ASSOCIATIVE CLASSIFICATION USING CBA ALGORITHM

Thiri Htein

Computer University (Mandalay)

Abstract

Applying the association rule into classification can improve the accuracy and obtain some valuable rules and information that cannot be captured by other classification approaches. However, the rule generation procedure is very time-consuming when encountering large data set. Besides, traditional classifier building is organized in several separate phases which may also degrade the efficiency of these approaches. Association rules have proved to be useful in building both partial and complete classification models. This system analyzes alternatives measures which could replace confidence in order to evaluate the suitability of a given association rule with respect to the classification problem by using classification based on association (CBA) algorithm.

The system integrates the classification and association rule mining. The system is intended to implement classification models by using associative classification algorithms. An algorithm is presented to generate all class association rules (CARs) and to build an accurate classifier. The system is implemented by using Java Programming Language and MySQL.

1.Introduction

Each year more operations are being computerized; all accumulate data on operations, activities and performance. All these data hold valuable information, e.g. trends and patterns, which could be used to improve business decisions and optimize success.

However, today's databases contain so much data that it becomes almost impossible to manually analyze them for valuable decision-making information. In many cases, hundreds of independent attributes need to be simultaneously considered in order to accurately model system behavior. Therefore, humans need assistance in their analysis capacity.

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing.

The term data mining has mostly been used by statisticians, data analysts, and the management information systems

(MIS) communities. It has also gained popularity in the database field.

Data mining techniques can be classified into the following categories: classification, clustering, association rules, sequential patterns, time-series patterns, link analysis and text mining [8, 25].

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to look up large database in order to find novel and useful patterns that might otherwise be unknown.

Association rules are used to build partial classification models in domains where conventional classifiers would be ineffective. For example, traditional decision trees are problematic when many values are missing and also when the class distribution is very skewed.

The system integrates the classification and association rule mining. The system is intended to implement classification models by using associative classification algorithms. An algorithm is presented to generate all class association rules (CARs) and to build an accurate classifier for supermarket sales amount on the sales transactional data.

2.Theorical Background

The rapid emergence of electronic data management methods has led some to call recent times as the "Information Age." Powerful database systems for collecting and managing are in use in virtually all large and mid-range companies -- there is hardly a transaction that does not generate a computer record somewhere. Each year more operations are being computerized; all accumulate data on operations, activities and performance.

However, today’s databases contain so much data that it becomes almost impossible to manually analyze them for valuable decision-making information. In many cases, hundreds of independent attributes need to be simultaneously considered in order to accurately model system behavior.

Therefore, humans need assistance in their analysis capacity. Data mining is techniques to discover strategic information hidden in very large databases [5, 20].

Association Rule Mining finds interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected or stored, many industries are becoming interested in mining association rules from their databases.

The discovery of interesting association relationship among huge amount of business transaction record can help in many business decision-making processes, such as catalog design, cross marketing and loss-leader analysis.

Multilevel association rules can be mined using several strategies, based on how minimum support thresholds are defined at each level of abstraction. When using reduced minimum support at lower levels, pruning approaches include level-cross filtering by single item and level-cross filtering by k-itemset. Redundant multilevel (descendent) association rules can be eliminated from presentation to the user if their support and confidence are close to their expected values, based on their corresponding ancestor rules.

Techniques for mining multidimensional association rules can be categorized according to their treatment of quantitative attributes. First, quantitative attributes may be discretized statically, based on predefined concept hierarchies. Data cubes are well suited to this approach, since both the data cube and quantitative attribute can make use of concept hierarchies. Second, quantitative association rules can be mined where quantitative attributes are discretized dynamically based on binning, where adjacent association rules may be combined by clustering. Third, distance-based association rules can be mined to capture the semantics of interval data, where intervals are defined by clustering.

One of the most important problems in database mining is mining association rules within a database so called the "basket data analysis" problem. Basket data type typically consists of a transaction identifier and the bought items par-transaction. By analyzing translation data, we can extract the association rule such as "90% of the customers who buy both A and B by C ".

Association Rule Mining (ARM) is concerned with how items in a transactional database are grouped together. It is commonly known as market basket analysis, because it can be likened to the analysis of items that are frequently put together in a basket by shoppers in a market.

Association rule mining searches for interesting relationships among items in a given data set. Market basket analysis is just one form of association rule mining. In fact, there are many kinds of association rules. Association rules can be classified in various ways, based on the following criteria [6, 229].

3.Association Rule in Classification

Classification models can be directly built from traditional association rules, using association rules where the class appears in their right-hand side. In the following section describes some of the approaches followed to build classification models from association rules.

