Associative Classification using Naïve Bayes Theorem

By

Fawze Abujaber

Supervisor

Dr. Rasheed Zoubidy

This Thesis was Submitted in Partial Fulfilment of the Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduate Studies

Philadelphia University

Feb,2014

جامعة فيلادلفيا

نموذج تفويض

أنا فوزي علي ابوجابر ، أفوض جامعة فيلادلفيا بتزويد نسخ من رسالتي للمكتبات أو المؤسسات أو الهيئات أو الأشخاص عند طلبها.

التوقيع :فوزي ابوجابر

التاريخ :2014/2/9

Philadelphia University

Authorization Form

I am Fawze Ali Abujaber, authorize Philadelphia University to supply copies of my thesis to libraries or establishments or individuals upon request.

Signature: Fawze Abujaber

Date: 9/2/2014

Associative Classification using Naïve Bayes Theorem

By

Fawze Abujaber

Supervisor

Dr. Rasheed Zoubidy

This Thesis was Submitted in Partial Fulfilment of the Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduate Studies

Philadelphia University

Feb,2014

Dedication

I dedicate this work to my Wife Lubna Abujaber, Who encouraged me all the way to reach this point, with her personal support, great patience at all times, and her endless love and support.

Fawze A. Abujaber

Acknowledgement

I would like to thank Dr. Rashid Al-zubaidi for his guidance and encouragement in completing the thesis. This would not have been possible if not for his belief in me. I am also grateful to Prof. Saed Algoul for his comments in shaping the thesis. Cannot thank enough my wife, Lubna, for all her support and encouragement in completing this work. Finally, I would like to dedicate this work to my parents who have been there all along to support me.

Fawze Abujaber

Table of content

List of Tables

List of Abbreviations

KDD	Knowledge Discovery in Databases
AC	Associative Classification
CBA	Classification Based Association
CACA	Class Based Associative Classification Approach
CMAR	Classification Cased on Multiple-Class Association Rule
CPAR	Classification based on Predictive Association Rule
CARs	Classification Association Rules
UCI	University of California Irvine
DM	Data Mining
MMCAR	Modified Multi-class Classification using Association Rule
LC	Looking at the Class
CPNAR	Classification based on Positive and Negative Association Rules
CAIG	Classification based on Attribute-value pair Integrate Gain
KNN	K-Nearest Neighbors
LUCS-KDD	Liverpool University Computer Science - Knowledge Discovery in Data
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
MCAR	Multi-class Classification based on Association Rule
NB	Naïve Bayes

List of Figures

Abstract

Associative classification usually generates a large set of rules. Therefore, it is inevitable that an instance matches several rules which classes are conflicted, several associative classification based their prediction on one rule and ignore all other rules, even high confident ones, In this research, a new approach called Associative Classification using naïve Bayes (AC-NB) is proposed, which uses an improved naïve Bayes theorem to address these issues.

Our experiments on five UCI datasets show that AC-NB outperforms both RIPPER and NB on accuracy, also compared with new associative classification approaches; our proposed approach was highly competitive.

Key words: Associative Classification, Classification based on Association, Prediction, Naïve Bayes theorem, Accuracy Power.

CHAPTER ONE

INTRODUCTION

1.1 General Introduction

Technology has aided in rapid advances in data capture, storage, and processing that resulting in creating large databases with high complexity and size, these data is stored internally in what is called “relational databases" that consist of rows and columns of data.

The traditional methods of analyzing data and pattern recognition manually are no longer feasible, this had been augmented by indirect and automatic data processing methods from the artificial intelligence such as decision trees (Kingsford and Salzberg,2008) and support vector machines (Fletcher,2009).

One of the exciting area in machine learning is data mining which it’s goal to find the useful and needed information for the user from large data or datasets, and finds relationships between these data, so that to be summarized in a novel way that is simple and useful for the data owner, It is often sets in the broader context of knowledge discovery in databases (KDD).

One of the recent data mining techniques is associative classification (AC) which combines two of the most familiar data mining tasks, association rule mining and classification to build classification system for prediction purposed, these two data mining tasks are analogues, with the exception that classification should assigns a class label for previously unseen records, while association rule mining goal is to find and describes relation between items in transactional datasets.

Several studies such as(Hadi,2013), (Zaixiang et al,2013),(Yuhanis and Refai, 2013),(Thabtah et al, 2010) (Ramasubbareddy et al, 2011) proved that association classification approaches is deriving more accurate classifiers than the traditional classification approaches such as C4.5 trees, and rule induction (Hilage and Kulkarni, 2012).

The first introduced association classification approach is called Classification based Association which uses the Apriori (Agrawal and Srikant, 1994) association rules algorithm to solve classification problems. It finds correlations (rules) based on two defined values minimum support and minimum confidence. The CBA algorithm operates in four stages as follows:

1. Finds frequent rule items based on predefined minimum support value.
2. Using the frequent rule items to generate association classification rule based on another predefined minimum confidence value.
3. Applying pruning technique to choose a subset of the generating association classification to build a classifier.
4. Classifying previously unseen records based on the classifier rules.

The focus of our study was to test data classification where we used a probabilistic measure based on naïve Bayes theorem to improve the predicting accuracy of CBA algorithm.

Several AC approaches like CACA (Tang and Liao, 2007)CMAR (Li et al, 2001),CPAR (Yin and Han, 2003) that utilize one rule for classifying test data which is considered, for example, assume we have a new object T we need to classify, and there are four rules that compromise the classifier (X1,X2,X3,X4), and the rules have the following confidence values of 96%, 95%, 94%, 93%, respectively, and the rule X1 associated with class Y1, and the three others rules associated with class Y2, most of the mentioned approaches assign the new object T to class Y1 because the rule X1 has the highest confidence value, using such prediction style is simple because only one rule used for the prediction decision and effective because the highest confident rule always has the highest impact in classifying the test objects, but, this technique can be criticized, because there could be more than one rule matched to a new object with high confidence values. In addition, the highest confidence rule may be misleading, especially for data where the class label percentages are unbalanced.

The proposed approach used group of rules based on naïve Bayes theorem to make the class prediction for the test data. In other words, this study investigated the possibility of using multiple rules in predicting test data instead of one rule, which enhanced the confidence in the prediction decision, since more than one rule contributed in such decision.

using more than one rule based on their frequency (weights) came up with a global weight for the rules classes that selecting the class with the largest weight to assign it to the test data improved the prediction rate of the derived classifiers.

1.2 Associative Classification

Thabtah et al, (2011) defined the association classification problem as follow: let a learning data D has a different attributes x1, x2… xn-1, xn, and Y different class values.

The attribute xi value could be a categorical value, or continues (real values), for categorical values to be used it must be mapped to positive integers, and for continues ones, a discretization methods must be used to transfer these values to categorical ones.

The problem of association classification is to extract strong rules with high support and confidence values, these strong rules then will be used to form an automated classifier that will be used in the prediction of new unseen objects.

Most of AC techniques based on two predefined values, minimum support which represents the frequency of the occurrence of specific value with specific class value in the learning data from all data.

Any combination of the attribute values with the class label that pass the predefined min support is known as frequent rule_item, the other important predefined value in AC is the minimum confidence, which is the percentage of the frequency of attribute, class value combination from the frequency of the attribute value.

Here we will produce the AC steps in brief, and will explain it in details in chapter 4 when we discuss our new proposed approach.
To build a classifier using AC there are 4 main steps:
Step1: discovering frequent rule_items
Step 2: generating classification association rules that pass the defined minimum confidence from the frequent rule_items which generated in step 1.
Step 3: Ranking and pruning the generated rules to select a subset of the generated classification association rules to form the classifier.
Step 4: Testing the classifier quality on new test data objects.

Figure 1.2 shows the main steps used in any AC approach, the first step is computational expensive and similar to the discovery of frequent items in association rule based, once all frequent rule_items generated, a subset for that group form the CARs in the form A→B, where B

must be a class and the rule_items must passed the minimum confidence, normally the generated CARs are large set of rules, so a rank and prune methods used to choose a subset of that rules to be used in the classification of the new test objects(Hadi ,2013).

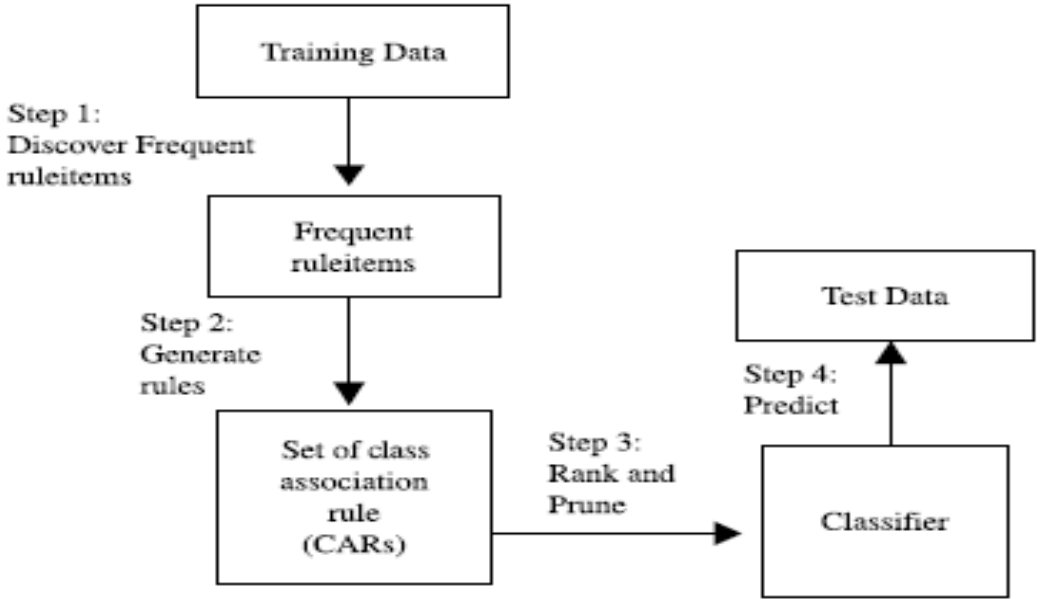

Figure 1.1 AC steps (Hadi, 2013).

