Comparison of Data Mining Algorithms in Credit Card approval

Wilson Muange Musyoka
St. Paul's University
Email: wmuange [AT] yahoo.com

Abstract--- **The use of credit scoring can be used to help the credit risk analysis in determining the applicant's eligibility. Data mining has been proven as a valuable tool for credit scoring. The last years have seen the development of many credit scoring models for assessing the creditworthiness of loan applicants. Traditional credit scoring methodology has involved the use of statistical and mathematical programming techniques such as discriminant analysis, linear and logistic regression, linear and quadratic programming, or decision trees. However, the importance of credit grant decisions for financial institutions has caused growing interest in using a variety of computational intelligence techniques. This paper concentrates on comparing several algorithms used by Weka, which is viewed as one of the most promising paradigms of computational intelligence. The aim of this paper is to evaluate data mining using Tree based algorithm, Rule based and Bayesian networks and see how either the data or the algorithms can be improved to enhance the output percentages of the algorithms.**

Keywords: Data mining, Computational Intelligence, pattern discovery

I. INTRODUCTION

Credit allows accessing to resources today with an agreement to repay over a period of time, usually at regular intervals. The resources may be financial, or they may consist of products or services. Credit has now turned into a very important component in everyday life. Although credit cards are currently the most popular form of credit, other credit plans include residential mortgages, auto loans, student loans, small business loans, trade financing and bonds, among others.

Data mining refers to computer-aided pattern discovery of previously unknown interrelationships and recurrences across seemingly unrelated attributes in order to predict actions, behaviours and outcomes. Data mining, in fact, helps to identify patterns and relationships in the data (Madan Lal Bhasin, 2006).

Data Mining also refers as analytical intelligence and business intelligence. Because data mining is a relatively new concept, it has been defined in various ways by various authors in the recent past. Some widely used techniques in data mining include artificial neural networks, genetic algorithms, K-nearest neighbour method, decision trees, and data reduction.

The data mining approach is complementary to other data analysis techniques such as statistics, on-line analytical processing (OLAP), spreadsheets, and basic data access. Data mining helps business analysts to generate hypotheses, but it does not validate the hypotheses. The main risk for banks and financial institutions comes from the difficulty to distinguish the creditworthy applicants from those who will probably default on repayments.

The recent world financial crisis has aroused increasing attention of financial institutions on credit risk prediction and assessment. The decision to grant credit to an applicant was traditionally based upon subjective judgments made by human experts, using past experiences and some guiding principles. Common practice was to consider the classic 3 C's, 4 C's or 5 C's of credit: character, capacity, capital, collateral and conditions (Abrahams and Zhang, 2008).

This method suffers, however, from high training costs, frequent incorrect decisions, and inconsistent decisions made by different experts for the same application.

These shortcomings have led to a rise in more formal and accurate methods to assess the risk of default. In this context, automatic credit scoring has become a primary tool for financial institutions to evaluate credit risk, improve cash flow, reduce possible risks, and make managerial decisions (Thomas et al, 2002).

Credit scoring is the set of decision models and their underlying methods that help lenders determine whether credit should be approved to an applicant. The ultimate goal of credit scoring is to assess credit worthiness and discriminate between 'good' and 'bad' debts, depending on how likely applicants are to default with their repayments (Lim and Sohn, 2007). Compared with the subjective methods, automatic credit scoring models present a number of interesting advantages (Rosenberg and Gleit, 1994; Thomas et al, 2002; Blo¨ chlinger and Leippold, 2006):

(i) reduction in the cost of the credit evaluation process and the expected risk of being a bad loan;
(ii) time and effort savings;

(iii) consistent recommendations based on objective information, thus eliminating human biases and prejudices;
(iv) facilities to incorporate changes in policy and/or economy into the system; and
(v) the performance of the credit scoring model can be monitored, tracked, and adjusted at any time.

The usage of credit scoring can be used to help the credit risk analysis in determining the applicant's eligibility. Data mining has been proven as a valuable tool for credit scoring. The last years have seen the development of many credit scoring models for assessing the creditworthiness of loan applicants. Traditional credit scoring methodology has involved the use of statistical and mathematical programming techniques such as discriminant analysis, linear and logistic regression, linear and quadratic programming, or decision trees. However, the importance of credit grant decisions for financial institutions has caused growing interest in using a variety of computational intelligence techniques. This paper concentrates on evolutionary computing, which is viewed as one of the most promising paradigms of computational intelligence.

