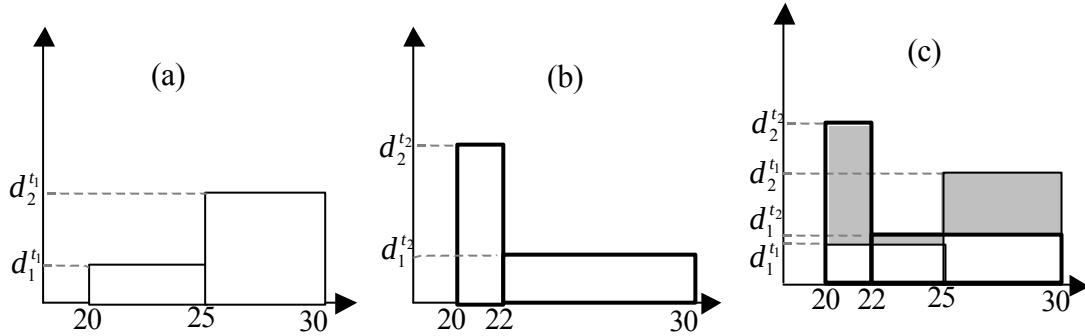
Generically, we define the distance between the premises of two rules $R_i^{t_1}$ and $R_j^{t_2}$ as:

$$d(P_i^{t_1}, P_j^{t_2}) = d(s_i^p, s_j^p) = \frac{\sum_{m=1}^{\# [P_i^{t_1} \cup NP_i^{t_1}]} d(a_m^{t_1}, a_m^{t_2})}{\# [P_i^{t_1} \cup NP_i^{t_1}]} \quad (2)$$

The m -th element $d(a_m^{t_1}, a_m^{t_2})$ is obtained comparing the distributions of the two elementary events belonging to the two symbolic objects s_i^p and s_j^p .

For example, let a_m be a continuous variable that has been differently categorized at time t_1 and t_2 : $a_m^{t_1} := \{[20-|25] 30\%; [25-|30] 70\%\}$ $a_m^{t_2} := \{[20-|22] 40\%; [22-|30] 60\%\}$. The previous distributions can be represented graphically through histograms in figure 1a and 1b, while in figure 1c the superimposition of the two histograms is shown and the grey area is proportional to the distance $d(a_m^{t_1}, a_m^{t_2})$.

Figure 1: A graphical representation of $d(a_m^{t_1}, a_m^{t_2})^2$



Let's denote with E the total number of distinct extremes of the categories of a_m both in t_1 and in t_2 (in the example E is equal to 4), with b_i ($i=1, \dots, E$) each of those distinct extremes (in the example $b_1=20$, $b_2=22$, $b_3=25$ and $b_4=30$) and with $d_{b_i}^{t_1}$ ($d_{b_i}^{t_2}$) the frequency densities associated to b_i . The distance $d(a_m^{t_1}, a_m^{t_2})$ can be expressed as:

$$d(a_m^{t_1}, a_m^{t_2}) = \sum_{i=1}^{E-1} \frac{(b_{i+1} - b_i) \cdot |d_{b_{i+1}}^{t_1} - d_{b_{i+1}}^{t_2}|}{(b_{i+1} - b_i) \cdot \max\{d_{b_{i+1}}^{t_1}; d_{b_{i+1}}^{t_2}\}} \bigg/ E - 1 \quad (3)$$

When a_m is not a categorized continuous variable, the differences $(b_{i+1} - b_i)$ ($\forall i=1, \dots, E$) are conventionally considered equal to 1 and the distance is based only on the comparison between the frequency distributions.

The distance between the premises is equal to 0 when the objects s_i^p and s_j^p describe the same subgroups and it is equal to 1 when the subgroups have no common characteristics.

Finally, the distance between the rules $R_i^{t_1}$ and $R_j^{t_2}$ takes into account the contribute coming both from the antecedent parts and from the consequent parts:

² On the vertical axes the frequency densities are represented.

$$d(R_i^{t_1}, R_j^{t_2}) = d(P_i^{t_1}, P_j^{t_2}) \cdot (1 - d(C_i^{t_1}, C_j^{t_2})) - d(C_i^{t_1}, C_j^{t_2}) \cdot (1 - d(P_i^{t_1}, P_j^{t_2})) \quad (4)$$

From (4) it results that $d(R_i^{t_1}, R_j^{t_2})$ is equal to $d(P_i^{t_1}, P_j^{t_2})$ when the two rules share the same consequence and it is equal to $d(C_i^{t_1}, C_j^{t_2})$ when the two rules share the same premise. In all the intermediate cases, the contribute provided by each partial distance is weighted by the complement to one of the dissimilarity of the other side of the association. The distance ranges from -1 to +1, where -1 refers to contradictory rules highlighting different behaviours of the same subgroups of individuals ($d(P_i^{t_1}, P_j^{t_2}) = 0$ and $d(C_i^{t_1}, C_j^{t_2}) = 1$) and +1 refers to rules describing similar behaviours of different groups of individuals ($d(P_i^{t_1}, P_j^{t_2}) = 1$ and $d(C_i^{t_1}, C_j^{t_2}) = 0$).

4. Concluding remarks

In this paper a distance measure between rules is proposed in the framework of symbolic data analysis. The proposed distance can be used in case of rules extracted in different time occasions but it can easily be extended to compare rules belonging to the same dataset. Further developments will refer to the possibility to take into account user a-priori knowledge and to match it with knowledge mined by the extracted rules. The proposed distance can also be used in an automatic clustering procedure in order to obtain groups of rules describing homogeneous behaviours across the time.

The flexibility of symbolic objects allows to treat different type of variables using an unified approach. Moreover the problem of different categorizations of the same variable in multiple time occasion is faced.

References

- Agrawal, R., Imielinski, T., Swami, A. (1993) Mining Association Rules between Sets of Items in Large Databases, *Proceedings of the 1993 ACM SIGMOD Conference*, Washington DC, USA, 207-216.
- Bocci, L., Rizzi A. (2000) Misure di prossimità nell'analisi dei dati simbolici, in *Atti della XL Riunione Scientifica della Società Italiana di Statistica, Sessioni Plenarie e Sessioni Specializzate*, Firenze, 26-28 aprile 2000, 91-102.
- Bock, H.H., Diday, E. (Eds) (2000) *Analysis of Symbolic Data*. Springer- Verlag.
- Bruzzese D., Davino C. (2001) Statistical Pruning of Discovered Association Rules, *Computational Statistics*, 16, 387-398.
- Gago, P., Bento, C. (1998) A metric for selection of the most promising rules, *Proceedings of the 1998 PKDD Conference*, Nantes, France, 19-27.
- Toivonen, H., Klemettinen, M., Ronkainen, P., Hatonen, K. and Mannila, H. (1995) Pruning and grouping of discovered association rules. *Notes of the Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, Heraklion, Greece, 47-52.