For example the CBA algorithm (Liu et al, 1998), the first association classification approach works as follows:

Once the classifier rules are selected and ordered according to its confidence value, and a new test data presented to predict its class label, the classification based on association passes on the classifier rules and chooses the class label of the highest confidence rule that match the test data body.

1.3 Research Problem:

Predicting of test data for certain real application related to classification is crucial, therefore, we want to investigate this important step to come with the most optimal classification method that reduces the number of error mistakes during classifying unseen instances.

There is no AC algorithm that used the Naïve Bayes classifier as prediction style, the CBA algorithm uses only one rule to build classifiers and ignore all other rules even high confidence

rules, moreover, the CBA algorithm suffers from high complexity resulted from the ranking stage of the rules according the confidence rate.

1.4 Research Objectives:

1. Enhancing the predictive accuracy of CBA by using Naive Bayes classifier as predicting method.
2. Reducing CBA complexity by ignoring the ranking step.

1.5 Research Approach to Meet the Research Objectives

In order to meet the mentioned objectives, the research will used mixed methods of both quantitative and qualitative techniques. One of the main project aims is to conduct Experimental results on common data mining benchmarks such as the UCI and comparison with the state of the art of classification methods in terms such as prediction accuracy, and the CPU time, quantitative methods will be used to conduct a literature review on the prediction methods used in recent the association classification approaches, while the quantitative approach will be a research strategy to accomplish this task successfully.

The approach of conducting this thesis can be segmented into four phases as follows:

- Providing literature review on the different classification techniques with the state of the art of classification methods.

- Customizing an open source code to normalize continues data sets, and building a java console application to integrate the naïve theorem principle as prediction methods in the CBA approach.

- Analyzing the proposed method results in terms as prediction accuracy, and the CPU time.

1.6 Thesis Outline

The thesis is structured as follows:

Chapter 1 provides a general introduction about Data mining and association classification, discussing the research goals and objectives, and research methods which will be used to meet these goal and objectives, and the most important discussing how we can be enhance the accuracy power of CBA by tackling its main problems.

Chapter 2 is designed to survey some of the new AC approaches and discussing its prediction phase, with emphasizing the difference between using a group of rules in the prediction phase versus relying only on one rule, finally summarizing the different AC approaches and the techniques used in each step of the approach.

Chapter 3 describes in more details the our proposed approach and it's steps with providing a complete example of how our proposed AC operates.

Chapter 4 shows our experimental results, with comparing our new AC with other AC approaches.

Chapter 5 is dedicated to summarize our proposed technique, with providing a conclusion for the research, in addition, provides directions towards the future researches using the probabilistic principle for incremental machine learning.

CHAPTER TWO

RELATED WORKS

2.1 Overview

The main goal of classification in the data mining is to predict the class label of a new unseen object instance.

In association classification the classification decision can be based either on one rule that applicable to the new unseen object instance, or making the classification decision based on multiple rules, In this chapter we will discuss different prediction methods employed by the current AC algorithms, finally we will summarize the current AC approaches,

2.2 Maximum Likelihood Matching Rule for Prediction

Several AC algorithm (Niu et al, 2009), (Li et al, 2008),(Tang and Liao, 2007) utilize the maximum likelihood matching rule for prediction, in these algorithms there is a classifier with a group of rules A and a new unseen object B that we need to predict its class label, in these algorithms only the highest precedence rule which matches the unseen object B example is considered, so that the test data class label of the new object instance will take the same class label of the matched rule.

If there is no matched rule to the new data item then the class label will take a predetermined default class label.

One of the famous AC algorithms used single rule approach is classification based on association approach (Liu et al, 1998), this approach works as follows: Once the classifier rules are selected and ordered according to its confidence value, and a new test data presented to predict it's class label, the classification based on association passes on the classifier rules and chooses the class label of the highest confidence rule that match the test data body.

If there is no rules match the new object body, the classification based on association assigns a predetermined default class label to the test case (Figure 2.1)

New developed algorithms such as L3G (Baralis et al, 2004), introduced new method to prevent the misclassifications that caused by using the predetermined default class label as in CBA algorithm do, it introduces two rule levels, the first level checked the classifier rules, if no rule match to the new object B, a second level will be checked, this method reduces the use of the default, but proves costly in the processing time.

Input: Classifier (*R*), test data set (*Ts*), array *Tr*

Output: error rate *Pe*

Given a test data (*Ts*), the classification process works as follow:

```
1 For each test case  ts Do
2 For each rule r in the set of ranked rules R Do
3    Find all applicable rules that match ts body  and store them in Tr
4     If Tr is not empty Do
5          If there exists a rule r that fully matches ts condition
6               assign r's class to ts
7           end if
8       else assign the default class to ts
9     end if
10 empty Tr
11 end
12 end
13 compute the total number of errors of Ts;
```

Figure 2.1 CBA prediction approach.

2.3 Group of Rule Class Prediction Method(s)

In this section we will introduce some of the new AC algorithms and emphasize and concentrate in its prediction phase that used a group of rule class allocation method.

In (Hadi ,2013), Expert Multi Class Based on Association Rule approach, Dominant Class Label prediction method was introduced; it classifies the test data with dominant class of the matched rules.

The idea of this classification method as shown in (figure 2.2) is to divide a set of applicable rules to the new object into groups according to its class label and assigns the class label of the major group to the new object.

In classifying a new object (line1), the proposed method uses a simple technique, that mark any matched rule to the new object even it if it partially match and add it to the group which match its class label and then count each group rule items (line 6), and assigns the class label of the highest count group to the new object (line 10). In cases there are no rules match the new object, the predetermined default class label will be assigned to the test case (line7), this method is similar to that of (Zaïane and Antonie, 2002) but doesn't take the dominance factor in consideration which is a statistical measure computed for each rules category, and only the rule category above the user predefined threshold of dominance factor can take part in the classification decision.

```
For each test case ts Do
2       Assign=false
3       Find all applicable rules that match ts body in the set of the ranked rules R and store them in Tr
4      If Tr is not empty
5        Divide the rules in Tr according to the class label in separate groups
6        Count the number of rules for each group.
7      else Give the default class to ts and Assign=true
8     end if
9      If Assign = false
10        Give the dominant class to ts
11        Assign=true
12        Empty Tr
13     end if
14 end For
```

Figure 2.2 Dominant Class prediction approach.

In (Thabatah et al, 2010) Looking at the class approach was introduced, in the prediction phase it used the Highest Average Confidence per Class Prediction Method, The idea of this prediction method as shown in (figure 2.3) is to divide a set of applicable rules to the new object into groups according to its class label, calculating the average confidence of the rules in each group and assigns the class label of the highest average confidence group to the new object.

In classifying a new object case (line 1), the proposed classification algorithm divides all the applicable rules to new object body into groups according to the class labels (line 5). Then, it computes the average confidence of all the rules per group (line 6), and finally assign the class label of highest average confidence group to the new object instance (line 10), In cases there are no rules match the new object, the predetermined default class label will be assigned to the test case (line 7).

```
For each test case ts Do
2       Assign=false
3       Find all applicable rules that match ts body and store them in Tr
4      If Tr is not empty Do
5.        Divide the rules in Tr according to the class label in separate groups
6.        Compute the average confidence for each group.
7       else Give the default class to ts and Assign=true
8      end if
9     If Assign = false
10       Give the class with highest average confidence to ts
11       Assign=true
12       Empty Tr
13     end if
14 end for
```

Figure 2.3 Highest Average Confidences per Class Prediction Method

In (Yuhanis and Refai, 2013) Modified Multi-class Classification using Association Rule Mining used the same principle in look at class approach except that if the two divided groups have the same confidence, the approach predict the class of the new Ts with the highest average support of the groups instead of average confidence.

In (Ramasubbareddy et al, 2011) Classification based on Positive and Negative Association Rules approach Introduces new classification methods to improve the accuracy of the classifier based on generating positive and negative rules, negative rule encapsulates relationship between the occurrences of one set of items with the absence of the other set of items, the algorithm generates four different negative rules forms and one positive rule form.

Simply a rule **A→ -B** has a support **S** if the percentage of transactions in T contain item A and at the same time doesn't contain item B.

In Classification based on Positive and Negative Association Rules The set of positive and negative rules are ordered by confidence and support, the sorted set of rules are representing the classifier, to classify a new Ts, a set of applicable rules within predetermined confidence margin is selected, the interval of selected rules is between the confidence of the first ranked rule minus the confidence margin, that's mean there is a high margin which is the top ranked rule confidence and a lower margin which is the top ranked rule confidence minus the confidence margin, the applicable rules to the Ts are divided according to its class label, then class groups are ordered according to the average confidence per class, the classification made by assigning the Ts to the class group with the highest score, the score is the summation of the confidences (with positive and negative confidences divided by the rules number,

In(Pal and Jain,2010) the proposed Combinatorial Approach of Associative Classification, firstly generates a binary combination of items including the classes using combinatorial mathematics, then it finds all frequent binary combination by eliminating all invalid combinations, to find the strong rules it used the confidence threshold (min_confidence), and produced the classifier by pruning all small strong items included in the large strong items, that mean any subset of a long strong rule will be eliminated, to classify an object for a test data, it match the attributes of the test item_set with that of the classifier rules and derived a matrix called classmat, where the last attribute in classmat stored the max number of matched attributes of the test object with the classifier rule, so we will have two dimensional array storing the rule and the max matched attributes between the rule body and the test item_set, another matrix classfreq contains the classes frequency with the maximum number of attributes has classified an object, another two dimensional array created he class index and the frequency of the max matched attribute, this data extracted from classmat matrix, the class index with the maximum value is the class of the object.

In (Tianzhong et al. 2010) classification based on attribute-value pair integrate gain (CAIG) collects the rules that match the new object, if all rules having the same class label, CAIG just simply assigns that label to the new object..

Otherwise, CAIG divides the rules in to groups according to the class label, where all rules in a group share the same class label.

