Classification using Association Rule Mining

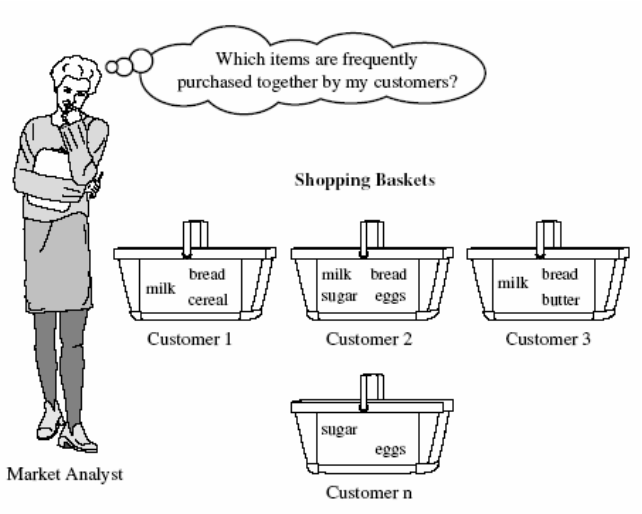
Sheela Singhal[2], Dr. G.N. Singh[2]
[1]Assistant Professor, Chandigarh Group of Colleges, Landran (Punjab)
[2]Department of Computer Science, Sudarshan Degree College, Lalgaon (M.P.)India
sheela_singhal@yahoo.com

Abstract: With wide applications of computers and automated data collection tools, massive amounts of data have been continuously collected and stored in databases, which creates an imminent need and great opportunities for mining interesting knowledge from data. Association rule mining is one kind of data mining techniques which discovers strong association or correlation relationships among data. The discovered rules may help market basket or cross-sales analysis, decision making, and business management.

Association Rule Mining

Association rule mining is the discovery of association relationships among a set of items in a dataset. Association rule mining has become an important data mining technique due to the descriptive and easily understandable nature of the rules. Al-though association rule mining was introduced to extract associations from market basket data, it has proved useful in many other domains (e.g. microarray data analysis, recommender systems, and network intrusion detection). In the do-main of market basket analysis, data consists of transactions where each is a set of items purchased by a customer. A common way of measuring the usefulness of association rules is to use the support-confidence framework.

Association rule mining finds interesting association or correlation relationships among a large set of data items. It first discovers frequent itemsets satisfying user-defined minimum support, and then from which generates strong association rules satisfying user-defined minimum confidence. The most famous algorithm for association rule mining is Apriori algorithm. Most of the previous studies on association rule mining adopt the Apriori-like candidate set generation-and-test approach. Apriori algorithm uses frequent $(k-1)$-itemsets to generate candidate frequent k-itemsets and use database scan and pattern matching to collect counts for the candidate itemsets. Recently, J. Han et al critiqued that the bottleneck of Apriori algorithm is the cost of the candidate generation and multiple scans of database. Han's group developed another influential method for discovering frequent pattern without candidate generation, which is called frequent pattern growth (FP-growth). It adopts divide-and-conquer strategy and constructs a highly compact data structure (FP-tree) to compress the original transaction database. It focuses on the frequent pattern (fragment) growth and eliminate repeated database scan. The performance study by Han's group shows that FP-growth is more efficient than Apriori algorithm.



Market Basket Analysis

Classification

Classification is another method of data mining. Classification can be defined as learning a function that maps (classifies) a data instance into one of several predefined class labels. The data from which a classification functions or model is learned is known as the training set. A separate testing set is used to test the classifying ability of the learned model or function. Examples of classification models include decision trees, Bayesian models, and neural nets. When classification models are constructed from rules, often they are represented as a decision list (a list of rules where the order of rules corresponds to the significance of the rules). Classification rules are of the form $P \rightarrow c$, where P is a pattern in the training data and c is a predefined class label (target).

Classification rule mining is to build a class model or classifier by analyzing predetermining training data and apply the model to predict the future cases. Besides other techniques for data classification such as decision tree induction, Bayesian classification, neural network, classification based on data warehousing technology, and etc, the associative classification or classification based on association rules is an integrated technique that applies the methods of association rule mining to the classification. It typically consists of two steps. The first step finds the subset of association rules that are both frequent and accurate using association rule techniques. The second step employs the rules for classification.

