Association Rules: Problems, solutions and new applications

María N. Moreno, Saddys Segrera and Vivian F. López
Universidad de Salamanca, Plaza Merced S/N, 37008, Salamanca
e-mail: mmg@usal.es

Abstract

Association rule mining is an important component of data mining. In the last years a great number of algorithms have been proposed with the objective of solving the obstacles presented in the generation of association rules. In this work, we offer a revision of the main drawbacks and proposals of solutions documented in the literature, including our own ones. The work is focused also in the classification function of the association rules, a promising technique which is the subject of recent studies.

1. Introduction

Association analysis has been broadly used in many application domains. One of the best known is the business field where the discovering of purchase patterns or associations between products is very useful for decision making and for effective marketing. In the last years the application areas have increased significantly. Some examples of recent applications are finding patterns in biological databases, extraction of knowledge from software engineering metrics or obtaining user's profiles for web system personalization.

Traditionally, association analysis is considered an unsupervised technique, so it has been applied in knowledge discovery tasks. Recent studies have shown that knowledge discovery algorithms, such as association rule mining, can be successfully used for prediction in classification problems. In these cases the algorithm used for generating association rules must be tailored to the particularities of the prediction in order to build more effective classifiers. However, while the improvement of association rules algorithms is the subject of many works in the literature, little research has been done concerning their classification aspect.

Most of the research efforts in the scope of the association rules have been oriented to simplify the rule set and to improve the algorithm performance. But these are not the only problems that can be found when rules are generated and used in different domains. Troubleshooting for them should consider the purpose of the association models and the data they come from.

The main drawbacks of the association rule algorithms are the following:

- Obtaining non interesting rules
- Huge number of discovered rules
- Low algorithm performance

In this work a review of the main contributions in the literature for the resolution of these problems is carried out. The paper is also focused on the predictive use of the association models due to it constitutes a promising technique for obtaining highly precise classifiers.

In the following section fundamentals of association rules are introduced. Section 3 is dedicated to the problem of obtaining interesting rules. Some interestingness measures are described and methods for reducing the number of discovered rules are presented. The section 4 deals with the classification use of the associative models. Finally, we present the conclusions.

2. Background

Since Agrawal and col. introduced the concept of association between items [2] [1] and proposed the Apriori algorithm [3], many other authors have studied better ways for obtaining association rules from transactional databases. Before considering such algorithms, we introduce the foundations of association rules and some concepts used for quantifying the statistical significance and goodness of the generated rules [23].

A set of discrete attributes $At=\{a_1, a_2, \ldots, a_m\}$ is considered. Let $D=\{T_1, T_2, \ldots, T_N\}$ be a relation consisting on N transactions $T_1, \ldots, T_N$ over the relation schema $\{a_1, a_2, \ldots, a_m\}$. Also, let an atomic condition be a proposition of the form $value_1 \leq attribute \leq value_2$ for ordered attributes and $attribute = value$ for unordered attributes, where *value*, $value_1$ and $value_2$ belong to the set of distinct values taken by attribute in D. Finally, an itemset is a conjunction of atomic conditions or items. The number of items in an itemset is called *length*. Rules are defined as extended association rules of the form $X \rightarrow Y$, where X and Y are itemsets representing the antecedent and the consequent part of the rule respectively.

The strength of the association rule is quantified by the following factors:

- *Confidence* or *predictability*. A rule has confidence *c* if *c*% of the transactions in D that contain X also contain Y. A rule is said to hold on a dataset D if the confidence of the rule is greater than a user-specified threshold.
- *Support* or *prevalence*. The rule has support *s* in D if *s*% of the transactions in D contain both X and Y.
- *Expected predictability*. This is the frequency of occurrence of the item Y. So the difference between expected predictability and predictability (confidence) is a measure of the change in predictive power due to the presence of X [17]. Usually, the algorithms only provide rules with support and confidence greater than the threshold values established.

The Apriori algorithm starts counting the number of occurrences of each item to determine the *large* itemsets, whose supports are equal or greater than the minimum support specified by the user. There are algorithms that generate association rules without generating frequent itemsets [13]. Some of them simplifying the rule set by mining a constraint rule set,

that is a rule set containing rules with fixed items as consequences [4] [5].

