Classification method based on Association Rule Mining

Divya Ramani[1] Chirag Pandya[2] Harshita Kanani[3]
[1, 2, 3]Assistant Professor
[1, 2, 3] Dept. of Computer Engineering
[1, 2, 3] LDRP, KSV, Gandhinagar, Gujarat

Abstract— In the field of data mining very large amount of data is processed in order to get small amount of useful data. There are two important data mining techniques to optimize efficiency, namely association rule mining and classification rule mining. In Data Mining, Classification is the process of finding and applying a model to describe and distinguish data classes, concepts and values. This work is a survey of major classification methods based on association rule mining. After this study better comparison of various classification methods can be done. After studying this all classification methods now, we develop a new method that name, BCAR in which we first apply fp-growth algorithm for rule generation and then calculate chi square value of each rule for subset selection after selecting this subset does classification based on chi square analysis. Main purpose is to build association rule classifier without loss of performance & accuracy of the resultant classifier.

I. Introduction

Data mining is a process or method of extracting interesting knowledge from large amounts of data. Data mining is one of the most important research fields that are due to the expansion of both computer hardware and software technologies, which has imposed organizations to depend heavily on these technologies. Data is considered as the number one asset of any organization, it is obvious that this asset should be used to predict future decisions. Consequently, and since organizations are continuously growing, their relative databases will grow as well; as a result their current data mining techniques will fail to cope up with large databases which are dynamic by nature .Data mining is the way to help organization make full use of the data stored in their databases, and when it comes to decision making, this is true in all fields, and is also true in all different types of organizations. [1]

II. Association rule mining

Association rule is a data mining technique which discovers the strong associations or correlation relationships among given data. Association Rule Mining aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. The major aim of ARM is to find the set of all subsets of items or attributes that frequently occur in many database records or transactions, and additionally, to extract rules on how a subset of items influences the presence of another subset. ARM algorithms discover high-level prediction rules in the form: IF the conditions of the values of the predicting attributes are true, THEN predict values for some goal attributes. In association rule mining there are so many methods. These methods work basically in two phases: the frequent itemset generation and the rule generation. Since the first phase is the most time consuming, all the association rule algorithms focus on the second phase. A set of attributes is termed as frequent set if the occurrence of the set within the dataset is more than a user specified threshold called minimum support. After discovering the frequent itemsets, in the second phase rules are generated with the help of another user parameter called minimum confidence. The task of mining association rules over market basket data is considered a core knowledge discovery activity. Association rule mining provides a useful mechanism for discovering correlations among items belonging to customer transactions in a market basket database. [2]

For an association rule$X \cup Y$, we can calculate

$$\text{Support } (X \cup Y) = \text{support } (XY) = \text{support}(X \text{ union} Y)$$
$$\text{Confidence } (X \cup Y) = \text{support } (XY) / \text{support}(X).$$

Support (S) and Confidence (C) can also be related to joint probabilities and conditional probabilities as follows.

$$\text{Support } (X \cup Y) = P(XY).$$
$$\text{Confidence } (X \cup Y) = P(Y/X).$$

III. Classification rules

Classification rules are one kind of conditional rules which can be used to discover data from large data sets.

A Classification Association Rule (CAR) is a rule of the form $X \rightarrow c$ where X is a set of attribute-values, and c is a class to which database records (instances) can be assigned. Mining of CARs usually proceeds in two steps. First, a training set of database records is mined to find all ARs for which one of the target classes is the consequent, and which satisfy specified thresholds of support and confidence. This stage is essentially similar to ARM in the more general case, with the class c treated as attribute-values, and the restriction that the only rules we need consider are those for which the consequent is one of these. A second stage then sorts and reduces the set of rules found, with the aim of producing a consistent set that will enable efficient and reliable classification of future instances. Identifying interesting rules from a set of discovered rules is not a simple task because a rule could be interesting to one user but not interesting to another.

There are two types of learning methods in classification.

Supervised learning (e.g. classification): The learning of the model is 'supervised' in that it is told to which class each training sample belongs.

Unsupervised Learning (e.g. clustering): In which the class

labels of the training samples are not known, and the number or set of classes to be learned may not be known in advance. [1]

IV. EXISTING METHODOLOGIES

Various Approaches Available for Association Rule Based Classification

As mentioned, association-rule based classification helps to discover classification rules more efficiently, we are in this paper also going to implement the association methods to discover classification rules. So let us take some light on the currently available approaches or methods concerning with association rule methods in depth.