CAIG use the Laplace expected error estimate and the support of rule to estimate the accuracy of rule, called the rule-strength.

in (Zaixiang et al,2013) the author proposed new AC approach called Association classification with KNN, the AC-KNN adopts the improved K-nearest neighbors principle to addresses the confliction between classifier rules, the AC-KNN chooses the K-nearest neighbors from selected training instances which covered by the best of predetermined n rules. To classify a new instance, AC-KNN (Figure 2.4), the approach chooses the top sorted n rules which matched that new instance (line 1).

If these rules predict the same class value, then the class value of the rules assigned to the new item_set (line 2-3). If the top n rules are conflicted, a KNN algorithm is applied (line 5-13), selecting all training instances which covered by the top n rules, calculate the distance between the test instance and each of the matched training instances, then ranking the matched training instances according to the lowest distance, Finally, divide these instances into groups according to class value and assign the class value of the group with the minimum average distance to the new instance.

Input: Training data set T; a set of rules R, Test instance O.

Output: class value assigned to O.

1: Select best n rules that matched test instance O from R.

2: If the best n rules predict the same class value C

3: Assign C to O.

4:Else

5: Collect all training instances covered by the best n rules.

6: For all t T1 do

7: calculate the distance between t and O.

8: end for.

9: Sort K-nearest neighbors in ascending order.

10: Select K nearest neighbors with lowest distance.

11: Divide K- nearest neighbors into groups according to class value.

12: Calculate average distance for each group.

13: Assign the class value C of the group with the lowest average distance to O.

14: end if.

Figure 2.4 AC- KNN prediction phase

2.4 Summary:

The main advantage of using multiple rules in the Ac prediction phase is that more than one rule contributing in the prediction, which limits the chance of favoring a single rule to predict all test objects satisfying its condition.

In this chapter we have surveyed some of the new AC approaches and discussing its prediction phase, as shown in Table (2.1).

In the next chapter we will describe in more the new association classification using naïve bayes theorem and its steps with providing a complete example of how the approach operates.

Table 2.1 AC Approaches

AC method	Data layout	Rule generation	Ranking	pruning	Prediction method	Reference
CBA	Horizontal	Apriori candidate generation	Support, confidence, rules generated first	Pessimistic error, database coverage	Maximum likelihood	Liu et al, (1998)
L3G	Horizontal	FP-growth approach	Support, confidence, rules cardinality, items lexicographical	Lazy pruning	Maximum likelihood	Baralis et al, (2004)
LC	Horizontal, Vertical	Apriori with common class label	Support, confidence	Matched rules to test item	Average confidence	Thabateh et al, (2010)
EMCAR	Horizontal	Apriori candidate generation	Support, confidence, rules cardinality, rules generated first	Database coverage	Dominant class	Hadi (2013)
CAIG	Horizontal	Apriori candidate generation	Support, confidence	Matched rules to test item	Laplace expected error	Tianzhong et al. (2011)
MMCAR	Vertical	Tid-list intersections, best related items	Support, confidence, cardinality	Database coverage	Average confidence, average support	Yuhanis and Refai (2013)
CAAC	Horizontal	Combinatorial Mathematics	Max matched attributes	Lazy pruning	Class frequency with Max matched attributes	Pal et al, (2010)
CPNAR	Horizontal	Apriori candidate generation	Support, confidence	confidence margin	Confidence summation	Ramasubbareddy et al,(2011)
AC-KNN	Horizontal	Apriori candidate generation	Confidence, mutual association between itemsets	information entropy	KNN	Zaixiang et al, (2013)

CHAPTER THREE

ASSOCIATIVE CLASSIFICATION USING NB

3.1 Introduction

The associative classification technique integrates two of the most well-known data mining tasks, association rule mining and classification, to build classification system for the purpose of allocation decision, Several studies such as: (Thabtah et al, 2010),(Hadi, 2013),(Baralis et al, 2008) provided evidences that association classification approaches are deriving more efficient classifiers than traditional classification techniques, such as decision trees (Kingsford and Salzberg,2008), and rule induction (Hilage and Kulkarni, 2012).

The proposed approach improved the prediction accuracy and efficiency of CBA techniques by dealing directly with two problems. The first one reducing CPU times by eliminating the ranking step (step 4 Figure 3.1), and most important issue is to develop an efficient method for building the classifier using naïve Bayes (step 5 Figure 3.1).

The proposed approach (Figure 3.1) deals with continuous attributes as well as categorical attributes and improves upon CBA method by eliminating the ranking step and using a probabilistic measure which based on the known Bayes theorem to improve the predicting power of CBA.

The proposed approach is presented in Section 3.2 where details about pre-processing the data, rule discovery, rule evaluation phase, and prediction of test data objects are discussed. Section 3.3 is devoted to pointing out the differences between the proposed method and other AC approaches.

3.2 The Proposed Approach

The proposed approach operates using two main stages to generate a classifier (Figure 3.1):

1. Generating a complete set of CARs (Classification Association Rules).
2. Using Naïve Bayes principle on CARs to classify new unseen item_sets.

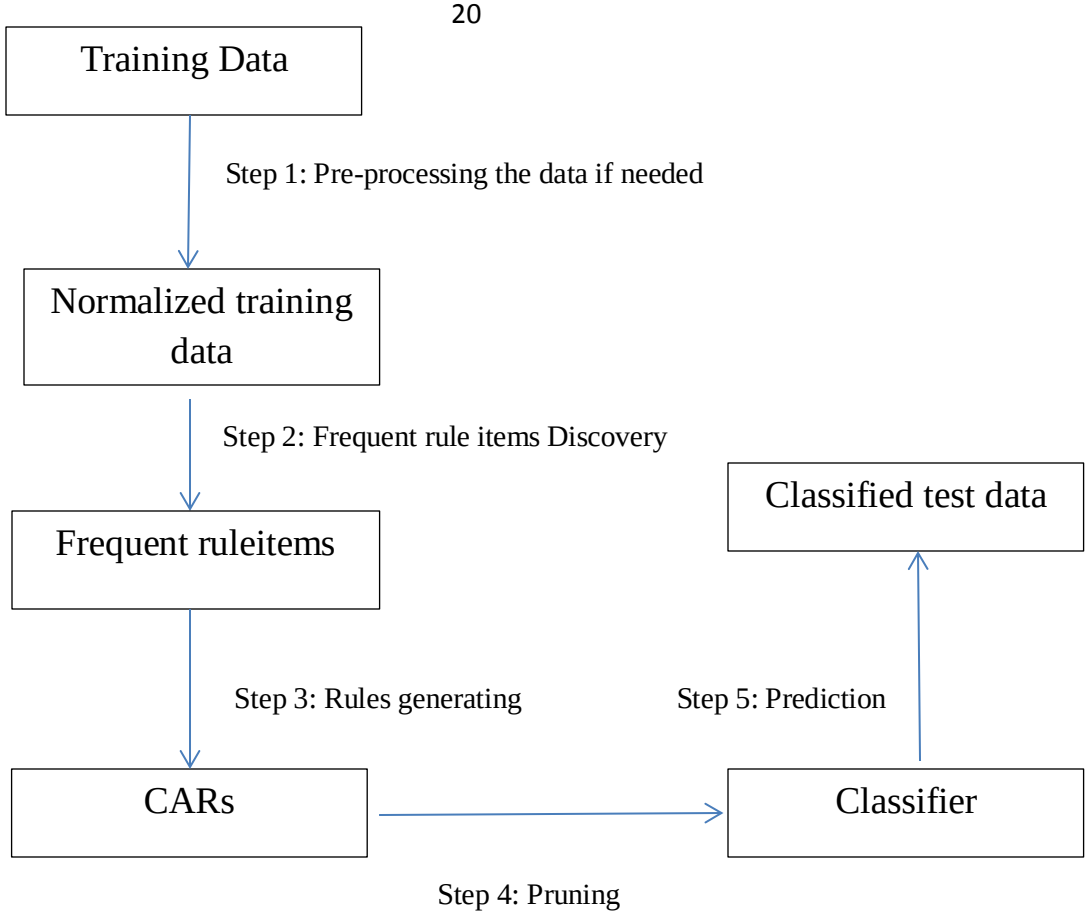

Figure 3.1 the steps of the proposed approach.

In the first stage(Steps 2 and 3), We used the known apriori approach in order to generate the strong rules, the training data is scanned and generate frequent rule_items, then combines the rule items to generate candidate rule items, any rule item with support and confidence larger than min support and min confidence, is labelled as potential rule.

In the second stage (steps 4 and 5), each potential rule match the new test item_set added to rules group which match its class label, finally the naïve Bayes principle applied to each rules group to classify the new instance, in section 3.2.4 we explained the prediction approach in more details.

Figure 3.1 represents the proposed algorithm, in step 1 the training data pre-processed if needed in order to generate normalized data that can be used in the proposed approach, in step 2 the normalized training data scanned and each item_set had support value bigger than the predefined min support considered as frequent item_set, in step 3 the frequent items_sets scanned and each

one with confidence value bigger than the predefined min confidence considered as strong association classification rule.

In step 4, each strong rule whose body matched the test data body added to the rules group which match its class label, in step 5 the naïve Bayes principle applied to classify the new test data.

The steps will be explained in details in the next sections with emphasizing how to deal with the CBA approach problems.

An example of how is our proposed method operate will be presented in section 3.3

3.2.1 Pre-Processing

For the discretization/normalization step in our proposed model, java open source code for The LUCS-KDD-DN software used, for more details see Appendix A, here we briefly explain how the pre-processing operates:

1. Calculating the range of the attributes values, and dividing them into user predefined sub groups count.
2. Count the attributes in the subgroup, and the percentage of each with respect to the class label.
3. For each sub group, identify the dominant class.
4. Combine sub-ranges with identical dominant classes to form a set if *divisions*.
5. If the number of division is less than or equal to the maximum desired number of divisions stop. Otherwise *merge* divisions until the maximum is reached

3.2.2 Frequent Rule_items Discovery

The frequent item_sets discovery method passes once over the data and counts the occurrences of 1-itemsets, from which it determines those that have support greater than the predetermined min support and identify them as frequent 1-itemsets.

