Statistica Applicata Vol. 12, n. 4, 2000

PRUNING, EXPLORING AND VISUALIZING ASSOCIATION RULES[1]

Dario Bruzzese, Cristina Davino

Dipartimento di Matematica e Statistica, Università di Napoli Federico II
Via Cintia Monte S. Angelo – I-80126 Napoli, Italia
bruzzese@dms.unina.it - cdavino@unina.it

Abstract

*Association Rules represent a valid and widespread tool in data mining framework. Nevertheless some important questions are still opened about their analysis and interpretation because of the huge number of rules that is usually mined. In this paper, an integrated strategy called PEV is proposed to face all the steps of the knowledge discovery process from the extraction of the most relevant rules (**P**runing) to the investigation of their common structure (**E**xploring) and to their graphical representation (**V**isualizing).*

1. INTRODUCTION

Association Rules are able to discover regularities in the data. Since they were first introduced (Agrawal *et al.*, 1993), they have been used in many fields of application, from credit scoring to business and marketing. Nowadays mining association rules in a database is a quite simple task but the analysis and the interpretation of the discovered rules are more difficult. Usually even if some constraints are defined by the user, a huge number of rules are discovered so that they cannot be analyzed and interpreted in order to find the useful ones.

A suitable exploitation of association rules requires facing the problem of reducing the number of rules *(Pruning)* and of summarizing *(Exploring)* and viewing the remaining rules *(Visualizing)*. In this paper we propose an approach named PEV (**P**runing-**E**xploring-**V**isualizing) that aims firstly to allow the user finding automatically a reduced subset of rules, secondly to capture the most

(1) *Tis research was supported by "Data Mining e Analisi Simbolica" PRIN2000 grant (Prof. C. Lauro).*

relevant structure inside the pruned set and, finally, to provide graphical tools able to visualize both single rules and higher order associations.

In literature, the main approaches facing the problem of the huge number of discovered association rules and of their relevance for the user are interestingness measures (Silberschatz and Tuzhilin, 1995) (Klemettinen *et al.*, 1994) and pruning methods (Toivonen *et al.*, 1995) (Weber, 1998). Both of them can be used either during the rule mining process or in the post-processing phase but, directly or indirectly, they require an active user involvement. In the PEV approach, the proposed pruning is an automatic procedure based on the use of statistical tests to evaluate the statistical significance of the discovered associations and to prune not significant rules (Bruzzese and Davino, 2001). Different solutions have been proposed to the problem of the visualization of the rules (Wong *et al.*, 1999) (Liu *et al.*, 1999) (Hofmann and Wilhelm, 2000) and most of them allows to have a global view on the set of discovered rules through their representation on 2-D or 3-D grids. In the PEV approach, the visualizing phase follows the exploring phase that exploits the advantages of Multidimesional Data Analysis that is able to synthetize the information stored in the rules and to represent in 2-D graphs the rules, the items and their interactions.

2. ASSOCIATION RULES

Let $I = i_1, i_2, \ldots, i_m$ be a set of items, called literals, (e.g. all products bought by a group of customers), and $T = t_1, t_2, \ldots, t_n$ be a set of n transactions, where each transaction t_i is a subset of I (e.g. all products in a customer's basket). An association rule R is an implication of the form: $IF[A]$ $THEN[C]$, where $A \subset I$ is the set of the antecedent items of the rule and $C \subset I$ is the set of the consequent items of the rule such that $A \cap C = \emptyset$. Each rule of the form $A \rightarrow C$ is characterized by two ratios:

- *Support:* $S_R = \frac{n_R}{n}$ where n_R is the number of transactions in T that contain $A \cup C$;
- *Confidence:* $C_R = \frac{n_R}{n_A}$ where n_A is the number of transactions in T that contain the antecedent items A.

The *Support* measures the proportion of transactions in T containing both A and C and it is not related to the possible dependence of C from A. On the other hand, the *Confidence* aims to measure the strength of the logical implication described by the rule.

3. THE PEV APPROACH

3.1 THE *PRUNING* PHASE

The PEV approach works on the *Rules Mart* defined as the set of association rules obtained by the mining process with very low support and confidence values. The huge number of discovered rules are sequentially pruned using three statistical tests performed on the significance of the consequence, of the antecedent and of the confidence. At each step, the rules are ranked according to the corresponding test statistic and a subset of rules is obtained pruning the rules out of a suitable threshold. This subset becomes the rules input set for the next step. While the first and the third step of the proposed pruning are applied rule by rule, the second step compares groups of rules with the same antecedent part and the pruning decision regards the whole set of rules. Moreover the third step can be used both to prune the rules and to establish a different ranking, alternative to the one related only to the confidence.