V. COMPARISON BETWEEN ALL EXISTING CLASSIFICATION METHODS

Techniques	**Description**	**Advantages/ Merits**	**Disadvantages**
CBA (Classification Based On Associations) CBA-RG (Classification Based On Associations-rule generator) CBA-CB (Classification Based On Associations-classifier builder) [3]	CBA generates all the association rules with certain support and confidence thresholds which are known as candidate rules. Then it selects a small set of the rules from them to form a classifier. At the time of the predication of the class label of the example having highest confidence is used for the classification known as the best rule. In CBA-RG algorithm the data is scanned multiple times. In these multiple pass all the frequent rule items are generated. In the first pass it counts the support and determines that whether it is frequent or not. In each subsequent pass it starts with the seed set of rules generated and found to be frequent in the previous pass. It uses this set to generate new possibly frequent rules called the candidate rules. The actual support for these candidate rules are calculated during the pass. At the end of the pass it determines which of the candidate rule items are actually frequent which can produces the CARs. The CBA-CB algorithm used to build a classifier by using CARs. To produce the best classifier evaluation of all the possible subsets of the training data is done and selection of the subset with the right rule sequence with the least number of errors is selected. This is a heuristic algorithm but the classifier it builds performs very well as compared to that built by C4.5. [3]	This algorithm is simple, but is inefficient because it needs to make many passes over the database but it working better than the C4.5 classification system. One algorithm performs 3 tasks 1) It can find some valuable rules that existing classification systems cannot. 2) It can handle both table form data and transaction form data 3) It doesn't require the whole database to be fetched into the main memory. [3]	The limitations of this approach are as follows 1) It generates huge amount of the mined rule. 2) This leads to computational overhead. 3) The classification is done based on single high confidence rule which can be biased. CBA also suffer some weakness as shown below. 1) it is not easy to identify the most effective rule at classifying a new case. 2) a training data set often generates a huge set of rules. 3) In CBA-RG algorithm the data is scanned multiple times. [3]
CMAR (Classification based on Multiple Association Rules) [4]	CMAR is proposed in which the classification is performed based on a weighted analysis using multiple strong association rules. The classification is performed based on a weighted X^2 analysis using multiple strong association rules. It derives a good measure on how strong the rule is under both conditional support and class distribution. CMAR consists of two phases: rule generation and classification. 1) CMAR prunes some rules and only selects a subset of high quality rules for classification.	The CMAR outperforms both C4.5 and CBA on accuracy and it is also scalable. It improves both accuracy and efficiency CMAR uses a novel data structure CR-tree to compactly store and efficiently retrieve a large number of rules for classification. CMAR prunes some rules and only selects a subset of high quality rules for classification. [4]	The limitations are as follows, • CMAR is significant advance compare to the CBA but still it is very slower. • The overall accuracy can be further improved. [4]