Many algorithms for obtaining a reduced number of rules with high support and confidence values have been proposed. However, these measures are insufficient to determine if the discovered associations are really useful. It is necessary to evaluate other characteristics that supply additional indications about the interestingness of the rules.

3. Mining interesting association rules

3.1. Interestingness measures

The interestingness issue refers to finding rules that are interesting and useful to users [16]. It can be assessed by means of objective measures such as support (statistical significance) and confidence (goodness), but subjective measures are also needed. Liu et al. [16] suggest the following ones:

- *Unexpectednes*: Rules are interesting if they are unknown to the user or contradict the user's existing knowledge.
- *Actionability*: Rules are interesting if users can do something with them to their advantage.

Actionable rules are either expected or unexpected, but the last ones are the most interesting rules due to they are unknown for the user and lead to more valuable decisions.

Most of the approaches for finding interesting rules in a subjective way require the user participation to articulate his knowledge or to express what rules are interesting for him.

In [16] a system that analyzes the discovered rules against user's knowledge is presented. It implements a pruning technique for removing redundant or insignificant rules by ranking and classifying them into four categories:

- *Conforming rules*: a discovered rule $A_i \in A$ conforms to a piece of user's knowledge U_j if both the antecedent and the consequent parts of A_i match those of $U_j \in U$ well.
- *Unexpected consequent rules*: a discovered rule $A_i \in A$ has unexpected consequents with respect to a $U_j \in U$ if the antecedent part of A_i matches that of U_j well.
- *Unexpected condition rules*: a discovered rule $A_i \in A$ has unexpected conditions with respect to a $U_j \in U$ if the consequent part of A_i matches that of U_j well, but not the antecedent part.
- *Both-side unexpected rules*: a discovered rule $A_i \in A$ is both-side unexpected with respect to a $U_j \in U$ if the antecedent and consequent parts of A_i do nor match those of U_j well.

Degrees into every category are used for ranking the rules.

In [21] new measures of the statistical significance are proposed in order to provide indicators of rule interestingness:

- *Any-confidence*: an association is deemed interesting if *any* rule that can be produced from that association has a confidence greater than or equal to the established minimum any-confidence value.
- *All-confidence*: an association is deemed interesting if *all* rules that can be produced from that association have a confidence greater than or equal to the established minimum all-confidence value.
- *Bond*: measure similar to the support but with respect to a subset of the data. The subsets are created considering the characteristics of the data.

Interestingness measures have been object of earlier works in the literature. Srikan and Agrawal [24] identify interesting rules by using a "greater-than-expected-value" measure based on deviation from expectation. Other authors consider alternative measures of interest as *gini index*, *entropy gain* or *chi-squared* for data-base segmentation [19] or a measure of implication called *conviction* [6]. Liu et al. [14] propose a technique for dealing with the *rare item problem* that allows the user to specify multiple minimum supports to reflect the natures of the items and their varied frequencies in the database.

These and other interestingness metrics are the base of many methods for reducing the number of discovered association rules.

3.2. Rule reduction methods

Extracting all association rules from a database requires counting all possible combination of attributes. Support and confidence factors can be used for obtaining interesting rules which have values for these factors grater than a threshold value. In most of the methods the confidence is determined once the relevant support for the rules is computed. However, when the number of attributes is large computational time increases exponentially. For a database of m records of n attributes, assuming binary encoding of attributes in a record, the enumeration of subset of attributes requires $m \times 2^n$ computational steps. For small values of n traditional algorithms are simple and efficient, but for large values of n the computational analysis is unfeasible. The best known algorithms, such as *Apriori*, which reduce the search space, proceed essentially by breadth-first traversal of the lattice, starting with the single attributes. They perform repeated passes of the database, on each of which a candidate set of attribute sets is examined. First, single attributes which have low support are discarded, after that, low frequent combination of two attributes are eliminated and so forth. Other algorithms have a similar form but differ in the way the candidate sets are derived. Coenen et al. [7] have developed a set of methods which begin by performing a single database pass to carry out a partial computation of the support count needed, storing these in a tree structure.

