EFFECTIVENESS OF CLASSIFIERS FOR THE CREDIT DATA SET : AN ANALYSIS

Soni P M[1], Varghese Paul[2], M.Sudheep Elayidom[3]

[1]*Assistant Professor, Department of Computer Applications, SNGIST N.Paravur, Kerala, India*
[2]*Professor, Department of IT, RSET, Kochi, Kerala, India*
[3]*Associate Professor, Computer Science, CUSAT, Kochi, Kerala, India*

Abstract

In today's world data mining is becoming an important area in terms of all business applications especially in the banking sector. In developing countries like India, bankers should be vigilant to fraudsters because they will create more problems to the banking organization. Application of data mining techniques helps the banks to look for hidden patterns in a group and discover unknown relationship in the data. Feature selection is a method used in data mining to select the most appropriate attributes for defining a relationship in a data set. It is very effective to build models based on these data mining techniques. There are several types of classifiers in data mining that helps to classify the records into two major groups based on the list of attributes. The proposed work is a comparative study of different types classifiers and evaluating the accuracy of the classifiers before and after applying the feature selection. After evaluating the results of experiment, it is easy to predict that feature selection is an important and necessary step during the process of data mining. From the results we can see that the performance metrics we obtained in different classifiers after applying feature selection is equal or better than that of before applying feature selection.

Keywords: *Feature Selection, Classification, Ranking; Feature Selection, Performance, Accuracy*

----------***----------

1. INTRODUCTION

The areas in which Data mining Tools can be used in the banking industry are customer segmentation, Banking profitability, credit scoring and approval, Predicting payment from Customers, Marketing, detecting fraud transactions, Cash management and forecasting operations, optimizing stock portfolios, and ranking investments [1]. Now a days predicting loan repayment from customer is a major task that can be handled by bank employees. It is very difficult for them to detect the fraud using their personal data at a glance. Detecting and preventing fraud is difficult, because fraudsters develop new schemes all the time, and the schemes grow more and more sophisticated to elude easy detection [2]. Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set [3]. An effective method is to detect the chance of loan repay ability or non repay ability using the most selective arrtibutes that obtained from the given data set. In the proposed work, before selecting the most important attributes the input data set is executed by InfoGainAttributeEval followed by Ranker search algorithm that helps to rank the attributes based on their information gain ratio. In order to predict the customer behavior we need not study about all the attributes and only the most appropriate attributes helps to classifies the data into two catagories such as customers who repay the loans promptly and they do not. After removing the least important attributes, various classifiers are applied to find out the resultant class. Different classifiers will get different performance metrics. The performance metrics selected here are Accuracy, Kappa Statistic and Mean Absolute Error. Aim of this paper is to make a comparative study about various classifiers. The performance metrics values of each classifier before applying feature selection and after applying feature selection are evaluated. The results obtained from the experiment shows that all the performance metrics after applying feature selection is better than that of before applying feature selection. It will be very helpful for the banking industry for easy detection of fraud and so that issues can be solved at an extent. This paper is organized as follows. The next section discusses about the dataset used for conducting the experiment. Section 3 explains about the concept used in this paper that gives an idea about Feature selection and the technology used. The proposed work is described in Section 4. Here the process of classification before and after applying feature selection is explained and a comparative study is done using various classifiers such as JRip, ZeroR, SMO, Adaboost, Random Forest, Kstar, Ridor and DTNB. The section 4 also discussed about the evaluation by different performance metrics and highlight the best classifiers. Conclusion is given in Section 5 followed by the references.

2. DATASET

The benchmarking datasets in WEKA were used for this study. The data available consists of 1000 records of bank loan transaction data including 21 data fields. Twenty attributes are considered here for the experiment and one attribute is considered as class attribute. The class attribute predict that whether a customer has the capability of making repayment or not. The input data set before applying feature selection is represented below:

Table I. List of Attributes

Sl	Name of attribute
1	checking_status
2	duration
3	credit_history
4	purpose
5	credit_amount
6	savings_status
7	employment
8	installment_commitment
9	personal_status
10	other_parties
11	residence_since
12	property_magnitude
13	Age
14	other_payment_plans
15	housing
16	existing_credits
17	Job
18	num_dependents
19	own_telephone
20	foreign_worker
21	Class

3. CONCEPTS USED

3.1 Feature Selection

Feature Selection is the preprocessing process of identifying the subset of data from large dimension data.[4]. G. Holmes et.al explained that in order to obtain useful results using super vised learning of real world datasets it is necessary to perform feature subset selection and to perform many experiments using computed aggregates from the most relevant features [5]. The study discuss about various classifiers that can be applied to the input data set and measure the performance accuracy of each to determine which one is better than the other. Feature ranking is the process of applying feature selection to the input data set so that the least important attributes can be removed and the classification process is performed on the new data set that was obtained after applying feature selection. From the experiment we can see that the classification performance will increase while reducing the number of features in the original data set.

