A Survey of Associative Classification Algorithms

Sohil Gambhir, Prof. Nikhil Gondliya

***Abstract*— In the field of data mining very large amount of data is processed in order to get small amount of useful data. To optimize efficiency two classical methods of data mining is merged, namely association rule mining and classical rule mining. The new method is called associative classification. This work is a survey of major associative classification methods. After this study better comparison of various associative classification methods can be done.**

***Index Terms*— Associative Classification, Data Mining, Survey**

I. Introduction

The data mining is known as methods to find small amount of useful data from very large amount of data [1]. To optimize efficiency two classical methods of data mining is merged, namely association rule mining and classical rule mining. The new method is called associative classification. [2].

Classification rule mining is used to discover a small set of rules in the database that forms an accurate classifier. Association rule mining finds all the rules existing in the database that satisfy some minimum support and minimum confidence constraints. For association rule mining, the target of discovery is not pre-determined, while for classification rule mining there is one and only one predetermined target.

This way great savings and conveniences to the user is achieved if the two mining techniques can somehow be integrated. The integration is done by focusing on a special subset of association rules whose right-hand-side is restricted to the classification class attribute. The integration is done by focusing on mining a special subset of association rules, called ***class association rules*** (CARs).

The new combined approach associative classification achieves higher accuracy but lesser speed than traditional classification approaches. There are various methods used for the associative classification [2, 3, 4]. This work is intended to do a comparative study of all major associative classification algorithms. The previous study of this type was done [5] but after the said work lot of new work has been done in this field. Hence this work will give a better understanding of the current state of associative classification techniques. A brief overview of associative classification is as follows.

A transactional database normally used in association rule mining does not have many associations. While classification data tends to contain a huge number of associations. Adaptation of the existing association rule mining algorithm to mine only the CARs is needed to reduce the number of rules to avoid combinatorial explosion. This adaptation involves discretizing continuous attributes based on the classification predetermined class target.

Data mining in the associative classification framework has three steps:

- Discretization of continuous attributes, if any
- Generating all the class association rules (CARs)
- Building a classifier based on the generated CARs

The associative classification has following three new things:

1. It shows a new way to build accurate classifiers. Results show that accuracy is more than by the state-of-the-art classification system like C4.5 classification system [6].
2. Association rule mining techniques can be applied to classification tasks.
3. It helps to solve a number of important problems with the existing classification systems.

The major problems of the existing systems are solved as below.

1. Understandability problem

The rules produced by standard classification systems are many times difficult to understand. Similarly many understandable rules are left undiscovered.

2. Interesting rule problem

In order to get a small set of rules of the existing classification systems results in many interesting and useful rules not being discovered.

3. Memory Problem

All the standard classification systems need to load the entire dataset into the main memory. But in this approach the database can reside in the disk rather than the main memory.

A survey of major associative classification techniques are as follows.

II. Survey of Historical Associative Classification Methods

A. *CBA (Classification Based On Associations)*

The CBA is an ordered rule algorithm based on convergence analysis. It consists of two parts. A *rule generator* ,namely CBA-RG, based on algorithm Apriori for

Sohil Gambhir is a Master's student at the B.V.M. Engineering College, Vallabhvidyanagar, Gujarat, India. (e-mail: gambhir.sohil@gmail.com)

Prof. Nikhil Gondliya is Associate Professor, Department of I.T. Engineering, at G. H. Patel College of Engineering & Technology, Vallabhvidyanagar, Gujarat, India. (e-mail: nikhilgondaliya@gcet.ac.in)

finding association rules. Second part a *classifier builder* , namely CBA-CB, generates the classifiers from the rules generated from the CBA-RG.

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 ruleitems 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.

This algorithm is simple, but is inefficient because it needs to make many passes over the database. The experimental results show that data set taken from UCI ML repository [7] 16 out of 26 data sets it working better than the C4.5 classification system [6].

The limitations of this approach are as follows

- It generates huge amount of the mined rule.
- This leads to computational overhead.
- The classification is done based on single high confidence rule which can be biased

B. CMAR (Classification based on Multiple Association Rules)

The associative classification suffers from the huge set of mined rules and sometimes biased classification or over fitting because the classification is done based on only single high-confidence rule. This associative classification method, CMAR (Classification based on Multiple Association Rules) [4] 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.

CBA also suffer some weakness as shown below.

First it is not easy to identify the most effective rule at classifying a new case.

Second a training data set often generates a huge set of rules.

To get better results instead of relying on a single rule for classification the class label is determined by a set of rules. To avoid bias a new technique called weighted X^2 is developed. It derives a good measure on how strong the rule is under both conditional support and class distribution.

To improve 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.

To speed up the mining of complete set of rules CMAR adopts a variant of recently developed FP-growth method which is much faster than Apriori-like methods.

CMAR consists of two phases: rule generation and classification. In 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 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.

The CMAR outperforms both C4.5 and CBA on accuracy and it is also scalable. 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.

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

The CARGBA generates the rules in two steps. In first, 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.

In the second step 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.

The experiments on 6 databases in UCI machine learning database repository show that 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.

D. Hyper Heuristic Approach

In this investigation is done for the possibility of associative classifiers by a general-purpose optimization heuristic called the hyper heuristic [20]. The hyper heuristic requires deciding which of several simpler search neighborhoods' to apply at each step while constructing a solution. After 16 different solution generated by a hyper heuristic called Peckish the results indicated that associative classification approach is the most applicable approach to such kind of problems with reference to accuracy.