Step 1: a test on the significance of the consequence

It is a matter of fact that a very high confidence can be related to the presence both of frequent items in the consequent part of the rule (S_C is very high) and of rare items in the antecedent part (S_A is very low) thus causing the described problem of meaningless associations.

Given a rule R, in order to consider significative the implication between a generic antecedent A and a generic consequent C, it is necessary to compare the proportion of transactions with that consequence in $R\left(C_R = \frac{n_{A\cup C}}{n_A} = \frac{n_R}{n_A}\right)$ with the proportion of transactions with the same consequence in the set of all transaction $T\left(S_C = \frac{n_C}{n}\right)$.

The proposed test is constructed starting from the following hypothesis:

H_0: $C_R = S_C$ $\qquad$ H_1: $C_R > S_C$

Under the null hypothesis C_R is a binomial random variable that can be approximate with a normal distribution:

$$\frac{n_R}{n_A} \sim N\left(\frac{n_C}{n}; \sqrt{\frac{\frac{n_C}{n}\left(1-\frac{n_C}{n}\right)}{n_A}}\right)$$

if n_A is sufficiently large[2].

The test statistic V_{cons} thus follows a standardized normal distribution:

$$V_{cons} = \frac{C_R - S_C}{\sqrt{\frac{S_C(1-S_C)}{n_A}}} \sim N(0;\ 1)$$

Using a significance level $\alpha = 0.05$[3] a consequence is significative if $V_{Cons} > 1.64$.

Step 2: a test on the significance of the antecedent

The second step aims to evaluate the significance of the antecedent part and, differently from the previous one, it is applied to a whole set of rules (R_A) with the same antecedent part.

An association rule is significative if it highlights a favorite association between the antecedent part and the consequent part of the rule while it is not interesting if the items in the antecedent part can be equally associated to any other consequence even if its confidence is high.

If a set of consequences is randomly associated to a given antecedent, the corresponding rules can be considered as the categories of a uniform distribution.

A statistical test is proposed in order to compare the observed frequencies n_R with the theoretical uniform frequencies. Really, as the preferences of each user can be described by more than one rule, it is necessary to normalize the observed frequencies so that they sum up to the number of transactions that share the items in the antecedent (n_A). We denote the normalized frequencies as n_R^*.

A classical Chi-2 statistic is then computed:

$$\chi^2_{oss} = \sum_{i=1}^{R_A} \frac{\left(n^*_{R(i)} - \hat{n}_{R(i)}\right)^2}{\hat{n}_{R(i)}}$$

with $R_A - 1$ degrees of freedom.

The observed value of the test statistic $\left(\chi^2_{oss}\right)$ must be compared with the theoretical value of the χ^2 statistic in case of a chosen significance level α and $n_A - 1$ degrees of freedom. If $\chi^2_{teo} > \chi^2_{oss}$ the logical implications with the antecedent part are not significative and they can be pruned.

---

[2] *This condition commonly happens in the data mining framework.*

[3] *An α value equal to 0.05 is a typical choice in statistical hypothesis tests.*

Step 3: a test on the significance of the confidence

If the previous steps allow to evaluate the logical meaning of the rules, the third and final step of the proposed strategy aims to evaluate the strength of the implications. At this regard we compare the support of the rule (S_R) with the proportion of transactions that shares the items in the antecedent part (S_A) and we look for rules with a support not significantly less than the support of the antecedent. This comparison allows to find rules with a confidence not significantly far from the hard implication ($C_R = 1$).

The proposed test is based on the following hypothesis:

$$H_0: S_R = S_A \qquad H_1: S_R < S_A$$

Under H_0, S_R is a binomial random variable that can be approximate with a normal distribution:

$$\frac{n_R}{n} \sim N\left(\frac{n_A}{n}; \sqrt{\frac{\frac{n_A}{n}\left(1-\frac{n_A}{n}\right)}{n}}\right)$$

if n is sufficiently large.