Classification

Classification is assigning an object to a certain class based on its similarity to previous examples of other objects. It can be done with reference to original data or based on a model of that data E.g: Me: "Its round, green, and edible" You: "It's an apple!"

Classification consists of assigning a class label to a set of unclassified cases.

1. **Supervised Classification**
   The set of possible classes is known in advance.
2. **Unsupervised Classification**

Set of possible classes is not known. After classification we can try to assign a name to that class. Unsupervised classification is called clustering.

In Supervised classification, the input data, also called the training set, consists of multiple records each having multiple attributes or features. Each record is tagged with a class label. The objective of classification is to analyze the input data and to develop an accurate description or model for each class using the features present in the data. This model is used to classify test data for which the class descriptions are not known.

II. RELATED WORK

This section reviews work on several algorithms used for computational intelligence when mining data for credit approval or other functions. The aim of this paper is to evaluate data mining using Tree based algorithm, Rule based and Bayesian networks and see how either the data or the algorithms can be improved to enhance the output percentages of the algorithms.

1. **P Salman Raju, Dr V Rama Bai, G Krishna Chaitanya (2014); Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014.**

This paper looks at data mining techniques and algorithm that are applicable to the banking sector. Customer retention plays vital role in the banking sector. The supervised learning method Decision Tree implemented using CART algorithm is used for customer retention. Preventing fraud being better than detecting the fraudulent transaction after its occurrence. Hence for credit card approval process using the data mining techniques Decision Tree, Support Vector Machine (SVM) and Logistic Regression are used. Clustering model implemented using EM algorithm can be used to detect fraud in banking sector.

It outlines the mining process as involving;

Knowledge discovery: The KDD process is outlined in Figure 1. This process includes several stages, consisting of data selection, data treatment, data pre-processing, data mining and interpretation of the results. This process is interactive, since there are many decisions that must be taken by the decision-maker during the process. The stages of KDD process are briefly described below.

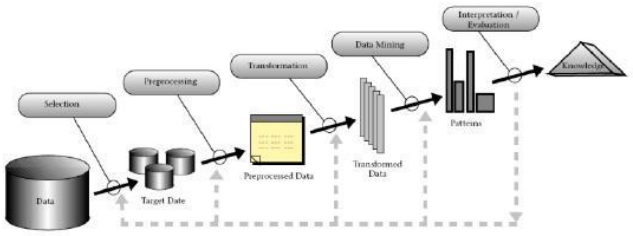

Figure 1: KDD process

Data selection: This stage includes the study of the application domain, and the selection of the data. The domain's study intends to contextualize the project in the company's operations, by understanding the business language and defining the goals of the project. In this stage, it is necessary to evaluate the minimum subset of data to be selected, the relevant attributes and the appropriate period of time to consider.

Data pre-processing: This stage includes basic operations, such as: removing noise or outliers, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data attributes, and accounting for time sequence information and known changes. This stage also includes issues regarding the

database management system, such as data types, schema, and mapping of missing and unknown values.

Data transformation: This stage consists of processing the data, in order to convert the data in the appropriate formats for applying data mining algorithms. The most common transformations are: data normalization, data aggregation and data discretization. To normalize the data, each value is subtracted the mean and divided by the standard deviation. Some algorithms only deal with quantitative or qualitative data. Therefore, it may be necessary to discredit the data, i.e. map qualitative data to quantitative data, or map quantitative data to qualitative data.

Data mining: This stage consists of discovering patterns in a dataset previously prepared. Several algorithms are evaluated in order to identify the most appropriate for a specific task. The selected one is then applied to the pertinent data, in order to find indirect relationships or other interesting patterns.

Interpretation/Evaluation: This stage consists of interpreting the discovered patterns and evaluating their utility and importance with respect to the application domain. In this stage it can be concluded that some relevant attributes were ignored in the analysis, thus suggesting the need to replicate the process with an updated set of attributes.

2. **Evaristus Didik Madyatmadja, Mediana Aryuni (2014); comparative study of data mining model for credit card application scoring in bank. Journal of Theoretical and Applied Information Technology. 20th January 2014. Vol. 59 No.2 © 2005 - 2014 JATIT & LLS. All rights reserved.**

The purpose of this study was to determine the proper data mining system for credit scoring credit card application in Bank in order to improve the performance and support the credit analysts' job. Classification being one of data mining finds a model or function that separates classes or data concepts in order to predict the class of an unknown object. For example, a loan officer requires data analysis to determine which loan applicants are "safe" or "risky". The data analysis task is classification, where a model or classifier is constructed to predict class (categorical) labels, such as "safe" or "risky" for the loan application data. These categories can be represented by discrete values, where the ordering among values has no meaning. Because the class labels of training data is already known, it is also called supervised learning.

