Rules generation by ARM and Classification based on Chi-square Analysis

Nilesh Solanki, Rikin Thakkar

***Abstract*-- Data mining become a very large area of research in past years. Several researches have been made in Classification and Association Rule Mining which are the techniques of Data mining. Associative classification is used to find a small set of rules in the database that forms an accurate associative classifier. Ensemble method combine multiple models into one usually more accurate than the best of its components. And association classifier improve the performance and accuracy of the resultant classifier. The paper surveys the most recent existing algorithm based on classification and some method which is based on Ensemble system.**

***Index Terms*-- Data Mining, Classification, Association Rule, Ensemble Method.**

I. INTRODUCTION

Data mining is a process of analyzing a large amount of data from different database and discover the relevant information It refers to extracting or mining knowledge from large amounts of data [1]. Data mining involve the following steps: cleaning and integration data from different data sources, pre-treatment of selecting and transformation target data, mining the required knowledge and in last steps evaluation and presentation of knowledge.

Association is defined as a relationship between the data. Association rule is the most important technique in data mining which discover strong relationships among given data. Association rule mining has a wide range of applicability such as market basket analysis, suspicious e-mail detection, library management and many areas[6]. Association rule is expressed in the form of X->y. X is called Antecedent and Y is called Consequent. it is a process of finding the all association rules that satisfy the minimum support and minimum support [7]. Support And confidence are two measure of rule interestingness.

Apriory Algorithm is used for Boolean association rules [5].it is used to find all frequent itemsets in a given database.it take a multiple over the passes over the database.it is based on breadth-first search through the search space, where k-itemsets are used to explore (k+1)-itemsets. In the first iteration, it will scan the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1.Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found.

Ensemble system combine a multiple models into one usually more accurate than the best of its component. The primary use of ensemble systems is the reduction of variance and increase in confidence of the decision. Due to Variations in a given classifier model the decision obtained by any given classifier may vary from one training trial to another even if the model structure is kept constant. So that time we should combine the output of several classifiers using ensemble system.

Associative classification is a combination of association rule mining and classical rule mining classical method. Classification rule mining is used to discover a small set of rules in the database that form an accurate classifier. association rule mining extract all rules which satisfy minimum support and minimum confidence constrain. The integration of two classical method is done by focusing on a special subset of association rules whose right-hand-side is restricted to the classification class attribute and on mining a special subset of association rules, called class association rules (CARs).

Data mining in the associative classification framework thus consists of three steps:[9]

• Discretization of continuous attributes, if any. Discretization can be done using any of standards discretization algorithms available in this standard literature.
• Generating all the class association rules (CARs), and
• Building a classifier based on the generated CARs.

II. PREVIOUS WORK

This algorithm use Frequent pattern growth for rule generation. In the FP-growth it will generate FP-tree. FP-tree does not generate a candidate set for the frequent item set. Classification is performed based on chi squared analysis using multiple association rules. CMAR use a CR-tree structure to store and retrieve mined association rules efficiently and prune rules effectively based on confidence , correlation and database coverage.[11]

MCAR is a new associative classification technique which extends the idea of association rule and integrates it with classification to generate a subset of effective rules that form a multi-class classifier. MCAR consists of two phases. In first MCAR filters the preparation information set to find regular single items, and after that recursively joins the items created to deliver items including more attributes. MCAR use

ranking method which is used to select the rules with high confidence are the part of the classifier. MCAR discover and generates frequent items and rules in one phase[14].

CARPT use Trie-tree which is used to remove the item that can not generate frequent rule directly by adding the count of class labels.it compress the storage of the database and reduce the number of database scan using two dimensional array of vertical data formate. Trie-tree reduce the cost of the query time to improve the efficiency [15].

III. PROPOSED ALGORITHM

Associative classification is an approach in data mining that utilizes the association rule discovery techniques to build classification systems, also known as associative classifiers. Ensemble methods have been called the most powerful development in data mining. They combine multiple classification models into one, usually more accurate one. Here in this thesis an efficient approach for classification using association rule ensemble is proposed. This presents an associative classification algorithm, which remove the frequent items that cannot generate frequent rules directly by adding the count of class labels. Main purpose is to ensemble association rule classifier without loss of performance & accuracy of the resultant classifier.

