International Journal of Advance Engineering and Research Development (IJAERD)
Volume 1,Issue 5,May 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

COMPARATIVE STUDY ON ASSOCIATIVECLASSIFICATION TECHNIQUES

Ravi Patel[1], Jay Vala[2], Kanu Patel[3]

[1]*Information Technology, GCET, patelravi32@yahoo.co.in*
[2]*Information Technology, GCET, jayvala1623@gmail.com*
[3] *Information Technology, BVM, kanu.patel@live.com*

Abstract: Most recent work has been focused on classification based on association rule mining algorithm. There is growing evidence that merging classification and association rule mining together can produce more efficient and accurate classification systems than traditional classification techniques. In the last few years, a new approach that integrates association rule mining with classification has emerged. Many experimental studies showed that \classification based on association rules mining is a high potential approach that constructs more predictive and accurate classification systems than traditional classification methods like decision trees Moreover, many of the rules found by associative classification methods cannot be found by traditional classification techniques. In this Paper we compare some of these techniques with traditional classifiers.

Keywords: Data mining, Classification, Association, Associative Classification, MMAC, CBA,TOPAC,CMAR,CPAR

I. INTRODUCTION

In Data Mining association rule mining and classification are most important task. Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery[1,26]..

There are many classification approaches such as statistical [2], divide-and-conquer [3] and covering [4] approaches. Based on these Numerous algorithms have been derived such as Naiave Bayes [2], See5 [5], C4.5 [6], PART [7], Prism [4] and IREP [8]. However, traditional classification techniques often produce a *small subset of rules*, and therefore usually miss detailed rules that might play an important role [9].

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.[32] Classification and association rule discovery are similar except that

classification involves prediction of one attribute, i.e., the class, while association rule discovery can predict any attribute in the data set. In the last few years, a new approach that integrates association rule mining with classification has emerged [10,11,12]. Few accurate and effective classifiers based on associative classification approach have been presented recently, such as CPAR [13], CMAR [12], MMAC [11] and CBA [10].

In this paper, the details of a recent proposed classification based on association rules techniques is surveyed and discussed, which extends the basic idea of association rule [31] and integrates it with classification to generate a subset of effective rules..Moreover, the integration of association rule-mining with classification is also investigated, The major algorithms we discover in this paper are: Topac[27], MMAC[11], CACA[28], CMAR[12]

II. ASSOCIATION RULE MINING

Association rule mining is the discovery of what are commonly called *association rules*.Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [14]. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*, identifies the frequent item sets. Another threshold, *confidence*, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket

analysis. For example, it could be useful for the Our Video Store manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form: P $\rightarrow$Q [s,c], where P and Q are conjunctions of attribute value-pairs, and s (for support) is the probability that P and Q appear together in a transaction and c (for confidence) is the conditional probability that Q appears in a transaction when P is present. For example, the hypothetic association rules:

Rent Type(X, "game") $\wedge$ *Age(X, "13-19")* $\rightarrow$*Buys(X, "pop") [s=2%, c=55%]* Would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a pop, and that there is a certainty of 55% that teenage customers who rent a game also buy pop.

Association rule mining is to find out association rules that satisfy the pre-defined minimum support and confidence from a given database.

1. Classification

In classification [23], by the help of the analysis of training data we develop a model which then is used to predict the class of objects whose class label is not known. The model is trained so that it can distinguish different data classes. The training data is having data objects whose class label is known in advance.

Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a *training set* where all objects are already associated with known class labels. Whilst single-label classification, which assigns each rule in the classifier the most obvious class label, has been widely studied [8] little work has been conducted on multi-label classification. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the Our Video Store managers could analyze the customers' behaviors of their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.

