An Efficient Method for Associative Classification using Jaccard Measure

Dharini B, Jayanthi M, Joyce A

Abstract— Classification is a data mining technique that categorizes the items in a database to target classes. The aim of classification is to accurately find the target class for each instance of the data. Associative classification is a classification method that uses Class Association Rules for classification. Associative classification is found to be often more accurate than some traditional classification methods. The major disadvantage of associative classification is the generation of redundant and weak class association rules. Weak class association rules results in increase in size and decrease in accuracy of the classifier. This paper proposes an efficient approach to build a compact and accurate classifier by using interestingness measures for pruning rules. Interestingness measures play a vital role in reducing the size and increasing the accuracy of classifier by pruning redundant or weak rules. Rules which are strong are retained and these rules are further used to build the classifier. The source of the data used in this paper is University of California Irvine Machine Learning Repository. The approach proposed in this paper is effective and the results show that the approach can produce a highly compact and accurate classifier.

Keywords—Associative Classification, Class Association Rules, Accuracy, Interestingness measure, Classifier

I. INTRODUCTION

Classification is one of the most important data mining tasks that can be applied for Business Intelligence, Decision Making, Handwriting Recognition, Pattern Recognition, Biological Classification, Document Classification, Medical diagnosis. A classification task begins with a data set in which the itemsets are mapped to the respective class labels. Data set is divided into two parts: one for building the model; the other for testing the model. Classifier is built by applying a classification technique to the part of data which is used for building the model. Classifier's performance can be tested by using the built classifier to classify the instances of data which is used for testing the model. Classification models are tested by comparing the output values to target values that are present in the test data. Accuracy refers to the percentage of correct predictions made by the classifier when compared with the actual classifications in the test data.

For example, a classification model that classifies credit risk could be developed based on observed data for many loan applicants over a period of time. Classifier built using the data can be used to classify the credit risk of new instances with unknown class labels. Many methods for classification such as Neural Network, Decision Tree, and Support Vector Machines have been proposed. In the most recent decade, a new classification technique called Associative Classification has picked up prominence because of its efficiency and simplicity. Associative classification (AC) is a promising data mining approach that combines classification and association rule discovery to build classification models called classifiers. In associative classification, classification is based on the set of rules known as Class Association Rules.

Association rules are conditional statements that help to find the hidden relationships between seemingly unrelated data in a relational database or other information repository. They play an important part in shopping basket data analysis, product clustering, catalogue design and store layout. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. The generation of an associative classifier consists of three steps. First step is to generate class association rules. Second step is to apply pruning technique and select high quality class association rules so that classification accuracy increases and size of classifier decreases. Third step is to build a classifier using the high quality rules generated.

Class Association Rules consists of two parts namely Antecedent and Consequent. Antecedent part consists of set of attribute and its values and the Consequent part consists of class labels. Classifier is constructed using this Class Association Rules and the classifier is used for finding the class labels of new instances. One of the main disadvantages of Associative classification is the large size of the classifier. Data is getting more and more now a days. Therefore whenever this enormous amount of data is used for the generation of Class Association rules naturally more numbers of Class Association rules are generated. As the classifier is built using Class Association rules, the size of classifier also increases to a large extent. If the size of classifier increases the time for classification also increases. Large size of classifier also decreases the accuracy of classification due to the large number of weak rules or less interesting rules in it. Large size classifier may also cause over fitting of data. Therefore decreasing the size of classifier by using proper pruning technique, results in decrease in running time and

Revised Version Manuscript Received on 10, September 2019.

Dharini B, Assistant Professor, Department of Computer science and Enginnering, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India (Email: dharini.1592@gmail.com)

Jayanthi M, Assistant Professor, Department of Computer science and Enginnering, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India(Email: jayantheemurugesan@gmail.com)

Joyce A, Assistant Professor, Department of Computer science and Enginnering, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India(Email: joyceccet.11@gmail.com)

increase in accuracy of the classification.