Methods for reducing the number of discovered rules based on support pruning are not always useful due to they do not consider interesting rules for infrequent items. For example, the *Apriory* algorithm generates high support sets of rules that are later checked for high confidence. Thus, high confidence rules with low support are not generated. In these cases, confidence based techniques are more appropriate.

Cohen et al. [7] proposed efficient algorithms for finding rules that have extremely high confidence but for which there is no or extremely weak support. The authors argue that high support rules are well known and they do not provide interesting new insights. The algorithms developed identify pairs of similar columns follow a three-phase approach: compute signatures,

generate candidates, and prune candidates. In the first phase a small hash-signature for each column is generated. In the second phase candidate pairs from the column signatures are produced. Finally, high similarity candidate pairs are extracted.

Generalization is an alternative way of reducing the number of association rules. Instead of specializing the relationships between antecedent and consequent parts and restricting rules to support values, in [20] and [12] new aggregates like SUM, MIN, MAX, AVG and other restrictions on market basket items are considered.

Imielinski et al. [11] have proposed a generalization method named *cubegrades*, were a hypercube in the multidimensional space defined by its member attributes is used to evaluate how changes in the attributes of the cube affect some measures of interest. *Cubegrades* deal with computing different aggregate summaries in cubes and evaluating changes in those aggregates due to changes in the structure of the cubes.

Huang and Wu [10] have developed the GMAR (Generalized Mining Association) algorithm which combines several pruning techniques for generalizing rules. The numerous candidate sets are pruned by using minimal confidence.

In [26] a new approach for mining association rules is proposed. It is based on the concept of frequent closed transactions. These groups can be ordered by magnitude, especially by set density. The method has presented an important reduction of the rules and they are generated in a short time.

3.3. Association rule refinement

Most of the methods commented before consider two factors, *support* and *confidence*, which capture the statistical strength of a pattern. However, these factors are useful neither for informing about rules convenience nor for detecting conflicts between rules. It is necessary to consider other factors in order to obtain consistent and interesting patterns. This is the motivation for rules refinement. The topic of knowledge refinement has been treated in the literature, but in the area of association rules little research has been done. In [22, 23] the concept of unexpectedness is introduced in an iterative process for refining association rules. The authors have proposed methods for generating *unexpected patterns* with respect to managerial intuition and use them to refine domain knowledge [22]. Recently they have proposed an iterative refinement process in which it is possible to search through all possible rules [23]. In [22] unexpectedness is defined by starting with a set of beliefs that represent knowledge about the domain. A rule $A \rightarrow B$ is defined to be unexpected with respect to the belief $X \rightarrow Y$ on the database D if the following conditions hold:

- B and Y logically contradict each other (B AND Y |= FALSE);
- A AND X holds on a ''large'' subset of tuples in D;
- The rule A, $X \rightarrow B$ holds.

For example, a belief $X \rightarrow Y$ is that professionals tend to shop more on weekends than on weekdays (Professional $\rightarrow$ Weekend). The rule December $\rightarrow$ Weekday is unexpected with respect to that belief if:

- Weekend AND Weekday |= FALSE.
- Professional and December holds on a large subset of tuples on the database.
- The rule Professional, December $\rightarrow$ Weekday holds.

Given a belief and a set of unexpected patterns, Padmanabhan and Tuzhilin refine the belief using the discovered unexpected patterns. In the same paper they demonstrate formally that the refined rules have more confidence than the original ones.

They use prior domain knowledge to reconcile unexpected patterns and to obtain stronger association rules. Domain knowledge is fed with the experience of the managers. This is a drawback for the use of the method in many application domains where the rules are numeric correlations between project attributes and they are influenced by many factors. It is very difficult to acquire experience in this class of problems. We have developed a refinement method [18] which does not need use managerial experience. It is also based on the discovery of unexpected patterns, but it uses "the best attributes for classification" in a progressive process for rules refinement. The best attributes are obtained by the technique of "importance of columns" [15] based on the amount of information (entropy) that the attributes provide in discriminating the classes. The rule refinement process is simplified due to the use of the best attributes for classification.

