Improving the Quality of Association Rules by Preprocessing Numerical Data

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

Abstract

Numerical attribute management is a usual pre-processing task in data mining. Most of the algorithms require the discretization of numerical attributes by splitting the continuous range of values into intervals. This task is critical in association rule induction since it has a significant influence in the quality of the induced rules. We have proposed a multivariate algorithm whose efficiency against other discretization methods is demonstrated in the comparative study presented in this paper.

1. Introduction

Many data mining problems require dealing with continuous numerical attributes, which must be split into intervals of values in order to obtain comprehensible results. This process, named attribute binning or discretization, is carried out either in preprocessing, or embedded within the induction algorithm, depending on the method. Some learning methods, such as decision tree, enclose the partitioning procedure; however, association rule algorithms do not usually include it. Therefore, it is necessary to carry out the discretization during the preprocessing stage.
for example, a decision tree learning algorithm.

Many works in the literature are focused on discovering accurate and high performance data mining algorithms without giving special attention to the binning process. However, an efficient attribute discretization can provide more benefits than the proper induction methods. Models induced from suitable discrete values are more understandable and concise. In the case of supervised algorithms a significant improvement in predictive accuracy can be achieved. For knowledge discovery models, such as association rules, shorter and more interesting rules can be obtained.

The rest of de paper is structured as follow. In the following section the basis of discretization methods is introduced. Section 3 introduces association analysis and main characteristics of association rules. Next section contains a shallow description of some papers in the literature concerning the improvement of association rule algorithms and specific procedures for managing quantitative attributes in association rule mining. Section 5 is dedicated to the proposed discretization algorithm. The experimental study and the analysis of results are presented in sections 6 and 7 respectively. Finally, we present the conclusions.

2. Numerical attribute preprocessing

Different classifications of attributes used in the data mining field can be done. A first division among categorical and quantitative (numerical) data is the most common. Categorical data can be subdivided in nominal and ordinal ones. Nominal data are named values without an order between them. Ordinal data are nominal data types with ordered values. Quantitative attributes can be continuous or discrete. While continuous data can take an infinite number of values, discrete attributes represent intervals in a continuous spectrum and therefore, they take a finite number of values [15].

Data mining process involves a preprocessing step in order to assure the data have the quality and the format required by the algorithms. Data transformation is a preprocessing task including any procedure that modifies the original form of the data. A common transformation consists on splitting a continuous numerical domain into intervals, obtaining discrete attributes required by the algorithms. This procedure, denominated binning or discretization, is critical in classification and knowledge discovery. Discretization algorithms should be adapted to the data mining methods in order to improve the induced models.

Several discretization approaches are available [15]. Two simple techniques commonly used are equal-width and equal-frequency, which consist on creating a specified number of intervals with the same size or with the same number of records respectively. The purpose of the discretized data and the statistical characteristics of the sample to be treated should be kept in mind when one of these algorithms is selected.

Other methods used mainly in classification problems consider entropy measures. These are often embedded within machine learning algorithms such as those of decision trees induction. In these cases, a supervised discretization is usually carried out. It considers class information for generating the intervals while unsupervised discretization does not. No class information is available when mining association rules, thus only unsupervised procedures can be used. Unless we want to give more weight to some attributes, for example, to those appearing in the consequent part (right side) of the rule.

On the other hand, discretization can be univariate or multivariate. Univariate discretization quantifies one continuous attribute at a time while multivariate discretization considers simultaneously multiple attributes. Though the univariate discretization is more used and simple than the multivariable one, the latter presents more advantages, particularly in discovery of association rules since all available attributes are simultaneously involved in the mining process.

3. Quality of association rules

Association analysis is a useful data mining technique exploited in multiple application domains. One of the best known is the business field where the discovering of purchase patterns or associations between products that clients tend to buy together is used for developing an effective marketing. The attributes used in this domain are mainly categorical data, which simplifies the procedure of mining the rules. In the last years the application areas involving other types of attributes have increased significantly. Some examples of recent applications are finding patterns in biological databases, extraction of knowledge from software engineering metrics [18] [19] or obtaining user's profiles for web system personalization [17].

Associative models have been even used in classification problems as the base of some efficient classifiers [13] [17].

Numerous methods for association rule mining have been proposed, however many of them discover too many rules, which represent weak associations and uninteresting patterns. The improvement of association rules algorithms is the subject of many works in the literature. Most of the research efforts have been oriented to simplify the rule set, to generate strong and interesting patterns as well as to improve the algorithm performance. When attributes used for inducing the rules take continuous values, these three objectives can be achieved by means of an efficient data discretization procedure such as the proposed in this paper.

In a set of transactions D the strength of an association rule in the form “If X then Y” is mainly 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.