The objective of this paper is to decrease the size of the classifier and increase the accuracy of classification by using proper rule ranking technique and using only strong rules to build the classifier. The remainder of this paper is organized as follows. Section 2 describes Problem Statement and assumptions. Section 3 gives a brief overview of previously related research work. Section 4 presents the preliminary concepts of associative classification. Section 5 and Section 6 describes the contributions of this paper which includes methodologies, computational experiments, experimental design, results and discussions.

II.OBJECTIVE

Associative classification has shown good results over many classification techniques. However due to enormous and ever increasing data, the number class association rules produced are large. Selecting the interesting and strong class association rules is a major challenge. This challenge is addressed in this paper by proposing a suitable technique for rule ranking and selection by using interestingness measures. This paper has contributed the technique for construction of classifier by combining the Jaccard measure with the Principality metric. The results show that the technique produces highly compact and accurate classifier.

III. RELATED WORKS

Associative classification has become very popular over a decade. Many papers have been published related to associative classification and building compact classifiers. Some methods include Classification Based on Associations (CBA), Classification based on Multiple Association Rules (CMAR), mining efficiently using Modified Equivalence Class Rule (MECR) tree, Class Association Rules with Interestingness Measures (CARIM), Principal Association Mining.

CBA is a technique in which classification is done by building a classifier using associative rules that satisfy the minimum confidence value. The rules that satisfy the minimum confidence value are then subjected to pruning methods and high quality rules are obtained to build the classifier. Classification Based on Association (CBA) used Apriori Candidate Generation Method. Classification Based on Multiple Association Rules used FP Tree method which adopts the divide and conquer method.

In Class Association Rules with Interestingness measure (CARIM), MECR tree structure is used for compressing the information from the dataset so that multiple scans of dataset can be avoided. Several interestingness measures such as Laplace, Confidence, Cosine, Jaccard, Rule Interest, Phi-Coefficient are proposed. In this method, any one measure is applied and based on the threshold values k rules are selected and these rules are used to build the classifier. Since interestingness measures are used weak rules are all eliminated using this method and the classifier built is highly compact and accurate.

In Principal Association Mining method a novel rule quality metric called Principality is proposed in which both the confidence and coverage aspects of dataset are considered. Depending on the weight that is given for the confidence and coverage metric the results will have variable accuracy value. The weight value should b chosen depending on the property of the dataset. The method for classifier construction is also proposed in this method. After finding the principality metric a rule list is prepared by arranging the rules in the decreasing order of Principality metric. Pick the rule with highest Principality. If it classifies atleast one instance from the test data correctly add it to the classifier. Remove the rule from the rule list. Delete all the instances that are covered by this rule from the dataset. This continues until all the rules from the dataset are covered.

Constrained Class Rule tree was proposed for mining Constrained Class Association Rules. In constrained Class Association Rule one attribute is kept as a constraint. That is the particular attribute which is kept as a constraint must occur in all class association rules. Constrained Class Rule tree consists of information about the dataset in a compressed way so that multiple scans of database can be avoided.

IV. PRELIMINARIES ON ASSOCIATIVE CLASSIFICATION

Itemset

Itemset is a set of attribute value pair <A,a1>,<B,b1>,…<K, k1> where A,B,...K represents the name of the attribute and a1,b1,…,k1 represents the value of the attribute. If there are n attributes then the set is called n-itemset.

Frequent Itemsets

The itemsets that occur frequently within the database are called Frequent Itemsets. In other words the itemsets whose number of occurrences is above the certain user given threshold are called frequent itemsets. The threshold is usually specified in terms of support value. That is if the support of an itemset is greater than the given support threshold then the itemset is considered to be afrequent itemset.

Support

Support is an indication of how frequently the item-set appears in the database. The support value of itemset X with respect to set of transactions T is defined as the proportion of transactions in the database that contains itemset X.

