Post-Analysis of Association Rules  
in a Symbolic Framework

D. Bruzzese, C. Davino  
Dipartimento di Matematica e Statistica  
Università Federico II, Via Cintia, I-80126 Napoli  
email: dbruzzes/cdavino@unina.it

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 $D_t$ be a set of databases in $T$ different occasions (where $t=1, \ldots, T$) 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 “$\cap$” and $C$ is the consequent item.

Let $R^i_t: P^i_t \rightarrow C^i_t$ and $R^j_t: P^j_t \rightarrow C^j_t$ be two rules extracted into two different time occasions ($t_1$ and $t_2$) where $P^i_t$ and $P^j_t$ are the item sets in the antecedent part and $C^i_t$ and $C^j_t$ those in the consequent. 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 $R^i_t$ or $R^j_t$.

It is possible to define with $NP^i_t$ ($NP^j_t$) the set of variables present in $P^i_t$ ($P^j_t$) but not in $P^j_t$ ($P^j_t$); in this way the item sets $P^i_t \cup NP^j_t$ and $P^j_t \cup NP^i_t$ will refer to the same set of variables. In the symbolic objects notation, each variable in the set $P^i_t \cup NP^j_t$ ($P^j_t \cup NP^i_t$) together with the related frequency distribution, defines a probabilistic elementary event of the modal symbolic object $s^i_t$ ($s^j_t$).

In the definition of a generic symbolic object $s^i_t$:

$$s = \bigwedge_{m=1}^{M} a_m = \bigwedge_{m=1}^{M} \left[ Y_m = \left\{ a_{mn}, a_{mn} \right\}_{n=1,\ldots,N} \right]$$

(1)

we consider that each $a_m$ is one of the variables in the set $P^i_t \cup NP^j_t$ (or in $P^j_t \cup NP^i_t$), $a_{mn}$ 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 $s^i_t$ ($s^j_t$) derives from the logical conjunction of the variables in the set $P^i_t \cup C^i_t \cup NC^i_t$ ($P^j_t \cup C^j_t \cup NC^j_t$) 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, $d\left(P^i_t, P^j_t\right)$, is based on the comparison between the two modal symbolic objects $s^i_t$ and $s^j_t$ while the distance between the consequences of two rules, $d\left(C^i_t, C^j_t\right)$, is based on the comparison between the two modal symbolic objects $s^i_t$ and $s^j_t$.

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 $R^i_t$ and $R^j_t$ as:

$$d \left( P^i_t, P^j_t \right) = d \left( s^i_t, s^j_t \right) = \frac{\sum_{m=1}^{\#[P^i_t \cup NP^i_t]} d \left( a^i_m, a^j_m \right)}{\#[P^i_t \cup NP^i_t]}$$

(2)

The $m$-th element $d \left( a^i_m, a^j_m \right)$ is obtained comparing the distributions of the two elementary events belonging to the two symbolic objects $s^i_t$ and $s^j_t$.

For example, let $a_m$ be a continuous variable that has been differently categorized at time $t_1$ and $t_2$: $a^i_m := \{[20\,\text{–}\,25]\,30\%;\,[25\,\text{–}\,30]\,70\%\}$, $a^j_m := \{[20\,\text{–}\,22]\,40\%;\,[22\,\text{–}\,30]\,60\%\}$. 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 $d \left( a^i_m, a^j_m \right)$.

**Figure 1**: A graphical representation of $d \left( a^i_m, a^j_m \right)^2$

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 $b_i$ ($i=1,\ldots,E$) each of those distinct extremes (in the example $b_1=20$, $b_2=22$, $b_3=25$ and $b_4=30$) and with $d^h_i$ ($d^v_i$) the frequency densities associated to $b_i$. The distance $d \left( a^i_m, a^j_m \right)$ can be expressed as:

$$d \left( a^i_m, a^j_m \right) = \sum_{i=1}^{E-1} \frac{(b_{i+1}-b_i) \cdot d^h_{b_{i+1}} - d^h_{b_i}) \cdot \max \left( d^v_{b_{i+1}} ; d^v_{b_i} \right)}{E-1}$$

(3)

When $a_m$ is not a categorized continuous variable, the differences $(b_{i+1}-b_i)$ ($\forall \, i=1,\ldots,E$) 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 $s^i_t$ and $s^j_t$ describe the same subgroups and it is equal to 1 when the subgroups have no common characteristics. Finally, the distance between the rules $R^i_t$ and $R^j_t$ 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.
From (4) it results that \(d\left( R_i^a, R_j^b \right) = d\left( P_i^a, P_j^b \right) \cdot \left( 1 - d\left( C_i^a, C_j^b \right) \right) - d\left( C_i^a, C_j^b \right) \cdot \left( 1 - d\left( P_i^a, P_j^b \right) \right)\)

\[ (4) \]

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


