REAL PERFORMANCE OF CATEGORIZATION-BASED ASSOCIATION RULE TECHNIQUES

FADI THABTAH
Modelling Optimisation Scheduling
And Intelligent Computing
Research Centre
Fabdelja@bradford.ac.uk

PETER COWLING
Modelling Optimisation Scheduling
And Intelligent Computing
Research Centre
P.i.Cowling@bradford.ac.uk

YONGHONG PENG
Department of Computing,
University of Bradford, BD7 1DP,
UK
Y.h.Peng@bradford.ac.uk

Abstract— **Integrating association rule mining and classification proved to be effective in producing classification systems with high prediction rates. In this research paper, four associative rule learning algorithms (CBA, CMAR, CPAR, MCAR) have been compared with regards to classification accuracy against twelve benchmark problems from the UCI data collection. The aim is to determine the most accurate technique in forecasting the future classes of unseen test objects. After experimentation with different data sets, the results indicate that none of the investigated techniques dominated the others with regards to accuracy. Moreover, MCAR proved to extract highly accurate classification systems than CBA, CMAR and CPAR. However, a post pruning method is recommended to reduce the size of MCAR classifiers especially for cases such as "Cleve" and "Germany" data sets.**

Keywords: Association Rule, Classification, Data Mining, Associative Classification, Prediction Accuracy

I. INTRODUCTION

The term data mining was defined in [5] as one of the main phases in Knowledge Discovery of Databases (KDD), which extracts useful information. Learning in data mining involves finding and describing structural patterns in data for many purposes such as prediction. Data mining could be used for many tasks, such as categorisation, clustering, association rules, regression and many others. Every task could be accomplished by using various data mining techniques that are adapted from different science fields, like Statistics and Artificial Intelligence. There is no single data mining technique that is applicable to all above tasks. Furthermore, when it comes to select a technique for a particular problem, the choice will be very crucial since one technique could work good for a task and poor else where. There are many factors that can be considered before taking such a decision, such as the size of the data set, the attribute types (text, real), the number of attributes in the data set, and the goal of the task.

Association rule mining is one of the most important tasks in data mining for discovering rules that pass certain user constraints in a data set. Association rule mining is a strong tool for market basket analysis that aims to find relationships among items in a sales transaction database [1]-[2]. In discovering association rules, one tries to find groups of items that are frequently sold together in order to infer items from the presence of other items in the customer's shopping cart. For instance, in a supermarket, if a customer buys soda, what is the probability that he/she buys an ice as well? Using such rules, marketing experts can develop strategic decisions concerning items shelving, sales promotions and planning.

Classification is another central task in data mining. Given a collection of records in a data set, each record consists of a group of attributes and one of the attributes is the class label. The classification task involves constructing a model from the classified objects, in order to classify previously unseen objects as accurately as possible [12]. This process involves prediction of future class labels, whereas association rule mining involves only the description of the relationships among items in a database. In addition, there is one and only one pre-specified target class in classification, however, the target classes for association rule are not pre-specified.

In the last few years, a new approach that integrates association rule mining and classification called associative classification has been proposed [4]-[6]-[8]. A few accurate and effective classifiers based on associative classification have been presented in last few years, such as CMAR [6], CPAR [13], CBA [7] and MCAR [11]. Many experimental studies [6]-[7] showed that associative classification is a promising approach, which builds more accurate classifiers than traditional classification techniques such as decision trees [9]. Moreover, many of the rules found by associative classification methods can not be discovered by traditional classification algorithms [7].

In this paper, we compare the-state-of-the-art associative classification techniques on different benchmark problems from UCI data collection [8]. In particular, we will compare between four associative classification techniques that are CMAR [6], CBA [7], MCAR [11] and CPAR [13]. The comparison will be based on error rate using cross validation.

Related research works on associative classification approach are surveyed in Section II. Basic concepts of associative classification are discussed in Section III and Experimental results are given in Section IV. Finally the conclusions are presented in Section V.