From these frequent 1-itemsets, we produce 2-item_sets, and in case the 2-itemsets passes the predefined support it will be defined as frequent 2-itemsets.

This method will run iteratively to produce all frequent item_sets.

To show how we determine a frequent item_sets are generated a detailed example is discussed in section 3.2.4.

3.2.3 Rules Generation

In this section we briefly explain how support and confidence for rule_items are calculated and show how rules are generated.

suppose we have classification association rule A, X--→Y, the support for this rule can be calculated by counting number of data sets which contain X and Y divided by the total records count, the confidence of the rule A is the ration of:

Confidence (A) = Support of A/ support of X.

Frequent rule_items are generated recursively from rule_items conditions having a smaller number of attributes, starting from frequent one- rule_items derived in a single pass through the training data set.

3.2.4 Prediction of Test Instances

The idea of the proposed prediction method as shown in Figure 3.2 is to choose the class label of the rules with the highest naive Bayes confidence that match the test data in order to allocate class label to the unseen object. In classifying a new unseen object (line 1), the algorithm divides the rules that matched the body of the new unseen object into groups according to the class labels (line 5). Then, it calculates the proportional confidence for each class group by multiplying the rules confidence values in each group (line 6), and finally calculate the class naïve value by multiplying the proportional confidence per class group and the class proportion in all the applicable rules (line 7) and classifies the new unseen object to the group class label with the highest class naïve value (line 11). Indeed, using more than one rule based on their frequency (weights) and confidence promise to come up with a global weight for the rules classes that give s each rule participating in the allocation decision by its confidence then selecting the class with the largest weight to assign it to the test data, this will keep us far from taking biased decision because there could be multiple applicable rules to the test instance, which makes selection of one rule unfair.

Input: strong generated rules (*R*), test dataset (*Ts*), array *Tr*

Output: Accuracy.

Given a test data (*Ts*), the classification process works as follow:

1 For each test case *ts* Do

2 Assign=false

3 Find all applicable rules that match *ts* body and store them in *Tr*

4 If *Tr* is not empty Do

5. Divide the rules in *Tr* according to the class label in separate groups

6. Compute the prop confidence for each group by multiple group rules confidences.

7. compute class naïve value by multiplying group prop confidence by class proportion

8. else Give the default class to *ts* and Assign=true

9. end if

10. If Assign = false

11. Give the class with highest class naïve value to ts

12 Assign=true

13 Empty *Tr*

14 end if

15 end for

16 Compute Accuracy

Figure 3.2 the proposed prediction method

In order to calculate the proportional confidence for each group, we multiple all the rules confidences in each other:

Accumulated Confidence(R/Ci) = confidence (r1/Ci) * confidence (r2/Ci) *…* confidence (rn/ Ci)

Suppose we have five matched rules to the unseen object which we need to predict its class label, we divide the rules according to the class label; all the rules with the same class label will be on the same group, as shown in table 3.1.

Table 3.1 calculating the proportional confidence

Group C1	Group C2
Confidence (r1/C1)=0.8, Confidence (r2/C1)=0.9, Confidence (r3/C1) =0.4.	Confidence (r4/C2)=0.3, Confidence (r5/C2) =0.5.

Accumulated confidence for group C1= Confidence (r1/C1)* Confidence (r2/C1)* Confidence (r3/C1).

Accumulated confidence for group C1=0.8*0.9*0.4=0.288.

Accumulated confidence for group C2= Confidence (r4/C2)* Confidence (r5/C2.

Accumulated confidence for group C2=0.3*0.5=0.15.

To compute the class naïve value we multiple the proportional confidence of each group by the class proportion.

Class naïve value (Ci) = Accumulated Confidence(R/Ci)* ci count/ (c1+c2_...+cn).

For example shown above:

Class naïve value (C1) = Accumulated Confidence(R/C1)* c1 count/ (c1+c2).

Class naïve value (C1) = 0.288* 3/ (3+2) =0.1728

Class naïve value (C2) = Accumulated Confidence(R/C2)* c2 count/ (c1+c2).

Class naïve value (C2) =0.15* 2/ (3+2) = 0.06

3.3 Example of how Associative Classification using NB operates

Assume we have the training data in Table 3.2 (UCI machine learning repository) and we want to generate frequent rules items.

Table 3.2 Subset of the contact-lenses data set

age	astigmatism	tear-prod-rate	contact-lenses
young	no	normal	soft
young	yes	reduced	none
young	yes	normal	hard
pre-presbyopic	no	reduced	none
pre-presbyopic	no	normal	soft
pre-presbyopic	yes	normal	hard
pre-presbyopic	yes	normal	none
presbyopic	no	reduced	none
presbyopic	no	normal	none
presbyopic	yes	reduced	none
presbyopic	yes	normal	hard

As we can see from the table, the training data consist of 4 attributes {age, astigmatism, tear-prod-rate, and contact-lenses}, the last attribute contact-lenses presents the class label.

To generate the frequent item sets we pass over the training data and generate 1 frequent item_set as shown in figure 3.3, suppose the predetermined min_support is 3, any 1 item_set with support< min_support will not take a rule in generating the 2 frequent item_sets, the 2 item_set simple an intersection between the 1 frequent item_set that has the same class label, again any 2 item_set with support below the min_support will not take a rule in generation the 3 item_sets and so on .

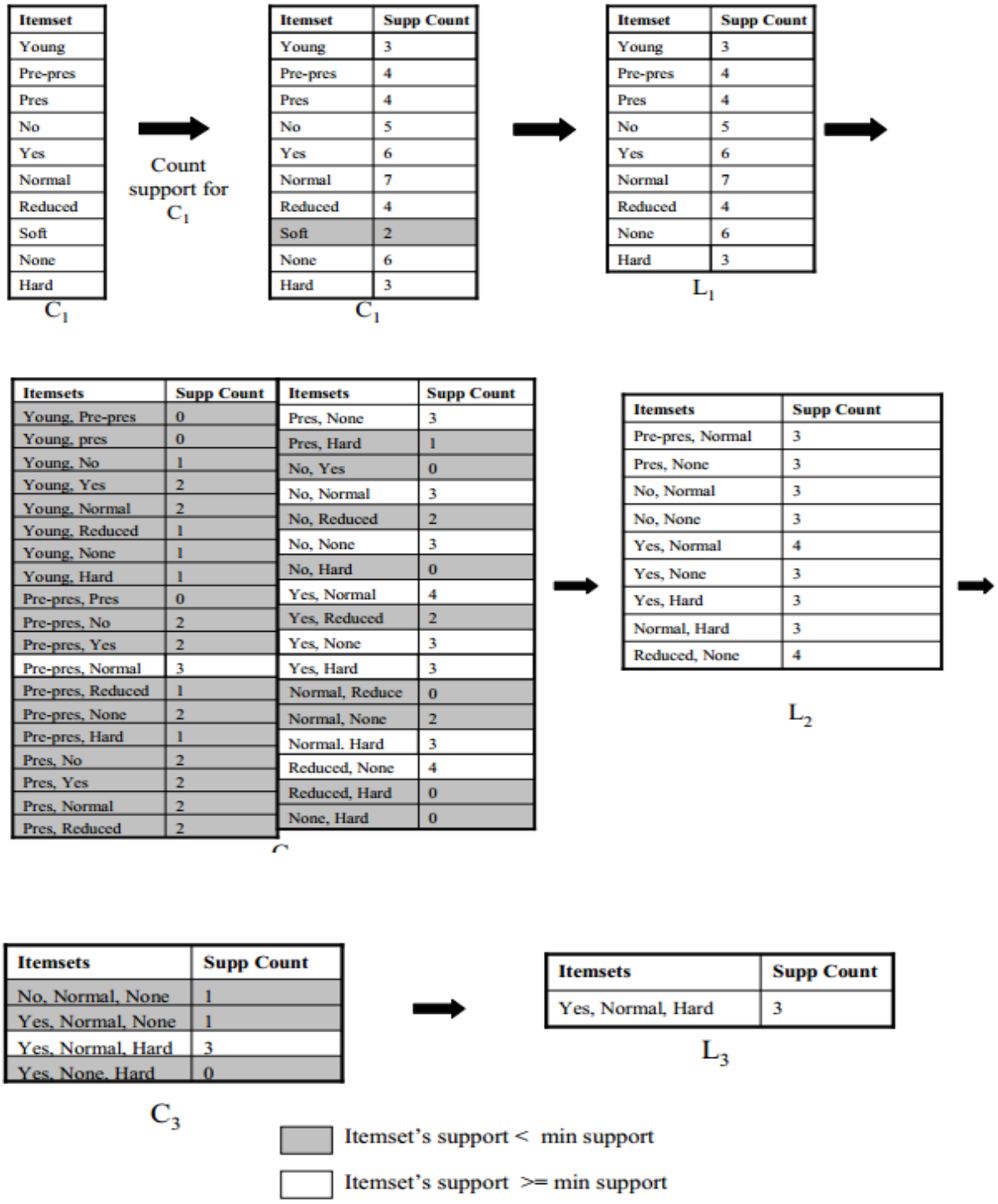

Figure 3.3 Frequent item_sets generation with min_support of 25%

To generate the Rules we calculate the confidence of each frequent item_set which its right hand side is class label, each rule with support and confidence greater than min support and min confidence respectively considered as strong rule.

Table 3.3 the generated Rules with min_support 25% and min confidence 50%

The rule	support	confidence
astigmatism(yes)→contact_lenses(none)	33.3%	50%
astigmatism (yes)→ contact_lenses (hard)	33.3%	60%
tear_prod_rate(reduced)→contact_lenses(none)	44.4%	100%
astigmatism(yes),tear_prod_rate(normal)→contact_lenses(hard)	33.3%	75%

As we can note that the class contact_lenses(soft) wasn't represented in any rule because it didn't pass the minimum support and confidence, in order to avoid such cases, we prefer always to choose lower minimum support value.

