International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 1
ISSN: 2454-4248
53 – 58

A Novel Hybrid Classification Model For the Loan Repayment Capability Prediction System

Soni P M
Research Scholar
Bharathiar University
Coimbatore, India
sonipm@sngist.org

Varghese Paul
Professor in IT
RSET
Kakkanad , India
varghesep@rajagiritech.ac.in

***Abstract*—** Classification is a powerful tool in Data mining to predict the loan repayment capability of a banking customer. This paper evaluates the performance of various classification algorithms and selects the most appropriate one for predicting the class label of the credit data set as good or bad. Feature selection is a data pre-processing technique refers to the process of identifying the most beneficial features for a given task, while avoiding the noisy, irrelevant and redundant features of the dataset. These irrelevant noisy features results in a poor accuracy for the selected classifier. In order to improve the accuracy of a classifier, the feature selection plays a vital role as a data preprocessing step. Feature selection technique reduces the dimensionality of the feature set of the dataset. This paper has two objectives. First objective is to find out the best classifier algorithm for the credit data set using two different tools such as weka and R. Here the experiment proved that Random Forest performs better for loan repayment credibility prediction system. The second objective is to evaluate the performance of various feature selection methods based on Random Forest classification method. Also a novel hybrid model is developed for the same.

Keywords- *feature selection, accuracy , performance,classification*

*****

I. INTRODUCTION

Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. [1]. Today, Customer Relationship Management in banking industry is purely based on Data Mining techniques. The different 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 and Cash management and Forecasting operations [2]. In developing countries like India, bankers should vigilant to fraudsters because they will create more problems to the banking organization. Banks hold huge volumes of customer behavior related data from which they are unable to arrive at a judgment if an applicant can be defaulter or not [2]. Applying classification techniques, it is very effective to build a successful predictive model that helps the bankers to take the proper decision.. Different classification algorithms produce different level of accuracy. Classification is one of the data analysis methods that predict the class labels [3]. There are more classification methods such as statistical based, distance based, decision tree based, neural network based, rule based[4]. Here we evaluate the accuracy of various classification algorithms using the two powerful data mining tools such as R and weka. The experiment proved that the accuracy obtained for the random forest classifier using weka is more perfect. The experiment shows that random forest performs better on the credit data set. Feature Selection plays a major role in the accuracy of classifier by removing the irrelevant features of the data set. For this experiment weka performs better and proved that OneRattributeEval feature selection leads to high accuracy random forest classification for the prediction of loan repayment capability of a customer.

This paper is organized as follows. The next section explains about the dataset and data mining tools used for conducting the experiment. Section III discuss about the proposed architecture and algorithm for the experiment. Section IV discusses with the concepts used such as feature selection, classification and random forest. Section V explains about results and discussions. A comparative study is performed here. Section VI demonstrates the proposed hybrid classification model. Conclusion is given in section VII followed by references.

II. DATASET AND TOOLS

Data was collected from a premier cooperative bank that provides loans to individuals, business firms, etc so as to meet the requirements of all type of customers. Data collection was completed through procedures including on site observation and interview with the concerned authority. A detailed study about the loan processing and banking transactions are also made for the same. The data available consists of 2500 records of bank loan transaction data including 25 data fields. Some of the fields are removed directly by manual data preprocessing and the database is termed as credit dataset.

In order to conduct the two stages of experiment, the two most popular data mining tools were used. They are R programming and weka. As the preprocessing step is the most important and time consuming one, classification and clustering techniques in R were used to make the data ready for further use [5]. Weka is also an efficient tool for evaluating different data mining techniques such as preprocessing, classification, clustering etc.

 ISSN: 2454-4248
53 – 58

III. PROPOSED EXPERIMENT

A. *Procedure*

The different steps that were carried out for conducting the experiment have been described in this section. The framework of the proposed work is given in the figure 1. The preprocessed data is considered as the input to the system. The various data preprocessing techniques were performed to remove noisy and redundant data from the database. Feature selection is also a data preprocessing technique to select the relevant attributes for the experiment. In order to remove unimportant attributes feature selection plays a vital role. Weka evaluates the performance of various feature selection methods. Weka provides various classification algorithms such as JRip, ZeroR, SMO, Adaboost, Ridor and Random Forest. R programming includes classification algorithms like Random Forest, Rpart, Nnet, Lda and Lmt. Comparison between these two were made and found that Random Forest will produce better performance. Then compared each feature selection method with Random Forest algorithm and found that OneAttributeEval produce better classification. In this proposed architecture Random Forest produce best accuracy and the suitable feature selection method found is OneRAttributeEval.