Here in this thesis using association rule mining build multiple classifiers and then vote for selecting best one boosting technique will apply on multiple classifiers.

1) Algorithm

Input

- Training dataset with Minimum Confidence

Output

- Classified Data

Procedure

1. Start
2. Load a training data from the database that fit in the memory.
3. Apply Apriori Algorithm to generate rule with the minimum threshold value.
4. Store all rules the in X, where X is set of rules generated by Apriori algorithm.
5. Calculate the support of the rule.
6. Calculate value of $\chi 2$ and assign class label based on minimum value of $\chi 2$.
7. Select subset of rules from rule list based on value of support and $\chi 2$
8. Calculate value of $\chi 2$ which gives co-relation between new data and selected rules.
9. Classify new data based on $\chi 2$ value.
10. If the desired number of classification model is not prepared,
    Then go to Step 7.
11. Vote for classification
12. Stop

Description of an Algorithm

First start it and Load the training data from the database. In the third step for generating rules with the minimum threshold apply Apriori method. The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.

Key Concepts :

• Frequent Itemsets: The sets of item which has minimum support (denoted by Li for ith-Itemset).

• Apriori Property: Any subset of frequent itemset must be frequent.

1 Join

- finding Lk, a set of candidate k-itemsets is generated by joining Lk-1 with itself
- The items within a transaction or itemset are sorted in lexicographic order For the (k-1) itemset: li[1]<li[2]<…<li[k-1]
- The members of Lk-1 are joinable if their first(k-2) items are in common
- Members l1, l2 of Lk-1 are joined if (l1[1]=l2[1]) and (l1[2]=l2[2]) and … and
- (l1[k-2]=l2[k-2]) and (l1[k-1]<l2[k-1]) – no duplicates
- The resulting itemset formed by joining l1 and l2 is l1[1], l1[2],…, l1[k-2],l1[k-1], l2[k-1]

2. Prune

- Ck is a superset of Lk, Lk contain those candidates from Ck, which are frequent Scanning the database to determine the count of each candidate in Ck – heavy computation
- To reduce the size of Ck the Apriori property is used: if any (k-1) subset of a
- candidate k-itemset is not in Lk-1, then the candidate cannot be frequent either,so it can be removed from Ck. – subset testing (hash tree)

In the fourth step store this generated rules which is generated by Apriori algorithm. Generate association rules. In the fifth step calculate value of support for each rule.

In the sixth step calculate value of $\chi 2$ and assign class label based on minimum value of $\chi 2$

Where support and confidence are calculated as below:

Support: It is the probability of item or item sets in the given transactional database:

$$Support(X)=n(X)/n$$

Where n is the total number of transactions in the database and n(X) is the number of transactions that contains the itemset X.

Confidence: It is conditional probability, for an association rule X=>Y and defined as

$$Confidence(X=>Y) = support(X \text{ and } Y)/support(X)$$

In the seventh step first arrange rules in descending order based on support value and ascending order of $\chi 2$ value then select best subset of rules from rule list.

In the eighth step calculate the value of $\chi 2$ which gives the co-relation between new data and selected subset of rules.

In the ninth step it does classification of new data based on $\chi 2$ value and select class label for a new data which has $\chi 2$ value minimum. Until the desired number of classification model not prepared repeat step seven to ten. After preparing multiple classification models votes for classification. The different classifiers are combined through a majority vote and select best one.Finally for classification call RCCA to combine multiple classifiers and create more accurate one.

Let the database of transactions consist of the sets {1, 2, 3, 4}, {1,2,3,4,5}, {2,3,4},{2,3,5}, {1,2,4}, {1,3,4}, {2,3,4,5}, {1,3,4,5}, {3,4,5}, {1,2,3,5}. Each number corresponds to a product such as "butter" or "water". The first step of Apriori is to count up the frequencies, called the supports, of each member item separately:

Item	Support
1	6
2	7
3	9
4	8
5	6

Fig.1.1 candidate 2-itemset

We can define a minimum support level to qualify as "frequent," which depends on the context. For this case, let min support = 4. Therefore, all are frequent. The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, Apriori prunes the tree of all possible sets. In next step we again select only these items (now 2-pairs are items) which are frequent (the pairs written in bold text):

Item	Support
{1,2}	4
{1,3}	5
{1,4}	5
{1,5}	3
{2,3}	6
{2,4}	5
{2,5}	4
{3,4}	7
{3,5}	6
{4,5}	4

Fig.1.2 frequent candidate -itemset

We generate the list of all 3-triples of the frequent items (by connecting frequent pair with frequent single item).