In Feature ranking, a number of ordered categories are used, representing the ranking relationship between instances [5]. Let A = $\{a_1, a_2, \ldots, a_m\}$ be the set of m attributes. Let r be a function r: $A_D \rightarrow R$ that assigns a value of merit to each attribute a ∈ A from D. A feature ranking is a function F that assigns a value of merit (relevance) to each attribute ($a_i \in A$) and returns a list of attributes ($a^*_i \in A$) ordered by its relevance, with i ∈ {1, . . .,m}:F($\{a_1, a_2 \ldots, a_m\}$) =< $a^*_1, a^*_2, \ldots, a^*_m$ > where $r(a^*_1) \geq r(a^*_2) \geq \ldots \geq r(a^*_m)$[6]. Before Feature selection begins, the relevance of the attributes is found out by using attribute evaluator InformationGain and ranker search algorithm. The tool selected for the experiment is WEKA 3.7. Using Information Gain we can determine the most useful attributes from a given data set. It is a numerical value that determines how important a given attribute from the list of original attribute list. It is achieved through an information ratio obtained for every attribute. Informain gain attribute evaluator is followed by a ranker search algorithm which will display the list of attributes in sorted order of their information gain ratio.

3.2 Technology Used

The Weka suite contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [10]. Weka is a powerful tool that acts as an aid for data mining process. It includes the basic activities in data mining such as data preprocessing, classification, clustering and Visualization. It is freely available as well as platform-independent software. The four applications such as Explorer, Experimenter, Knowledge Flow and simple CLI are the main components of this tool. The files opened in weka should be in .arff or .csv format. Weka can apply several algorithms and techniques in data mining and it is possible to compare the result of a process in different techniques. In this experiment, feature selection and classification algorithms are mostly concentrated. Microsoft Excel is a powerful tool to manage data in tabular form and chart format. The performance metrics obtained before applying feature selection and after applying feature selection can be easily represented in tabular as well as chart form in Excel.

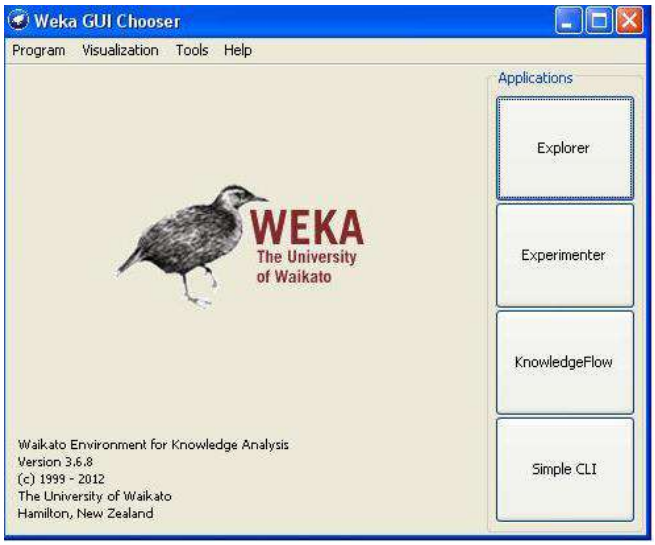

Fig. 1 User interface of weka

4. PROPOSED WORK

The different steps that were carried out for conducting the comparative study have been described in this section.

4.1 InfoGainAttributeEval and ranker search

The attributes of the weka dataset is ranked using InfoGainAttributeEval followed by ranker search algorithm. The process of selecting attributes using Information Gain and ranker search algorithm in WEKA 3.7 is represented in Fig.2