It follows that the test statistic V_{Conf} is a standardized normal random variable:

$$V_{Conf} = \frac{S_R - S_A}{\sqrt{\frac{S_A(1-S_A)}{n}}} \sim N(0;1)$$

and using a significance level $\alpha = 0.05$, the acceptance region represents our region of interest and it is limited by the values greater or equal than -1.64. If the test statistic values are used to establish a ranking of the rules instead of pruning them, it can happen that the resulting ranking reverts the ranking based on the confidence values.

3.2 THE *EXPLORING* PHASE

The *Exploring* phase is performed by a factorial method that allows to capture the structural information inside the data and to visualize it on 2-dimensional graphs. The rules stored in the *Rules Mart* after the pruning process must be recoded in order to be described by a *cases* × *variables* matrix **E** that represents the input data of the factorial method. The number of rows of **E** is equal to the number n of

rules survived to the pruning step and the number of columns $p = p_{if} + p_{then}$ corresponds to the total number of different items, both in the antecedent parts and in the consequent parts of the n rules. Each rule is coded by a binary array assuming value 1 if the corresponding column item is present in the rule and value 0 otherwise. In order to take into account also the well known confidence and support measures, the **E** matrix is partitioned in three column groups: $\mathbf{E}_{if}$ refers to the antecedent items, $\mathbf{E}_{then}$ refers to the consequent items and, finally, $\mathbf{E}_{meas}$ refers to the confidences and supports (Fig. 1). The final **E** matrix has thus $n \times (p_{if} + p_{then} + 2)$ dimensions and it can be analysed through the Multiple Correspondence Analysis (MCA) (Benzècri, 1973) that allows to represent the relationships among the observed variables, the similarities and differences among the rules and the interactions between them.

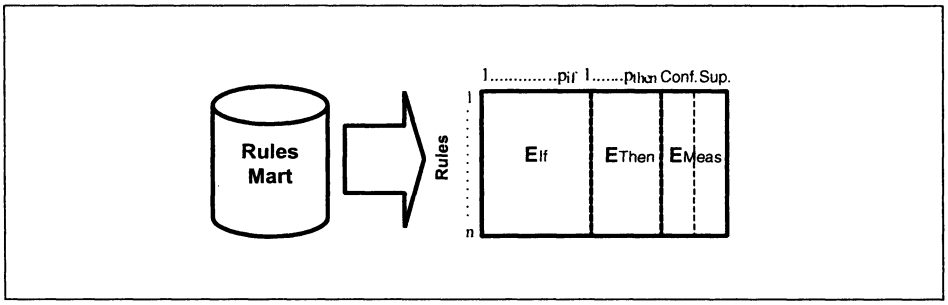


Fig. 1: The coding of the *Rules Mart*.

MCA allows to reduce the number of original variables finding linear combinations of them, the so called factors, that minimize the loss of information deriving. A typical feature of MCA and of all factorial methods is the possibility to assign different roles to different sets of variables: the variables that intervene directly in the analysis and define the factorial planes, are called *active* while the variables depending by the former and projected later on the defined factorial planes are called *supplementary*. In the *Exploring* phase the groups of the **E** matrix play different roles: the antecedent items are the *active* variables while the consequent items and the two measures, support and confidence, are the *supplementary* variables. This choice is related to the logical link that exists between the two components of a rule where the antecedent part is the premise, the cause of the consequent part.

Once the MCA is performed it is possible to represent the rows and the columns of the *Rules Mart* on reduced dimension subspaces. The results of the *Exploring* phase can thus provide different starting points for the *Visualizing* phase: it is possible to consider either the factorial planes allowing to explain at least a user defined threshold of the total variability either a user defined factorial plane or the factorial plane best defined by a user chosen item.

3.3 THE *VISUALIZING* PHASE

The *Visualizing* phase exploits the interpretative power of MCA in order to provide the user different *views* on the set of rules.

1. **Items Visualization.**
   Both the antecedent and the consequent items are represented on the factorial plane by points with a dimension proportional to their supports. The support and the confidence measures are represented by oriented segments linking the origin of the axes to their projection on the plane as their coordinates are the correlation coefficients with the axes. The previous expedient allows to identify privileged regions in the plane with high supports and confidences. The proximity between two antecedent items shows the presence of a set of rules sharing them while the proximity between two consequent items is related to a common causal structure. Finally, the closeness between antecedent items and consequent items highlights the presence of a set of rules with a common dependence structure.