Item	Support
{1,3,4}	4
{2,3,4}	4
{2,3,5}	4
{3,4,5}	4

Fig.1.3 frequent candidate 2-itemset

The algorithm will end here because the pair {2, 3, 4, 5} generated at the next.

The number of rules generated by class-association rule mining can be huge. To make the classification effective and also efficient, it needs to prune rules to delete redundant and noisy information The rationale of this pruning is that we use the rules reflecting strong implications to do classification. By removing those rules not positively correlated, it prunes noise.

After a set of rules is selected for classification, we are ready to classify new objects. Given a new data object, we collect the subset of rules matching the new object from the set of rules for classification. Then determine the class label based on the subset of rules. Trivially, if all the rules matching the new object have the same class label, it just simply assigns that label to the new object.

If the rules are not consistent in class labels, we divide the rules into groups according to class labels. All rules in a group share the same class label and each group has a distinct label. We compare the effects of the groups and yields to the strongest group. To compare the strength of groups, we need to measure the “combined effect” of each group. Intuitively, if the rules in a group are highly positively correlated and have good support, the group should have strong effect.

There are many possible ways to measure the combined effect of a group of rules. For example, one can use the strongest rule as a representative. That is, the rule with highest x2 value is selected. However, simply choosing the rule with highest x2 value may be favorable to minority classes.

once the classifiers are generated, a strategy is needed to combine their outputs. In simple majority voting, a commonly used combination rule, each classifier votes on the class it predicts, and the class receiving the largest number of votes is the ensemble decision. It can be shown that if classifier outputs are independent, and each classifier predicts the correct class with a probability of one half or higher, the correct classification performance of the ensemble approaches one as the number of classifiers increases.

IV. Experiments And Results

In this thesis to test the performance of RCCA, we compared it with CBA, CMAR & C4.5 The experimental datasets using 8 datasets refer table 6.1 that come from the UCI machine learning database.

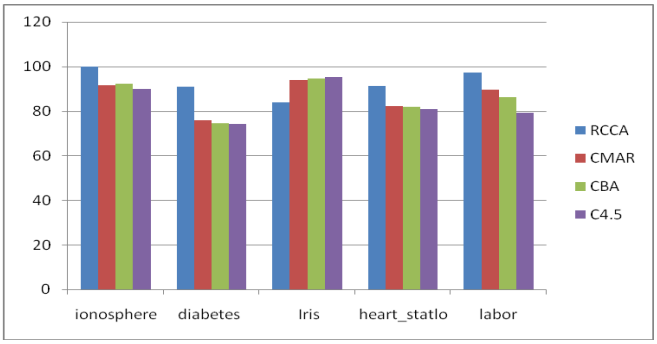

Fig.1.4 Comparison of CBA, CMAR, C4.5 & RCCA on Accuracy

So, at last there is average accuracy of RCCA algorithm will be 91.71. There is increase ratio of RCCA on accuracy will be minimum 8% to maximum 15%.

Datasets	#attribute	#class	#instance	RCCA
Iris	5	3	150	84.00
Ionosphere	35	2	351	100
Diabetes	9	2	768	91.04
Heart_statlog	14	2	270	91.25
Labor	17	2	57	97.29
Ecoli	8	8	348	83.92
Balance_scale	625	3	5	96.8
Hayes_roth	132	4	5	89.39

Fig.1.5 RCCA on Accuracy of 8 different datasets

Now, we need to measure all objective measures like TP-Rate, FP-Rate, Precision, Recall, F-Measure. And also measure comparison between NaiveBayes and RCCA on objective measure of 5 different datasets.