Confidence

Confidence is the widely used metric for the generation of Association Rules. Confidence is an indication of how often the rule has been found to be true. The confidence value of a rule, X=>Y with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.

Confidence is defined as:

Conf(X=>Y) = supp(XUY) / supp(X)

Association Rules

Association Rule is of the form X=>Y where X and Y are the itemsets within a dataset T. Association rule generation is usually split up into two separate steps:

1. A minimum support threshold is applied to find all frequent item-sets in a database.
2. A minimum confidence constraint is applied to these frequent item-sets in order to form rules.

Association rule mining

Association rule mining is a method for discovering interesting relations between variables in large databases.

It is intended to identify strong rules discovered in databases using some measures of interestingness. Association Rule Mining involves two steps.

(1) Find all frequent itemsets in the given set of transactions.

(2) Generate strong association rules from the frequent itemsets: For each frequent itemset X and its proper subsets Y, generate all rules Y X=>Y, with confidence no less than the given threshold minconf.

Class Association Rules

Class Association Rules consist of antecedent and consequent part. Antecedent part contains itemsets and consequent part contains class label [2]. For example a1, b1, c3=>Y is a class association rule where a1, b1, c1 are values of attributes present in the dataset and Y is the value that the target attribute can take.

Interestingness measures

Interestingness measures are used to select and rank rules according to the interest of the user [6]. They can be used to increase the efficiency by decreasing the time, cost and improving the accuracy of the classification. Some of the interestingness measures are confidence, cosine, jaccard, ruleinterest, laplace, lift. [1]

Patterns

A pattern is of the form <A, Y> where A denotes the itemset and Y denotes the class label. A pattern implies that A =>Y. Therefore mining association rules involves finding frequent itemsets and using that itemsets to produce interesting patterns.

Jaccard Measure

Jaccard measure is used for measuring the interestingness of the association rule. Consider the association rule X=>Y , the Jaccard measure can be calculated as count(XY)/(count(X)+count(Y)-count(XY)) where count(XY) denotes the number of transactions that contains the itemset X and class label as Y, count(X) denotes the number of transactions that contains the itemset X and count(Y) denotes the number of transactions whose class label is Y.

For example the Jaccard measure of the rule x1,y1=> High in the Table 1 can be calculated as

Jaccard(x1,y1=>High)=
count(XY)/(count(X)+count(Y)-Count(XY))
=1/((3)+(2)-1)
= 1/4
Jaccard(x1, y1=> High) = 0.2

V. METHODOLOGIES

Tree structure

Modified Equivalence Class Rule Tree (MECR) is used for mining Class Association Rules that satisfy a certain minimum support value.[8-10] The tree can be used to mine Class Association Rules effectively by avoiding multiple database scans since the tree stores all the information from the database in a compressed way. The components of the tree are given below:

1. Obidset:

These are set of identifiers that represent the records in a database. Each row in a database is represented by an identifier. For a particular itemset obidset denotes the set of identifiers of the row that contains the itemset.

2. Bitwise Representation:

Each itemset in the database can be represented uniquely using bitwise representation. This can be done by making the bit corresponding to the itemset as 1 whenever it is represented.

Example: In a table of 3 attributes, the itemset a1 can be represented as 001 which in decimal format is equal to1.The itemset b1 is equal to 010 which in decimal format is equal to 2.The itemset a1b1 can be represented as 011 which in decimal format is equal to 3.

3. Count

The count of number of instances of a particular itemset belonging to the each class in the database is stored in the tree.

4. Position

Position value with respect to an itemset denotes the class to which the itemset is mapped for large number of times when compared with other classes.

OID	X	Y	Z	CLASS
1	x1	y1	z1	High
2	x1	y2	z1	Low
3	x2	y2	z1	Low
4	x3	y3	z1	High
5	x3	y1	z2	Low
6	x3	y3	z1	High
7	x1	y3	z2	High
8	x2	y2	z2	Low

Table I. Example Dataset

Completeness

Completeness can be defined as the proportion of instances that are predicted by the rule. For example consider the rule X=>Y, completeness can be defined as the ratio of number of instances in database that contains the itemset X with the

class label Y to number of instances in the database that contains class label Y.