2. **Rules Visualization.**
   The association rules are represented on the factorial planes by points with a dimension proportional to their confidence.
   The proximity among two or more rules shows the presence of a common structure of antecedent items associated to different consequences. The set of close rules can be changed into a higher order macro-rule obtained linking the common behaviour described by the antecedent items to the logical disjunction of the different consequent items.

3. **Conjoint Visualization.**
   The factorial planes of the items and of the rules can be overlapped because of the features of MCA. The coordinates of each item can be expressed as the mean of the coordinated of the rules containing that item. This feature happens simmetrically for the set of rules.

4. A REAL DATA SET APPLICATION

The proposed approach has been applied on a real data set regarding the italian national channels television preferences. The data regard italian public television (RAI) preferences and refer to the typologies of television programs chosen by a group of 330 users during a day. The 33 considered typologies group the possible programs in macro-genders such as variety show, comics, television news, cartoons, film, holy mass, musical, spot, sports time, etc. The group of 330 users has been randomly drawn by the official users panel that RAI collects to study television customers behaviour (called *auditel*).

The data mart is a 330×33 binary matrix, each cell is equal to 1 if the corresponding user has watched the corresponding macro-genders for more than five minutes during a day (a threshold equal to five minutes allows to avoid users that frequently change channels). The considered data set can be associated to a set T of transactions where each transaction is a subset of macro-genders chosen by the user during a day. The aim of the application is to explain the television customers behaviour through the discovery of the logical associations among the watched typologies of television programs.

The number of rules obtained applying the mining process to the set of 330 television preferences is 6901. This huge number is obtained considering only three order rules[4] and fixing very low support and confidence values (0.01). The main results of the pruning phase (Tab. 1) show that after the first step, 1000 rules were pruned without influencing significatively the support and confidence ranges. The second step allows to prune a huge number of rules (4885) as it works on groups of rules with the same antecedent items. Using also the third step as a pruning tool, 79 final rules result with a minimum confidence equal to 0.89 and a minimum support equal to 0.34.

Tab. 1: Information about the pruning process.

	Befor Pruning	**After Step 1**	**After Step 2**	**After Step 3**
Number of rules	6901	5901	1016	79
Minimun Confidence	0.06	0.08	0.08	0.89
Maximum Confidence	1	1	1	1
Minimum Support	0.05	0.05	0.05	0.34
Maximum Support	0.76	0.76	0.76	0.76

[4] *The rules generated contain at most two items in the antecedent and one item in the consequence.*

The 79 survived rules still represent a huge number of rules to be inspected manually by the user. These rules are stored in the *Rules Mart* with 9 different antecedent items and 6 different consequent items. From the previous *Rules Mart*, the **E** matrix is generated with $n = 79$ rows and $p = (9 + 6 + 2) = 17$ columns. The 9 antecedent items, with 18 modalities associated, are used as active variables of the exploring phase while the 6 consequent items and the values of confidence and support are the supplementary variables. Once the MCA is performed, we consider the first factorial plane that represents the most relevant information and the most significative relationships among the items. The different *views* provided by the Visualizing phase on the set of rules are the following.

1. Items Visualization.

In figure 2 the antecedent items, the consequent items and the support and confidence measures are represented on the first factorial plane.

The proximity between the antecedent items *Cartoons* and *Sport Time* highlights the presence of rules with a common antecedent structure. These rules are characterized by a different consequent structure as the closest "THEN" points show *(Spot, Advertisement)*. Similarly, the proximity between the consequent items *TV Film* and *Sport Time* is related to a common causal structure that can be easily recognized in the antecedent item *Sport*. The two arrows point to regions of the plane characterized respectively by high values of support and confidence.