The interestingness issue refers to finding rules that are interesting and useful to users [14]. It can be assessed by means of objective measures such as support (statistical significance) and confidence (goodness), defined before, but subjective measures are also needed. Liu et al. [14] 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. Unfortunately these subjective factors cannot be easily obtained in some application areas such as project management, especially when a large number of quantitative attributes are involved and, so, it is very difficult to acquire domain knowledge. These applications present additional problems such as the discretization of continuous quantitative attributes, which can take a wide range of values. In order to reduce the number of rules generated it is necessary to split the range of values into a suitable number of intervals.

In this paper a multivariate discretization method is proposed. The procedure was applied in the discovery of association rules from a project management data base, yielding a reduced number of strong association rules which cover a large percentage of examples.

4. Related works

The concept of association between items [1] [2] was first introduced by Agrawal and col. Since they proposed the popular Apriori algorithm [3], the improvement of the algorithms for mining association rules have been the target of numerous studies. Many other authors have studied better ways for obtaining association rules from transactional databases. Most of the efforts have been oriented to simplify the rule set and improve the algorithm performance.

The best known algorithms, such as Apriori, which reduce the search space, proceed basically 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. Cohen et al. [5] proposed efficient algorithms for finding rules that have extremely high confidence but for which there is no or extremely weak support.

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 [11] and [10] new aggregates and other restrictions on market basket items are considered. Imielinski et al. [9] 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. Huang and Wu [8] 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 based on the concept of frequent closed transactions is proposed.

The topic of knowledge refinement is used in some methods in order to obtain a reduced number of consistent and interesting patterns. In [20] and [21] the concept of unexpectedness is introduced in an iterative process for refining association rules. It uses 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. It is an effective procedure for classification problems that is very suitable for applications that manage quantitative attributes where domain knowledge cannot be easily obtained. The aim is to provide managers with a convenient number of good association rules for prediction, which help them to make right decisions about the software project. However, in many cases an efficient binning of project data can be more effective than complex 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. Nevertheless, 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 x 2n computational steps. For small values of n, traditional algorithms are simple and efficient, but for large values of n the computational analysis is unfeasible. When continuous attributes are involved in the rules, the discretization process is critical in order to reduce the value of n and to obtain high confident rules at the same time.

Attribute discretization methods for mining association rules have been treated in the literature. Nearly everyone take the support factor of the rules as the main feature for splitting the attribute values into intervals, that is, they consider the weight of the records in the interval in relation to the total number of records. In [24] the measure of "partial completeness", based on confidence and support factors, is defined for quantifying the information lost by partitioning quantitative attributes. Values are fine-partitioned and then, adjacent intervals are combined if it is necessary.

Lian et al. [12] introduce the concept of "density" to capture the characteristics of quantitative attributes and propose an efficient procedure to locate "dense regions" for mining association rules. The density of the regions representing rules is a measure used to evaluate the interestingness of the rules instead of using support and confidence.

An alternative to the conversion of continuous attributes into discrete data is to consider the distribution of the continuous data, via standard statistical measures such as mean and variance [4]. In [4] a rule is considered a subset of a population. A validity testing is proposed taking account if the mean of the subset is significantly different to the mean of its complement in the database. The rules are also evaluated depending on other measure of probability distributions such as variance.

Recently, several partition methods based on the fuzzy set theory have been proposed [7] [25]. The mined rules are expressed in linguistic terms, which are more natural and understandable. In these works either both the antecedent and consequent parts of the rules are formed by a single item or the consequent part is not fixed. In our case the consequent part must be fixed because there are input and output attributes and both consequent and antecedent parts are itemsets. So, it is more suitable a multivariate discretization that consider all the attributes.

5. Discretization procedure

All the attributes that are used in this work to generate association rules are continuous, that is, they can take a wide range of values. In order to reduce the number of rules generated it is necessary to discretize the attributes by splitting the range of values into a manageable number of intervals.

A clustering technique was applied for discretizing multiple attributes simultaneously. Clusters of similar records were built by using the iterative k-means algorithm with a Euclidean distance metric [6]. This distance D(p,q) between two points p and q in a space of n dimensions is:

$$(1) \quad [D(p,q)]2 = \| p - q \|2 = \Sigma_i \ (p_i - q_i)^2$$

where p_i and q_i are the coordinates of the points p and q respectively. In our case, the points are the records to be compared, and the coordinates are the n attributes of each record.