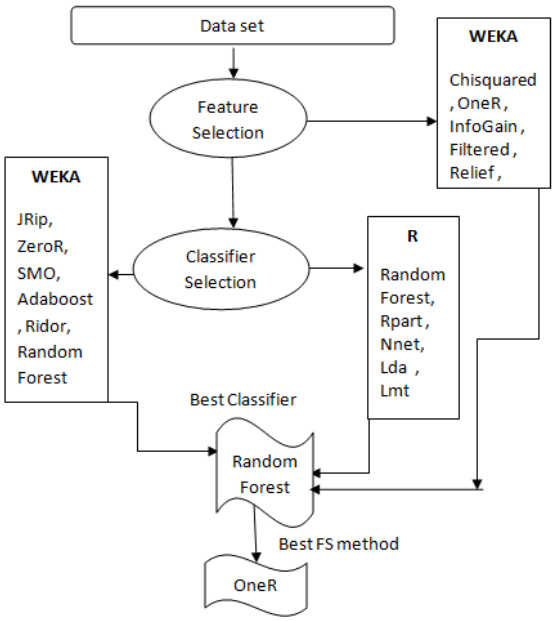


Figure 1:. Proposed architecture for the experiment.

B. *Algorithm*

The figure 2 explains about the algorithm for the proposed experiment. . Arrays are used to store the feature selection methods and classifications methods under we selected for the experiment. FS is an array to store the feature selection methods such as Chisquared, OneR, Infogain, Filtered and Relief. CLW is an array to store all classification methods in weka. In this experiment CLW contains JRip, ZeroR, SMO, Adaboost, Ridor and Random Forest. CLR is an array to store all classification methods in R. The classification methods in this experiment under R are Rpart, Nnet, Lda, Lmt and Random Forest. Each classification methods are evaluated and find the best classifier in both tools. If the best classification method obtained is same in both tools we can select it as a best classifier. Here it is stored in the variable BCL. In order to find the best feature selection method, apply each feature selection method of FS to the best classifier. The best feature selection method is stored into the variable BFS. Finally the algorithm returns BCL and BFS as best classifier and best feature selection method suitable for the best classifier.

1. Start
2. Select 'm' feature selection methods and store in an array FS[ ]
3. Select 'n1' classification methods in weka and store in an array CLW[ ]
4. Select 'n2' classification methods in R and store in an array CLR[ ]
5. For i=1 to n1
   Compare accuracy of each CLW[i] and store the best CLW [i] into the variable WCL.
   End
6. For i = 1 to n2
   Compare accuracy of each CLR[i] and store the best CLR [i] into the variable RCL.
   End
7. If WCL=RCL the best classifier is same in both weka and R. Assign WCL or RCL to BCL.
8. For i= 1 to m
   Apply each FS[i] to the BCL and find the best FS [i] and store into the variable BFS.
9. Return BCL and BFS.
10. End

Figure 2:. Algorithm for Proposed experiment

IV. CONCEPTS USED

The concepts used in the experiment are feature selection, classification and especially random forest algorithm. This section also explains about various parameters that affect the performance of random forest classifier.

A. *Feature Selection*

Feature selection is the process of removing redundant or irrelevant features from the original data set. So the execution time of the classifier that processes the data reduces, also accuracy increases because irrelevant features can include noisy data affecting the classification accuracy negatively[6] Feature selection has become interest to many research areas which deal with machine learning and data mining, because it provides the classifiers to be fast, cost-effective, and more accurate[1]. Feature selection is a preprocessing technique used for reducing the number of attributes by removing the irrelevant and noisy attributes. This process leads to increases the learning accuracy of the classifier and the predictions also more accurate. Reducing the dimensionality of the data reduces the size of the hypothesis space and thus results in faster execution time [6]. The accuracy obtained before feature selection and after feature selection had examined and found that feature selection produces good classification accuracy.