Classification consist two processes: (1) training and (2) testing. The first process, training, builds a classification model by analysing training data containing class labels. While the second process, testing, examines a classifier (using testing data) for accuracy (in which case the test data contains the class labels) or its ability to classify unknown objects (records) for prediction.

In the study A naïve (or simple) Bayesian classifier based on Bayes' theorem is used as a probabilistic statistical classifier, which the term "naïve" indicates conditional independence among features or attributes. Its major advantage is its rapidity of use because it is the simplest algorithm among classification algorithms. Hence, it can readily handle a data set with many attributes.

It also uses Decision tree classifiers to construct a flowchart-like tree structure in a top down, recursive, divide-and-conquer, manner. Using The Attribute Selection Method (ASM), it selects a splitting criterion (attribute) that best splits the given records into each of the class labels, then selected attributes become nodes in a decision tree.

The output from the two algorithms is shown below;

Table 1: The Confusion Matrix of Naïve Bayes Classifier

	true Approved	true Rejected	Class Precision
Predicted Approved	163	35	82.32%
Predicted Rejected	37	165	81.68%
Class Recall	81.50%	82.50%	

Table 2:The Confusion Matrix of ID3 Classifier

	true Approved	true Rejected	Class Precision
Predicted Approved	160	56	74.07%
Predicted Rejected	40	144	78.26%
Class Recall	80.00%	72.00%	

The accuracy of Naïve bayes classifier is (81.50%+82.50%+82.32%+81.68%)/4 =82%, and ID3 has (80.00%+72.00%+74.07%+78.26%)/4 =76% of accuracy.

The authors presented a data mining model which applied classification methods using Naïve Bayes and ID3 algorithm for credit scoring in credit card application. The best accuracy is achieved by Naïve bayes classifier (82%), while ID3 has 76% of accuracy. So the paper concluded that Naïve Bayes classifier has better accuracy than ID3

classifier. In addition, the proposed data mining model able to improve the performance and support the credit analyst's job. The study also explains that the selection of the important features is a challenge. It proposes further work to conduct some feature selection methods to know which method can give best classification performance.

3. **A. I. Marque´s, V. Garcia and J. S. Sanchez (2013); A literature review on the application of evolutionary computing to credit scoring. Journal of the Operational Research Society (2013) 64, 1384–1399 © 2013 Operational Research Society Ltd. All rights reserved. 0160-5682/13.**

This study focused on Computational intelligence, which is defined as the study of adaptive mechanisms to enable or facilitate intelligent behaviour in complex and changing environments (Bezdek, 1994; Engelbrecht, 2007). These mechanisms include artificial intelligence concepts, paradigms, algorithms, and implementations that exhibit an ability to learn or adapt to new situations, to generalize, abstract, discover, and associate.

Currently these techniques are applied to a variety of problems, ranging from scientific research to finance, industry, and commerce to name but a few. A. I. Marque´s, V. Garcia and J. S. Sanchez explain that the two main families of methods that primarily comprise this field are evolutionary computing and swarm intelligence.

In particular, evolutionary computation constitutes a subfield of computational intelligence, which involves combinatorial optimization problems. Thus evolutionary computing is the generic term for a set of problem-solving techniques based on the Darwinian principles of natural selection and evolution (Eiben and Smith, 2007). It is inspired by biological processes of inheritance, mutation, natural selection, and the genetic crossover that occurs when parents mate to produce offspring. It makes use of the concept of survival of the fittest by progressively accepting better solutions to the problem.

The common underlying idea behind all variants of evolutionary algorithms is that, given a population of individuals, the environment causes natural selection and this produces a rise in the fitness of the population. Given a quality function to be maximized, we can randomly generate a set of candidate solutions (a group of IF-THEN rules) and apply the quality function as an abstract fitness measure. Based on this, some of the better candidates are chosen to seed the next generation by applying selection, mutation, and crossover operators. These operators lead to a set of new candidates that compete with the old ones for a place in the next generation. This process can be iterated until a candidate with sufficient quality is found or a previously fixed computational limit is reached.

