Quantitative Association Rules over Incomplete Data

Vincent Ng and John Lee

Department of Computing, Hong Kong Polytechnic University,
Kowloon, Hong Kong

ABSTRACT

This paper explores the use of principle component analysis (PCA) to estimate missing values during the mining of quantitative association rules. An example of such association may be "15% of customers spend \$100 - \$300 every month will have 2 cable outlets at home". In our algorithm, instead of imputing missing values before the mining process, we propose to integrate the imputation step within the process. The idea is to reduce the unnecessary imputation effort and to improve the overall performance. First, only attributes with enough support counts and with missing values are required to perform imputations. Thus, effort will not be wasted on unimportant attributes. Further, rather than estimating the actual value of a missing data, the possible range of the value is guessed. This will not affect the resultant quantitative association rules much but will cut down the guessing effort.

1 INTRODUCTION

Many excellent studies on knowledge discovery have been conducted and reported [1, 2, 5, 6]. Piatetsky-Shapiro, et. al., defined the goal of studying KDD and its general requirement [4]. In general, data mining algorithms can be classified into two categories: concept generalization-based discovery and discovery at the primitive concept levels. The former relies on the generalization of concepts (attribute values) stored in databases and then summarization of the data regularities at a higher conceptual level, such as the DBLearn system developed by Han [6]. The latter relies on the discovery of strong regularities (rules) from the databases without concept generalization. Association rule is a good example of this approach.

Agrawal, et. al., first defined the problem of mining association rules in their pioneer work [1]. The algorithm proposed there can be applied to sales transaction records from large retailing companies. Rules like "[Children's Hairliners] => [Infants and Children's Wear] (66.55%)" were discovered which is interpreted as the rule "among all the customers, who had bought something in the Children's Hardliner Department, 66.55% of them also bought something in the Infants and Children's Wear Department". However, data mining algorithms are difficult to apply directly when data is incomplete. Further, in addition to binary and categorical data, quantitative data are often analyzed, which cannot be dealt with effectively by the boolean association rules. This paper explores the use of principle component analysis (PCA) to estimate missing values during the mining of quantitative association rules. An example of such quantitative association may be "15% of customers spend \$100 - \$300 every month will have 2 cable outlets at home".

In the next section, we will first review the recent work in discovering quantitative association rules. Section 3 will discuss what imputation techniques have been suggested and used in data mining. In Section 4, we will present our definition of the missing information problem in the mining of quantitative data. In Section 5, we present three algorithms. The *Pre-Guess* algorithm will perform the imputation before mining. The *Post-Guess* algorithm will utilize principle component analysis to estimate the missing values after the mining process. In the last algorithm, *Deviation* algorithm, it does not estimate the actual value of a missing data, rather the possible range of the value is guessed. This will not affect the resultant quantitative association rules much but

The work was supported in part by the Polytechnic University Research Grant 0351/638.

0-7803-4778-1/98 \$10.00 © 1998 IEEE

the resultant quantitative association rules much but will cut down the guessing effort. Preliminary experimental results are presented in Section 6 and Section 7 will conclude our work.

2 QUANTITATIVE ASSOCIATION RULES

Srikant and Agrawal, has reported their work on mining quantitative association rules (QAR) in 1996 [10]. A QAR extends the classical association rules to include range predicates like ($value_1 \leq Item \leq value_2$). A quality metric, *partial-completeness*, is developed to find out what are good intervals for the rules. In their work, quantitative attributes are first partitioned according to the values of the attributes and adjacent partitions are then combined as necessary. The partial-completeness is used to ensure that intervals are neither too big, nor too small with respect to the set of rules they generate.

In developing the measure, quantitative properties of the intervals, such as the relative distance between values, or the density of an interval, are not considered. The algorithm works well for ordinal data where there is no semantic meaning between the values. It does not work well on highly skewed data since it may separate close values that have the same behavior.