The steps of the specific refinement process to be taken are described below:

1. Obtain the best attributes for classification and create the sequence: seqA = $< A_k >$, $k = 1 \ldots t$ (t: number of attributes). The attributes in the sequence are ordered from greater to lesser purity.

2. Split the continuous values of each attribute into discrete intervals. The intervals of values of the attribute A_k are represented as $\{V_{k,l}\}$, $l = 1 \ldots m$ (m: number of intervals).

3. Set $k = 1$ and establish the minimal *confidence* c_{min} and minimal *support* s_{min}.

4. Generate initial beliefs with confidence $c \geq c_{min}$ and support $s \geq s_{min}$.

5. Select beliefs with *confidence* near c_{min} or with conflicts between each other:

   Let $X_i \rightarrow Y_i$ and $X_j \rightarrow Y_j$ be two beliefs, R_i and R_j respectively. There is a conflict between R_i and R_j if $X_i = X_j$ and $Y_i \neg= Y_j$.

6. With the selected beliefs create the rule set setR = $\{R_i\}$, $i = 1 \ldots n$ (n: number of selected beliefs)

7. For all beliefs $R_i \in$ setR do:

   7.1. Use the values $\{V_{k,l}\}$ of the attribute A_k for generating unexpected pattern fulfilling conditions of unexpectedness and *confidence* $\geq c_{min}$. The form of the patterns is: $V_{k,l} \rightarrow B$.

   7.2. Refine the beliefs by searching for rules R' like:

   $X_i , V_{k,l} \rightarrow B$

   $X_i , \neg V_{k,l} \rightarrow Y_i$

7.3. Let setR' be the set of refined rules, then the beliefs refined in step 7.2 should be added to it:

setR' = setR' ∪ { R'$_u$}, $u = 1 \ldots f$ (f: number of refined rules obtained in the iteration i).

8. Set $k = k + 1$ and setR = setR'.

9. Repeat steps 7 and 8 until no more unexpected patterns can be found.

The principal feature of our approach is the gradual generation of the unexpected patterns by taking a single attribute in each iteration. We take advantage of knowledge of good attributes for classification and use them progressively, beginning with the best. This simplifies the selection of patterns and the refinement process.

4. Association rules for classification

Supervised and unsupervised techniques have been used to solve different kind of problems. However, recent studies show that knowledge discovery algorithms, such as those for discovering association rules, can be successfully used for classification tasks [13] [9] [25]. Since the improvement of the knowledge discovery algorithms, especially association rules, is the subject of many works in the literature their classification aspect has hardly been treated. Association rule algorithms discover patterns in the form of rules X → Y. If the consequence part (Y) of the rule is a class, these patterns can be used to predict the class of unclassified records.

A proposal of this category is the CBA (Classification Based on Association) algorithm [15] that consists of two parts, a rule generator for finding association rules and a classifier builder based on the discovered rules. In [25] the weight of the evidence and association detection are combined for flexible prediction with partial information. The main contribution of this method is the possibility of making prediction on any attribute in the database. Moreover, new incomplete observations can be classified. The algorithm uses the weight of the evidence of the attribute association in the new observation in favour of a particular value of the attribute to be predicted. This approach uses all attributes in the observation, however in many domains some attributes have a minimal influence in the classification, so the process can be unnecessary complicated if they are taken in consideration.

We have proposed a procedure which evaluates the importance of the attributes on the classification. It is based on the algorithm of refinement presented before. A significant advantage of the algorithm is the gradual generation of the refined rules. Good attributes in discriminating the classes are taken one by one in the iterative process, beginning with the best. This simplifies the selection of patterns, the refinement process and generates the best rules for class prediction. In other approaches a great set of rules is generated, which is pruned later without considering classification criteria.

The associative model obtained after refining the association rules is used for making predictions. This model is composed by rules at different levels of refinement. The most refined rules are used in the first instance. If an observation to be classified has attributes coincident with the antecedent part of the rule, the label class assigned to the observation is the consequent part. If there are not rules that match the observation, previous level of refinement is used for searching suitable patterns. The problem of finding more than one rule matching the observation is solved taking the more confident rule.