This study focused on analyzing the behavior of low-level heuristics that were selected by the hyper heuristic and improved upon the quality of the current. These rules can be used to quickly predict the appropriate low-level heuristics to call next. The experimental tests showed a better performance for associative classification techniques (MCAR, MMAC, CBA) over decision trees (C4.5), rule induction (RIPPER) and PART algorithm.

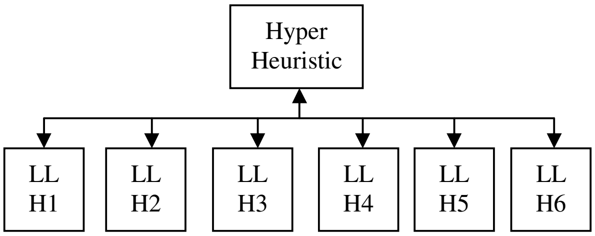

Figure 1 Hyper Heuristic General Framework

E. CPAR (Classification based on Predictive Association Rules)

The CPAR[4] combines the advantages of both associative classification and traditional rule-based classification. Instead of generating a large number of candidate rules 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.

In comparison with other CPAR has the following advantages:

- 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.

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.

F. A Parameter-Free Associative Classification Method

In this method Parameter-Free associative classification is done [8]. The associative classification [2] is based on the classical objective interestingness measures for association rules – frequency and confidence – for selecting candidate classification rules. Since then, the selection procedure has been improved leading to various CBA-like methods. The support-confidence-based methods show their limits on imbalanced data sets. Indeed, rules with high confidence can also be negatively correlated. Even considering a correlation measure is not satisfactory for a n-class imbalanced context.

A rule can be positively correlated with two different classes what leads to conflicting rules. The common problem of these approaches is that they are one-vs-all methods. Means they split the classification task into n two-class classification tasks (positives vs negatives) and look for rules that are relevant in the positive class and irrelevant for the union of the other classes. First an OVE (one-vs-each) is method used that avoids some of the problems observed with typical CBA-like methods. Next a constrained hill climbing technique is designed that automatically tunes the many parameters (frequency thresholds) that are needed.

It computes class association rules that are frequent in the positive class and infrequent in every other class. Tuning the large number of parameters required by this approach may become a problem hence an automatic tuning method that relies on a hill-climbing strategy was applied. The result shows that accuracy of this approach is quite promising.

G. A Lazy Approach to Associative Classification

In associative classification approach [2] there is a limitation that it generates very large amount of data. Due to large correlated data set a huge set of rules may be generated. In this approach [9] a lazy pruning technique called L^3 is employed to selectively prune the rules. This pruning will lead to smaller number os rules. The L^3 associative classifier is based on the idea that all the knowledge extracted from the training set may be useful for the classification. To do this L^3 couples a lazy pruning approach with a compact representation of the classification rule set. The lazy pruning technique discards only “harmful” rules. Here harmful means the rules that only misclassify training data. It has two levels. Level 1 includes few high-quality rules . Rules usually discarded by previous approaches are included in Level 2 and only considered when rules in Level 1 did not classify a test case. Experimental results show thatL^3 appropriately classifies data that were usually not covered or erroneously assigned to the default class by previous associative classifiers due to this Level 2. L^3 is based on the concepts of closed and generator itemsets and a macroitem. This form avoids information loss and allows the regeneration of the complete rule set. It also allows representing very large rule sets by compact sets of limited size. This way lower support thresholds can be considered during the mining phase, and large rule sets can be exploited to build the classifier. The availability of a larger rule selection allows a significant increase of the classification accuracy.

III. Comparison

The accuracy of various associative classification algorithms for UCI ML Repository Data Set, as claimed by respective work, is given in Table 1.

IV. Conclusion

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. However there are no significant improvements to merge two or more associative classification methods for better accuracy.

REFERENCES

[1] Data Mining Concepts and Techniques by Jiawei Han & Micheline Kamber.Pulication Elseiver.

[2] Integrating classification and associative rule mining by Liu B., Hsu W., and Ma W. In KDD'98, New York, NY, Aug. 1998..

[3] CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules by Li W., Han J., Pei J. In ICDM'01, pp. 369{376, San Jose, CA, Nov.2001.

[4] Classification based on Predictive Association Rules by Yin X., Han J. Proc. 2003 SIAM Int.Conf. on Data Mining (SDM'03), San Fransisco, CA, May 2003.

[5] A survey of associative classification algorithms by Dhirendra Kumar Swami, R. C. Jain. ADIT JOURNAL OF ENGINEERING, VOL. 2, NO.1, DECEMBER 2005.

[6] C4.5: Programming for machine learning, Morgan by J. R. Quilan.

[7] A Parameter- Free Associative Clasification Method by I.-Y. Song, J. Eder, and T.M. Nguyen (Eds.): DaWaK 2008, LNCS 5182, pp. 293–304, 2008. _c Springer-Verlag Berlin Heidelberg 2008.

[8] A Lazy Approach to Associative Classification by Elena Baralis, Silvia Chiusano, and Paolo Garza, IEEE TRANSACTIONS ON NOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 2, FEBRUARY 2008

[9] Data Set provided by UCI Machine Learning repository web reference is http://www.ics.uci.edu/~mlearn/MLRepository.html.