The study generally concluded that genetic algorithms and genetic programming are efficient, flexible methods to find optimal or near-optimal solutions from the search space, what makes them especially appealing to a wide variety of economic and financial applications. In this sense, we have found that evolutionary computation has mainly been applied to variable selection and parameter optimization as two important preprocessing steps for further classification with other prediction models, especially artificial neural networks and support vector machines.

An important subject in credit scoring is the comprehensibility or transparency of the model, which has become a key factor for the lending industry. In general, evolutionary algorithms should preferably be combined with (neuro)-fuzzy rules or decision trees to achieve a higher degree of transparency. Another promising method in this direction is the so-called multi-objective genetic algorithm because it is suitable for systems with conflicting goals, which in the case of credit scoring are to increase the predictive power and to reduce the complexity of the models. On the other hand, the use of evolutionary computation in variable selection leads to a reduction of the complexity, what also allows to increase the comprehensibility of the model.

The study also observed that there is no single best algorithm across different credit databases. A technique may be the best on some particular data sets, but it will perform worse than the other algorithms on other different data sets. Therefore it is not possible to conclude that genetic algorithms and genetic programming are better or worse than other models, but are simply an alternative to conventional methods.

However, the major interest of using evolutionary methods probably comes from their ability to properly optimize the internal parameters of those classifiers that have demonstrated to perform well in credit scoring applications.

4. **Elma Zannatul Ferdousy, Md. Mafijul Islam & M. Abdul Matin (2013); Combination of Naïve Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models. Computer and Information Science; Vol. 6, No. 3; 2013 ISSN 1913-8989 E-ISSN 1913-8997 Published by Canadian Center of Science and Education.**

This study introduces a new classifier that combines the distance-based algorithm K-Nearest Neighbor and statistical based Naïve Bayes Classifier. That is equipped with the power of both but avoid their weakness. The performance of the proposed algorithm in terms of accuracy is experimented on some standard datasets from the machine-learning repository of University of California and compared with some of the art algorithms. The experiments show that in most of the cases the proposed algorithm outperforms the other to some extent. The algorithm is applied for predicting profitability positions of some financial institutions of Bangladesh using data provided by the central bank.

The idea of the algorithm proposed is very simple. To classify a new object first use the
KNN algorithm to find the K Nearest Neighbor from the training dataset. While implementing the KNN, do not include the categorical attributes. The distance will be measured only by the numerical attributes. After selecting the K nearest object, build a model using the Naïve Bayes algorithm, but using only the categorical attributes. From the model, classify the new object. So, this is a two-step process. In the first step, only the numerical attributes are used to select the closest data of the new object. This makes sense, as numerically close objects are supposed to have the same characteristics. Now, after getting K nearest object of the new object, instead of taking simple voting scheme as in KNN, look deeper in the characteristic of the categorical data and the relation to the class. Use the Naïve Bayes for the purpose. In this way, both the numerical and categorical attributes are used for the classification of a new object without any alteration in the data. Hence, the proposed method keeps the data intact. No discretization or complex similarity measurement is required. Thus, the proposed method is nothing but a 'combination of Naïve Bayes and K Nearest Neighbor', cNK.
cNK algorithm is described (see also Figure 2) as follows:
Step 1. Obtain the K-Nearest Neighbor of a new observation based on the numerical attributes.
Step 2. Use the set of K observations, found in step 1 as training data and use it to build a model exploiting the
Naïve Bayes algorithm based only on the categorical attributes.
Step 3. Use the model built in step 2 to classify the new observation.

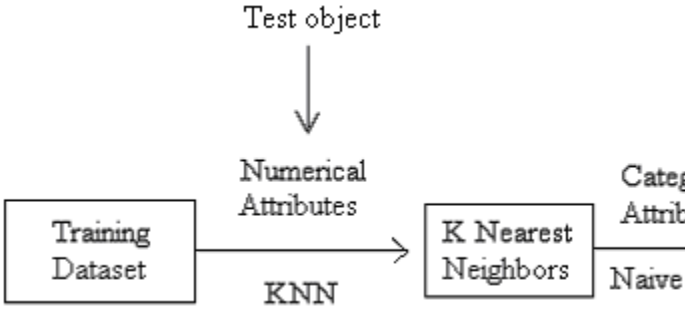

Figure 2: Graphical representation of the proposed algorithm

III. METHODOLOGY

Introduction

This section describes how test data has been used to evaluate credit scores through experiments using Weka with Tree based algorithms, Rule based algorithm and Bayesian based algorithm.