The iterative k-means algorithm takes as input the minimum and maximum number of clusters (k). The selected values in this work were 1 and 10 respectively. This clustering method groups the records in a way that the overall dispersion within each cluster is minimized. The procedure is the following:

1. The value of the minimum number of clusters is assigned to k.
2. The k cluster centers are situated in random positions in the space of n dimensions.
3. Each record in the data is assigned to the cluster whose center is closest to it.
4. The cluster centers are recalculated based on the new data in each cluster.
5. If there are records which are closer to the center of a different cluster than the cluster that they belong to, then, these records are moved to the closer cluster.

Steps 4 and 5 are repeated until no further improvement can be made or the maximum number of clusters is reached.

The distribution of attribute values in the clusters was used for making the discretization according to the following procedure [19]:

1. The number of intervals for each attribute is the same of the number of clusters. If m is the mean value of the attribute in the cluster and σ is the standard deviation, the initial interval boundaries are (m - σ) and (m + σ).
2. When two adjacent intervals overlap, the cut point (superior boundary of the first and inferior boundary of the next) is placed in the middle point of the overlapping region. These intervals are merged into a unique interval if one of them includes the mean value of the other or is very close to it.

3. When two adjacent intervals are separated, the cut point is placed in the middle point of the separation region.

This procedure was applied for creating intervals of values for every one of the attributes in order to generate the association rules.

6. Experimental study

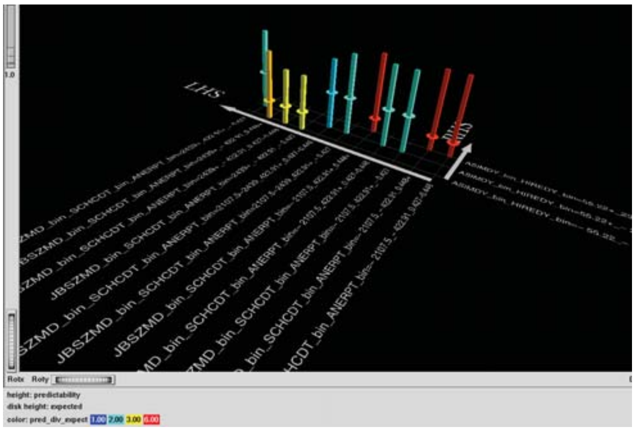
The data used in this study come from a dynamic simulation environment developed by Ramos et al. [22] [23]. This environment manages data from real projects developed in local companies and simulates different scenarios. It works with more than 20 input parameters and more than 10 output variables. The number of records generated for this work is 300 and the variables used for the data mining study are those related with time restrictions, quality and technician hiring.

The aim of the work is to obtain an associative model that allows studying the influence of the input variables related to the project management policy on the output variables related to the software product and the software process.

The clusters were created with a weight for the output variables three times greater than for input attributes. This is a supervised way of producing the most suitable clusters for the prediction of the output variables, which appear in the consequent part of the rules. In this study the clustering algorithm produced three clusters.

Rules representing the impact of project management policies on software quality, development time and effort were generated and visualized by using Mineset, a Silicon Graphics tool [16]. Figure 1 is a graphical representation of the rules on a grid landscape with left-hand side (LHS) items on one axis, and right-hand side (RHS) items on the other. A rule (LHS RHS) displayed at the junction of its LHS and RHS itemset relates the itemset containing the input attributes with the itemset formed for the output attributes. The display includes bars, disk and colors whose meaning is given in the graph. Rules generator does not report rules in which the predictability (confidence) is lesser than the expected predictability (frequency of occurrence of the item RHS), that is, the result of dividing predictability by expected predictability (pred_div_expect) should be greater than one. Good rules are those with high values of pred_div_expect. We have also specified a minimum predictability threshold of 60%.

Under the exposed conditions, eleven rules were generated. Their confidence and support factors are presented in the table 1.



height: predictability
disk height: expected
color: pred_div_expect 1.00 2.00 3.00 6.00

Fig. 1. Association rules

Rule	%Confidence	%Support
1	100	1.14
2	86.67	4.94
3	69.77	11.41
4	69.23	3.42
5	88.24	5.70
6	100	16.35
7	100	7.60
8	100	7.60
9	100	10.27
10	100	4.18
11	100	2.66
SUM		75.27
AVERAGE	92.18	

Table 1. Support and confidence factors for the association rulesAnalysis of results

The proposed procedure generated a low number of association rules. Their support and confidence factors showed in table 1, which capture the statistical strength of the patterns, have high values. In our study domain, the more confident a rule is, the more reliable it will be when it will be used to take project management decisions. Seven discovered rules have the maximum confidence value (100%) and the remaining rules have high values of this factor, yielding an average value of 92.18%, therefore they are good for taking decision in future projects. On the other hand, the induced associative model is useful if it is constituted by a controllable number of rules and the rule set covers a large percentage of examples (records). The coverage measure is provided by de total support of the rules, that is, the sum of individual supports. In our case study the proposed method gives a model that covers the 75% of the examples with just eleven association rules (see table 1).