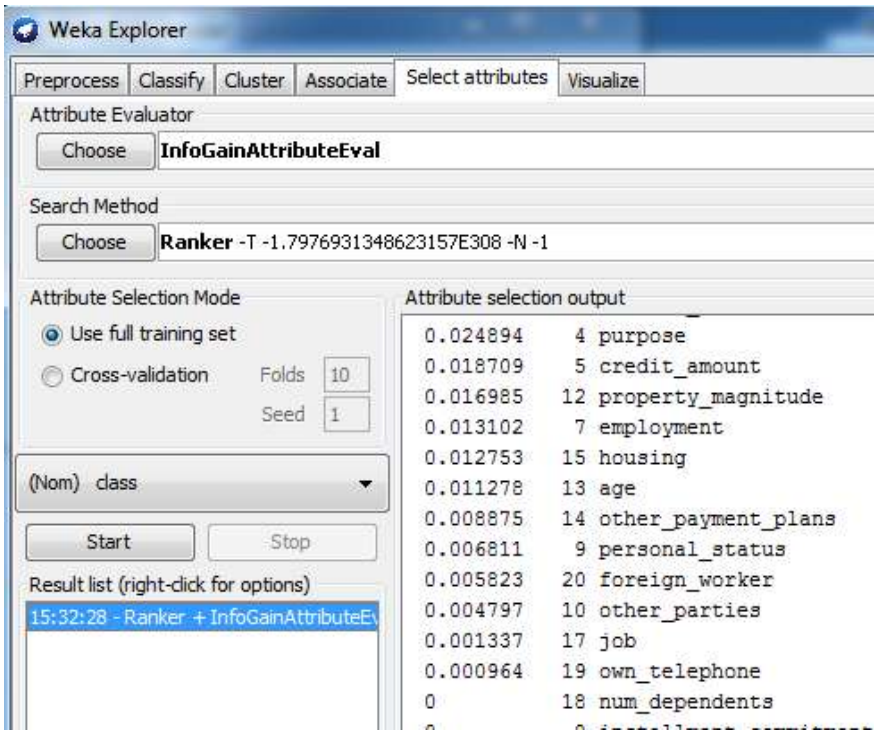

Fig.2. Process of selecting attributes

The resulting weights of the attributes obtained after applying InfoGainAttributeEval followed by ranker search algorithm are normalized into the interval between 0 and 1 as shown in the Table 2. From this list of attributes remove the least important attributes manually and apply classification process repeatedly until there is no change in the performance metrics. The classification procedure is represented in the Fig, 3