Completeness (X=>Y) = |XY|/|Y|

From Table1

Completeness(x3, y3=>High) =2/4=0.5

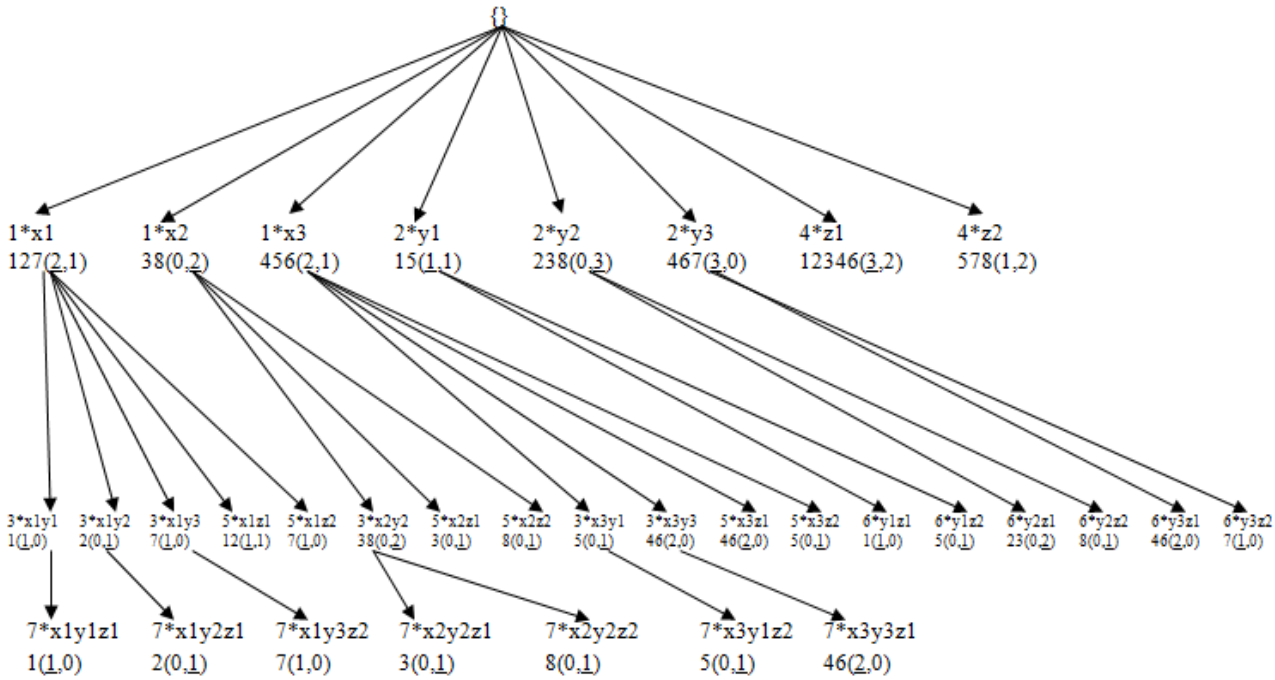

Fig 1. MECR Tree Structure of the given example dataset

Principality

Principality metric is a combination of confidence and completeness measure. For accurate classification and to avoid over fitting both confidence and completeness should be considered.

Principality= α *confidence + (1-α)* completeness

where α is a user defined value which ranges from 0 to1.

From Table1, Principality is calculated for the rule

x3, y3=>High

Assumption: User defined α value = 0.6

Principality(x3, y3=>High) = 0.6*(2/2) +0.4(2/4)

=0.6+0.2

Principality(x3, y3=>High) =0.8.