Some fundamental differences exist between classification and association rules discovery. Association rule do not involve prediction, nor do they provide any mechanism to avoid underfitting and overfitting apart from the crude minimum support threshold. In classification problem, inductive bias is also needed to solve classification problems, i.e. a basis for favoring one hypothesis over another. This bias, like any other bias, must be domain-dependant. Association rules have, however, been used to solve classification problems directly.

Association rules are used to build partial classification models in domains where conventional classifiers would be ineffective. For example, traditional decision trees are problematic when many values are missing and also when the class distribution is very skewed.

A tree of rules is built from an arbitrary set of association rules without using an ad-hoc minimum support threshold.

The method extends an efficient frequent pattern mining method, FP-growth, constructs a class distribution-associated FP-tree, and mines large databases efficiently. Moreover, it applies a CR-tree structure to store and retrieve mined association rules efficiently, and prunes rules effectively based on confidence, correlation and database coverage. The classification is performed based on a weighted analysis using multiple strong association rules. CMAR is consistent, highly effective at classification of various kinds of databases and has better average classification accuracy in comparison with CBA and C4.5. Moreover, our performance study shows that the method is highly efficient and scalable in comparison with other reported associative classification methods.

CPAR integrates the features of associative classification in predictive classification. CPAR employs the following feature to further improve its accuracy and efficiency: (1) CPAR uses dynamic programming to avoid repeated calculation in rule generation and (2) when generating rules, instead of selecting only the best literal, all the close-to-the best

literals are selected so that important rules will not be missed. CPAR generates a smaller set of rules, with higher quality and lower redundancy in comparison with associative classification. As a result, CPAR is much more time-efficient in both rule-generation and predication but achieves a high accuracy as associative classification.

Emerging patterns (EPs) are item sets whose supports change significantly from one dataset to another; they were recently proposed to capture multi-attribute contrasts between data classes, or trends over time. new classifier, CAEP, using the following main ideas based on EPs: (i) Each EP can sharply differentiate the class membership of a (possibly small) fraction of instances containing the EP, due to the big difference between its supports in the opposing classes; we define the differentiating power of the EP in terms of the supports and their ratio, on instances containing the EP. (ii) For each instance t, by aggregating the differentiating power of a fixed, automatically selected set of EPs, a score is obtained for each class. The scores for all classes are normalized and the largest score determines t's class. CAEP is suitable for many applications, it does not depend on dimension reduction on data; and it is usually equally accurate on all classes even if their populations are unbalanced.

CBA, Classification Based on Association [3], builds complete classification models. All class association rules are extracted from the available training dataset (i.e. all the association rules containing the class attribute in their consequence, and the most adequate rules are selected to build on " associative classification model " which uses a difficult class to make it complete. This classifier builder uses a brute-force exhaustive global search, and yields excellent results. CBA performance was improved allowing multiple minimum support thresholds for the different problems classes and reverting to traditional classifiers when no accurate rules are found.

A similar strategy to that of CBA is used to classify text documents into topic hierarchies. All the generalized association rules with the class attributes in their consequence are extracted, these rules are ranked, and some of them are selected to build a classifier which takes context into account, since class proximity is important when classifying documents into topics.

4.System Design and Implementation

The system implements classification system with the following associative classification model.

The associative classification model classifies the class label with three steps. Firstly, it generates rules for building classifier according to the user input. And then, these rules are classified by classifier and then it generates the class label for classification. The proposed system implements the above diagram. The working steps are described as follows. First of all, the user inputs the desired attribute values for classification. Depending on user's attributes, the CBA algorithm will produce frequent item sets with minimum support and confidence by using training dataset. Then, CBA algorithm produces class association rules and classifies the final class label for input attributes. With the result, the system will show the output class label to the user.

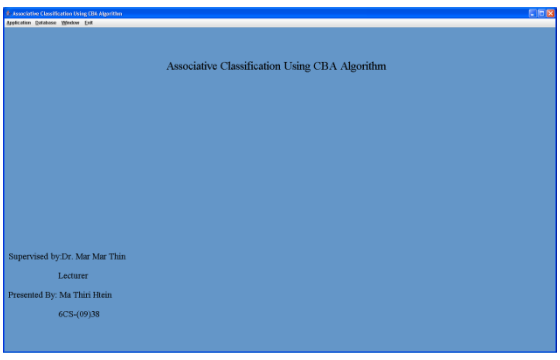

Figure 4.3 Welcome Screen

From the Application MenuItem from Menu Bar, the user can choose associative classification to start the calculation. If the user chooses associative classification menu item, the following interface will appear with database.

From the above interface, the user can choose the database that would like to classify. In the above example, the system has three tables to predict class label. After clicking the desired database, the user can start the calculation by clicking 'Start Process' Button. If so, Figure 4.4(A) Screen will appear.

From the above database, the user can choose desired attributes values, support and confidence to predict the class label by pressing Predict Button.

