

Asymmetric Correspondence Analysis for the Selection of Association Rules

Analisi delle Corrispondenze su Matrici Asimmetriche per la Selezione di Regole di Associazione

Alfonso Iodice D'Enza
Dip. di Matematica e Statistica
Università Federico II - Napoli
iodicede@unina.it

Francesco Palumbo
Dip. di Istituzioni Economiche e Finanziarie
Università di Macerata
palumbo@unimc.it

Riassunto: Le Regole di Associazioni, così come sono state definite da Agrawal *et al.* (1993), rappresentano un potente strumento di Data Mining utile per l'individuazione di *pattern* e regolarità all'interno dei database. Tuttavia, i metodi di ricerca di tali regole tendono a riportare un numero eccessivo di regole, che rendono difficile la loro interpretazione. Per identificare solo la parte di informazione utile, *a posteriori* vengono utilizzati algoritmi di riduzione del numero di regole. Questo lavoro, invece, propone un approccio, basato su metodi propri dell'Analisi Multidimensionale dei Dati, volto alla identificazione *a priori* di regole interessanti.

Keywords: Association Rules, Correspondence Analysis.

1. Introduction

Since association rules were introduced by Agrawal *et al.* (1993), many algorithms were developed to mine rules and regularities in data stored in databases. In the Data Mining framework, *association rules* (AR's) Agrawal *et al.* (1993) represent a suitable tool to identify frequently occurring patterns of information in a database. AR's are mainly applied in *Market Basket Analysis*; in this context data are stored as transactions: all the transactions put down to the same customer represent a market basket. A rule, in its simplest form, involves a pair of items that represent, respectively, the antecedent part of the rule (*body*), and the consequent part of the rule (*head*): the information contributed by the rule is expressed in two parts: *support* and *confidence*. The support is the strength of the association between the considered items; the confidence measures the strength of the logical dependence expressed by the rule. The reader can easily understand that the number of potential rules grows very quickly according as the number of items increases, so AR's are mined by fixing thresholds for both support and confidence: there will only be selected the rules associated to combinations of events having support and confidence greater than the thresholds. Threshold choice is a ticklish question, since it is related to a trade off: if the user sets tight thresholds, the risk is to lose interesting information; otherwise, it leads to an excessive number of uninteresting rules. Hugeness of data stored in data bases determines such a mass of generated rules that makes hard, or even impossible, to analyze them, thus in literature, many proposals aim at grouping and pruning AR's to eliminate the redundant and trivial information *a posteriori*.

An example will illustrate the AR's in their simplest form: let a and b be a pair of goods, *beer* and *chips* respectively, the following notation represents the achieved AR:

$$a \implies b = \{\text{support} = 20\%, \text{confidence} = 80\%\}.$$

The rule shows that the 20 percent of customers buy both beer and chips, and if a customer buys beer, in 80 percent of cases buys chips too. Complex rules are able to analyze the regularities among sets of items and represent a more powerful tool, however, complex rules are even harder to handle.

The present paper proposes a strategy to identify *a priori* items that are antecedent or consequent parts of interesting rules; it aims at avoiding the generation of thousands of non-significant rules. The proposed strategy exploits the graphical and analytical capabilities of Multidimensional Data Analysis (MDA), in particular of Correspondence Analysis and Clustering.

Since the supplied definition of rule is related to a data bases terminology, in statistical terms we define AR's exploiting the following concepts: given a set of binary random variables $A_j : \Omega \rightarrow \{0, 1\}$ with $j = 1, \dots, p$, defined on the sample space Ω , a subset of the set of positive integers, and let these variables be associated, for example, to the purchase of a given item; furthermore, let us consider a power space $\mathcal{P}(\Omega)$, and some probability measure P on the measurable space $(\Omega, \mathcal{P}(\Omega))$ Hofmann H. (2001). We are interested in those cases $\omega \in \Omega$, in which is $A_j = 1$, event associated to the purchase of the item j . On the basis of the previous definitions we set $P(A_j) = P\{\omega \in \Omega : A_j(\omega) = 1\}$. We associate the event $(A_i = 1)$ to the item A and the event $(A_j = 1)$ to the item B with $i \neq j$ and $i, j = 1, \dots, p$. This probabilistic formalization, under the hypothesis $n \rightarrow \infty$, allows to present a generic rule as:

$$A \implies B = \{P(A \cap B), P(B|A)\},$$

where $P(A \cap B)$ and $P(B|A)$ represent the support and the confidence, respectively. It is worth noticing that the support is the joint probability that the events A and B occur at once, while confidence is the conditional probability of event B given A .

Let $\mathbf{Z}(n \times p)$ be a presence/absence matrix characterized by n transactions considered over p items and let $\mathbf{S} = n^{-1}\mathbf{Z}'\mathbf{Z}$ be a symmetric contingency table. The generic extra-diagonal term of $\mathbf{S}$, $s_{ij} = s_{ji} = P(A_i \cap A_j)$ (support), with $i \neq j$ and $i, j = 1, \dots, p$; the generic diagonal term $s_{ii} = P(A_i)$.

The asymmetric square matrix $\mathbf{C}$ is defined as $\mathbf{C} = \mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1}$, where $\mathbf{D}$ is a diagonal matrix such that, the generic term d_{ii} = 'number of purchases of the item i ', with $(i = 1, \dots, p)$. The generic term c_{ij} , with $i \neq j$ and $i, j = 1, \dots, p$ of the matrix $\mathbf{C}$ corresponds to $P(A_i | A_j)$, with $i \neq j$ and $i, j = 1, \dots, p$; it corresponds to the definition of confidence.