II. LITERATURE REVIEW

The authors of [3] have adapted the popular step-wise association rule mining algorithm (Apriori) [2], for extracting class association rules that represent characteristics of the data classes in two applications, i.e. telecommunication and medical diagnoses. Their aim was to discover a set of overlapping rules that are individually accurate and not prediction of future class labels. Thus, the presented method cannot be considered a fully classification method since the ultimate aim for classification in data mining is prediction. The results of the two case studies indicated that association rule can be used for partial classification in which many useful rules for general practitioners have been derived for the medical diagnoses study. Finally, the authors speculated whether association rule mining can be used for complete classification.

One of the first algorithms to use association rule approach for classification was proposed in [7]. It has been named CBA. CBA implements the famous Apriori algorithm [2] that requires multiple passes over the training data set in order to discover frequent items. Once the discovery of frequent items finished, CBA proceeds by converting any frequent item that passes the minimum user confidence into a rule. In doing that, only one subset of the generated rules will be considered in the final classifier. Evaluating all the generated rules against the training data set does the selection of the subset. The frequent items discovery and rules generation are implemented in two separate phases in CBA.

An associative classification algorithm that selects and analyses the correlation between high confidence rules, instead of relying on a single rule, has been developed in [6]. It uses a set of related rules to make a prediction by evaluating the correlation among them. The correlation measures how effective are the rules based on their support values and class distributions. In addition, a new prefix tree data structure named CR-tree to handle the set of rules generated and to speed up the retrieval process of a rule has been introduced. The CR-tree has proven to be effective in saving storage since many conditions of the rules are shared in the tree.

A new approach for building classification systems based on both positive and negative rules has been introduced in [4]. The interestingness of the rules for the proposed algorithm is based on the correlation coefficient that measures the strength of the linear relationship between a pair of variables. Besides confidence and support thresholds, correlation coefficient has been used for pruning the final classifier, giving a much reduced rules set if compared with support and confidence pruning methods. The algorithm generates the rules similar to Apriori approach [2] and ranks the rules similar to CBA rules ranking method [7]. Experimental tests on six UCI data sets showed that negative association rules are useful when used with positive ones for producing competitive classification systems.

A greedy associative classification algorithm called CPAR was proposed in [13]. CPAR adopts FOIL [10] strategy in generating the rules from data sets. It seeks for the best rule condition that brings the most gain among the available ones in the data set. Once the condition is identified, the weights of the positive examples associated with it will be deteriorated by a multiplying factor, and the process will be repeated until all positive examples in the training data set are covered. The searching process for the best rule condition is time consuming process for CPAR since the gain for every possible item needs to be calculated in order to determine the best item gain. Thus, CPAR uses an efficient data structure, i.e. PNArray, to store all the necessary information for calculation of the items gain. In the rules generation process, CPAR derives not only the best condition but all close similar ones since there are often more than one attribute items with similar gain. It has been claimed that CPAR improves the efficiency of the rule generation process if compared with popular associative classification methods like CBA and CMAR.

III. ASSOCIATIVE CLASSIFICATION APPROACH

Let *D* be the training data set with *n* attributes (columns) A_1, A_2, … , A_n and |*D*| rows. Let *C* be a list of class labels. Specific values of attribute A_i and class *C* will be lower case *a* and *c*, respectively.

Table 1. Training data 1

RowId	A1	A2	Class
1	x1	y1	c1
2	x1	y2	c2
3	x1	y1	c2
4	x1	y2	c1
5	x2	y1	c2
6	x2	y1	c1
7	x2	y3	c2
8	x1	y3	c1
9	x2	y4	c1
10	x3	y1	c1

Definition 1: An item, or condition is defined as a set of attributes A_i together with a specific values a_i for each attribute in the set, denoted < (A_{i1}, a_{i1}), (A_{i2}, a_{i2}), … (A_{im}, a_{im})>.

Definition 2: A rule *r* maps an item to a

specific class label, denoted:
<(A_{i1}, a_{i1}), (A_{i2}, a_{i2}),…, (A_{im}, a_{im})>ϕ*C*.

Definition 3: The actual occurrence *ActOcc*(*r*) of a rule *r* in *D* is the number of rows of *D* that matches *r's* condition.

Definition 4: The support count *SuppCount*(*r*) of *r* is the number of rows of *D* that matches *r's* condition, and belong s to *r's* class.