Data as it is

Using J48

In the starting interface of Weka, click on the button ***Explorer***.
In the ***Preprocess*** tab, click on the button ***Open File***. In the file selection interface, select the file credit.arff.

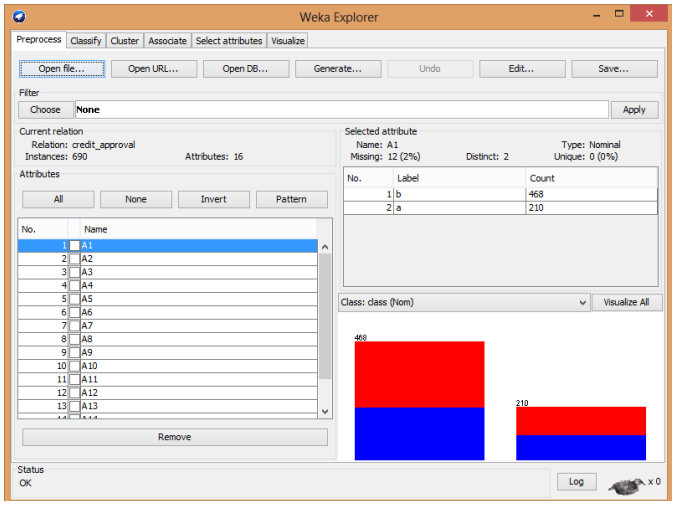

Table 1

The dataset is characterized in the ***Current relation*** frame: the name, the number of instances (compounds), the number of attributes (descriptors + activity/property). We see in this frame that the number of compounds is 690, whereas the number of descriptors is 16, which is the number of attributes (16) minus the activity field. The ***Attributes*** frame allows user to modify the set of attributes using *select* and *remove* options. Information about the selected attribute is given in the ***Selected attribute*** frame in which a histogram depicts the attribute distribution. One can see that the value of the currently selected descriptor fp_1 (the first bit in the corresponding fingerprint) is "on" in 468 compounds and "off" in 201 compounds in the dataset.

Nonactive compounds are depicted by the blue color whereas active compounds are depicted by the red color in the histogram.
• Click on the tab ***Classify***.

The ***ZeroR*** method is already selected by default. This needs to be changed to the method we are using for our cased which is **tree based**, specifically **J48**. For assessing the predictive performance of all models to be built, the 10-fold cross-validation method has also be specified by default.
• Click on the ***Start*** button to build a model.

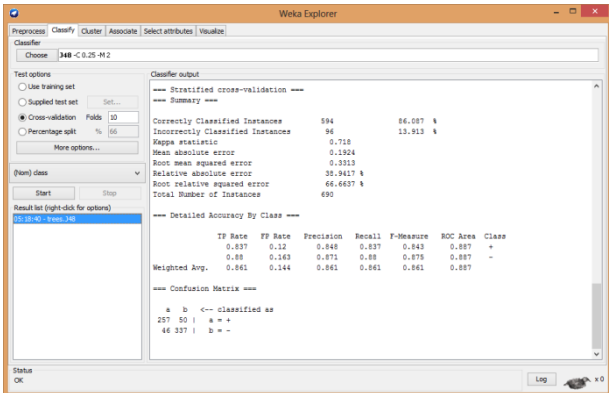

Table 2

The predictive performance of the model is characterized in the right-hand Classifier output frame. The Confusion Matrix for the model is presented at the bottom part of the Classifier output window. It can be seen from it that all compounds have been classified as "nonactive". It is clear that such trivial model is useless and it cannot be used for discovering "active" compounds. However, pay attention that the accuracy of the model (Correctly Classifieds Instances) of this trivial model is very high: 86.087 %. This fact clearly indicates that the accuracy cannot be used for assessing the usefulness of classification models built using unbalanced datasets. For this purpose a good choice is to use the "Kappa statistic", which is zero for this case. "Kappa statistic" is an analog of correlation coefficient. Its value is zero for the lack of any relation and approaches to one for very strong statistical relation between the class label and attributes of instances, i.e. between the class of biological activity of chemical compounds and the values of their descriptors.

Using Decision Table

The output is shown as follows;

In the ***classifier*** frame, click ***Chose***, then select the *decision table* method from the *tree* submenu.

Click on the ***Start*** button to build a model.