	2) CMAR extracts a subset of rules matching the object and predicts the class label of the object by analyzing this subset of rules. If all the rules give same class label then it is classified. Otherwise the combined group effect will be taken into consideration. [4]		
CARGBA (Classification based on Association Rule Generated in a Bidirectional Approach) [5]	CARGBA generates the rules in two steps. 1) It generates a set of high confidence rules of smaller length with support pruning. Then augments this set with some high confidence rules of higher length with support below minimum support. The purpose is not knowledge extraction but to obtain better accuracy. 2) Rules are generated as specific as possible. They have higher length and therefore lower support and thus they easily capture the specific characteristics about the data set. So if there is a classification pattern that exists over very few instances or there are exceptions to the general rule, then it will be covered by the specific rules. Since these instances are small in number, specific rules are produced without any support pruning. This result is a better mixture of class association rules. All the rules generated by CARGBA rule generator will not be used in the classification. So, the second part builds a classifier with the essential rules and is called CARGBA Classifier Builder. [5]	CARGBA is consistent, highly effective at classification of various kinds of databases and has better average classification accuracy in comparison with C4.5, CBA and CMAR. 1) they easily capture the specific characteristics about the data set. 2) specific rules are produced without any support pruning. [5]	Overall accuracy can be less. Data will be scanned multiple times so, time complexity will be increased. [5]
CPAR (Classification based on Predictive Association Rules) [5]	CPAR adopts a greedy algorithm to generate rules directly from training data. To avoid over fitting it uses expected accuracy to evaluate each rule and uses the best k rules in prediction. CPAR inherits the basic idea of FOIL in rule generation also integrates the features of associative classification in predictive rule analysis. CPAR generates a smaller set of rules with higher quality and lower redundancy. So CPAR is much more time efficient in both rule generation and prediction. It also achieves as high accuracy as associative classification. To avoid generating redundant rules it generates each rule by comparing with the set of “already-generated" rules When predicting the class label it uses the best k rules. It uses dynamic programming to get better results. In rule generation instead of selecting only the best literal all the close-to-the-best literals are selected. [5]	• It generates a much smaller set of high-quality predictive rules directly from the dataset. • It generates a much smaller set of high-quality predictive rules directly from the dataset • To avoid generating redundant rules it generates each rule by comparing with the set of already-generated" rules when predicting the class label it uses the best k rules. • It uses dynamic programming to get better results. • In rule generation instead of selecting only the best literal all the close-to-the-best literals are selected. [5]	It is more complex for understand as well as implementation. It is needed to learn greedy algorithm before implementing CPAR.

CARPT (classification algorithm based on trie tree of associative rule) [6]	In this method first scene whole data base ones and convert it into two-dimensional array, in which the horizontal position said the item number and types of properties, the vertical position said the transaction number. According to the definition and the construction method of Trie-tree. **Property 1 of Trie-tree:** If a sub-tree takes a non- frequent bucket for root node, then all the buckets of the sub- tree are not frequent. **Property 2** In order <i1, i2, in>, there cannot be frequent itemset which contains two or more items take in for a prefix. When p>q, frequent item ip cannot take Iq for a prefix. Using property 1 and property 2, remove the frequent item that cannot generate frequent rules directly when transform the database into vertical bitmap of two-dimensional array to improve the achievement of Trie-tree and reduce the number of its nodes. [6]	CARPT cannot generate frequent rules directly by adding the count of class labels. By using this method the storage of database will be compress using the two-dimensional array of vertical data format also reduce the number of scanning the database. So, time and space can be saved effectively.	It removes frequent items. It is required to use two dimensional array of vertical data format.

Table. 1: Comparison Results

VI. PROPOSED METHOD

A. BCAR: Boosting on Multiple Classifiers Based on Association Rule Mining

BCAR, i.e. Boosting on multiple classifiers based on association rule mining. The method extends an efficient frequent pattern mining method, FP-growth, and mines large database efficiently. Select the subsets of rules and then does classification. This completes the first iteration then for the next iteration for selecting subset of rules change value of support and x2 and makes a new subset and classification is performed again after taking different subset. The classification is performed based on an x2 analysis using multiple strong association rules. At last Boosting will be applied on this multiple classifier and collect votes.

B. Develop a BCAR for accurate and efficient classification and make the following contributions.

First, instead of relying on a single rule for classification, BCAR determines the class label by a set of rules.

Second, to speed up the mining of complete set of rules, BCAR use FP-growth method.

C. Mining Class Association Rules Passing Support and Confidence Thresholds

BCAR first mines the training data set to find the complete set of rules passing certain support and confidence thresholds. To make mining highly scalable and efficient, BCAR adopts a variant of FP-growth method.

D. Algorithm of a proposed method

Step. 1 : Start

Step. 2 : Load a training data from the database that fits in the memory.

Step. 3 : Apply FP-Growth to find the frequent itemsets with the minimum threshold value.

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Step. 4 : Store frequent itemsets in X, where Suppose X is set of the frequent item set generated by FP-Growth algorithm.

Step. 5 : Calculate value of x2 for each rule.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N},$$

Step. 6 : Select minimum value of support and x2 based on that select subset of rules.

Step. 7 : Classification based on x2.

Step. 8 : If the desired number of classification model is not prepared.

Then go to Step 6.

Step. 9 : Vote for classification using boosting.

Step. 10 : Stop

– *Description of an Algorithm*

First start it and Load the training data from the database. In the third step for finding frequent itemsets with the minimum threshold apply FP-Growth method.