Definition 5: The support of *r* is defined as the *SuppCount*(*r*)/|*D*|.

Definition 6: The minimum support which a rule in our rule base may have is denoted *MinSupp*.

Definition 7: The confidence of *r* is defined as *SuppCount*(*r*)/*ActOcc*(*r*).

Definition 8: The minimum confidence which a rule in our rule base may have is denoted *MinConf*.

Consider for instance the training data set shown in Table 1 and assume that *MinSupp* is 20% and *MinConf* is 50%. The support of rule $<(A_1, x_1)>\rightarrow c1$ is 3/10, which satisfies the *MinSupp* threshold. The confidence of rule $<(A_1, x_1)>\rightarrow c1$ is 3/5, and thus this rule also satisfies the *MinConf* threshold and is a candidate rule in the classifier.

Generally, in association rule mining, any item that passes *MinSupp* is known as a **frequent item**. If the frequent item consists of only a single attribute value, it is said to be a frequent single item. For example, with *MinSupp* = 20%, the frequent single items in Table 1 are < (A_1, x_1)>, < (A_1, x_2)>, < (A_2, y_1)>, < (A_2, y_2)> and < (A_2, y_3)>. Most of the current associative classification techniques search for frequent items by making multiple passes over the training data set. In the first pass, they find the support of each single item, and then in each subsequent pass, they start with items found to be frequent in the previous pass in order to produce new possible potential frequent items involving more attribute values, known as candidate items.

In other words, frequent single items are used for the discovery of potential frequent items that involve two attribute values, and frequent items that involve two attribute values are input for the discovery of candidate items involve three item values and so on. After frequent items have been discovered, only one subset of them will form the final classifier. The selection of the final subset is accomplished in various ways. CBA for instance, select high confidence rules after evaluating the complete set of class association rules on the training data. On the other hand, CPAR uses a greedy method to select the best rules for the classifier. Generally, associative classification methods derive a complete set of rules for those frequent items that pass *MinConf* and select only a subset to represent the classification system.

IV. EXPERIMENTAL RESULTS

Twelve different data sets from UCI data collection [8] have been used in the experiments using stratified ten-fold cross-validation [12]. Cross-validation is a standard evaluation measure for calculating error rate on data in machine learning. Four popular associative classification techniques that are CBA, CMAR, CPAR and MCAR have been compared in terms of classification accuracy. The choice of such learning methods is based on the different strategies they use to generate the rules.

All experiments were conducted on a Pentium IV 1.6 GHz machine. Since we were unable to obtain the source codes for CPAR and CMAR to conduct the experiments, therefore their experimental have been provided by their authors. However, CBA experiments were conducted using a VC++ implementation version provided by [14] and MCAR experiments were conducted using a Java version provided by [11].

Many studies have shown that the support threshold plays a major role in the overall classification accuracy of the set of rules produced by existing associative classification techniques [7]. Moreover, the support value has a larger impact on the number of rules produced in the classifier and the processing time and storage needed during the algorithm rules discovery and generation. From

Table 2. Accuracy of CBA, CMAR, CPAR and MCAR algorithms using Ten-fold Cross Validation

Data	**Size**	**Classes**	**CBA**	**CMAR**	**CPAR**	**MCAR**
Cleve	303	2	82.80	82.20	81.50	81.62
Breast	699	2	96.30	96.40	96.0	96.10
Diabetes	768	2	74.50	75.80	75.10	78.96
Glass	214	7	73.90	70.10	74.40	77.57
Iris	150	3	94.70	94.00	94.70	95.49
Pima	768	2	72.90	75.10	73.80	77.80
Wine	178	3	95.00	95.00	95.50	95.00
Austral	690	2	84.90	86.10	86.20	86.90
German	1000	2	73.40	74.90	73.40	73.28
Labor	57	2	86.30	89.70	84.70	89.92
Tic-Tac	958	2	99.60	99.20	98.60	99.23
Led7	3200	10	71.40	72.50	73.60	71.96

our experiments, we noticed that the support rates that ranged between 1% to 5% usually achieve the best balance between accuracy rates and the size of the resulted classifiers, therefore, as in [6]-[7], the *MinSupp* was set to 1% in the experiments. The confidence threshold, on the other hand, is less complex and does not have a larger effect on the behaviour of any associative classification method as support value, and thus it has been set to 50%.