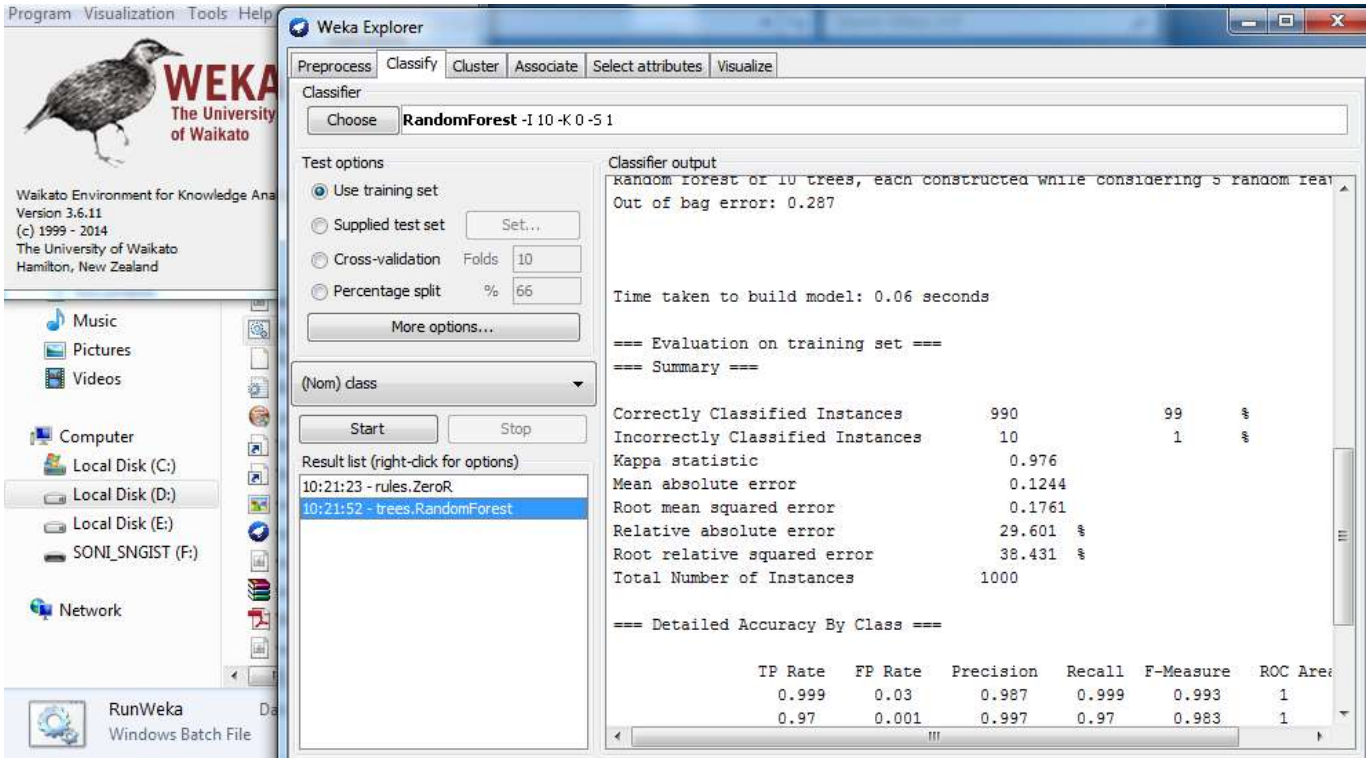

Fig. 3 Process of Classification

Table 2. Ranked List of Attributes

Sl	Weight	Attribute
1	0.094739	checking_status
2	0.043618	credit_history
3	0.0329	Duration
4	0.028115	savings_status
5	0.024894	Purpose
6	0.018709	credit_amount
7	0.016985	property_magnitude
8	0.013102	Employment
9	0.012753	Housing
10	0.011278	Age
11	0.008875	other_payment_plans
12	0.006811	personal_status
13	0.005823	foreign_worker
14	0.004797	other_parties
15	0.001337	Job
16	0.000964	own_telephone
17	0	num_dependents
18	0	installment_commitment
19	0	residence_since
20	0	existing_credits

In this experiment, the least six attributes existing credits, residence since, installment commitment, num_dependents, own telephone and Job are manually removed before applying the classification process.

4.2 Classifiers

Classification is an application of Data mining that helps to classify a new data record into one of the many possible classes which are already known. Classification is a data mining technique to accurately predict the target class within a given data set. Here the classification model can be used to identify loan candidates as good or bad credit risks. The different types of classifiers are decision tree classifier, neural network, naïve bayes classifier, support vector machine etc. The experiment is conducted here using the datamining tool WEKA 3.7. The different classifiers under the study are JRip, ZeroR, SMO, Adaboost, Random Forest, Kstar, Ridor and DTNB.

4.3 A Comparative Study

The performance metrics of each classifier can be improved by reducing the number of attributes of input data set. The reduction can be done by removing the irrelevant and noisy attributes. The performance metrics are evaluated based on the three main factors such as Accuracy, Kappa Statistic and Mean Absolute Error. A comparison of the classification accuracy was performed using WEKA 3.7, before and after attribute selection using the above stated classifiers. The Different performance metrics are used to compare and evaluate which one is better and whether the feature selection is mandatory before applying the classification during the mining process.

4.3.1 Accuracy

The percentage of correctly classified instances is known as accuracy. Accuracy is a widely used metric for measuring the performance of a classifier.

4.3.2 Kappa Statistic

Cohen's kappa is another widely used measure. It measures the extent to which the agreement between observed and predicted is higher than that expected by chance alone [9]. A Kappa Statistic value greater than 0 means that the classifier selected is better than chance.

4.3.3 Mean Absolute Error

The Mean Absolute Error measures the average magnitude of the errors in a set of forecasts. The error rates are used for numeric prediction rather than classification. The Mean Absolute Error values are in the range from 0 to ∞. They are negatively-oriented scores. That means that lower mean absolute error values are better.

After removing the least important seven attributes the classification results are explained below. The Table 3 displays the experiment results before applying the feature selection. The Table 4 shows the results after applying the feature selection to the input data set.