A distance-based measure has been recently developed to consider the quantitative properties of an interval [9]. The *distance-based association* rules can be discovered by first finding out clusters which represent the interesting intervals and then applying a standard association technique. The Birch algorithm [11] is used as the clustering algorithm because it can incrementally identify and refine clusters in a single pass over the data.

3 IMPUTATION TECHNIQUES

Imputation has been an interesting area in Statistics for many years [3]. It has caught much attention, particularly, when people are working on multivariate analysis. In general, there are three approaches to deal with missing data. The simplest and most direct approach is to include only those records with complete information. This approach is fine if data are missing at random; otherwise the results from any analysis will be biased. A second simple alternative for missing data is to delete the offending attributes. In many situations where a nonrandom pattern of missing data is present, this may be the most efficient solution. However, after the deletion, the analyst may only leave with a handful of attributes and the analysis's results are uninteresting.

A third approach to deal with missing data is through various imputation techniques. Imputation is the process of estimating the missing value based on valid values of other attributes and/or records in the data. There are four common techniques used to replace missing values.

1. Record substitution: this is to replace a record with missing data by another record not in the current data set.

2. Mean substitution: this replaces the missing value of an attribute by the mean value of that attribute based on all complete records.

3. Cold deck imputation: constant values obtained from previous studies are used to replace missing values of attributes in the current data set.

4. Regression imputation: regression analysis is used to predict the missing values of an attribute based on its relationship to other attributes in the data set.

In data mining, method (1) and (3) are not appropriate as there may not be any external data nor previous study. Even when there is external data available, the discovered patterns will be heavily biased towards the external information. Method (2) is feasible, but one has to deal with non-integral values in the data mining algorithms and the interpretations of the discovered patterns afterwards. Amongst the four methods, method (4) demands more effort but will have better predictions of the missing values. The concern, though, is related to the performance and applicability when working with huge data set.

Lakshminarayan et al. have proposed to use machine-learning based methods to deal with missing data [8]. The idea is similar to method (4) above, except that two well-known machine-learning techniques are adopted. The first technique adopts a Bayesian approach to cluster the data into classes by using an unsupervised clustering strategy. Autoclass is used to generate a set of resultant classes which are then be used to predict the missing values of attributes. The second method predicts the missing values via a decision tree-based classifier, C4.5. Before the predictions, supervised induction is used to train the classifier with a sample set first. However, for both techniques, they are only dealing with a relatively small set of data. Their performance and applicabilities on large volume of data are not tested.

A more promising approach is the use of principle component analysis (PCA) to estimate the missing values [7]. Korn et. al. proposes a single-pass mining algorithm to find linear rules from a set of quantitative data. Rules like "Customers that spend between 7 to 15 dollars on coffee beans and 3 dollars on milk will likely to spend 1 dollar on sugar" are found. The algorithm is based on the results of performing a PCA on the data. The linear rules found can also be used to guess any missing values in the data. When comparing with mean substitution, the linear rules can achieve up to 5 times smaller guessing error.

4 PROBLEM STATEMENT

In the work of [7, 8], it has been assumed that there are two states of a quantitative item within a given transaction. The item can either be having a numeric value, or its value is missing. However, if we consider the popular supermarket scenario, we can observe that, in fact, there is one more possible state – "the item is not in the transaction". For example, suppose there are only three items available in a supermarket: bread, milk and butter. If we bought $3 of milk and $5 of bread, it does not imply that we have bought some butter as well.

Let $V = I_1, I_2, \ldots, I_M$ be a set of quantitative items, and T be the transactions of a database D. For each transaction t, t[k] > 0 means that t contains item I_k with the value t[k]. t[k]=0 means that I_k does not exist in t while t[k]= -1 means that I_k's value is missing. In our work, we are interested in finding all quantitative association rules, $X => Y$, that satisfy the following conditions:

- X and Y are intersections of predicates like ($V_k = value1$) or ($value1 \leq V_k \leq value2$).

- The clusters of items represented by the predicates in X and Y are disjoint.

