A SURVEY ON CLASSIFICATION AND ASSOCIATION RULE MINING

N. D. Solank, Mansi patel
Computer Engineering Department
Silver Oak College of Engineering and Technology, Ahmedbad

***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.

II. ASSOCIATION RULE MINING

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.

Support : Support is the number of transitions that contain all the item in the antecedent and consequent part of rule.it is defined as follow.

Supp(Y->Y) = P(X union Y)

Confidence : Confidence is the ratio of the number of transaction that contain all items in the consequent as well in the antecedent(called support) to the number of transaction that contain all item in antecedent .it is defined as follow.

Conf(X->Y) = Supp(X union Y) / Supp(X)

Consider Example :

Buys(X, ”Laptop”) ->Buys(X, ”software”) [support = 2% , confidence = 60%]

X represent person. A confidence is 50% that means if person buys a Laptop , there is 50% chance that she will buy software as well.2% support means that 2 % of all the transaction under analysis showed that laptop and software were purchased together.

Association Rule Mining Technique

A. Apriori Algorithm

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.

Apriori property: All nonempty subsets of a frequent itemset must also be frequent [5]. This property belongs to a special category of properties called antimonotone. It is called antimonotone because the property is monotonic in the context of failing a test.

Advantages of Apriori Algorithm is as follow :

- Uses large item set property
- Easily parallelized
- Easy to implement
- The Apriori algorithm implements level-wise search using frequent item property

Disadvantages of Apriori Algorithm is as follow :

- There is too much database scanning to calculate frequent item .so it will reduce the performance.
- It assumes that transaction database is memory resident .
- Generation of candidate itemsets is expensive.

B. FP-Growth Algorithm

FP-Growth is a technique of ARM which is used for mining all frequent itemsets without candidate's generation. It use a horizontal and vertical database layout to store the database into main memory. When the database is larger ,that means if database cannot fit in to main memory then partition database and then construct an FP-tree and mine it in each projected database. FP-growth method shows that it is efficient and scalable for mining both long and short frequent patterns. It is a faster then an Apriory algorithm.

Advantages of FP-Growth Algorithm

- Uses compact data structure.
- Eliminates repeated database scan.

Disadvantages of FP-Growth Algorithm

- FP-Tree may not fit in memory.
- FP-Tree is expensive to build.

III. ENSEMBLE SYSTEM

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.

Advantages of Ensemble System is a follow :

- Ensemble system split a large amount of dataset into small dataset and each used to train a separate classifier. This is more efficient to use several model rather than single model.

- If we combine the output of several classifiers than it may reduce average risk of individual classifier whose performance is very poor.it may not beat performance of the classifiers but it reduce the overall risk of a poor selection.

- Classification method use divide and conquer method in which data space is divided into smaller and assign a classifier to the smaller space. And last it combine the output of different models.
- If we collect the data from different data source, then the nature of feature are different, a single classifier cannot be used to learn the information from different data source. Applications in which data from different sources are combined to make a more informed decision are referred to as data fusion applications, and ensemble based approaches have successfully been used for such applications.[8]

Ensemble Methods

A. Bagging

Bagging method is derived from bootstrap.it create classifier using training sets that are bootstrapped(that means training datasets are randomly drawn with replacement from whole training data).

Bagging method build several instance of estimator on random subset of the original training set and then aggregate their individual predictions to form a final

prediction. Bagging method is used for strong and complex models(fully developed decision tree).Neural networks and decision trees are good candidates for this purpose, as their instability can be controlled by the selection of their free parameters. [8].

Suppose that you are a patient and would like to have a diagnosis made based on your symptoms. Instead of asking one doctor, you may choose to ask several. If a certain diagnosis occurs more than any of the others, you may choose this as the final or best diagnosis. That is, the final diagnosis is made based on a majority vote, where each doctor gets an equal vote. Now replace each doctor by a classifier, and you have the basic idea behind bagging. Intuitively, a majority vote made by a large group of doctors may be more reliable than a majority vote made by a small group.[1]

B. Boosting

Boosting involve Sequential production of classifier. In Boosting each classifier is dependent on the previous one, and focuses on previous one's errors.

In some cases, boosting has been shown to yield better accuracy than bagging, but it also tends to be more likely to over-fit the training data. The focus of boosting is to produce a *series* of classifiers. The training set is chosen based on the performance of the earlier classifier(s) in the series.