Association Rule Based Classification

The associative classification algorithm can be divided into two fundamental parts: association rule mining and classification. The mining of association rules is a typical data mining task that works in an unsupervised manner. A major advantage of association rules is that they are theoretically capable of revealing all interesting relationships in a database.

Recently, Bing Liu et al proposed Classification Based on Association rules (CBA) algorithm as an integration of classification rule mining and association rule mining. The integration was done by finding a special subset of association rules called class association rules (CARs) and building a classifier from the CARs. The main strength of CBA algorithm is its ability to use the most accurate rules for classification, which explains its better performance compared with some original classification algorithms such as C4.5. Liu's research group also proposed some methods to deal with the problems of the original CBA algorithm such as single minimum support and not being able to generate long rules for many datasets. The performance of the algorithm was improved by using multiple minimum support (S_{min}) instead of a single S_{min}, and combining CBA algorithm with other techniques such as decision tree method. More recently, Wenmin Li et al critiqued some weakness of Liu's approach as follows: (1) simply selection a rule with a maximal user-defined measure may affect the classification accuracy, (2) the efficiency problem of storing, retrieve, pruning, and sorting a large number of rules for classification when there exist a huge number of rules, large training data sets, and long pattern rules. They proposed a new associative classification algorithm: Classification based on Multiple Association Rules (CMAR). The experimental results show that CMAR provides better efficiency and accuracy compared with CBA algorithm. The accuracy of CMAR is achieved by using multiple association rules for classification. The efficiency of CMAR is achieved by extension of efficient frequent pattern method, FP-growth, construction of a class distribution-associated FP-tree, and applying a CR-tree structure to store and retrieve mined association rules.

The main issues on the integration of association and classification are:

(1) efficiency and accuracy:
how to efficiently find out the high quality rules using association rule mining and how to generate more accurate classifier,

(2) scalability:
it is important when there exist large training data sets, huge number of rules and long pattern rules. The efficiency and accuracy typically affect each other. We need to balance these two issues.

Several parallel mining algorithms for association rules exist in literatures. T. Shintani et al suggested a parallel algorithm for mining association rules with classification hierarchy on a shared-nothing parallel machine. It partitions the candidate itemsets over the processors. It uses a hierarchic structure to minimize inter-processor communication. In the classification hierarchy, all the candidate itemsets whose root items are identical will be

allocated to the identical node, which eliminates communication of the ancestor items. It identifies the frequently occurring candidate itemsets and copy them over all the processors, through which frequent itemsets can be processed locally without communication, which saves memory space.

Conclusion:

In this paper, we have proposed an association rule based classification, which is to be separable in two components:

1. Generation of Association Rules
2. Classification based on these rules.

References:

1. *Data Mining: Concepts and Techniques*, Jiawei han and Micheline Kamber, Morgan Kaufmann Publishers, San Francisco, CA, 2000.
2. Rakesh Agrawal, Ramakrishnan Srikant: "Fast Algorithms for Mining Association Rules in Large Databases." VLDB 1994.
3. J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", SIGMOD, 2000.
4. Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining." KDD-98, 1998.
5. Bing Liu, Yiming Ma, C-K Wong, "Classification Using Association Rules: weaknesses and Enhancements." To appear in Vipin Kumar, et al, (eds), *Data mining for scientific applications*, 2001.
6. Bing Liu, Yiming Ma, Ching Kian Wong, "Improving an Association Rule Based Classifier", PKDD-2000, 2000.
7. W. Li, J. Han, and J. Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules", ICDM'01, 2001.
8. Takahiko Shintani and Masaru Kitsuregawa "Parallel Mining Algorithm for Generalized Association Rules with Classification Hierarchy", SIGMOD, 1998.
9. "Meta-Learning in Distributed Data Mining Systems: Issues and Approaches", Andreas L. Prodromidis, Salvatore J. Stolfo and Philip Chan, "Advances of Distributed Data Mining" book, editors Hillol Kargupta and Philip Chan, AAAI press, August 2000.
10. Aleksandar Lazarevic and Zoran Obradovic "The Distributed Boosting Algorithm", KDD01, 2001.