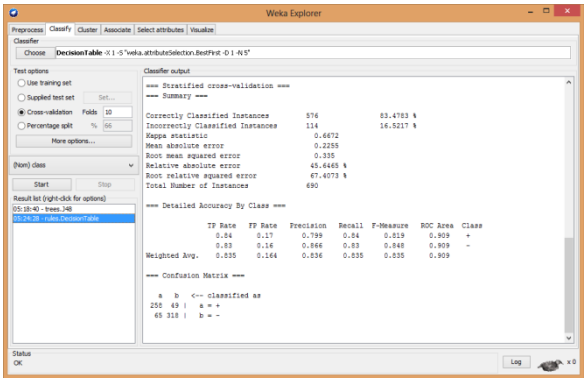

Table 3

Using Bayesian Networks

The output is shown as follows;

In the ***classifier*** frame, click ***Chose***, then select the *bayesNet* method from the *bayes* submenu.

Click on the ***Start*** button to build a model.

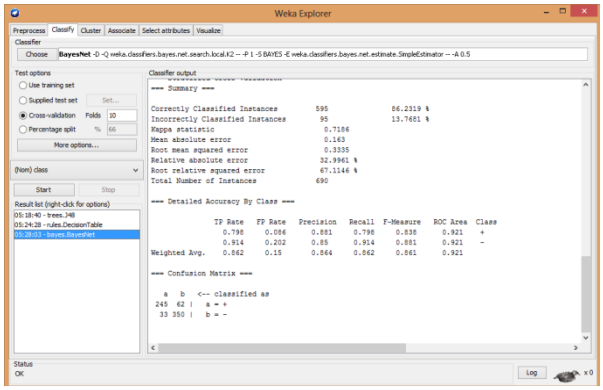

Table 4

The output of the three algorithms is summarised as follows;

J48 – 86.087%
0.16 seconds

Decision table – 83.4783%
0.25 seconds

Bayesian Network – 86.2319%
0.06 seconds

From the above summary, it shows that Bayesian networks has better performance for data as it is.

Improvements

Data

Data could be improved by filling in the missing values during pre-processing using mode and mean. When this is done the following outputs are obtained.

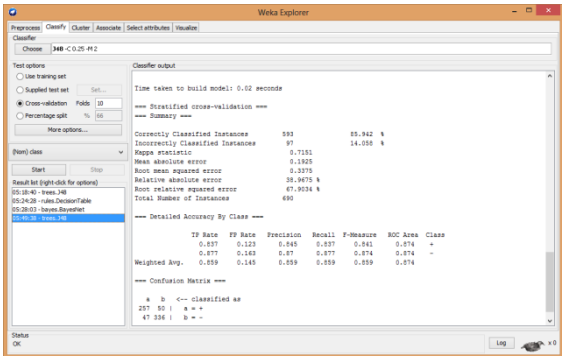

Table 5

The output using J48 shows that the time taken improves to 0.02 seconds but the percentage reduces to 85.942%

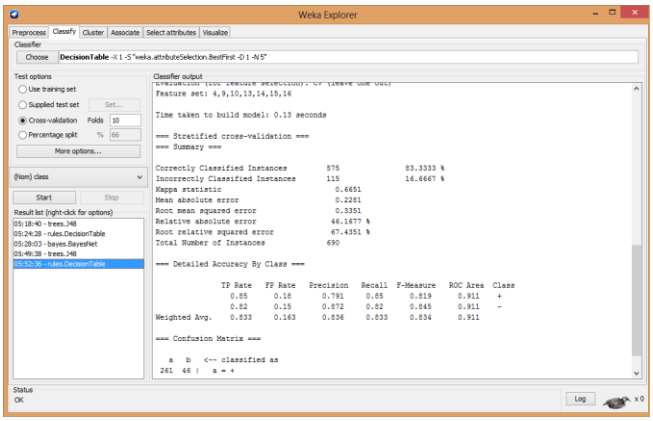

Table 6

When using decision tree when missing data has been replaced, the time taken is 0.13 seconds and the performance percentage is 83.3333%.