CAR Mining

From the input dataset MECR tree is constructed. Using the tree structure the class association rules that satisfy the minimum support values are generated. These class association rules may have many redundant and uninteresting rules.[3-5] These rules are called weak rules. Jaccard measure is applied for these class association rules and the rules that satisfy the minimum Jaccard threshold are selected. These rules are further used for classifier construction.

Applying the Principality on the rule set containing weak rules is not much useful. The rules that are weak causes a strong class association rule to be projected as a weak association rule. So the rule will not be included in the classifier and the classification accuracy decreases. Therefore applying Jaccard measure resolves this problem.

Classifier Construction

For each of the rules obtained after applying the Jaccard measure, Principality measure is calculated. The rules are then ordered in the descending order of the Principality. The rule with the highest principality is picked. If it classifies atleast one instance correctly, then the rule is added to the classifier. Delete the instances that can be classified by the rule from dataset. This has to be continued until the rule set becomes empty or the database contains no items such that it cannot be classified by the classifier.

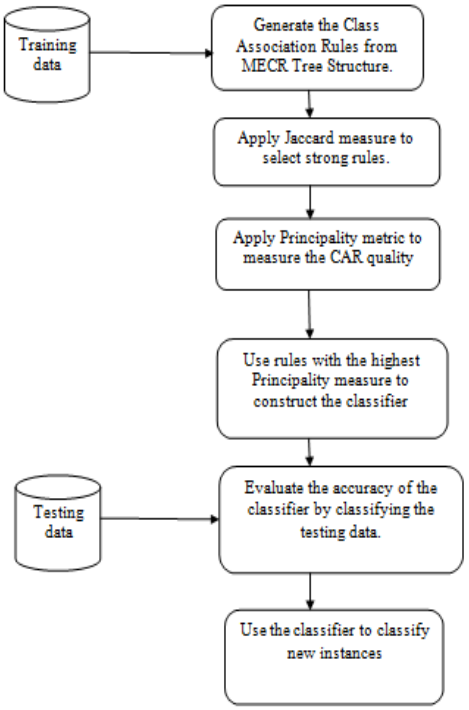

Fig 2. Phases of proposed method

Procedure 1: Classifier Construction
Input: Dataset T
Output: Associative Classifier AC (Class Association Rules)

Construct the MECR tree structure from the dataset T
Enumerate the Association Rules from MECR Tree Structure.
Apply interestingness measures on the rule.
Select k rules of highest strength. That is the k rules that has the highest interestingness value.
Calculate Principality of all the k rules.
for each rule in the rule set
while(ruleset!=empty ! dataset T!=empty)
pick up the rule R with highest Principality.
if it classifies atleast one data instance correctly
add the rule R to the classifier
delete the rule from the rule set
delete all the objects covered by R from T.
end if
end while
end for

VI. EXPERIMENTAL RESULTS

The experiment is conducted using Windows 7 64bit 4GB memory i3 processor. The dataset taken for this experiment is Hungarian dataset. The dataset is taken from UCI Machine Learning Repository.

The characteristics of the dataset are given below
Number of instances: 293
Number of distinct items: 17
Number of attributes: 6
Number of classes: 2

Various comparisons are made based upon the threshold values, size of the classifier, number of rules generated, accuracy of the classifier. No one threshold value or α value is suitable for all the datasets. Depending upon the characteristics of the datasets, optimal value should be selected.

With increase in threshold value, the number of rules generated decreases. That is the weak rules are eliminated and the strong rules are retained. Principality metric is applied only among the strong rules. This increases the accuracy of the classifier.