3.1 Classification Techniques

3.1.1 Rule Based Classifiers

Rule based classifiers deals with the the discovery of high-level, easy-to-interpret classification rules of the form if-then. The rules are composed of two parts mainly rule antecedent and rule consequent. The rule antecedent, is the if part, specifies a set of conditions referring to predictor attribute values, and the rule consequent, the then part, specifies the class predicted by the rule for any example that satisfies the conditions in the rule antecedent. These rules can be generated using different classification algorithms, the most well known being the decision tree induction algorithms and sequential covering rule induction algorithms.[15]

3.1.2 Bayesian Networks

A Bayesian network (BN) consists of a directed, acyclic graph and a probability distribution for each node in that graph given its immediate predecessors . A Bayes Network Classifier is based on a Bayesian network which represents a joint probability distribution over a set of categorical attributes. It consists of two parts, the directed acyclic graph G consisting of nodes and arcs and the conditional probability tables. The nodes represent attributes whereas the arcs indicate direct dependencies. The density of the arcs in a BN is one measure of its complexity. Sparse BNs can represent simple probabilistic models (e.g., naïve Bayes models and hidden Markov models), whereas dense BNs can capture highly complex models. Thus, BNs provide a flexible method for probabilistic modeling [16].

3.1.3 Decision Tree

A Decision Tree Classifier consists of a decision tree generated on the basis of instances. The decision tree has two types of nodes: a) the root and the internal nodes, b) the leaf nodes. The root and

the internal nodes are associated with attributes, leaf nodes are associated with classes. Basically,each non-leaf node has an outgoing branch for each possible value of the attribute associated with the node. To determine the class for a new instance using a decision tree, beginning with the root, successive internal nodes are visited until a leaf node is reached. At the root node and at each internal node, a test is applied. The outcome of the test determines the branch traversed, and the next node visited. The class for the instance is the class of the final leaf node [17].

3.1.4 Artificial Neural Network

An artificial neural network[24], often just called a neural network is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase [5].A Neural Network Classifier is based on neural networks consisting of interconnected neurons. From a simplified perspective, a neuron takes positive and negative stimuli (numerical values) from other neurons and when the weighted sum of the stimuli is greater than a given threshold value, it activates itself. The output value of the neuron is usually a non-linear transformation of the sum of stimuli. In more advanced models, the non-linear transformation is adapted by some continuous functions.

2. Associative classification

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 [18], CAEP [19], and ADT [20]. 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 [18, 19, 20] show that association based classification may have better accuracy in general.
Some associative classifiers are explained here:

2.1. CBA: integrates association rule mining with classification

Liu et al. proposed an algorithm called CBA that integrates association rule mining with classification. CBA operates in three main steps. First, it discretises real/integer attributes and second, it uses the apriori approach[32,31]of Agrawal and Srikant [14] to discover frequent itemsets and generate the rules. Finally, a subset of the rules produced is selected to represent the classification system. The discovery of frequent itemsets is the most resource- and time-consuming step in CBA, since it requires multiple passes over the training data. In each pass, the seed of the rule items found in the previous pass are used to generate potential rule items in the current pass. The experimental results show that CBA scales well with regard to error rate if compared with decision trees .

2.2. TOPAC

TOPAC[27] is proposed to mine classification rules without candidate rules generated[25]. TOPAC produces only classification rules whose confidence pass a given minimum confidence threshold. Therefore, it does not use effort to generate unnecessary rules. Moreover, TOPAC produces the classification rules based on closed itemsets to give a small number of high quality predictive rules with no redundancy.

THE TOPAC ALGORITHM

A. Generation Closed Itemsets

TOPAC is proposed to generate rules from closed itemsets, because the number of rules is smaller than the rules generated from frequent itemsets. Moreover, the rule can express more information than those generated from frequent itemsets. TOPAC discovers closed itemsets

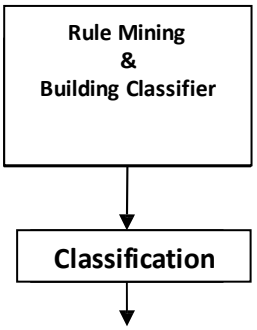
B. Generating Rules with 100% Confidence

To find the high quality rules for prediction, the rules with high confidence should be found first. In the first step, TOPAC divides datasets according to classes into sub-datasets and then mines the sub-datasets. At the first level, TOPAC discovers the largest itemsets contained in the transactions. Then the largest itemsets of all sub-transactions of the class are generated in descending order of the number of transactions. The largest itemsets at the first level are generated by using the items union method.