Other experiments were carried out in order to compare results. The same algorithm for mining association rules was used but the discretization procedures were different. We applied equal-width and equal-frequency discretization methods for splitting attributes into five intervals (fewer intervals gave worse results). The comparative study of results from the three methods is showed in the table 2.

Discretization Method	**Number of rules**	**% Mean confidence**	**%Total support**
Equal-frequency	19	82.79	28.88
Equal-width	23	81.25	57.00
Our method	***11***	***92.18***	***75.27***

Table 2. Comparative results of several discretization methods

We can observe that our method, with a significant smaller number of rules, gets a quite bigger mean confidence factor than the other methods. The results about the support are more interesting, while the other methods cover a very small percentage of examples, 28.88% and 57.00% respectively, with a high number of rules, 19 and 23, our method gets a total support of 75,27% with just 11 rules.

In the study carried out, a reduced number of strong rules have been generated. In addition, the rule induction process was very fast, due to the association rule algorithm works with a reduced number of intervals of values of the attributes, which are provided by the discretization method. Then, the obtained associative model, which relates management policy factors with quality, time and effort, provides managers with a useful tool for taking decisions about current or future projects.

7. Conclusions

The paper deal with the problem of finding useful association rules from software project management data. The main drawbacks in this application field are the treatment of continuous attributes and the difficulty to obtain domain knowledge in order to evaluate the interestingness of the association rules. We have proposed an association rule mining algorithm for building a model that relates management policy attributes with the output attributes quality, time and effort. The success of the algorithm is mainly due to the supervised multivariate procedure used for discretizing the continuous attributes in order to generate the rules. The result is an association model constituted by a convenient number of high confident rules representing relevant patterns between project attributes. This allows estimating the influence of the combination of some variables related to management policies on the software quality, the project duration and the development effort simultaneously.

In addition, the proposed method avoids three of the main drawbacks presented by the rule mining algorithms: production of a high number of rules, discovery of uninteresting patterns and low performance.

References

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

[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., 1993b, pp. 207-216.

[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, 1994, pp. 487-489.

[4] Aumann, Y. and Lindell, Y. A Statistical Theory for Quantitative Association Rules. *Journal of Intelligent Information Systems*, 20 (3), 2003, 255-283.

[5] Coenen, F., G. Goulbourne and P. Leng. Tree Structures for Mining Association Rules. *Data Mining and Knowledge Discovery*, 8, 2004, pp. 25-51.

[6] Grabmeier, J. and Rudolph, A. Techniques of Cluster Algorithms in Data Mining , *Data Mining and Knowledge Discovery*, 6, 2002, pp. 303-360.

[7] Hong, T.P., Kuo, C.S. and Chi, S.C. Mining association rules from quantitative data, *Intelligent Data Analysis* (1999) 363-376.

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

[9] 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, 1996, pp. 256-261.

[10] 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.*, 1999, pp. 158-168.

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

[12] Lian, W, Cheung, D.W. and Yiu, S.M. An Efficient Algorithm for Dense Regions Discovery from Large-Scale Data Streams. *Computers & Mathematics with Applications,* , vol 50, no. 3-4, 2005, pp. 471-490.

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

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

[15] Liu, H., Hussain, F., Tan, C.L., Dash, M. Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery*, 6, 2002, 393-423.

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

[17] Moreno, M.N., García, F.J., Polo, M.J. and López, V. Using Association Analysis of Web Data in Recommender Systems. *Lectures Notes in Computer Science*, LNCS 3182, 2004, pp. 11-20.

[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, 2004, pp. 305-317.

[19] Moreno, M.N., Ramos, I., García F.J., Toro, M. An association rule mining method for estimating the impact of project management policies on software quality, development time and effort. *Expert Systems with Applications*, 34 (2008) pp. 522–529.

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

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

[22] Ramos, I., Riquelme, J. and Aroba, J.C. Improvements in the Decision Making in *Software Projects: Application of Data Mining Techniques*, IC-AI`2001, 2001.

[23] Ruiz, M.; Ramos, I. and Toro, M., A Simplified Model of Software Project dynamics, *The Journal of Systems and Software*, 59 (2001), 2001, pp. 299-309.

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

[25] Verlinde, H., De Cock, M. and Boute R. Fuzzy Versus Quantitative Association Rules: A Fair Data-Driven Comparison. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 36 (3), 2006, pp. 679-684.

[26] Zaki, M.J. Mining Non-Redundant Association Rules. *Data Mining and Knowledge Discovery*, 9, 2004, pp. 223-248.