Iris	TP-Rate	FP-Rate	Precision	Recall	F-Measure
RCCA	0.84	0.08	0.93	0.84	0.85
Naïve Bayes	0.33	0.33	0.11	0.33	0.16

Fig.1.6 Comparison of RCCA & NaiveBayes on objective measure of iris dataset

From the table 1.6 we can see that accuracy is more than the existing methods. Accuracy is higher because we prepare more than one classification models and at last collect result of each classification models and vote for the classifiers. So, due to voting method and multiple classifiers accuracy is more.

V. CONCLUSION

In this thesis, we examined two major challenges in ensemble of classifiers based on association rule mining. 1] We can handle huge amount of rule effectively& 2]
We can successfully predict new class labels with high classification accuracy We proposed classification method RCCA i.e. boosting on multiple classifiers based on association rule mining. The method has several features 1] it leads to better overall classification accuracy. 2] it highly efficient at classification of various kinds of databases.

REFERENCES

Books:

[1] Jaiwei Han and Micheline Kamber,"Data MiningConcepts and Techniques", Second Edition , Morgan Kaufmann Publishers..

Web References:

[2] Adaboost
http://en.wikipedia.org/wiki/AdaBoost

[3] http://www.datasciencecentral.comprofiles/blogs/random-forests-algorithm

[4] http://en.wikipedia.org/wiki/Gradient_boosting

Research Papers:

[5] Jeetesh Kumar Jain, Nirupama Tiwari, Manoj RamaiyaInternational Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 ,Vol. 3, Issue 1, January -February 2013

[6] K.Saravana Kumar, R.Manicka Chezian – "A Survey on Association Rule Mining using Apriori Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 45– No.5, May 2012

[7] Sanjeev Rao, Priyanka Gupta- "Implementing Improved Algorithm Over APRIORI DataMining Association Rule Algorithm" ISSN : 0976-8491 IJCST Vol. 3, Issue 1, Jan. - March 2012

[8] Robi polikar ―Ensemble based system in decision making‖ IEEE circuits and systems magazine-2006

[9] Sohil Gambhir, Prof. Nikhil Gondliya "A Survey of Associative Classification Algorithms" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 1 Issue 9, November – 2012

[10] Bing Liu Wynne Hsu Yiming Ma "Integrating classification and association rulemining" Department of Information Systems and Computer Science NationalUniversity of Singapore Lower Kent Ridge Road, Singapore 119260-1998

[11] Wenmin Li Jiawei Han Jian Pei_ "CMAR: Accurate and Efficient Classification Basedon Multiple Class-Association Rules" School of Computing Science, Simon FraserUniversity Burnaby, B.C., Canada. ICDM 2001, Proceedings IEEE International
Conference on Data Mining, 2001.

[12] Gourab Kundu1, Sirajum Munir1, Md. Faizul Bari1,Md. Monirul Islam1?, and K. Murase2- A Novel Algorithm for Associative Classification1 Department of Computer Science and Engineering Bangladesh University of Engineering and Technology (BUET)Dhaka-10002 HAIS Department, University of Fukui, Fukui 910-8507, Japan

[13] Yingqin Gu, Hongyan Liu, Jun He, Bo Hu and Xiaoyong Du "MrCAR: A Multirelational Classification Algorithm based on Association Rules" Key Labs of DataEngineering and Knowledge Engineering, MOE, China Information School, Renmin University of China, Beijing, 100872, China School of Economics and Management,Tsinghua University,

Beijing,100084, China. International Conference on Web Information Systems and Mining, 2009. WISM 2009

[14] Fadi Thabtah, Peter Cowling, Yonghong Peng "MCAR: Multi-class Classification based on Association Rule" Modeling Optimization Scheduling And Intelligent Control Research Centre University of Bradford, BD7 IDP, UK Modeling Optimization Scheduling And Intelligent Control Research Centre University of Bradford, BD7 IDP, UK Department of Computing, University of Bradford, BD7IDP, UK. The 3rd ACS/IEEE International Conference on Computer Systems andApplications, 2005.

[15] Yang Junrui, Xu Lisha, He Hongde "A Classification Algorithm Based on Association Rule Mining" College of Computer Science and Technology Xi'an University of Science and Technology Xi'an, China-2012. International Conference on Computer Science & Service System (CSSS), 2012