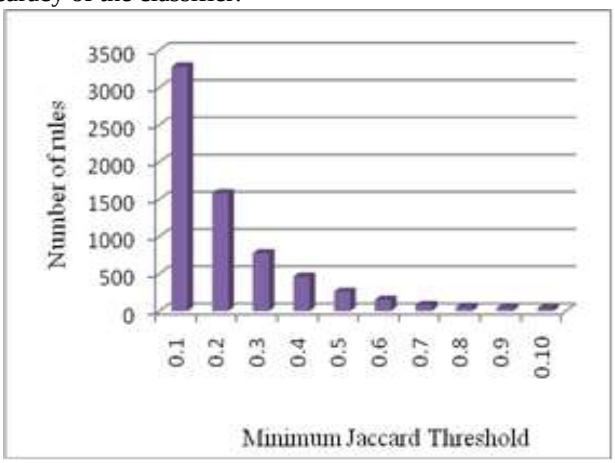

Fig 3.Variation of number of CARs generated with respect to minimum Jaccard threshold value (Before applying Principality metric)

With increase in threshold values the running time for the generation of rule also decreases. This occurs because number of rule decreases with increase in threshold values.

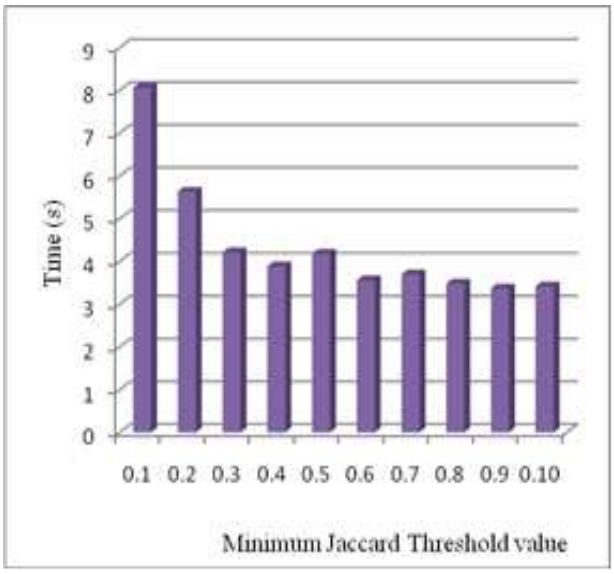

Fig 4: Execution time with respect to minimum Jaccard threshold value
(Before applying Principality metric)

For the dataset taken, the accuracy of the classifier increases whenever higher weightage values are given for completeness of the rule.

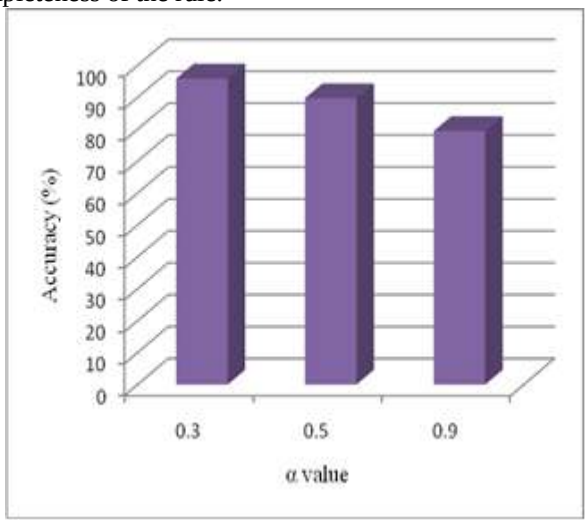

Fig 5. Accuracy with respect to α values

As α value increases, the classifier become more and more compact for the dataset under experiment.

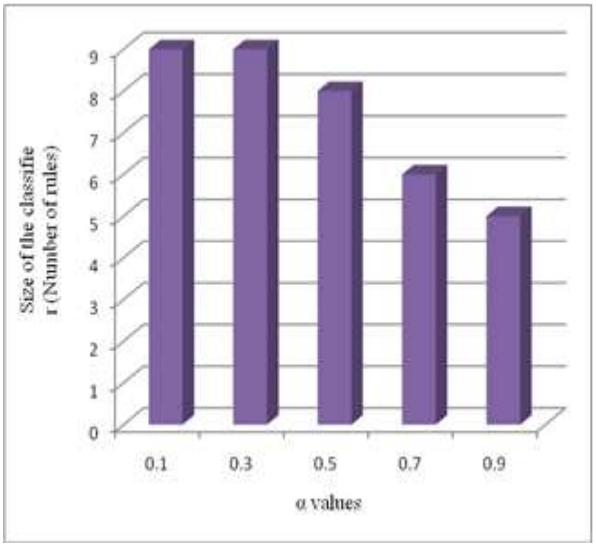

Fig 6: Size of the classifier with respect to α values

The accuracies are compared before and after applying Jaccard measure. At each case, it is inferred that the accuracy increases when the interestingness measures are applied at the initial stages.