Suppose we have new itemset to predict {age(young),astigmatism(yes), tear_prod_rate(normal)}.

We match the body of rules (rules in table 3.3) with the new item_set data, any rule body partially or fully match is marked.

In our example: {age(young),astigmatism(yes), tear_prod_rate(normal)}.

Table 3.4 partially and fully match rules to the test data.

Rules	Support	confidence
astigmatism(yes)→contact_lenses(none)	33.3%	50%
astigmatism (yes)→ contact_lenses (hard)	33.3%	60%
astigmatism(yes),tear_prod_rate(normal)→contact_lenses(hard)	33.3%	75%

Then the rules divided according the class label to groups:

Table 3.5 partially and fully match rules of class label contact_lenses(hard)

astigmatism (yes)→ contact_lenses (hard)	33.3%	60%
astigmatism(yes),tear_prod_rate(normal)→contact_lenses(hard)	33.3%	75%

Table 3.6 partially and fully match rules of class label contact_lenses(none).

astigmatism(yes)→contact_lenses(none)	33.3%	50%

Now we calculate the prop confidence for each group:

in contact_lenses (hard) group we have 2 rules(table 3.4) with confidence values 60% and 75%, we multiple the confidence value of the rules in each other

Hard group prop confidence: 0.6*.75=0.45.

None group= 0.5

To calculate the class naïve value, we multiply each group prop confidence by the class proportion (number of the rules in the group divided by the rules in all the groups)

Class naïve value= prop confidence *class proportion

Class contact_lenses (hard) has 2 rules (table 3.4) and Class contact_lenses (none) has 1 rule.

Class naïve value for hard group=0.45*(2/3) =0.3

Class naïve value for none group=0.5*(1/3) =0.165

So the new test item_set will be classified as contact_lenses (hard) because the Class naïve value for hard group is higher than the class naïve value for none group.

3.4 Evaluation Methods

Another essential step in association classification models is measuring the classifier quality on the test data, if the rules produced in the learning phase accurately predict the test objects class, we accept it, on the other hand, if there are several misclassification, we reject them, So, how we can measure the effectiveness of our proposed model?

There are many evaluation methods proposed in classification such as error-rate (Witten and

Frank, 2000), recall-precision (Van, 1979) and others,

In the proposed approach we used the accuracy measure to evaluate the effectiveness of our classifiers. Using this method, the classifier simply predicts the class of test data objects; if it is correct, this will be counted as a success, and otherwise it will be counted as an error. The number of success cases divided by the total number of cases in a test data set gives the overall accuracy on this data. The accuracy of a classifier on a test data set measures its predictive accuracy.

3.5 proposed method Features

- The proposed model overcome the ranking rule by giving the rules to participate in classification by its rule confidence value disregards to its location in the array, in a direction toward reducing CBA complexity.

- In our proposed approach we are investigated the possibility of using multiple rules in prediction, which consequently should enhance the confidence in the prediction decision, since more than one rule contributed in such decision.

CHAPTER FOUR

EXPEREMENTAL RESULTS

4.1 Introduction

In this chapter we will provide details of the used data in the experiments and comparing the proposed approach with the traditional classification methods and the new AC approaches in terms of accuracy and CPU execution time, moreover will evaluate of the selected data sets using another known evaluation methods confusion matrix.

4.2 Data Collection

At this stage, data sets have been acquired through the UCI machine learning repository which can be accessed at http://archive.ics.uci.edu/ml/ datasets.html. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for conducting empirical studies on machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets. As an indication of the impact of the archive, it has been cited over 1000 times, making it one of the top 100 most cited "papers" in all of computer science; currently it contains 239 different data sets as a service to the machine learning community, we choose five of the most popular data sets (Bache and Lichman ,2013)

4.3 Experimental Results

Experiments on five of the most used data sets from the UCI machine learning repository were conducted using our proposed model, 5 popular classification techniques: decision Naïve bayes, RIPPER, CBA, MCAR, LC, and some of the new AC techniques have been compared to the proposed model in terms of accuracy, CPU time

The experiments were conducted on IBM laptop CORE i5 machine with 4GB RAM, the proposed approach and CBA implemented using JAVA Console programming language with a *min support* 3% and *min confi*dence of 30%. The *min suppo*rt has been set to 3% because several experiments reported in (Li et al, 2001, Thabteh et al, 2005) suggested that min support between 2-5 % is one of the rates that would achieve an excellent balance between accuracy power and the classifiers size, but, The confidence threshold has a lesser impact on the behavior of

association classification approaches and it has been set to 30%. In the next sections, we will provide a comparative analysis between the results of our proposed prediction approach and other known algorithms using several data sets from the UCI Machine Learning Repository.

4.3.1 The accuracy Power

Figure 4.1 shows us the accuracy rates of classification based on association, decision tree RIPPER, MMCAR, MCAR, LC, our proposed algorithms obtained on the chose 5 data sets. The experiments of C4.5 and RIPPER algorithms were conducted using the *Weka* software system (Weka, 2001). CBA experiments were conducted using an implementation version provided by the authors of (CBA, 1998) and MCAR, MMCAR, LC results were obtained from the models published papers.

The results shown in Table 4.1 that our proposed method outperforms the other rule learning techniques on the majority of the data sets in terms of accuracy.

Table 4.1 Accuracy of the different approaches

	Iris	Baloons	Glass	Pima	Lenses
NB	96		48.59	74.3	83.33
RIPPER	94.66		68.69	73.3	75
CBA	93.25	100	69.89	75.49	80
MCAR	95.32		69.67	78.54	75
LC	94.25	100	69.89		79
MMCAR	94.26		74.2	74.44	
AC-NB	96	100	73.2	74.3	100

In the following few graphs we are going to compare the accuracy of our proposed model with our classification approaches.

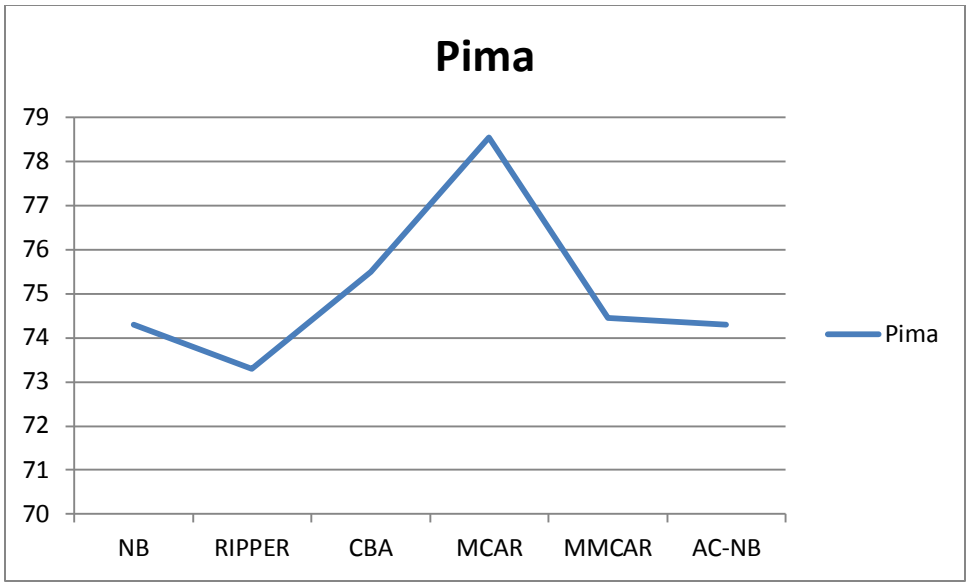

Figure 4.1 Accuracy of different approaches on Pima data set.

As we can see from figure 4.1, the proposed model has high accuracy (74.3) relatively to the other approaches.

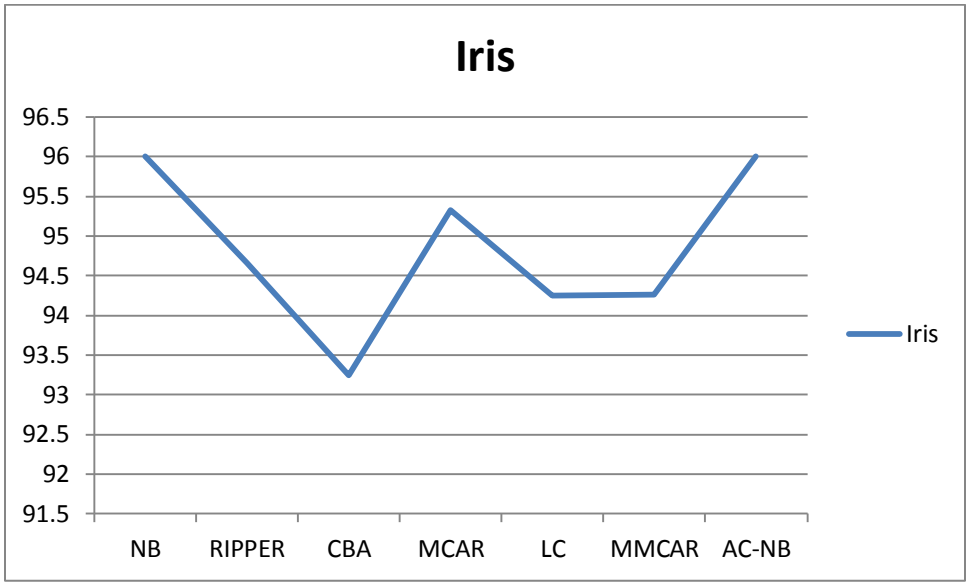

Figure 4.2 Accuracy of different approaches on Iris data set

As we can see from figure 4.2, the proposed model has the highest accuracy (96) among all the other approaches, This indicates that using naïve bayes increase the accuracy.