- The frequencies (support counts) of the transactions satisfying the predicates in X and Y are greater than a given threshold.

- X and Y may contain 1 or more item with missing information.

5 ALGORITHMS

In this section, we will discuss three different approaches to support mining on missing information. Their differences are mainly due to how we apply the imputation technique before, after or during the mining step. We have adopted the distance-based association (DBA) algorithm [10] for the mining of quantitative association rules because of its efficiency. Only a single pass of the data is needed in its first phase of clustering the database. In its second phase, a clustering graph is formed and maximal cliques of the graph will correspond to large itemsets in the data. From the large itemsets, the quantitative association rules are then derived. Due to the space limitation, we will not cover the details of the algorithms, but will highlight their major ideas.

5.1 Pre-Guess Algorithm

The first algorithm is based on the pre-processing approach. That is, we will first impute any missing values and then perform the mining step afterwards. The imputation method utilizes the principle component analysis as proposed in [7]. Eigenvectors are first computed from the covariance matrix of the database D. They are then used to predict any missing value. However, this method only handles transactions with a constant number of items. In our database, the number of quantitative items varies for different transactions. This implies that we may need different covariance matrices for different types of transactions. In order to save the computational and book-keeping effort, in the *Pre-Guess Algorithm*, we propose to construct one covariance matrix only.

In Figure 1, we show the steps in finding out the covariance matrix. After forming the matrix, we will scan the database D again. When a transaction with missing values is encountered, columns and rows corresponding to its non-empty items in the matrix are pulled out. The sub-matrix is used to compute the corresponding eigensystem to do the predictions afterwards. The system is saved for other transactions with the same items containing missing values later. Several eigensystems may be resulted but only two passes of D are required.

After the prediction of all the missing values, we use the DBA algorithm to find the quantitative association rules with no modification.

5.2 Post-Guess Algorithm

It is realized that not all items in a database would have sufficient support counts. Therefore, there is a waste to guess values of those items which will not be included at the end. The *Post-Guess Algorithm* is designed not to guess the missing values at the

```
Given a database $D$ with $T$ transactions and $M$ quantitative items. Find out the covariance matrix $C$.
1. For $i = 1$ to $M$
    • Initialize colavg[i] = 0.0
    • For $j = 1$ to $M$
        – Initialize C[i][j] = 0.0
2. For each transaction $t$ in $D$
    • colavg[i] = colavg[i] + t[i] for all (t[i] > 0)
    • For $j = 1$ to $M$
        – C[i][j] = C[i][j] + t[i]*t[j]
          if ((t[i] > 0) and (t[j] > 0))
3. For $i = 1$ to $M$
    • colavg[i] = colavg[i]/ $|D|$
4. For $i = 1$ to $M$
    • For $j = 1$ to $M$
        – C[i][j] = C[i][j] - $|D|$ * colavg[i]*colavg[j]
```

Figure 1: Pre-Guess Algorithm: Finding the Covariance Matrix.

beginning, but to do it after the initial clustering step.

Suppose we consider each transaction as a multi-dimensional point and each point can be of different dimensions. In this case, the direct application of the clustering distances cannot be used [11]. We propose to normalize the distance measurements in order to allow clustering points (transactions) of different dimensions. For example, the *Mahanttan distance* D1 between two clusters C1[X] and C2[X] of dimension k is re-defined as:

$$\text{D1}(\text{C1}[\text{X}], \text{C2}[\text{X}]) = |X0_1 - X0_2|/k$$

where $X0_1$ and $X0_2$ are the centroids of the clusters C1 and C2 respectively.

As in the DBA algorithm, there are two phases in the mining step for the Post-Guess Algorithm. During the clustering phase, when transactions are added to the CF-tree, the single covariance matrix similar to that in the Pre-Guess Algorithm is computed. In addition, the support counts of individual items are recorded. When a transaction with missing values is encountered in this phase, it is put aside in a separate storage area (S). After the CF-tree is built, each transaction in S is taken out. Its missing values are predicted by utilizing the covariance matrix as described before, and the transaction is inserted into the closest cluster in the CF-tree finally. Note that for items whose support counts are less than the threshold, their values will not be estimated. After the final clustering, we continue with the second phase as described to discover the quantitative association rules.