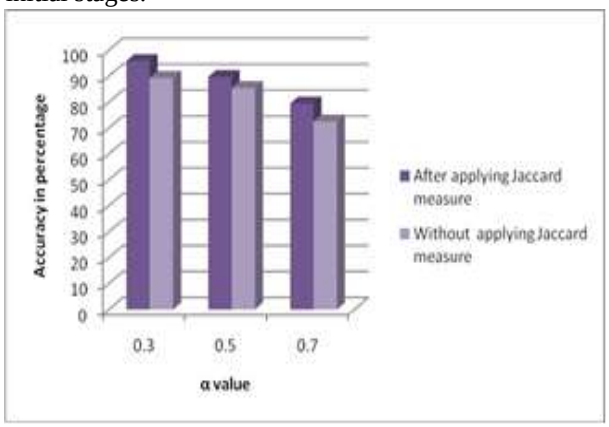

Fig 7: Comparison of accuracies without and after applying Jaccard measure

VII. CONCLUSION

With the increase in size of data, analysing the data and building a good classifier is a challenging task. With the large number of rules, the size of classifier increases and also the accuracy decreases as most of the rules are weak. So it is better to use an interestingness measure and select strong rules and then follow the process for building classifier. In this paper the impact of Jaccard measure on accuracy of classification is analyzed. The experimental results have shown that the accuracy of the classification is greatly increased whenever the strong rules are selected, processed and used for building the classifier.

REFERENCES

1. Loan Nguyen, Bay Vo, and Tzung-Pei Hong. CARIM: An Efficient Algorithm for Mining Class-Association Rules with Interestingness Measures. The International Arab Journal of Information Technology, Vol. 12., No. 6A, 2015.
2. W. Li, J. Han, J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. Proceedings IEEE International Conference on Data Mining (ICDM 2001), IEEE, 2001, pp. 369–376.
3. Chen,Yanlan Wang, Minqiang Li, Harris Wu, Jin Tian. Principal Association Mining: An efficient classification approach in Knowledge-Based Systems 67 (2014) 16–25
4. L.T. Nguyen, B. Vo, T.P. Hong, H.C. Thanh. Classification based on association rules: a lattice-based approach. Expert Syst. Appl. 39 (13) (2012) 11357–11366.
5. B.Liu, Y. Ma, C.K. Wong. Improving an association rule based classifier. Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Springer, Berlin Heidelberg, 2000, pp. 504–509.
6. Shekar B. and Natarajan R. A Transaction-based Neighbourhood-Driven Approach to Quantifying Interestingness of Association Rules. Proceedings of IEEE International Conference on Data Mining pp. 194-201, 2004
7. Coenen F., Leng P., and Zhang L.. The Effect of Threshold Values on Association Rule based Classification Accuracy.Data and Knowledge Engineering, vol. 60, no. 2, pp. 345-360, 2007.
8. Vo B. and Le B., "Interestingness Measures for Association Rules: Combination between Lattice and Hash Tables," Expert Systems with Applications, vol. 38, no. 9, pp. 1630-11640,2011.
9. J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent patterns without candidate generation: a frequent-pattern tree approach, Data Min. Knowl. Disc. 8 (1) (2004) 53–87.
10. Liu B., Hsu W., and Ma Y..Integrating Classification and Association Rule Mining.Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, USA, pp. 80-86, 1998.