The TOPAC algorithm produces interesting rules for prediction without candidate rules generated. Firstly, the rules with high confidence are produced and then the closed itemsets having lower than 100% confidence will be extended to find other interesting rules. If the closed itemset has a lower minimum confidence, it will be stopped from extending. The rules with lower minimum confidence are not produced. Therefore, TOPAC is efficient in terms of time and memory space.

2.3. CACA

The CACA[28] algorithm synchronize the rule generation and classifier building phases, shrinking the rule mining space when building the classifier to help speed up the rule generation. CACA is making better performances in accuracy and efficiency in Associative classification approaches.



Approach of CACA

There are 4 Phase integrated in CACA:
1. use the class based strategic to cut down the searching space of frequent pattern;
2. design a structure call Ordered Rule-Tree to store the rules and their information which may also prepare for the synchronization of the two steps;
3. redefine the compact set so that the compact classifier is unique and not sensitive to the rule reduction;
4. synchronize the rule generation and building classifier phases.

According to the characteristic of associative classification, a new class based frequent pattern mining strategic is designed in CACA to cut down the searching space of frequent pattern. OR-Tree structure enables the synchronization of the traditional phases which may not only simplify the associative classification but help to guide the rule generation and speed up the algorithm. And the redefinition of the redundant rule and compact set guarantee the usage of the compact set to help improve the classification efficiency and rule quality won't affect the accuracy of CACA.

4.4 CMAR

The method extends an efficient frequent pattern mining method, FP-growth[32], constructs a class distribution-associated FP-tree, and mines large database 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 x^2analysis using multiple strong association rules.

Association- based classification may have better accuracy in general. However, this approach may also suffer some weakness: it is not easy to identify the most effective rule at

classifying a new case., 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.

Here developed CMAR, for accurate and efficient classification and make the following contributions.

First, instead of relying on a single rule for classification, CMAR determines the class label by a set of rules. Given a new case for prediction, CMAR selects a small set of high confidence, highly related rules and analyzes the correlation among those rules. To avoid bias, we develop a new technique, called weighted X^2 which derives a good measure on how strong the rule is under both conditional support and class distribution.

Second, to improve both accuracy and efficiency, CMAR employs a novel data structure, CR-tree, to compactlystore 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.

Third, to speed up the mining of complete set of rules, CMAR adopts a variant of recently developed FP-growth method. FP-growth is much faster than Apriori-like methods used in previous association-based classification especially when there exist a huge number of rules, large training data sets, and long pattern rules.

CMAR consists of two phases: rule generation and classification.

In the first phase, rule generation, CMAR computes the complete set of rules in the form of R : P -> C where P is a pattern in the training data set, and C is a class label such that Sup(R) and Conf(R) pass the given support and confidence thresholds, respectively. Furthermore, CMAR prunes some rules and only selects a subset of high quality rules for classification.

In the second phase, classification, for a given data object obj, CMAR extracts a subset of rules matching the object and predicts the class label of the object by analyzing this subset of rules.

CMAR[33,29] is consistent, highly effective at classification of various kinds of databases and has better average classification accuracy in comparison with CBA and C4.5[6,21], and is more efficient and scalable than other associative classification methods.

.4.5. MMAC

A new approach for multi-class multi-label classification rules has been proposed that has many distinguishing features over traditional and associative classification methods: (1) It produces classifiers that contain rules with multiple labels

(2) It presents four evaluation measures for determining accuracy that are applicable to a wide range of applications

(3) It employs an efficient method for discovering rules that requires only one scan over the training data

(4) It employs a detailed ranking method, which prunes redundant rules, and ensures only effective ones are used for classification.

MMAC[30] consists of three phases: rule generation, recursive learning and classification. In the first phase, the method scans the training data to discover and generate a complete sort of CARs. In the second phase, MMAC proceeds to discover more rules that pass the MinSupp and MinConf thresholds from the remaining unclassified instances, until no further frequent items can be found. In the third phase, the rules sets derived during each iteration will be merged to form a global multi-label classifier that will then be tested against test data.