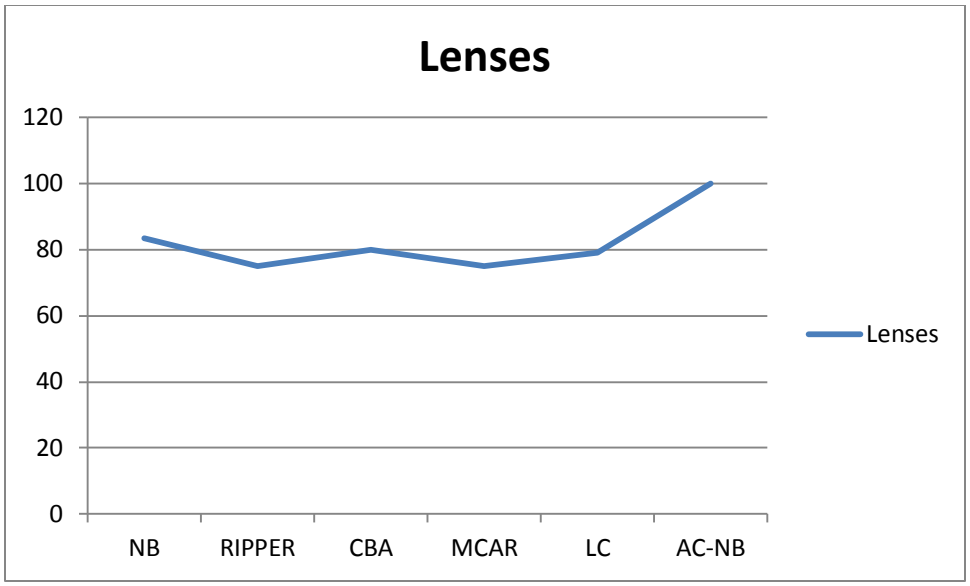

Figure 4.3 Accuracy of different approaches on Lenses data set

As we can see from figure 4.3, the proposed model has the highest accuracy (96) among all the other approaches; This indicates that using naïve bayes increase the accuracy in great way.

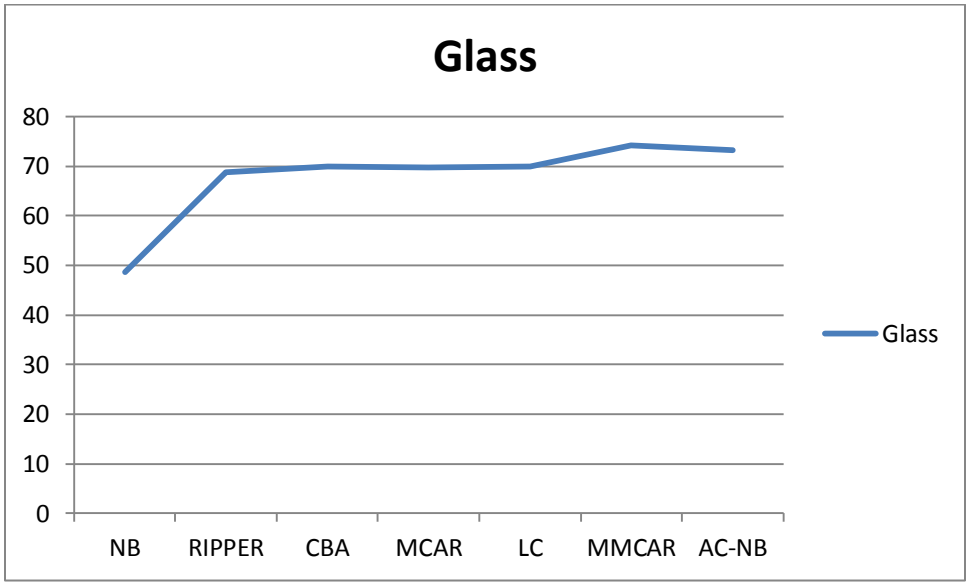

Figure 4.4 Accuracy of different approaches on Glass data set

As we can see from figure 4.4, the proposed model has the second highest accuracy (73.2) among all the other approaches; this indicates that using naïve bayes increase the accuracy in great way.

As noticed from table 4.6 that only 2 approaches other than our approach show the accuracy on balloons data sets, it's also shown that the 3 approaches reached 100% accuracy.

In table 4.6 we show the average accuracy of the various used algorithms among the used UCI data sets,

The average accuracy was calculated by:

$$\text{average(k)} = \sum_{k=1}^{k=n} (the\ approah\ accuracy\ on\ k\backslash n$$

Table 4.2 Average accuracy of different approaches on the five UCI data sets

The Classification approach	Average accuracy
NB	75.55
RIPPER	77.9
CBA	83.7
MCAR	79.6
MMCAR	81.0
LC	85.8
AC-NB	88.8

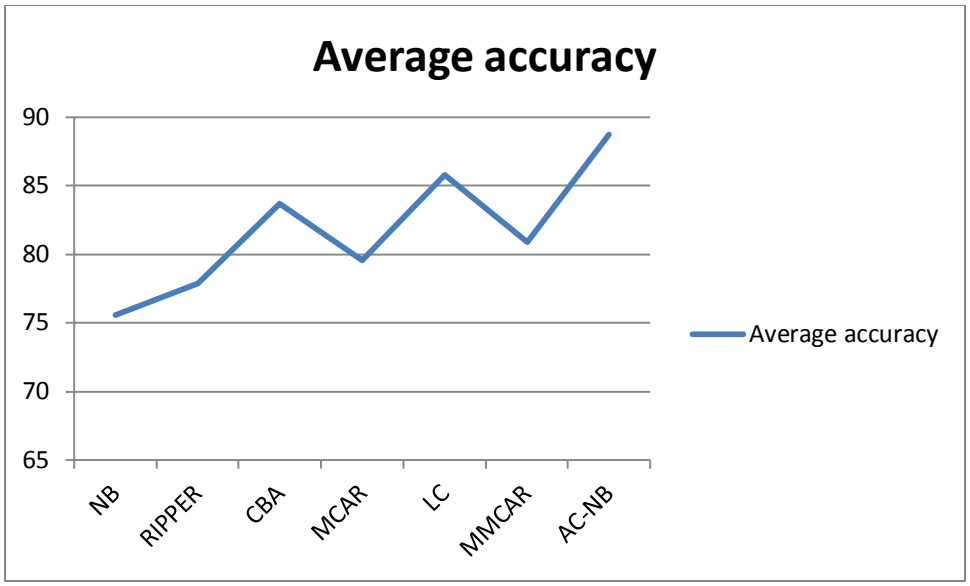

Figure 4.5 Average accuracy of different approaches on the five UCI data sets.

As we can see from figure 4.5, the proposed model has the highest average accuracy (88.7) among all the other approaches; this indicates that using naïve bayes increase the accuracy in great way.

The processing time for CBA and our proposed model is recorded and presented in table 4.7, for example for glass dataset, the execution time has declined from 3054 ms in the CBA to 1850 ms in the proposed model with a significance difference of 39.4%, this obvious difference generated since the ranking step omitted.

It is obvious from the numbers displayed in table 4.3 that the proposed model save a large amount of processing time in compared to CBA.

Table 4.3 Execution time (milliseconds) of CBA and the proposed model

Dataset	CBA	AC-NB	Difference (%)
Ballons	992	350	64.7
Contact	241	220	8.7
Iris	190	600	-68.3
Glass	3054	1850	39.4
Average CPU	1119	755	32.5

After analyzing the Iris dataset, it turn out that more than 70 generated rules have 100% accuracy, and the rules generated for the different classes are mutually exclusive which indicate that the data set data are highly asymmetric, and the classification in the CBA based the top ranked rule which almost with 100 % confidence value.

4.4 Approach Implementation.

The AC-NB tool is a user-friendly application developed in order to classify the relational dataset, the application was customized from an open java source code that developed in Liverpool University, and this system is essentially processes as follow:

1) The user loading the relational (table) dataset, and specifying the destination file storage by using browse buttons.

2) The user enters the two-threshold values of minimum support and minimum confidence.

3) The default cross validation method is 10 fold, but it can be performed using 50:50(50% training/50 testing).

4) The user pressing the AC-NB button will trigger the approach to operate.

5) The Import result file, allow the user to import the out file to the tool GUI.

Figure 4.6 shows the snapshot of the main screen of our tool GUI.