Table 2 represents the accuracy of the classification systems generated by CBA, CMAR, CPAR and MCAR on the twelve benchmark problems. The results indicate that there was no dominant algorithm. The classification accuracy figures derived show consistency among the four learning algorithms in term of accuracy. However, MCAR classification systems have slightly better accuracy than CBA, CMAR and CPAR ones. In particular, MCAR outperformed CBA, CMAR and CPAR on six data sets. One of the principle reasons for this appears to be that MCAR often generates few more rules than the rest. The increase in accuracy suggests that this is not simply overfitting and would likely justify the small increase in classification rate for MCAR over the rule learning techniques in applications [11]. However, in some cases, like the “German” and “cleve” data sets, the number of rules is large, even though every rule represents at least one training object. Thus, a post pruning method, like pessimistic error pruning [9] may be useful in such cases.

V. CONCLUSIONS

Prediction accuracy is one of the main crucial factors in association rule and classification tasks in data mining. In this paper, four different associative classification techniques have been compared in term of accuracy in order to indicate the one that produces highly effective classification systems. Performance studies on twelve data sets from the UCI data collection indicate that there is consistency between CBA, CMAR, CPAR and MCAR with regards to accuracy. In particular, MCAR produced slightly more accurate classification systems than CBA, CMAR and CPAR on six data sets. The increase in accuracy suggests that this is not simply overfitting and would likely justify the small increase in classification rate for MCAR over the rule learning techniques in applications. However, in practice, humans may scarify part of the accuracy to end up with an optimised classification system that may contain small but effective number of rules. Our further work will investigate the extraction of multiple class labels using association rule discovery for a wide range of application problems. Moreover, we will look into the issues of rules features and runtime of existing associative classification techniques.

REFERENCES

[1] R. Agrawal, T. Amielinski, and A. Swami. Mining association rule between sets of items in large databases. *In Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data,* Washington, DC, May 26-28 1993, pp. 207-216.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rule. *Proceedings of the 20th International Conference on Very Large Data Bases***.** 1994, pp. 487 - 499.

[3] K. Ali, S. Manganaris, and R. Srikant. Partial Classification using Association Rules. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, The AAAI Pre*ss, pp. 115-118.

[4] M. Antonie, O. Zaïane and A. Coman. Associative Classifiers for Medical Images. *Lecture Notes in Artificial Intelligence 2797, Mining Multimedia and Complex Data*, Springer-Verlag, 2003, pp. 68-83.

[5] U.M. Fayyad; G. Piatetsky-Shapiro; P. Smyth, 1996, *Advances in knowledge discovery and data mining,* MIT Press.

[6] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple-class association rule. In *ICDM’01,* San Jose, CA, Nov. 2001.

[7] B. Liu, W. Hsu, and Y. Ma. Integrating Classification and association rule mining. *Proceedings of the KDD,* 1997, New York, NY, pp. 80-86.

[8] C. Merz and P. Murphy. UCI *Repository of Machine Learning Data- bases*. Irvine, CA, University of California, Department of Information and Computer Science, 1996.

[9] J. Quinlan. C4.5: *Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[10] J. Quinlan and R. Cameron-Jones. FOIL: A midterm report. *Proceedings of 1993 European Conference on Machine Learning*, Vienna, Austria, 1993, pp. 3-20.

[11] F. Thabtah, P. Cowling and Y. Peng. MCAR: Multi-class Classification based on Association Rule. *Proceedings of the 3rd ACS/IEEE International Conference on Computer Systems and Applications,* Cairo, Egypt, January 2005.

[12] I. Witten and E. Frank. *Data mining: practical Machine learning tools and techniques with Java implementations*, San Francisco: Morgan Kaufmann, 2000.

[13] X. Yin, and J. Han. CPAR: Classification based on predictive association rule. *Proceedings of the SDM.* San Francisco, CA, pp. 369-376

[14]CBA:http://www.comp.nus.edu.sg/~dm2/p_download.html.