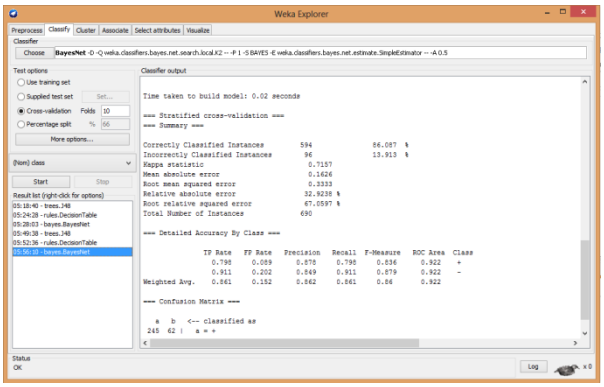

Table 7

When using Bayesian networks, the time taken is 0.02 seconds and the percentage is 86.067%. With improvement on the data pre-processing it is clear that the time taken improves but the percentages reduce in terms of performances.

Attribute Selection

The process of attribute selection involves going to the starting interface of Weka, click on the button ***Explorer***. In the ***Preprocess*** tab, click on the button ***Open File***. In the file selection interface, select the file credit.arff.
On the filter tab, select the attribute selection option as shown below;

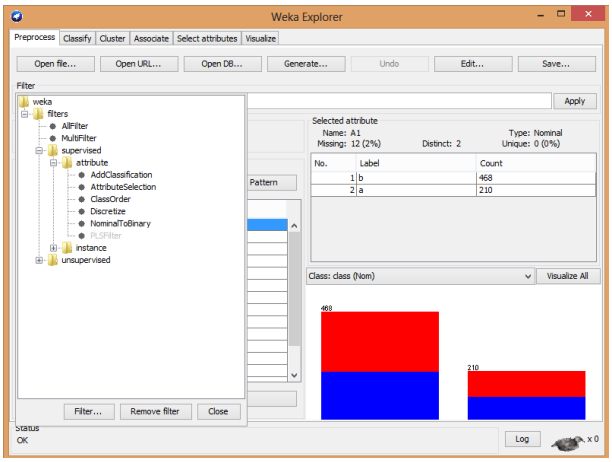

Table 8

Then after selecting attribute selection, the output of the credit data appears as follows;

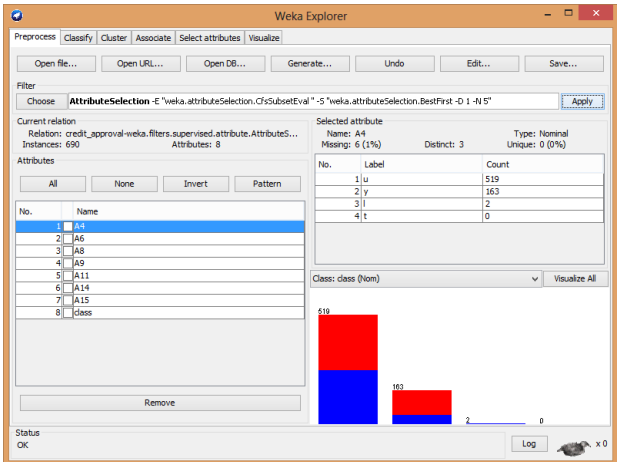

Table 9

With attribute selection the outputs are shown below;

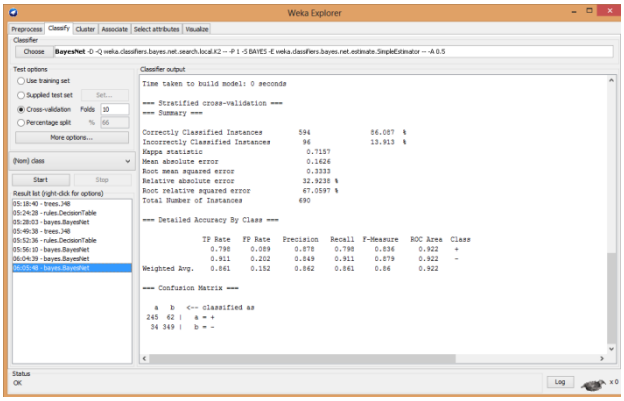

Table 10

With Bayesian networks the time taken to build the model is 0 seconds while the percentage is 86.067%.

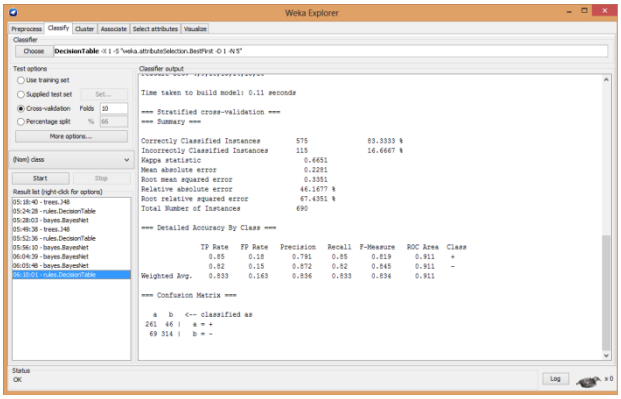

Table 11

Using a decision table the time taken to build the model is 0.11 seconds and the percentage is 83.333%.