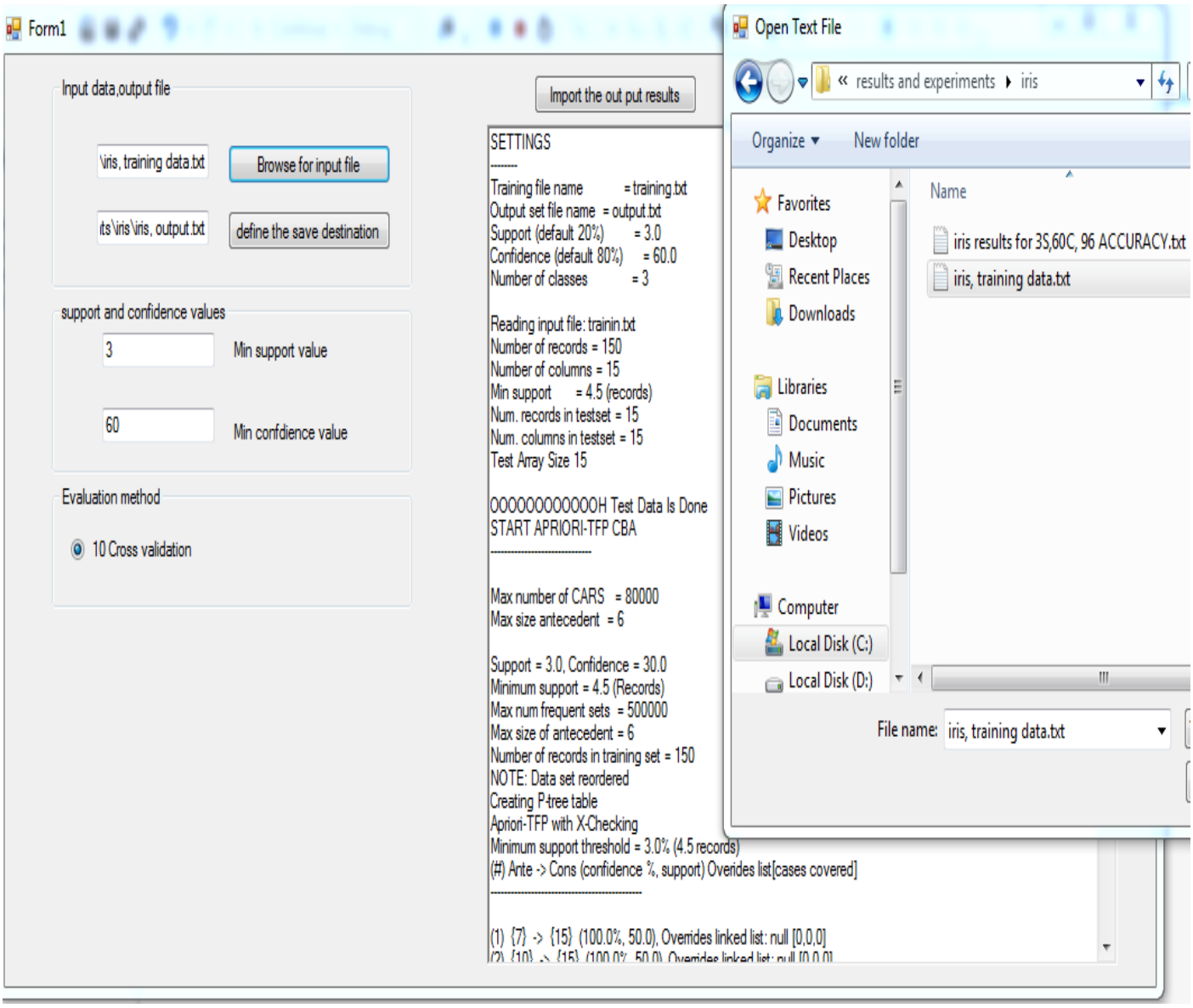

Figure 4.6 AC-NB main screen.

CHAPTER FIVE

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

In this chapter, we have introduced the evaluation process of the proposed prediction phase, the outcome is a new effective prediction phase in the AC approaches, the proposed method has number of new features such as lowering the complexity level of CBA by overcoming the ranking step by giving the rule to participate in the prediction regardless his location in the array and using multiple rules in the prediction phase which consequently enhanced the confidence in the prediction decision.

Performance studies on five of the most used data sets from UCI data collection indicated that our proposed method is highly competitive when compared with traditional classification algorithms such as RIPPER and C4.5 in term of prediction accuracy. Furthermore, our proposed model scales well if compared with popular AC approaches like CBA with regards to prediction power and the CPU execution time.

5.2 Future Works

The proposed work can be extended in many directions. These include:

1. Applying the naïve Bayes theorem in the rule generating step in order to reduce number of generating rules
2. Investigating the algorithm's scalability. Scalability measures the solution's ability to deal with large scale problems, without losing its accuracy. This is an important attribute for any deployable solution.
3. A large empirical study with different data sets can be performed to confirm the obtained results. Data sets can be obtained from other corpus.

Appendix

Appendix A: The Data used for the Experimental Purposes

A.1 Iris Data Set

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day.

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Predicted attribute: class of iris plant.

Table A.1 **Iris Data Set Information**

Data Set Characteristics	Multivariate	Number of Instances	150	Area	Life
Attribute Characteristics	Real	Number of attributes	4	Date donated	1988-07-01
Associate Task	Classification	Missing values?	No	Number of web hits	434651

The attribute Information are:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

A.2 Lenses Data Set

The examples are complete and noise free. The examples highly simplified the problem. The attributes do not fully describe all the factors affecting the decision as to which type, if any, to fit.

Table A.2 Lenses Data Set Information

Data Set Characteristics	Multivariate	Number of Instances	24	Area	N/A
Attribute Characteristics	Categorical	Number of Attributes	4	Date Donated	1990-08-01
Associated Tasks	Classification	Missing Values?	No	Number of Web Hits	39171

The Attribute Information are:

1.age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic
2. spectacle prescription: (1) myope, (2) hypermetrope
3. astigmatic: (1) no, (2) yes
4. tear production rate: (1) reduced, (2) normal

5. Classes:
-- the patient should be fitted with hard contact lenses,
-- the patient should be fitted with soft contact lenses,
-- the patient should not be fitted with contact lenses.

A.3 Pima Indian Diabetes Data

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm; see the paper for details.

Table A.3 Pima Indian Diabetes Data

Data Set Characteristics	Multivariate	Number of Instances	768	Area	Life
Attribute Characteristics	Integer, Real	Number of Attributes	8	Date Donated	1990-05-09
Associated Tasks	Classification	Missing Values?	Yes	Number of Web Hits	86458

The attribute Information are:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg).
4. Triceps skin fold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)^2).
7. Diabetes pedigree function.
8. Age (years).
9. Class variable (0 or 1)

A.4 Balloons Data Set

There are four data sets representing different conditions of an experiment. All have the sattributes.

Table A.4 Balloons data set Data

Data Set Characteristics	Multivariate	Number of Instances	16	Area	Social
Attribute Characteristics	Categorical	Number of Attributes	4	Date Donated	N/A
Associated Tasks	Classification	Missing Values?	No	Number of Web Hits	57581

the attribute Information are:

(Classes Inflated T or F)
Color: yellow, purple
size: large, small
act: stretch, dip
age: adult, child
inflated: T, F

A.5 Glass Data Set

The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence...if it is correctly identified!

Table A.5 Glass data set

Data Set Characteristics	Multivariate	Number of Instances	214	Area	Physical
Attribute Characteristics	Real	Number of Attributes	10	Date Donated	1987-09-01
Associated Tasks	Classification	Missing Values?	No	Number of Web Hits	80762

The attribute Information are:

1. Id number: 1 to 214
   2. RI: refractive index
   3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
   4. Mg: Magnesium
   5. Al: Aluminum
   6. Si: Silicon
   7. K: Potassium
   8. Ca: Calcium
   9. Ba: Barium
   10. Fe: Iron
   11. Type of glass: (class attribute)
   -- 1 building_windows_float_processed
   -- 2 building_windows_non_float_processed
   -- 3 vehicle_windows_float_processed
   -- 4 vehicle_windows_non_float_processed (none in this database)

-- 5 containers

-- 6 tableware

-- 7 headlamps

Appendix B: Data Discretization:

in order to normalize a continuous data, the discretization tool needs to know the schema of the converted data, the schema file must contains 3 lines[Figure B.1]:

Line 1: describes the data type of each field, there are only 4 permitted values to be used: integer, double, nominal, and unused.

Line 2: the attributes names, it's not used in the normalization process, but it may be useful for clarification purposes.

Line 3: the legal values for each data type, for integer, double, and unused data types the legal value is null for nominal values are separated by '/'.

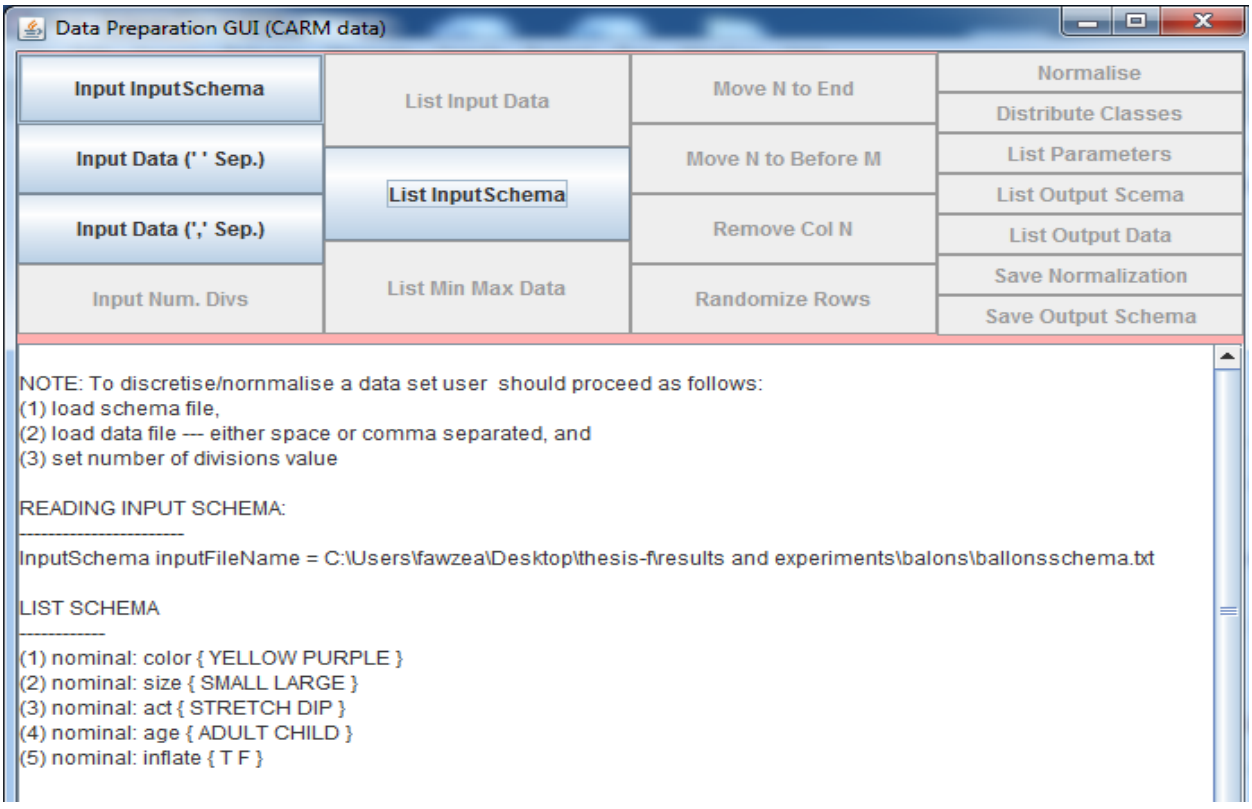

Figure B.1: The main screen of the discretization tool

For example the schema file for balloons UCI data set, is as follows:

nominal nominal nominal nominal nominal

color size act age inflate

YELLOW/PURPLE SMALL/LARGE STRETCH/DIP ADULT/CHILD T/F.