5.3 Deviation Algorithm

Estimating the missing values require the construction of the covariance matrix, finding of the eigensystem and then the final prediction steps. All these steps demand non-trivial computational effort. In discovering quantitative association rules, we are not really interested in the actual values associated with the items. Instead, a good guess of its possible range is often sufficient.

In the *Deviation Algorithm*, we propose to drop the PCA but to exploit the clustering property offered by Birch in the first phase of the mining step. The idea is similar to the work done by Lakshminarayan et al. except that we are not using decision trees nor the Bayesian approach. We assume that there will not be a lot of missing values in the database; otherwise it will be impossible to do any meaningful data mining. Further, if values are missing systematically, it would be easier to discover and correct them accordingly. On the other hand, if only a small number of values are missing randomly, then they can be easily approximated by their nearby information (clusters).

In the clustering phase, when a transaction with missing values is encountered, it is put aside in a separate storage area (S) as in the previous algorithm. However, after the CF-tree is built, each transaction in S is inserted into the closest cluster in the CF-tree directly. Any missing values of the transaction will be imputed by the centroids the corresponding dimensions. The method is similar to the mean substitution, but it only considers transactions within the same cluster rather than the complete database. After the final clustering, we continue with the second phase to discover the quantitative association rules afterwards.

6 EXPERIMENTS

In order to assess the performance of the algorithms, we have implemented them in C++. The preliminary experiments have been run on a Sun Ultra workstation with 64 Mbytes of main memory. Initially, a data set of 100,000 transactions is created. Each transaction contains two groups of data items. Each item in the first group contains a pair of num-

- Let P be the predication accuracy and set it to 0 initially.
- For each rule R_i in A
    - Find a rule R_j in B with the largest overlapping of items.
    - For each pair of predicates involving the same item in R_i and R_j, calculate the ratio of their interval overlaps. Add up all these ratios and divided the sum by the number of predicates in R_j.
    - Add the resultant value to P.
- $P = P/|B|$

Figure 2: Prediction Accuracy.

bers generated randomly. The first number represents the item number and the second value represent the item's value over the interval [0,1000]. The items in the second group is similar to those in the first group except their item values are calculated with the item values in the first group by using some linear formulae.

During the experiments, we randomly select 10% (i.e. 10,000) of the transactions to have missing values. In the first experiment, the 10% transactions will have only 1 item with missing value. The second experiment will have two items with missing values and so on, until there are 5 items with missing values at the end. During the experiments, we measure the total times needed for the mining and the prediction accuracies of the quantitative rules when compared with those discovered from the complete data set. Suppose we have two rule sets A and B discovered from the incomplete and complete data sets respectively. We can then compute the prediction accuracy as shown in Figure 2.

The results of our experiments are shown in Figure 3 and Figure 4. From these results, we can observe that the Deviation algorithm takes less time in mining the quantitative association rules, but its accuracies are comparable with the PCA algorithm.

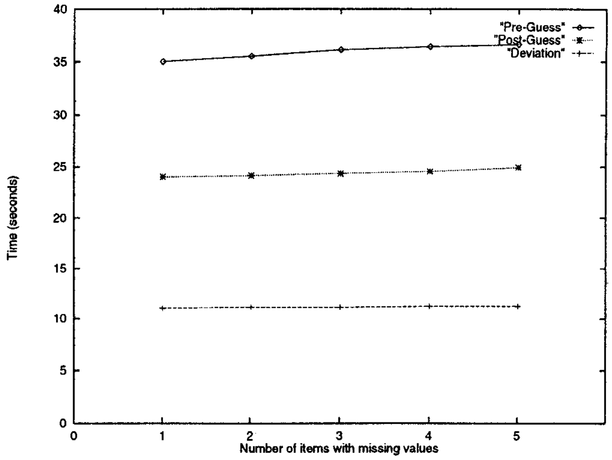

Figure 3: Total mining times.