MMAC 3 phases are:

– Phase 1:

1. Scan the training data T with n columns to discover frequent items
2. Produce a rules set by converting any frequent item that passes MinConf into a rule
3. Rank the rules set according to (confidence, support, . . . , etc.)
4. Prune redundant rules from the rules set

– Phase 2:

1. Discard instances Pi associated with the rules set just generated in phase 1

2. Generate new training data T _ ← T − Pi
3. Repeat phase 1 on T _ until no further frequent item is found

– Phase 3:

1. Merge rules sets generated at each iteration to produce a multi-label classifier
2. Classify test objects and calculate error rate using an accuracy measure

One principle reason for extracting more rules is due to the recursive learning phase that MMAC employs. This discovers hidden information that most of the associative classification techniques discard, as they only extract the highest confidence rule for each frequent item.

MMAC is an accurate and effective classification technique, highly competitive and scalable if compared with other traditional and associative classification approaches.

4.6. CPAR: Classification based on Predictive Association Rules

A greedy associative classification algorithm called CPAR, which adopts the FOIL strategy [34] to generate rules, was proposed by Yin and Han [35]. CPAR seeks the best rule condition that brings most FOILgain among the available ones in the data set. FOIL gain is used to measure the information gained from adding a condition to the current rule. Once the condition is identified, the weights of the positive examples associated with it are reduced by a multiplying factor, and the process repeats until all positive examples in the training data set are covered. The search for the best rule condition is the most time-consuming process of CPAR, since the gain for every possible item needs to be calculated to determine the best overall gain. In the rule-generation process, CPAR derives not only the best condition but also all similar ones, since there is often more than one attribute item with similar gain. It is claimed that CPAR improves the efficiency of the rule-generation process when compared with popular associative classification methods such as CBA.

III. CONCLUSION

CBA and MMAC are classification algorithms that are based on association rule mining techniques that use statistical constraints and depend on the co occurrence of items in the database.The recently proposed classification based on association rules algorithm, i.e. MMAC, performed the best with regards to classification accuracy. A possible reason for the accurate classification systems produced by MMAC is the fact that MMAC employs a more detailed rules' ranking method that looks for high confidence detailed rules to play part of the classification system.To test the scalability of *CMAR*, we compare the runtime of CBA and *CMAR*. *CMAR* is faster than CBA in many cases.CPAR improves the efficiency of the rule-generation process when compared with popular associative classification methods such as CBA. CPAR generates far fewer rules than CMAR, it shows much better efficiency with large sets of training data. Unlike traditional algorithm, The TOPAC algorithm produces interesting rules for prediction without candidate rules generated

Therefore, TOPAC is efficient in terms of time and memory space.The searching spaces of the CACA are smaller than that of MCAR. The time cost and searching space size of CACA increase slower than those of MCAR, when the *minsupp* becomes smaller.

REFERENCES

1. Shelly Gupta , Dharminder kumar ,Anand Sharma "data mining classification techniques applied for breast cancer diagnosis and prognosis", ijcse vol. 2 no. 2 ,2011
2. John, G. H. and Langley, P. (1995). *Estimating Continuous Distributions in Bayesian Classifiers*. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345. Morgan Kaufmann, San Mateo.
3. Furnkranz, J. (1996). *Separate-and-conquer rule learning*. *Technical Report* TR-96-25, Austrian Research Institute for Artificial Intelligence, Vienna.
4. Cendrowska, J. (1987). *PRISM: An algorithm for inducing modular rules*. International Journal of Man-Machine Studies. Vol.27, No.4, pp.349-370.
5. Quinlan, J. R. See5.0 (http://www.rulequest.com) viewed on May 2010.

6. Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, San Francisco
7. Frank, E. and Witten, I. (1998). Generating accurate rule sets without global optimization. In Shavlik, J., ed., *Machine Learning: Proceedings of the Fifteenth International Conference*, pp. 144-151,Madison,Wisconsin, Morgan Kaufmann, San Francisco.
8. Furnkranz,J. and Widmer, G. (1994).*Incremental reduced error pruning*. In *Machine Learning: Proceedings of the 11th Annual Conference*, New Brunswick, New Jersey, Morgan Kaufmann.
9. Pazzani, M., Mani, S., and Shankle, W. R. (1997).*Beyond Concise and colorful:learning intelligible rules*. In *KDD-97*.
10. Liu, B., Hsu, W., and Ma, Y. (1998). *Integrating Classification and association rule mining*. In *KDD '98*, New York, NY, Aug. 1998
11. Thabtah, F., Cowling, P. and Peng, Y. H. (2004).*MMAC: A New Multi-Class, Multi-Label Associative Classification Approach*. Fourth IEEE International Conference on Data Mining (ICDM'04).
12. Li, W., Han, J., and Pei, J. (2001). *CMAR: Accurate and efficient classification based on multiple-class association rule*. In *ICDM'01*, pp. 369-376, San Jose, CA
13. Yin, X., and Han, J. (2003). CPAR: Classification based on predictive association rule. In *SDM 2003*, San Francisco, CA.
14. Agrawal, R., Amielinski, T., and Swami, A. (1993). *Mining association rule between sets of items in large databases*. In Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, DC, May 26-28.
15. G.L. Pappa and A.A. Freitas, Automating the Design of Data Mining Algorithms. An Evolutionary Computation Approach, Natural Computing Series, Springer, 2010
16. G.F. Cooper, P. Hennings-Yeomans, S.Visweswaran and M. Barmada, "An Efficient Bayesian Method for Predicting Clinical Outcomes from Genome-Wide Data", AMIA 2010 Symposium Proceedings, 2010, pp. 127-131
17. M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, "Building Decision Trees with Constraints", Data Mining and Knowledge Discovery, vol. 7, no. 2, 2003, pp. 187 – 214
18. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *KDD'98*, New York, NY, Aug.1998.
19. G. Dong, X. Zhang, L. Wong, and J. Li. Caep: Classification by aggregating emerging patterns. In *DS'99 (LNCS1721)*, Japan, Dec. 1999.
20. K. Wang, S. Zhou, and Y. He. Growing decision tree on support-less association rules. In *KDD'00*, Boston, MA,Aug. 2000.
21. J. R. Quinlan. *C4.5: Programsfor Machine Learning*. Morgan Kaufmann, 1993
22. P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3:261–283, 1989.
23. R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. JohnWiley & Sons, 1973.
24. Y. Sngh Y, A.S. Chauhan, "Neural Networks in Data Mining", Journal of Theoretical and Applied Information echnology, 2005, pp. 37-42
25. J. Han, H. Pei, and Y. Yin." Mining Frequent Patterns without Candidate Generation." In: Proc. Conf. on the Management of Data SIGMOD'00, ACM Press
26. Tipawan Silwattananusarn and Dr. KulthidaTuamsuk "Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012" IJDKP Vol.2, No.5, September 2012
27. Panida Songram.,"Mining associative classification without candidate rules" IJFCC'2012
28. Zhonghua Tang and Qin Liao .," A new class based associative classification algorithm" IJAM'2007
29. Wenmin Li Jiawei Han Jian Pei.,"CMAR: accurate and efficient classification based on multiple class-association rules "ICDM -2004
30. Fadi Abdeljaber Thabtah, Peter Cowling ,Yonghong Peng "Multiple labels associative classification" Springer'2005
31. Kanu patel, Vatsal Shah.,"Implementation of classification using association rule mining" IJETCAS'2013
32. B.Santhosh Kumar, K.V.Rukmani .,"Implementation of web usage mining using APRIORI and FP growth algorithms" IJANA'2010
33. "Alaa Al Deen" Mustafa Nofal and Sulieman Bani-Ahmad .,"Classification based on association-rule mining techniques: a general surveyand empirical comparative evaluation " ACIT'2010
34. Quinlan, J., Cameron-Jones, J.: FOIL: a midterm report. In: Proceedings of 1993 European Conference on Machine Learning, pp. 3–20. Vienna, Austria (1993)
35. Yin, X., Han, J.: CPAR: classification based on predictive association rule. In: Proceedings of the SDM 2003, pp. 369–376. San Francisco, CA (2003)