Table 3. Classifiers Performance Before Feature Selection Based On Accuracy And Kappa

Classifiers	Accuracy	Kappa	Mean Absolute Error
JRip	74.3	0.346	0.3666
ZeroR	70	0	0.42
SMO	78.4	0.45	0.216
Adaboost	73.7	0.225	0.342
Random Forest	99	0.976	0.124
Kstar	100	0	0.009
Ridor	76	0.2701	0.24
DTNB	71.1	0.394	0.362

Table 4. Classifiers Performance After Feature Selection Based On Accuracy And Kappa

Classifiers	Accuracy	Kappa	Mean Absolute Error
JRip	75.6	0.364	0.359
ZeroR	70	0	0.42
SMO	78.5	0.444	0.215
Adaboost	73.7	0.226	0.342
Random Forest	99.6	0.99	0.124
Kstar	100	1	0.002
Ridor	78.3	0.374	0.271
DTNB	74.9	0.437	0.346

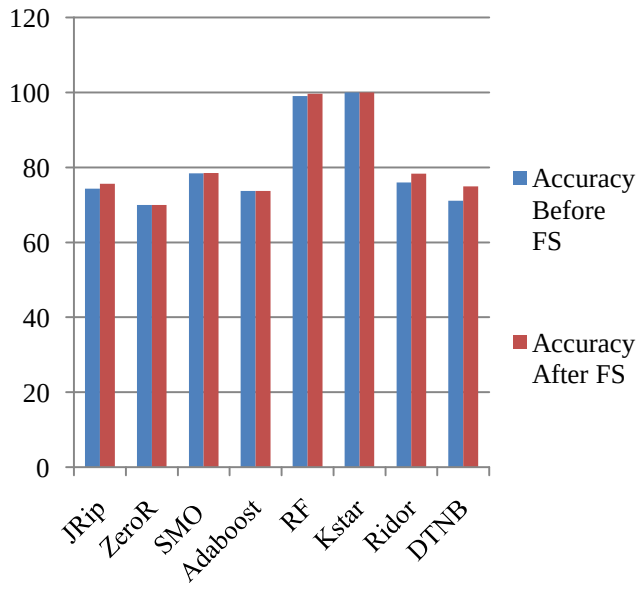

Fig 4. Graph showing performance accuracy of different classifiers before and after feature selection

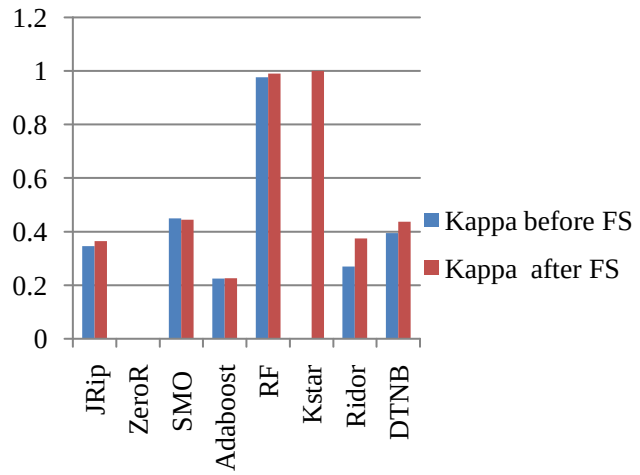

Fig.5. Graph showing comparison of classifiers before and after feature selection based on kappa.

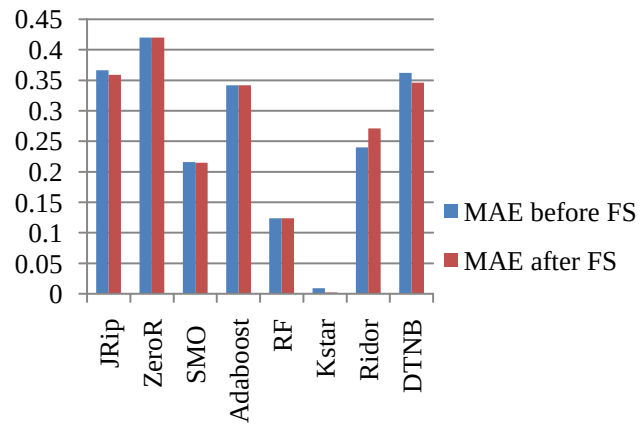

Fig. 6. Graph showing comparison of classifiers before and after feature selection based on Mean Absolute Error

From this comparative study, it is evident that the feature selection is very important. It is clear from this work that all the attributes in a dataset is not required for classification. Noisy and unimportant attributes can be removed from the data set. Data set contains 20 attributes including the class attribute. So after feature selection 20 attributes has been reduced to 14 attributes including the class attribute. Table 3 shows the classifiers performance before feature selection based on accuracy, kappa and mean absolute error . Table 4 shows the classifiers performance after feature selection based on accuracy, kappa and mean absolute error. Random Forest and Kstar classifiers have highest accuracy and least mean absolute error value. From this study we can see that Random Forest and Kstar are the best classifiers with respect to their performance metrics.