If the user wants to predict next support and confidence, the user can input next new support and confidence data and press 'Repeat Predict' button. If so, the system will show the class label with message box. The above 'Predict' and 'Repeat Predict' button show only the class label result exclusive of detailed calculation. If the user wants to see the real calculation, the user can see the result by clicking 'Display Result' button. At that time, the system will show the detailed calculation for previous chosen attributes, support and confidence. See Figure 4.4(D). If the user wants to quit from the system, the user can choose exit menu item from

the menu bar. If so, the user will confirm if he/she exits or not. At that time, the user can make the choice.

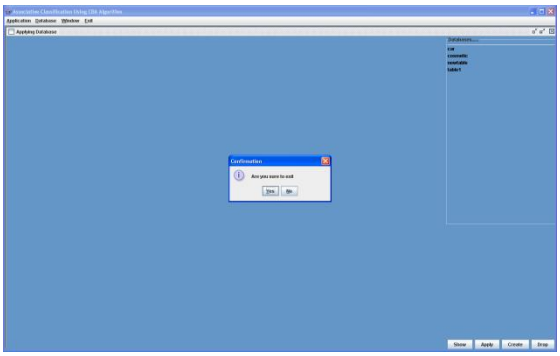

Figure 4.6 Exit Screen

5.Conclusions

Associative classification is a promising approach in data mining. Associative classifiers produce more accurate classification models than traditional classification algorithm such as decision trees and Naïve Bayesian.

The CARs produced by associative algorithm are simple if then rules which are easily understood by human. One challenge in associative classification is the exponential growth of rules; therefore, pruning becomes essential.

In this system, Classification Based on Association (CBA) is used to classify the sale amount (high, medium, low) depending on the transaction of supermarket sales items.

Depending on the supermarket's sales transaction, the system classifies the class label, sales amount by using CBA algorithm mixed with association and classification.

The significant factor of this CBA algorithm is that it classifies the transactional data with associative rules. So, instead of using traditional classification algorithm, CBA is more effective because it can predict the class label with multiple user-constraints support and confidence.

The CBA algorithm will predict the class label more in efficient manner other than other ordinary classification methods because of taking the advantages of association rules.

There are some limitations in this system. The system will only classify on the supermarket dataset. The associative item sets for one transaction may not be more than 20 items. The system will work well if the transactions are not more than ten thousands. If more, the system will produce the output with low speed because of the calculation time that will take more than traditional classification model because of associative rules.

This association classification can be implemented with many kinds of algorithms. In this system, classification

based on association algorithm is used to classify the input data item sets.

The further research can use other associative classification algorithms such as CMAR, Classification based on Multiple Association Rules, CPAR, Classification based on Predictive Association Rules, CAEP, and Classification by aggregating emerging patterns.

This application can be implemented with classification methods such as CMAR, CPAR, CAEP, and so on.

Moreover, CBA algorithm can apply on dynamic database that is any kind of database. In other ways, this CBA algorithm can undertake the advantages of all kinds of database and datasets.

6.References

[1] R. Agrawas, T. Imielinski, and A Swarni, Mining associative rules between sets of items in large database ACM SIGMOD Record, 22(2): 207-216, June1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc, 20th Int. Conf Very Large Data Bases, VLDB, pages 487-499, September 12-15 1994.

[3] K.ALi, S. Manganars, and R. Srikant. Partial classification using association rules. In Proceedings of the 3rd International Conference on Knowledge Discovery in Database and Data Ming, Neoport beach, Colifornia, USA, pages 115-118, August 14-17 1997.

[4] Fernando Berzal, Juan-Carlos Cubero, Nicolas Marin Daniel Sanchez, Jose-Maria Serrano, Amparo Vila. Association Rule Evaluation for Classification Purposes ETS Information. Universidad de Grananda. 18071 Granda.

[5] Michael Goebel, Le Gruenwald. A Survey of Data Mining and Knowledge Discovery Tools School of Computer Science, University of Auckland Private

Bag 92019 Anckland, New Zealand, University of Oklahoma 200 Felgar Street, Room 114, Norman, ok, 73019.

[6] Jiawei Han & Micheling Kamber. Data Mining Concepts and Techniques. ISBN 1-55860-489-8, Morgan Kaufmann Publishers.

[7] B. Lia, W. Hsu and Y. Ma. Integrating Classification and associative rule mining. In proceedings of the Fourth International Conference (KDD-98), New York City, USA, pages 80-86, August 27-31 1998.

[8] Chin-Ping Wei, Yen-Hsien Lee and Che-Ming Hsu Empirical Comparison of Fast Clustering Algorithm for Lage Data Sets. 33rd Hawaii International Conference on System Sciences-2000 National Sun Yat-Sen University.