This classifier has been successfully applied in the software project management field as well as for making recommendations in personalized web systems. Figures 1 and 2 show its application in the prediction of software size from attributes of earlier phases of the software life cycle. The figures are *Mineset* [17] graphical representations of the initial and refined rules on a grid landscape with left-hand side (LHS) items on one axis, and right-hand side (RHS) items on the other. Attributes of a rule (LHS → RHS) are displayed at the junction of its LHS and RHS item. The display includes bars, disk and colours whose meaning is given in the graph.

The simplicity of the models obtained with the refined rules lead to their efficient application for prediction. Another benefit is the lesser number of descriptive attributes needed with respect to traditional classification methods.

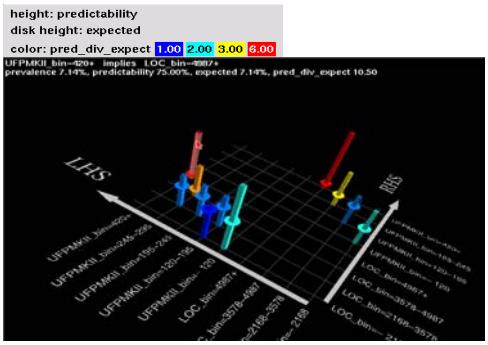

Figura 1. Rules representing the initial beliefs

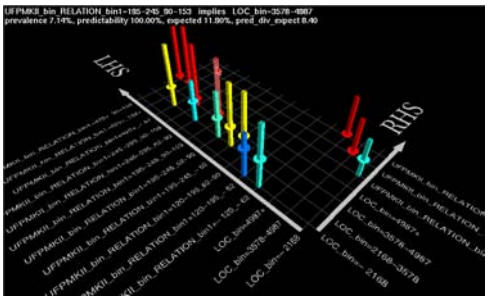

Figura 2. Refined beliefs in the first iteration

5. Conclusions

In this work a revision on the main problems presented by the association rules and proposals of solution has been made. We also include our proposal about a classifier which uses an associative model obtained by means of an algorithm for refining association rules. This is based on incremental knowledge discovery that addresses the problem of producing a reduced number of useful high confident rules without need of domain

knowledge. The identification of weak rules representing beliefs and conflicts between them is the starting point to the iterative refinement process.

Our proposal evaluates the importance of the attributes on the classification. The algorithm produces the most useful rules for prediction, due to the fact that it uses the most important attributes in discriminating the different values of the class attribute, which is the target of the prediction. Other rule refinement methods do not consider this issue, thus they reduce the number of rules, but they do not obtain the most suitable rules for the desired objectives. With respect to the advantages of using association instead of classification, the first is the lower number of attributes required and the second, the greater efficiency of the association algorithm.

Referencias

[1] Agrawal, R., Imielinski, T., Swami, A. Database Mining: A performance Perspective. IEEE Trans. Knowledge and Data Engineering, vol. 5, 6, pp. 914-925, 1993.

[2] Agrawal, R., Imielinski, T., Swami, A. Mining associations between sets of items in large databases. Proc. of ACM SIGMOD Int. Conference on Management of Data, Washinton D.C., pp. 207-216, 1993.

[3] Agrawal, R., Srikant, R. Fast Algorithms for mining association rules in large databases. Proc. of 20th Int. Conference on Very Large Databases, Santiago de Chile pp. 487-489, 1994.

[4] Bayardo, R., Agrawal, R., Gunopulos, D. Constraint-based rule mining in large, dense database, Proc. 15th Int. Conference on Data Engineering, pp. 188-197, 1999.

[5] Bayardo, R., Agrawal, R.: Constraint-based rule mining in large, dense database. Proc. Mining the most interesting rules. Proc. ACM SIGKDD Int. Conf. Knowledge Discovery in Databases, ACM Press, NY, 145-154, 1999.