The second in the discretization process is loading the input data file, it could be space separated or comma separated.[figure B.2]

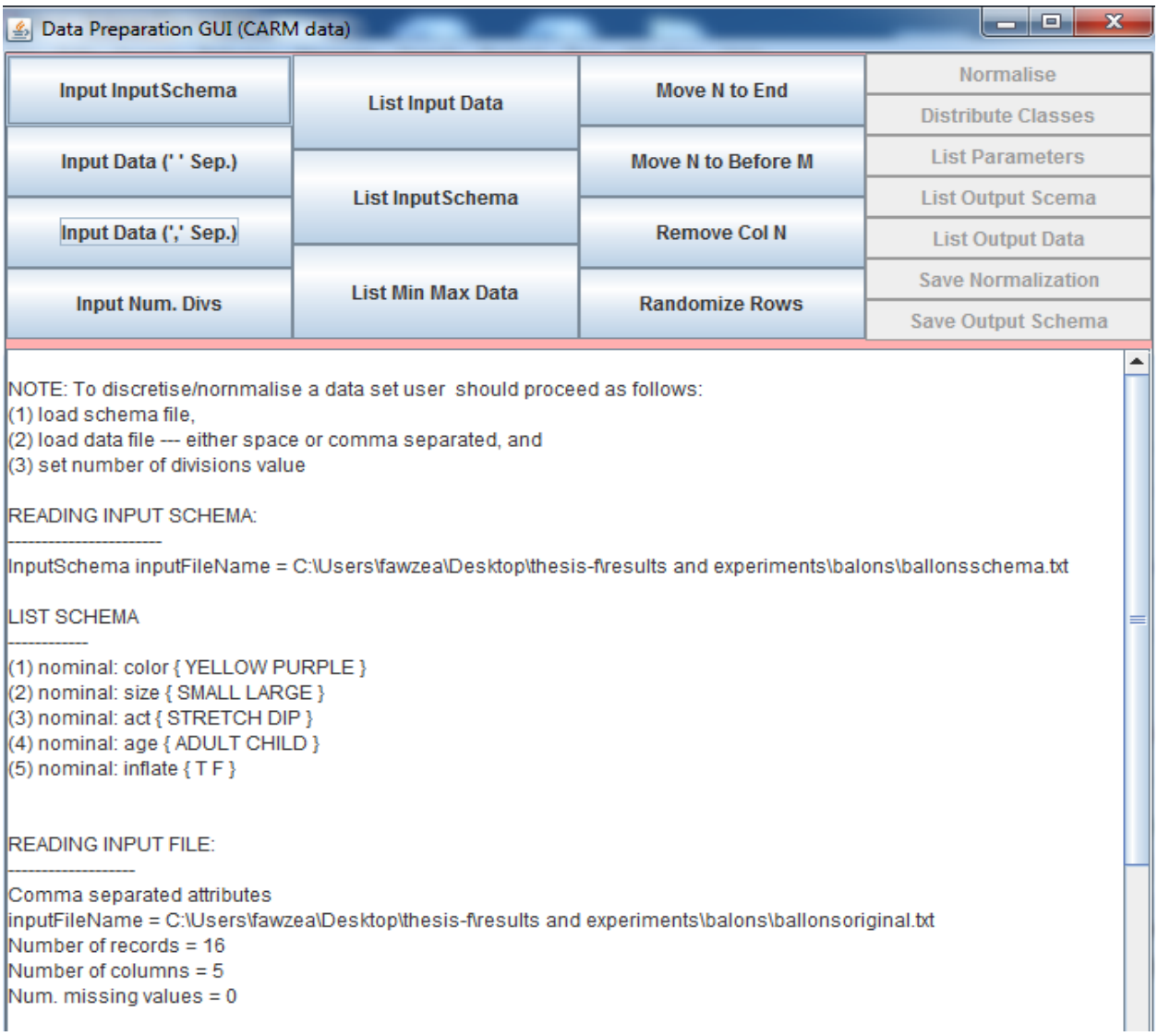

Figure B.2 Reading input file to be discretized

The third step is choosing the number of divisions, which mostly equal to the number of classes, then we can normalize the input data, for the balloons data the normalize data are show in figure below.[Figure B.3]

Figure B.3 discretized data

Then we can use this data as an input for the algorithm.

REFERNCES

Agrawal R., and Srikant R. (1994). Fast algorithms for mining association rule. In Proceedings of the 20th International Conference on Very Large Data Bases. Santiago, Chile, pp. 487_499.

Bache K., and Lichman M.,(2013).UCI Machine Learnin Repsitory, http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Baralis E., Chiusano S., and Garza P.,(2004). On support thresholds in associative classification, Proceedings of the 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus, (pp. 553-558).

Baralis E.,Chiusano S., and Garza, P. (2008). A Lazy Approach to Associative Classification, IEEE Trans. Knowledge Data Engineering 20,pp 156-171.

Fletcher Tristan (2009). Support Vector Machines Explained, UCL, pp 1-15.

Hadi Wael (2013). Expert Multi Class Based on Association Rule, I.J.Modern Education and Computer Science, Volume 3, pp 33-41

Hilage T., and Kulkarni R.(2012). Review of Literature on Data Mining, IJRRAS Journal Volume 10, pp 107-114.

Kingsford C., and Salzberg S.(2008). What are decision trees? nature biotechnology 26,pp 1011-1013.

Kumar V., and Rathee, N.,(2011). Knowledge discovery from database Using an integration of clustering and classification, international Journal of Advanced Computer Science and Applications 2, 29-33.

Li X., Qin D., and Yu C. (2008). Associative Classification Based on Closed Frequent Itemsets. FSKD 2008, 380-384.

Liu B., Hsu W., and Ma Y.,(1998) Integrating classification and association rule mining. Proceedings of the KDD, New York, NY 1998.

Liu Y.,Jiang Y.,Liu X., and Yang S.(2008). A combination strategy for multi-class classification based on multiple association rules. Elsevier, Knowledge-Based Systems, pp 786–793.

Merz C., and Murphy P., (1996). UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA.

Niu Q., Xia S.,and Zhang L., (2009). Association Classification Based on Compactness of Rules, Second International Workshop on Knowledge Discovery and Data Mining 2009, 245-247.

Pal R., and Jain C.,(2010). Combinatorial Approach of Associative Classification, Int. J. Advanced Networking and Applications Volume: 02, pp 470-474.

Ramasubbareddy B., Govardhan & Ramamohanreddy A. (2011). Classification Based on Positive and Negative Association Rules, International Journal of Data Engineering, (IJDE), Volume 2 ,pp. 84-92.

Tang Z., and Liao Q., (2007). A New Class Based Associative Classification Algorithm. IMECS 2007: pp 685-689.

Tianzhong H., Zhongmei Z., Zaixiang H., and Xuejun W. (2011). Classification Based on Attribute-Value Pair Integrate Gain, Database Theory and Application, Bio-Science and Bio-Technology Communications in Computer and Information Science Volume 258, pp 31-40 .

Thabateh F., Qazafi M.,McCluskeyy L.,and Abdel-Jaber H.(2010). A New Classifcation Based on Association Algorithm, Journal of Information & Knowledge Management, Vol. 9,pp 55-64.

Thabtah F., Cowling, P., and Peng, Y. (2005). MCAR: Multi-class classification based on association rule approach. Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications, Cairo, Egypt, (pp. 1-7).

Thabtah F., Cowling P., and Peng Y. (2004) MMAC: A new multi-class, multi-label associative classification approach. Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM ’04), (pp. 217-224). Brighton, UK.

Van Rijsbergan, (1979). Information Retrieval, 2nd edn. London: Buttersmiths.

Weka, 2001, Data mining software in Java. http://www.cs.waikato.ac.nz/ml/weka.

Witten I. ,and Frank E.,(2000).Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco, CA: Morgan Kaufmann.

Yuhanis Y., and Refai M.,(2013). Modified Multi-Class Classification using Association Rule Mining, Pertanika J. Sci. & Technol. 21, pp 205 - 216

Yin, X., and Han, J. (2003) CPAR: Classification based on predictive association rule, Proceedings of the SDM, San Francisco, CA, (pp. 369-376).

Zaïane O., and Antonie A., (2002). Classifying text documents by associating terms with text Categories, Proceedings of the Thirteenth Australasian Database Conference (ADC'02), Melbourne, Australia, (pp. 215 - 222).

Zaixiang H., Zhongmi Z., Tianzhong H.,(2013). Association Classification with KNN, Journal of Theoretical and Applied Information Technology Vol. 49 No.3, pp 1013-1019

Zemirline A., Lecornu, L., Solaiman, B., and Ech-cherif A.,(2008). An Efficient Association Rule Mining Algorithm for classification, Springer-Verlag Berlin Heidelberg,pp 717-721

ملخص الدراسه

ان التصنيف المبني على الروابط دائماً ما يقوم بإنتاج عدد كبير من القوانين، لذلك لا مفر من ذلك بان تكون احدى البيانات تتلائم مع عده قوانين متناقضه، العديد من خوارزميات التصنيف المبنيه على الروابط تعتمد في تصنيفها على قانون واحد وتتجاهل باقي القوانين حتى وان كانت درجه الثقه بها مرتفعه، من خلال هذا البحث، نقدم خوارزمية جديده والتي تستخدم نظريه بايز لتخطي هذه المشكلة.

أظهرت النتائج التي قمنا بها بان ألخوارزميه المقترحة فاقت خوارزميات التصنيف التقلديه والحديثيه منها أيضاً.

التصنيف المبني على الروابط باستخدام نظريه بايز

بواسـطة

فوزي علي ابوجابر

بإشـراف د. رشيد الزبيدي

قدمت هذه الرسالة استكمالاً لـمتطلبات الحصول على درجة

الـمـاجـستير في علـم الـحـاسوب

عـمـادة البحث العلمي والدراسات العليا

جامعة فيلادلفيا

شباط, 2014