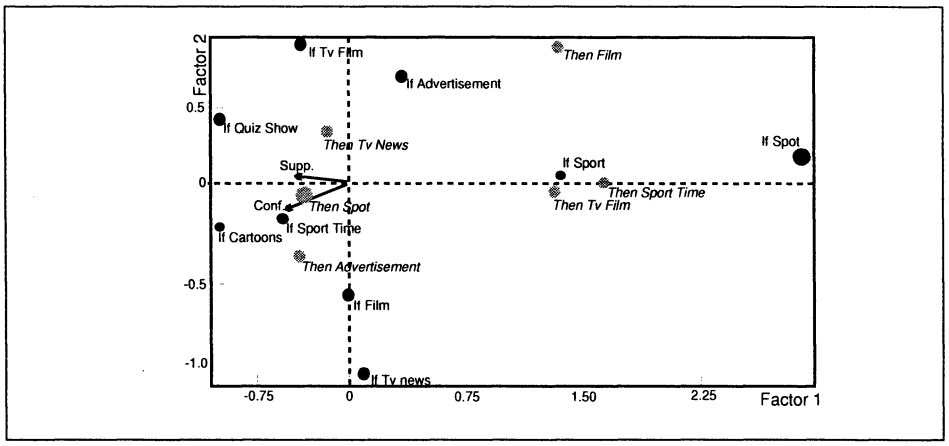
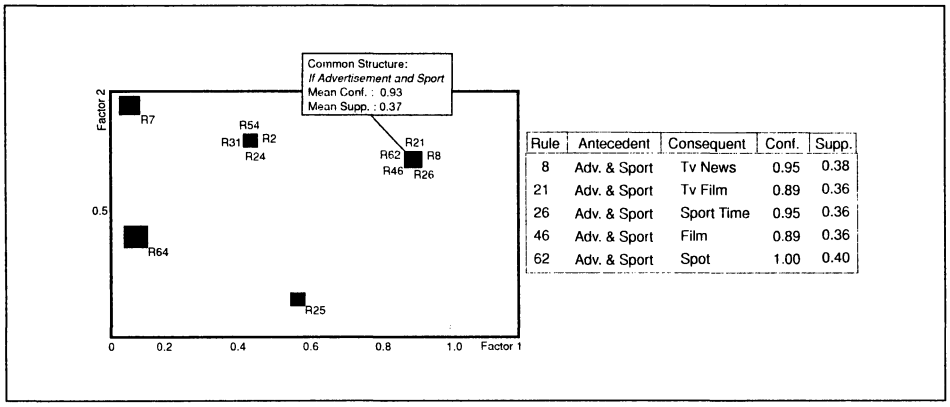


Fig. 2: Item Representation.

2. Rules Visualization.

In figure 3 a zoom of the representation of the association rules on the first factorial plane is shown. The group of rules highlighted are detailed in the right table. It can be noted that they overlap on the plane because they have the same antecedent items *(Advertisement and Sport)* that have defined the plane during the *Exploring* phase. Even if the antecedent structure is the same, each rule has a different consequence and it is possible to deduce an higher order macro-rule of the form:

IF [*Advertisement*] and [*Sport*] THEN [*TV News*] or [*TV Film*] or ...



Rule	Antecedent	Consequent	Conf.	Supp.
8	Adv. & Sport	Tv News	0.95	0.38
21	Adv. & Sport	Tv Film	0.89	0.36
26	Adv. & Sport	Sport Time	0.95	0.36
46	Adv. & Sport	Film	0.89	0.36
62	Adv. & Sport	Spot	1.00	0.40

Fig. 3: Rules Representation.

3. Conjoint Visualization.

In figure 4 a conjoint representation of an antecedent item *(Sport Time)* and of all the rules containing that item is shown. This kind of representation allows to investigate graphically on a specific item and to capture similarities and dissimilarities among the rules.

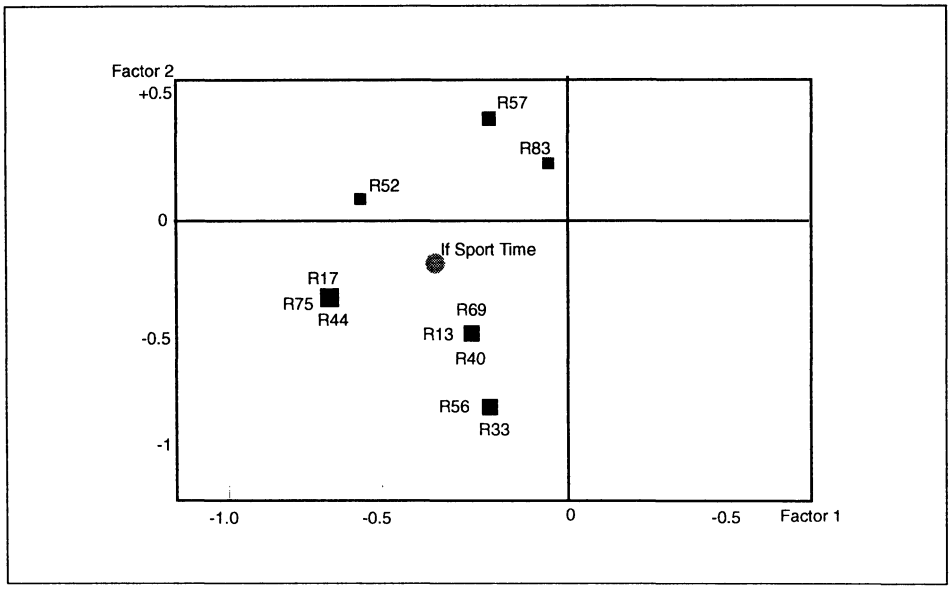


Fig. 4: Conjoint Representation.