The graph given in the figure 4 shows the performance accuracy of different classifiers before and after the application of feature selection. The graph given in the figure 5 shows the kappa performance of different classifiers before and after the application of feature selection. The graph given in the figure 6 shows the mean absolute error performance of different classifiers before and after the application of feature selection. Some classifiers show better performance than the other classifiers after feature selection. Some classifiers are giving almost the same performance. These graphs confirm the fact that feature selection produces the same or improved classification accuracy.

5. CONCLUSION

In this paper, the comparative study of different classifiers before and after feature selection was performed. It is concluded that that feature selection i.e. removal of noisy and redundant attributes, indeed leads to better classification accuracy.

As a future scope, we can experiment the same techniques over different data sets and use this scheme to decide the best classifier as well as the best feature selection method that may suit to any particular dataset.

6. REFERENCES

[1] Dileep B. Desai, Dr. R.V.Kulkarni "A Review: Application of Data Mining Tools in CRM for Selected Banks", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 199 – 201.).

[2] Dr. K. Chitra1, B. Subashini , "Data Mining Techniques and its Applications in Banking Sector " , International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8, August 2013) K.Arutchelvan, Dr.R.Periyasamy,' Cancer Prediction System Using Datamining Techniques ', International Research Journal Of Engineering And Technology (Irjet) E-Issn: 2395-0056 Volume: 02 Issue: 08 | Nov-2015 Www.Irjet.Net P-Issn: 2395-0072

[3] Ioannis Anagnostopoulos,Christos Anagnostopoulos, Angelos Rouskas1,George Kormentzas,Dimitrios Vergados, The Wisconsin Breast Cancer Problem: Diagnosis and DFS time prognosis using probabilistic and generalised regression neural classifiers, special issue Computational Analysis and Decision Support Systems in Oncology, last quarter 2005

[4] R.P.L.Durgabai, "Feature Selection using ReliefF Algorithm" International Journal of Advanced Research in Computer and Communication Engineering , Vol. 3 , Issue 10 October 2014

[5] M.A. Hall and G. Holmes, ―Benchmarking Attribute Selection
Techniques for Discrete Class Data Mining,‖ IEEE Trans.
Knowledge and Data Eng., vol. 15, no. 6, pp. 1437-1447, Nov./Dec.2003.

[6] Xiubo Geng, Tie-Yan Liu, Tao Qin, Hang Li, 'Feature Selection for Ranking', SIGIR'07, July 23–27, 2007, Amsterdam,Holland.Copyright 2007 ACM 1-58113-000-0/00/0004...$5.00

[7] Roberto Ruiz, Jes´us S. Aguilar–Ruiz, Jos´e C. Riquelme, and Norberto D´ıaz–D´ıaz, 'Analysis of Feature Rankings for Classification', A.F. Famili et al. (Eds.): IDA 2005, LNCS 3646, pp. 362–372, 2005. _c Springer-Verlag Berlin Heidelberg 2005

[8] Kira, K., Rendell, L.: A practical approach to feature selection. In: 9th Int. Conf.on Machine Learning, Aberdeen, Scotland, Morgan Kaufmann (1992) 249–256

[9] Rodrigo Morgon, Silvio do Lago Pereira,' Evolutionary Learning of Concepts', Journal of Computer and Communications
Vol.02 No.08(2014), Article ID:47412,10 pages 10.4236/jcc.2014.28008

[10] Swasti Singhal, Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering" , International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013

[11] L´aszl´o A. Jeni1 , Jeffrey F. Cohn1, 2, and Fernando De La Torre, 'Facing Imbalanced Data Recommendations for the Use of Performance Metrics', Proceeding ACII '13 Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction Pages 245-251IEEE Computer Society Washington, DC, USA ©2013

[12] Kehan Gao, Taghi M. Khoshgoftaar, Huanjing Wang and Naeem Seliya,' Choosing software metrics for defect prediction: an investigation on feature selection techniques', SOFTWARE – PRACTICE AND EXPERIENCE Softw. Pract. Exper. 2011; 41:579–606 Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.1043

[13] Mitchell, T.: Machine Learning. McGraw Hill (1997)

[14] L.Breiman.Random forests.Machine Learning,45(1):5–32, 2001.

[15] Jehad Ali ,Rehanullah Khan, Nasir Ahmad, Imran Maqsood, 'Random Forests and Decision Trees', IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012, ISSN (Online): 1694-0814.