2. A Multidimensional Data Analysis approach for the identification of interesting rules

Taking into account the information in the matrices $\mathbf{S}$ and $\mathbf{C}$, this section describes a multi-step procedure to identify sets of items that are antecedent or consequent part of interesting rules. The proposal consists in partitioning the n baskets in k homogeneous classes: the analysis of the obtained groups of baskets leads to detecting the two sets of items that will be the antecedent and consequent part of the interesting rules. The proposed strategy consists of the following steps: *i*) a factorial analysis over $\mathbf{S}$, in order to obtain an orthogonal sub-space; *ii*) a clustering is applied to the configuration of points

representing the baskets on the supports subspace; *iii*) matrices S and C are defined for each group: the comparisons between the AR's computed inside the groups with respect to the global AR's allow to identify interesting AR's and the items, as a consequence.

Tables S and C are contingency tables, with the peculiarity that rows and columns are associated to the same set of items. Thus we apply a particular Correspondence Analysis (CA) over S and C and it is more correct to talk about a linear transformation similar to a CA. Let us start analyzing the supports matrix S : the factorial sub-space resulting from this kind of CA shows the relative position of the items according to the considered transactions. The following consists in projecting, over the obtained subspace, points representing the baskets. Over the resulting configuration of points/baskets it is finally applied a *classification* algorithm: the aim is to obtain groups of homogeneous baskets in order to find out regularities that the observation of the whole data set should not reveal. The set of *interesting* items is given by the comparison of the supports related to the single groups with the supports of the whole data set. The analysis of S matrix leads to the selection of interesting *antecedents* items. To select the interesting *consequents* items of the rules, we analyze the table C , with given foresight due to the characteristics of C : as the generic element represents a confidence, that is a conditional probability, this table is square and asymmetric. This aspect leads to adopt a CA transformation properly modified to take into account of the asymmetry of C . Different proposals in literature generalize CA over square asymmetric tables Bove (1989); Greenacre (2000), and generally these methods are based on the decomposition of the asymmetric table in two components: *symmetric* and *skew symmetric*. Applying this decomposition over table C , it results:

$$C = C_s + C_{sk}, \quad (1)$$

where C_s represents the *symmetric* component of C , $C_s = \frac{1}{2} (C + C')$, and C_{sk} represents the *skew symmetric* component of C , $C_{sk} = \frac{1}{2} (C - C')$.

The separate analyses of the two components allows to obtain a joined representation of the symmetric and skew-symmetric characteristics of table C . The items are represented through *circles* on a factorial plan: the position of the circle centers on the factorial plan represents the symmetric component of C ; the ranges of the circles represents the skew-symmetric component of C .

We are interested in those items characterized by a high skew-symmetric component; because the aim is to select interesting consequence items. Repeating the analysis on the confidence tables related to each group, we select the items/head of the rules that will be generated at the end of the procedure.

The interesting rules are finally generated among the two set of selected items, resulting from the analysis of S and C .

The proposed results can be viewed in probabilistic terms: let us consider the definition of k groups of baskets, $G_1, G_2, \dots, G_k$, as a partition of the sample space Ω . Let $(A_i \cap A_j)$ be the generic event such that $(A_i \cap A_j) \subset \Omega$, recalling the definition of conditional probability:

$$P(G_s | (A_i \cap A_j)) = \frac{P(G_s) P((A_i \cap A_j) | G_s)}{P(A_i \cap A_j)} \quad (2)$$

where $s = 1, \dots, k$ and $i \neq j, i, j = 1, \dots, p$. Since $P(G_s | (A_i \cap A_j))$ is the probability that a basket belongs to group s given the event $(A_i \cap A_j)$, we can find, for each group, the most characterizing items that determine the presence of a basket in that group.

Let us consider, as an example, a data set regarding the TV preferences of 330 spectators with respect to 45 emissions of Italian television. In this case each single transaction represents the preference of a spectator and each item is associated to a different TV emission. In this framework, a basket is the set of the TV-emissions watched. The items included in the data set regard the whole areas of television supply, from information programs to movies, as well as entertainment and kids programs. Following the presented strategy we obtain a subspace of the supports. By the comparison between the obtained classes and the whole set of baskets we select the set of interesting antecedent items (TV-emissions). The consequent items are selected according to the visualization of confidences of each single group. The items, represented through circles, are selected according to the size of the related circle; furthermore, if two circles are close, it means that the corresponding items are consequent parts that recall similar antecedent items.

Figure 1 consists of two scatter plots, labeled 'a)' and 'b)', showing the distribution of 40 television programs on two dimensions, factor 1 and factor 2. The programs are represented by black dots and labeled with their names. The left plot (a) shows programs like 'Sole a picco' and 'Sole a mezzogiorno' clustered together, while the right plot (b) shows programs like 'Sole a picco' and 'Sole a mezzogiorno' clustered together. The programs are labeled with their names and are represented by black dots.

The future developments of our studies are associated to the determination of an optimization criterion of the classes number determined by the procedure, as well as a sophistication of the logical distance measure among group supports and supports of the whole data set. An additive task is given by the comparison of the process performance with respect to methods based on pruning and grouping algorithms.

Agrawal R., Imielinski T. and Swami A.N. (1993) Mining association rules between sets of items in large databases, in: *ACM SIGMOD*, ACM Press, N.Y., volume 22, 207–216.

Bove G. (1989) *Nuovi metodi di rappresentazione di dati di prossimità*, Tesi di Dottorato, Univ. “La Sapienza”, Roma.

Greenacre M. (2000) Correspondence analysis of square asymmetric matrices, *Applied Statistics*, 49, 3, 297–310.

Hofmann H.W.A. (2001) Visual comparisons of association rules, *Comp.Stat.*, 16, 399–416