5. CONCLUDING REMARKS

This paper proposes an integrated strategy in the association rules framework: a Rules Mart is generated from the classical mining process and it is submitted to a pruning procedure based on statistical tests. The survived rules are inspected through a Multidimensional Data Analysis technique. The exploration of the rules put out 2-d representations where it is possible to visualize the items, the rules and their interactions.

The proposed pruning faces some of the main drawbacks of mining association rules: reduction of the huge number of rules, avoiding arbitrary thresholds for support and confidence, pruning of meaningless associations, statistical evaluation of the implications strength. Each step of the proposed pruning has a very low computational cost as it is based on the use of measures already computed during the mining process. Further developments of the pruning strategy will involve the identification of the best order of the rules. Actually it is necessary to decide the number of items in each rules to avoid the computational explosion during the mining process.

The exploration and visualization of the rules exploits the power of MCA to analyse simutaneously big data matrices with many rows (rules) and many columns (items) and to provide different and simple views on the data.

REFERENCES

AGRAWAL R., IMIELINSKI T. & SWAMI A.: Mining Association Rules between Sets of Items in Large Databases, *Proceedings of the 1993 ACM SIGMOD Conference,* May, Washington DC, USA, (1993) 207–216.

BENZÈCRI J.-P.: *L'Analyse des Donnèes,* Dunod, Paris (1973).

BRUZZESE D. & DAVINO C.: Statistical Pruning of Discovered Association Rules, *Computational Statistics,* Vol. 16, (2001).

HOFMANN H. & WILHELM A.: Validation of Association Rules by Interactive Mosaic Plots. In Bethlehem, J.G., van der Heijden, P.G.M. (eds.): *Compstat 2000 - Proceedings in Computational Statistics,* Physica-Verlag, Heidelberg (2000) 499–504.

KLEMETTINEN M., MANNILA H., RONKAINEN P., TOIVONEN H. & VERKAMO A.I.: Finding interesting rules from large sets of discovered association rules, *Proceedings of the Third International Conference on Information and Knowledge Management CIKM–94,* (1994) 401–407.

LIU B., HSU W., WANG K. & CHEN S.: Visually Aided Exploration Interesting Association Rules. *Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining* (PAKDD-99). Springer Eds., April 26–28, Beijing (1999).

SILBERSCHATZ A. & TUZHILIN A.: On subjective measures of interestingness in knowledge discovery. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining,* (1995) 275–281.

TOIVONEN H., KLEMETTINEN M., RONKAINEN P., HATONEN K. & MANNILA H.: Pruning and grouping of discovered association rules. *Workshop Notes of the ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases,* Heraklion, Greece, April 1995 (1995) 47–52.

WEBER I.: On Pruning Strategies for Discovery of Generalized and Quantitative Association Rules. *Proceedings of Knowledge Discovery and Data Mining Workshop,* Singapore (1998).

WONG P.C., WHITNEY P., THOMAS J.: Visualizing Association Rules for Text Mining. In Wills, G., Keim, D. (eds.): *Proceedings of IEEE Information Visualization '99,* IEEE CS Press, Los Alamitos, CA (1999).

PRUNING, ESPLORAZIONE E VISUALIZZAZIONE DELLE REGOLE DI ASSOCIAZIONE

Riassunto

Le Regole di Associazione rappresentano uno strumento di analisi molto diffuso nei processi di estrazione della conoscenza. Nonostante la loro popolarità, l'analisi e l'interpretazione delle Regole di Associazione resta acora un problema aperto dal momento che il numero di regole estratte è di solito elevatissimo. In questo lavoro si propone una strategia denominata PEV, che consente di affrontare tutte le fasi del processo di Knoweldge Discovery, dall'estrazione delle regole più rilevanti (Pruning) all'investigazione della loro struttura comune (Esplorazione) fino alla loro rappresentazione grafica (Visualizzazione).