In Boosting, examples that are incorrectly predicted by previous classifiers in the series are chosen more often than examples that were correctly predicted. Thus Boosting attempts to produce new classifiers that are better able to predict examples for which the current ensemble's performance is poor. If we are talking about weak model that time boosting method is better than a Bagging method. Boosting create classifier by resampling the data which are then combined by majority voting.[8]

C. Ada-Boosting

It can be used in conjunction with many other types of learning algorithm to improve their performance. The output of the other learning algorithm is combined into a weighted sum that represents the final output of the boosted classifier [8]. Ada-Boost is a capable of handling multiclass and regression problems. It maintains a set of weights over the original training set and adjusts these weights after each classifier is learned by the base learning algorithm. The adjustments increase the weight of examples that are misclassified by the base learning algorithm and decrease the weight of examples that are correctly classified.

D. Gradient Tree Boosting

Gradient Tree Boosting or Gradient Boosted Regression Trees (GBRT) is a generalization of boosting. Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees [4]. Gradient Tree Boosting is an accurate procedure that can be used for regression and classification problem.it is used in web search ranking and ecology.

E. Random Forests Algorithm

Random Forests are an ensemble learning method which is used for classification and regression. Random Forests construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees . Random Forests Algorithm are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest [3]. The basic principle is that a group of "weak learners" can come together to form a "strong learner". This algorithm is accurate and run efficiently for a large database.

IV. ASSOCIATIVE CLASSIFICATION

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.

A survey of major associative classification techniques are as follows.

A. CBA(Classification Based on Association)

This algorithm is an integration of Classification rule mining and association rule mining. The process of CBA Algorithm is divided into two phase .(1)CBA-RG(Classification Based on Association-rule Generation) It generate all the frequent rule item by making multiple pass over the data.(2)CBA-CB(Classification Based on Association Building a classifier)in this phase it build a classifier using CAR(class association rules). It evaluates all the subset from the training data and selects the subset with right rule sequence that give the least number of errors [10]. CBA is simple but it is inefficient when Database is not inside the main memory.

B. CMAR(Classification based on multiple association rule)

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] Cr-tree is a compact and index base structure.it save a lot of space on storing rules. Advantage of CMAR is it finds frequent patterns and generates rules in one steps.

C. CARGBA(Classification based on Association Rules Generated in a Bidirectional Approach)

Generally all associative classification algorithms use a support threshold to generate association rules from dataset.so it may be possible that high quality rules that have high confidence, but lower support will be missed.so that time CARGBA was introduced. CARGBA rule generator is used to generate rules. It generates rules in two steps. At first , it produce a set of high certainty standards of more modest length with help pruning and after that enlarge this set with some high certainty principles of higher length with backing underneath least backing. CARGBA Classifier Builder is used to build the classifier. In this step, we select a subset of the rules from the pruned rules to cover the dataset. Then selected rules are sorted in descending order based on rule length, confidence, and support. FinalRuleSet is a list that contains those rules which can correctly classify at least one training example. CARGBA is predictable, profoundly viable at arrangement of different sorts of databases and has better normal characterization precision in examination with C4.5, CBA and CMAR [12].

D. MCAR(Multi –Class Classification based on Association rule)

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 is the part of the classifier. MCAR discover and generates frequent items and rules in one phase [14].

E. MrCAR(A Multi – relational Classification Algorithm based on Association Rules)

Current classification algorithm based on association rules is a single table that means data is stored in a single relational table.so if we directly apply this entire algorithm in multi-relational table then it will improve the performance of the algorithm. MrCAR is based on Multi-relational table. MrCAR algorithm divides into three steps. [13] 1] Mine Multi-relational

CFCIs. CFCI stand for class frequent closed item set that contain a class label and an item set which is used to generate the classification rules.2] Generating multi-relational classification rules.3] Predicting class labels bases on the rules

F. CARPT(Classification Algorithm based on association rule mining)

CARPT use Trie-tree which is used to remove the item that cannot 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 format. Trie-tree reduces the cost of the query time to improve the efficiency [15].

V. CONCLUSION AND FUTURE WORK

In this paper we learned some well-known algorithm which is based on classification. As we see all existing algorithm will not give high accuracy in large database. we will use Ensemble method which is used to combine the multiple model and improve the accuuuracy.So, Our goal of research is to find out new algorithm for classification so we can improve the efficiency and accuracy in terms of running time, number of database scan, memory consumption.

REFERENCES

Books:

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

Web References:

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

[3] http://www.datasciencecentral.com profiles/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 rule mining" Department of Information Systems and Computer Science National University of Singapore Lower Kent Ridge Road, Singapore 119260-1998

[11] 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. 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, RenminUniversity 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, BD7 IDP, UK. The 3rd ACS/IEEE International Conference on Computer Systems and Applications, 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