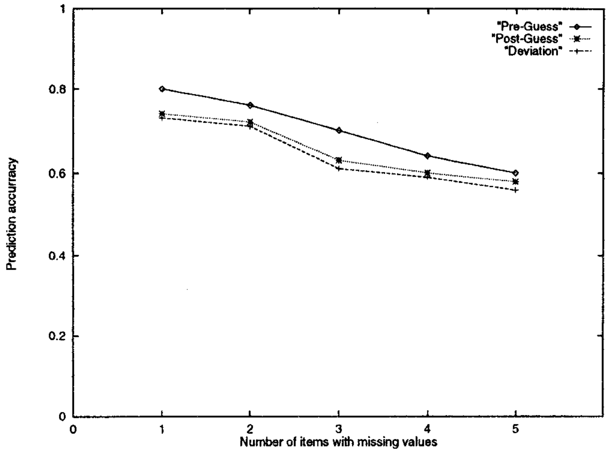

Figure 4: Prediction accuracies.

7 CONCLUSIONS

This paper explores the use of principle component analysis (PCA) to estimate missing values during the mining of quantitative association rules. We have presented three algorithms. The *Pre-Guess* algorithm performs the imputation before mining. The *Post-Guess* algorithm utilizes principle component analysis to estimate the missing values after the mining process. The last algorithm, *Deviation* algorithm, does not estimate the actual value of a missing data, but the possible range of the value is guessed. Preliminary experimental results have shown the comparable accuracies of the algorithm and its great savings in the computational effort.

References

[1] R. Agrawal, and R. Srikant. "Fast Algorithms for Mining Association Rules in Large Databases", *Proc. of the 20th International Conference on Very Large Data Bases*, pp. 478-499, September 1994.

[2] R. Agrawal, T. Imielinski, and A. Swami. "Mining Association Rules between Sets of Items in Large Databases", *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 207-216, May 1993.

[3] J.F. Hair Jr., R.E. Anderson, R.L. Tatham, and W.C. Black. *Multivariate Data Analysis*, 4th Edition, Prentice Hall, 1995.

[4] G. Piatetsky-Shapiro, and W.J. Frawley. *Knowledge Discovery in Databases*, AAAI/MIT Press 1991.

[5] J. Han, Y. Fu, Y. Huang, Y. Cai, and N. Cercone. "DBLearn: A System Prototype for Knowledge Discovery in Relational Databases", *ACM-SIGMOD*, Minneapolis, MN, May 1994.

[6] J. Han, Y. Yu. "Discovery of multiple-level association rules from large databases", *Proc. of International Conference on Very Large Databases*, Zurich, Switzerland, pp. 420-431, September 1995.

[7] F. Korn, A. Labrinidis, Y. Kotidis, C. Faloutsos, A. Kaplunovich, and D. Perkovic. "Quantifiable Data Mining Using Principle Component Analysis", *Technical Report*, Department of Computer Science, University of Maryland at College Park, 1996.

[8] K. Lakshminarayan, S.A. Harp, R. Goldman and T. Samid. "Imputation of Missing Data using Machine Learning Techniques", *Proc. of KDD*, Portland, USA, pp. 140-145, August 1996.

[9] R.J. Miller and Y. Yang. "Association Rules over Interval Data", *SIGMOD 1997*, pp. 452-461.

[10] R. Srikant and R. Agrawal. "Mining Quantitative Association Rules in Large Relational Tables", *ACM SIGMOD*, 1996.

[11] T. Zhanf, R. Ramakrishnan and M. Livny. "BIRCH: An Efficient Data Clustering Methods for Very Large Databases", *ACM SIGMOD*, 1996.