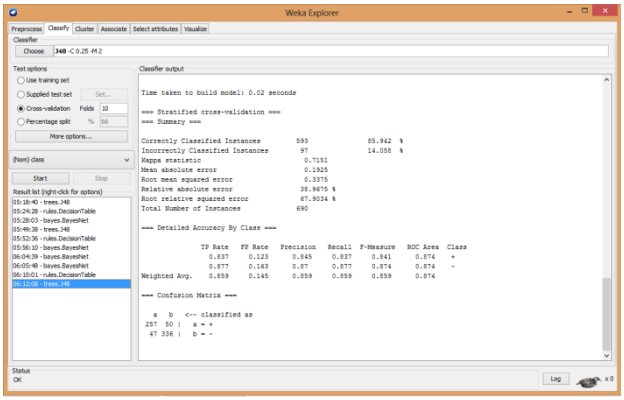

Table 12

With J48, the time taken to build the model is 0.02 seconds and the percentage is 85.942%.

IV. DISCUSSION AND CONCLUSION

The simulation results show that the highest correctly classified instances is 595 (86.2139%) out of 690 instances by Bayesian Networks and the lowest correctly classified instances is 575 (83.3333%) by Decision Table algorithm.

The total time required to build the model is also a crucial parameter in comparing the classification algorithm. From the simulation result of table 8, we can say that a Bayesian Networks classifier with attribute selection requires the shortest time which is around 0 seconds compared to the others. From all the tables showing outputs, the confusion matrix of each classifier can be observed which shows sensitivity, specificity and accuracy.

Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity. The average Kappa score from the selected algorithm is 0.786563. Based on the Kappa Statistic criteria, the accuracy of this classification purposes is substantial.

It is also possible to observe the differences of errors resultant from the training of the selected algorithms. This experiment implies a very commonly used indicator which is mean of absolute errors and root mean squared errors. Alternatively, the relative errors are also used. Since, we have two readings on the errors, taking the average value will be wise.

An algorithm which has a lower error rate and maximum accuracy will be preferred as it has more powerful classification capability.

V. REFERENCES

A. I. Marque´s, V. Garcia and J. S. Sanchez (2013); A literature review on the application of evolutionary computing to credit scoring. Journal of the Operational Research Society (2013) 64, 1384–1399 © 2013 Operational Research Society Ltd. All rights reserved. 0160-5682/13.

Bezdek JC (1994). What is computational intelligence? In: Zurada JM,Marks II RJ, Robinson CJ (eds). Computational Intelligence: Imitating Life. Chapter 1. IEEE Press: Piscataway, NJ, pp 1–12.

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36:4626– 4636.

Eiben AE and Smith JE (2007). Introduction to Evolutionary Computing. Springer: Heidelberg, Germany.

Elma Zannatul Ferdousy, Md. Mafijul Islam & M. Abdul Matin (2013); Combination of Naïve Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models. Computer and Information Science; Vol. 6, No. 3; 2013 ISSN 1913-8989 E-ISSN 1913-8997 Published by Canadian Center of Science and Education.

Engelbrecht AP (2007). Computational Intelligence: An Introduction. Wiley: Chichester, UK.
Evaristus Didik Madyatmadja, Mediana Aryuni (2014); comparative study of data mining model for credit card application scoring in bank. Journal of Theoretical and Applied Information Technology. 20th January 2014. Vol. 59 No.2 © 2005 - 2014 JATIT & LLS. All rights reserved.

Li J, Li G, Sun D and Lee C-F (2012). Evolution strategy based adaptive Lq penalty support vector machines with Gauss kernel for credit risk analysis. Applied Soft Computing 12(8): 2675–2682.
Lim MK and Sohn SY (2007). Cluster-based dynamic scoring model. Expert Systems with Applications 32(2): 427–431.
Madan Lal Bhasin, 2006. Data Mining: A Competitive Tool in the Banking and Retail Industries

P Salman Raju, Dr V Rama Bai, G Krishna Chaitanya (2014); Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014.