FP-Growth: allows frequent itemset discovery without candidate itemset generation. Two step approach:
Step 1: Build a compact data structure called the FP-tree
Built using 2 passes over the data-set.
Step 2: Extracts frequent itemsets directly from the FP-tree
FP-Tree is constructed using 2 passes over the data-set:
Pass 1:

Scan data and find support for each item.
Discard infrequent items.
Sort frequent items in decreasing order based on their support.
Use this order when building the FP-Tree, so common prefixes can be shared.

Pass 2:

Nodes correspond to items and have a counter
FP-Growth reads 1 transaction at a time and maps it to a path
Fixed order is used, so paths can overlap when transactions share items (when they have the same prefix).
In this case, counters are incremented Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines).

The more paths that overlap, then higher the compression. FP-tree may fit in memory.
Frequent itemsets extracted from the FP-Tree.

In the fourth step store this generated itemsets which is generated by FP-Growth algorithm. Generate association rules. In the fifth step calculate value of x2 for each rule. In the sixth step select minimum value of support and x2 and then select set of rules based on selected support and x2.
Where support and confidence are calculated as below:
Support: It is the probability of item or item sets in the given transactional data base:
Support(X) =n(X)/n

where n is the total number of transactions in the database and n(X) is the number of transactions that contains the item set X.
Confidence: It is conditional probability, for an association rule X=>Y and defined as
Confidence(X=>Y) = support(X and Y)/support(X).

In the eighth step it does classification on selected subset based on x2 value. Until the desired number of classification model not prepared repeat step six to nine.
After preparing multiple classification model votes for classification using boosting. In boosting, resampling is strategically geared to provide the most informative training data for each consecutive classifier. In essence, boosting creates three weak classifiers: the first classifier C 1 is trained with a random sub- set of the available training data. The training data subset for the second classifier C 2 is chosen as the most informative subset, given C 1. That is, C 2 is trained on a training data only half of which is correctly classified by C 1, and the other half is misclassified. The third classifier C 3 is trained with instances on which C 1 and C 2 disagree. The three classifiers are combined through a three-way majority vote. Finally for classification call BCAR to combine multiple classifiers and create more accurate one.

VII. Conclusion

In this paper, Finally by this study it can be understood that during this course of time new features are added in the original proposed approach in order to get better results. So, it is needed to build classification method based on association rule mining which gives better performance without loss of performance. BCAR has several features than existing methods 1] it gives better classification accuracy because of their iteratively subset selection. 2] it does effective classification among all type of datasets.

Acknowledgment

We would like to thanks my research guide Mrs. Harshita Kanani & Mr. Chirag Pandya for many discussions and for providing guidelines throughout my research work.

References

[1] Jiawei han, micheline kamber “Data mining – concept and techniques”
[2] Agrawal R, Imielinski T, and Swami “A., Mining association rules between set of items in large databases” In Proceedings of ACM SIGMOD, pages 207-216, May 1993
[3] Bing Liu, Wynne Hsu, Yiming Ma “Integrating classification and association rule mining” Department of Information Systems and Computer Science National University of Singapore Lower Kent Ridge Road, Singapore 119260-1998
[4] Wenmin Li Jiawei, Han Jian Pei_ “CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules” School of Computing Science, Simon Fraser University Burnaby, B.C., Canada
[5] Sohil Gambhir, Prof. Nikhil Gondliya “A Survey of Associative Classification Algorithms” International Journal of Engineering Research & Technology (IJERT)]
[6] Yang Junrui, Xu Lisha, He Hongde “A Classification Algorithm Based on Association Rule Mining” College
[7] of Computer Science and Technology Xi’an University of Science and Technology Xi’an, China-2012
[8] Yingqin Gu, Hongyan Liu, Jun He, Bo Hu and Xiaoyong Du “MrCAR: A Multi-relational Classification Algorithm based on Association Rules” Key Labs of Data Engineering and Knowledge Engineering, MOE, China Information School, Renmin University of China, Beijing, 100872, China School of Economics and Management, Tsinghua University, Beijing,100084
[9] Fadi Thabtah, Peter Cowling, Yonghong Peng “MCAR: Multi-class Classification based on Association Rule” Modelling Optimization Scheduling And Intelligent Control Research Centre University of Bradford, Department of Computing, University of Bradford, BD7 IDP, UK