[6] Brin, S., Motwani, R., Ullman, J. and Tsur, S., Dynamic itemset counting and implication rules for Market Basket Data. Proc. ACM SIGMOD Conf., pp. 255-264, 1997.

[7] Coenen, F., G. Goulbourne and P. Leng. Tree Structures for Mining Association Rules. Data Mining and Knowledge Discovery, 8, 25-51. Kluwer Academic Publishers, 2004.

[8] Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indik, P. Motwani, R., Ullman, J. and Yang, C. Finding interesting associations without support pruning. Proc. 16th.Int. Conf. on Data Engineering, pp. 489-500, 2000.

[9] Hu, Y.C., Chen, R.S., Tzeng, G.H.: Mining fuzzy associative rules for classifications problems. Computers and Industrial Engineering, 43 (4) pp. 735-750 , 2002.

[10] Huang, Y. F., C. M. Wu. Mining Generalized Association Rules Using Pruning Techniques. Proceedings of the IEEE International Conference on Data Mining (ICDM'02), Japan, pp. 227-234, 2002.

[11] Imielinski, T., A. Virmani and A. Abdulghani. DataMine: Application Programming Interface and Query Language for Database Mining. *Proceedings ACM Int'l Conference Knowledge Discovery & Data Mining*, ACM Press, pp. 256-261, 1996.

[12] Lackshmanan, L.V.S., Ng, R., Han, J. and Pang, A. Optimization of constrained frequent set queries with 2-variable constraints. Proc. of ACM SIGMOD Conf., pp. 158-168, 1999.

[13] Li, J., Shen, H., Topor, R. Mining the smallest association rule set for predictions. Proc. IEEE International Conference on Data Mining (ICDM'01), 2001.

[14] Liu, B., Hsu, W. Ma, Y. Mining association rules with multiple minimum supports. Proc. Int. Conference on Knowledge Discovery and Data Mining, pp. 337-341, 1999.

[15] Liu, B., Hsu, W. Ma, Y. Integration Classification and Association Rule Mining. Proc. 4th Int. Conference on Knowledge Discovery and Data Mining, pp. 80-86, 1998.

[16] Liu, B., Hsu, W., Chen, S., Ma, Y. Analyzing the subjective Interestingness of Association Rules. IEEE Intelligent Systems, september/October 47-55, 2000.

[17] Mineset user's guide, v. 007-3214-004, 5/98. Silicon Graphics, 1998.

[18] Moreno, M.N., Miguel, L.A., García, F.J., Polo, M.J.: Building knowledge discovery-driven models for decision support in project management. Decisión Support Systems , 38, pp. 305-317, 2004.

[19] Morimoto, Y., Fukuda, T., Matsuzawa, H., Tkuyama, T. and Yoda, K. Algorithms for mining association rules for binary segmentation of huge categorical databases. Proc. of Very Large Databases Conf., pp. 380-391, 1998.

[20] Ng, R., Lackshmanan, L.V.S., Han, J. and Pang, A. Exploratory mining and pruning optimizations of constrained association rules. Proc. of ACM SIGMOD Conf., pp. 13-24, 1998.

[21] Omiecinski, E.R. Alternative interest measures for mining associations in databases. IEEE Transaction on Knowledge and Data Engineering, 15 (1) pp. 57-69, 2003.

[22] Padmanabhan, B., Tuzhilin, A.: Knowledge refinement based on the discovery of unexpected patterns in data mining. Decision Support Systems 27, 303-318, 1999.

[23] Padmanabhan, B., Tuzhilin, A.: Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems 33, pp. 309-321, 2002.

[24] Srikant, R. and Agrawal, R. Mining quantitative association rules in large relational tables. Proc. of ACM SIGMOD Conf., pp. 1-12, 1996.

[25] Wang, Y., Wong, A.K.C. From Association to Classification: Inference Using Weight of Evidence. IEEE Transactions on Knowledge and Data Engineering, 15 (2003) 764-767, 2003.

[26] Zaki, M.J. Mining Non-Redundant Association Rules. Data Mining and Knowledge Discovery, 9, 223-248, Kluwer Academic Publishers, The Netherlands, 2004.