IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), PP 50-51
www.iosrjournals.org

Improving Efficiency of Association Rules by Means of Classification

L Kiran Kumar Reddy[1], Dr.S.Phani Kumar[2]

[1]Research Scholar Department of Computer Science and Engineering GITAM University, Hyderabad, India
[2]Prof & Head Department of Computer Science and Engineering GITAM University, Hyderabad, India

Abstract: *Data Mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis. Information mining in enterprise applications facing challenges due to the complex data distribution in large heterogeneous sources. In such scenario, a single approach or method for mining, limited the information needs and it also will be high processing and time consuming. It is necessary to develop an effective mining approach which can be useful for the real time business requirements and decision making tasks.*

Keywords: *Data mining, Association rules, Classification, Associative classification.*

I. Importance of the problem

Data mining is used to deal with large amounts of data which are stored in the database, to find out desired information and knowledge [2]. It commonly consists of various algorithms namely clustering, classification, association rule mining and more. Among these algorithms, Association Rule Mining (ARM) is one of the most important techniques in the data mining.

Since its introduction, Association Rule Mining [1], has become one of the core data mining tasks, and has attracted tremendous interest among data mining researchers and practitioners. It has an elegantly simple problem statement, that is, to find the set of all subsets of items (called itemsets) that frequently occur in many database records or transactions, and to extract the rules telling us how a subset of items influences the presence of another subset. The prototypical application of associations is in market basket analysis, where the items represent products and the records the point-of-sales data at large grocery or departmental stores. These kinds of database are generally sparse, i.e., the longest frequent itemsets are relatively short. However there are many real-life datasets that very dense, i.e., they contain very long frequent itemsets.

Building accurate and efficient classifiers for large databases is one of the essential tasks of data mining and machine learning research. Given a set of cases with class labels as a training set, classification is to build a model (called classifier) to predict future data objects for which the class label is unknown.

II. Status on the Problem in Literature

Previous studies have developed heuristic/greedy search techniques for building classifiers, such as decision trees [10], rule learning [2], naive-Bayes classification [4], and statistical approaches [8]. These techniques induce a representative subset of rules (e.g., a decision tree or a set of rules) from training data sets for quality prediction.

Recent studies propose the extraction of a set of high quality association rules from the training data set which satisfies certain user-specified frequency and confidence thresholds. Effective and efficient classifiers have been built by careful selection of rules, e.g., CBA [9], CAEP [3], and ADT [11]. Such a method takes the most effective rule(s) from among all the rules mined for classification. Since association rules explore highly confident associations among multiple variables, it may overcome some constraints introduced by a decision-tree induction method which examines one variable at a time.

Extensive performance studies [6, 9, 3, 11] show that association based classification may have better accuracy in general. However, this approach may also suffer some weakness. On one hand, it is not easy to identify the most effective rule at classifying a new case. Some method, such as [9, 3, 11], simply selects a rule with a maximal user-defined measure, such as confidence.

A training data set often generates a huge set of rules. It is challenging to store, retrieve, prune, and sort a large number of rules efficiently for classification. Many studies [1, 5] have indicated the inherent nature of a combinatorial explosive number of frequent patterns and hence association rules that could be generated when the support threshold is small (i.e., when rare cases are also be included in the consideration). To achieve high accuracy, a classifier may have to handle a large set of rules, including storing those generated by association mining methods, retrieving the related rules, and pruning and sorting a large number of rules.

DOI: 10.9790/0661-1806055051

III. Objectives of the Research

Previous studies propose that associative classification has high classification accuracy and strong flexibility at handling unstructured data. However, it still suffers from the huge set of mined rules and sometimes biased classification or over fitting since the classification is based on only single high-confidence rule. The objective of the research work is to develop a new associative classification method for accurate and efficient classification. The method extends an efficient frequent pattern mining method, FP-growth, constructs a class distribution-associated FP-tree, and mines large database efficiently and make the following contributions.

First, instead of relying on a single rule for classification, we determine the class label by a set of rules. Given a new case for prediction, we select a small set of high confidence, highly related rules and analyzes the correlation among those rules.

Second, to improve both accuracy and efficiency, we employ a novel data structure, CR-tree, to compactly store and efficiently retrieve a large number of rules for classification. CR-tree is a prefix tree structure to explore the sharing among rules, which achieves substantial compactness. CR-tree itself is also an index structure for rules and serves rule retrieval efficiently.

IV. Methodology

The research methodology initiates with revisiting the general idea of associative classification to the generation of rules for classification. Using the generated rules we perform how to classify a new data object. The experimental results on classification accuracy and the performance study on efficiency and scalability will be made on completion. The steps of methodology process are as follows.

A. Associative Classification
B. Generating Rules for Classification
i. Mining Class Association Rules Passing Support and Confidence Thresholds
ii. Storing Rules in CR-tree
iii. Pruning Rules
C. Classification Based on Multiple Rules
D. Experimental Results and Performance Study

Status of the work

Based on the illustrated process methodology we currently focused on the part of Associative Classification. In Associative Classification, with a given training data set, the task of classification is to build a classifier from the training data set such that it can be used to predict class labels of unknown objects with high accuracy.

Besides many different approaches for classification, such as decision tree approach, naive Bayesian approach, K-nearest neighbors approach, neural network approach, a new approach is to explore association relationships between object conditions and class labels [9]. The idea is new since it utilizes frequent patterns and association relationships between cases and class labels in training data set to do classification. If strong associations among some frequent patterns and class labels can be observed in training data set, the future object of similar patterns can be classified.

Currently we are working on the Generating Rules for Classification for generating association rules.

V. Future Plans

In future work, to speed up the mining of complete set of rules, we adopt a variant of recently developed FP-growth method. FP-growth is much faster than Apriori-like methods used in previous association-based classification, such as [9, 3, 11], especially when there exist a huge number of rules, large training data sets, and long pattern rules.

References

[1]. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In VLDB'94, Chile, Sept. 1994.
[2]. P. Clark and T. Niblett. The CN2 induction algorithm. Machine Learning, 3:261–283, 1989.
[3]. G. Dong, X. Zhang, L. Wong, and J. Li. Caep: Classification by aggregating emerging patterns. In DS'99 (LNCS 1721), Japan, Dec. 1999.
[4]. R. Duda and P. Hart. Pattern Classification and Scene Analysis. John Wiley & Sons, 1973.
[5]. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In SIGMOD'00, Dallas, TX, May 2000.
[6]. B. Lent, A. Swami, and J. Widom. Clustering association rules. In ICDE'97, England, April 1997.
[7]. W. Li. Classification based on multiple association rules. M.Sc. Thesis, Simon Fraser University, April 2001.
[8]. T.-S. Lim, W.-Y. Loh, and Y.-S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning, 39, 2000.
[9]. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In KDD'98, New York, NY, Aug. 1998.
[10]. J. R. Quinlan. C4.5: Programsfor Machine Learning. Morgan Kaufmann, 1993.
[11]. K. Wang, S. Zhou, and Y. He. Growing decision tree on support-less association rules. In KDD'00, Boston, MA, Aug. 2000.