 ISSN: 2454-4248
53 – 58

B. *Classification*

Classification is a Data mining technique that classifies a new data record into one of the many possible classes which are already known. For example, a classification model can be used to identify loan applicants as low, medium, or high credit risks [7]. This supervised learning technique accurately predicts the target class based on training and test data sets. Classification is the most commonly applied data mining technique, which employs a set of pre classified examples to develop a model that can classify the population of records at large [8]. Here the classification model can be used to predict the loan repayment capability of customers by splitting the records into two class labels called good or bad. Data mining techniques can be implemented mainly using weka or R programming. The different types of classifiers are decision tree classifier, neural network, naïve bayes classifier, support vector machine etc. The different classifiers under the tool weka are JRip, RF, SMO, Ridor, ZeroR and Adaboost. RF, rpart, nnet, lda and lmt are some classifiers chosen in R programming. The experiment evaluates the accuracy of the credit data set under each of the classifier in weka or R programming. Also a comparison was performed to evaluate which tool is better than the other based on their accuracy produced. To classify if the applicant is a defaulter or not, the best data mining approach is the classification modeling using Decision Tree [3]. From this experiment it is also found that random forest performs better in both weka and R programming tools .

C. *Random Forest*

Using Random Forests for prediction has many advantages such as their immunity to over fitting, an appropriate selection of randomness type leads to accurate classification or regression, the correlation and strength of predictors makes a good estimate of the ability for prediction, faster than boosting and bagging, better estimation of internal errors, not complicated, and can perform well in parallel processing [10].
The important parameters of random forest algorithm in R programming are[11]

1. mtry: mtry refers to the number of variables selected at each split. mtry is calculated as the floor of square root of the number of independent variables. For regression model mtry is calculated as the floor of number of variables divided by 3.

2. ntree : ntree refers to the number of trees to develop . By default value of ntree is 500.

3. nodesize :nodesize refers to the minimum size of terminal nodes. By default the value is 1.

4. replace : It is flag to check whether sampling with/without replacement. TRUE implies with replacement. FALSE implies without replacement. By default it is TRUE .

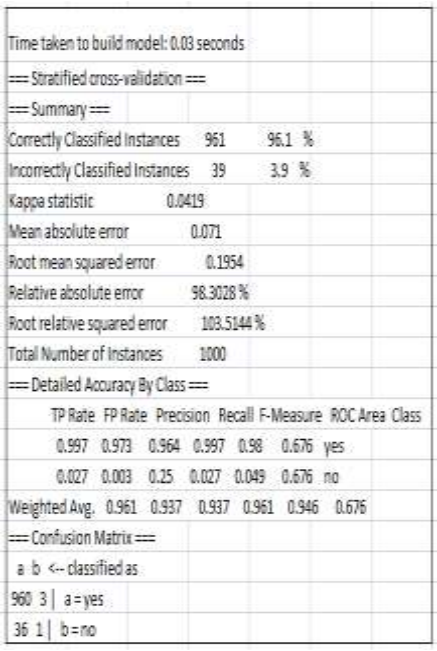
5. sampsize : Sample size to be drawn from the data for growing each decision tree. By default, it takes 63.2% of the data

6. importance: It is flag to determine whether importance metrics of variables to be required or not.

The various parameters that affect the performance of random forest algorithm in weka are

1. m_numTrees :Number of trees in forest.
2. m_numFeatures : Number of features to consider in random feature selection.
3. m_randomSeed :The random seed.
4. m_KValue : Final number of features that were considered in last build.
5. Seed: Random number seed to be used.

The number of features in the data set is an important parameter that affects the classification accuracy and it leads to the fact that feature selection has a major role in data mining process. Seed is actually not a parameter and it is used to generate pseudo-random numbers. Two parameters are important in the random forest algorithm are Number of trees used in the forest and Number of random variables used in each tree . Information gain is the function by which we split the data into daughter nodes in a particular tree of the random forest
The summary of the classification procedure is listed in Figure 10. It list the number of correctly classified attributes , incorrectly classified attributes, Kappa Statistic , mean absolute error, root mean squared error, relative absolute error, root relative absolute error and total number of instances.
TP Rate, FP Rate , Precision , Recall , F-Measure , ROC Area , Class and weighted averages of each are represented as a table. The confusion matrix easily pointed out the number of good and bad customers.



Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	961	96.1 %
Incorrectly Classified Instances	39	3.9 %
Kappa statistic	0.0419	
Mean absolute error	0.071	
Root mean squared error	0.1954	
Relative absolute error	98.3028 %	
Root relative squared error	103.5144 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.997	0.973	0.964	0.997	0.98	0.676	yes
	0.027	0.003	0.25	0.027	0.049	0.676	no
Weighted Avg.	0.961	0.937	0.937	0.961	0.946	0.676	

=== Confusion Matrix ===

a	b	<-- classified as
960	3	a = yes
36	1	b = no

Figure 3 : Summary of Random Forest

 ISSN: 2454-4248
53 – 58

V. RESULTS AND DISCUSSION

Table 1 shows the accuracy obtained by random forest algorithm after applying various feature selection methods using Weka .

TABLE 1 : Feature selection performance in weka

Feature Selection	Time	Correctly classified	Incorrectly classified
Chisquared	0.03	78.4	21.6
Filtered	0.06	74.7	25.3
InfoGain	0.03	74.7	25.3
OneR	0.03	96.1	3.9
Relief	0.06	60.4	39.6

TABLE II : Classification Accuracy on credit dataset in weka

Classifiers	Accuracy
JRip	74.3
ZeroR	70
SMO	78.4
Adaboost	73.7
Random Forest	99
Ridor	76

TABLE III : Classification Accuracy on credit dataset in R

Classifiers	Accuracy
Random Forest	0.81237
rpart	0.80358
nnet	0.80280
lda	0.80202
Lmt	0.81048

Table 2 shows the classification accuracy of various classifiers in weka after OneR feature selection method. It also shows the time taken to conduct the experiment. From the table it is clear that Random Forest performs better. Table 3 shows the classification accuracy of various classifiers in R From the table it is clear that Random Forest classification produce high accuracy than others.

The performance metrics considered in the experiment are accuracy and Time. Accuracy is the percentage of the correctly classified positive and negative examples. [8] Accuracy is a widely used metric for measuring the performance of a classifier. The credit data set is classified into two categories such as persons who are capable of repaying the loan amount and persons who don't have the capability to repay the loan amount. This classification performed better in Random Forest algorithm using weka and also the same in R . Weka produce better accuracy than R. Considering the time required for completing the feature selection process OneRAttributeEval, ChisquaredAttributeEval and InfoGainAttributeEval performed with less time than FilteredAttributeEval and RelieF. So with respect to time and accuracy we can consider OneRAttributeEval is the best feature selection method for Random Forest Classification algorithm. Figure 4 depicts the graph of classification accuracy in weka. Figure 5 depicts the graph of classification accuracy in R.

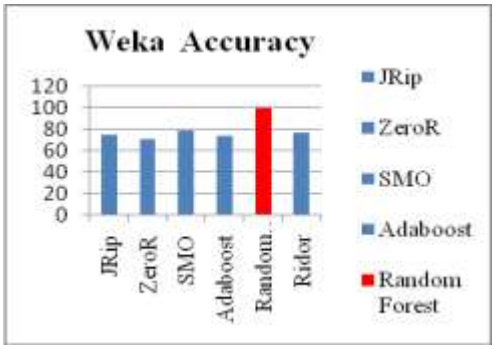


Figure 4: Classification Accuracy in weka Selection

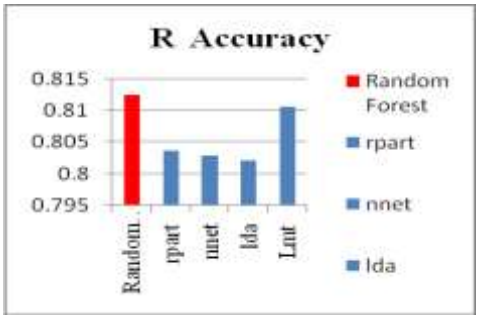


Figure 5: Classification Accuracy in R

Feature selection performance such as correctly classified and incorrectly classified records are explained in Figure 6. The Time required for each Feature Selection method is explained in Figure 7.

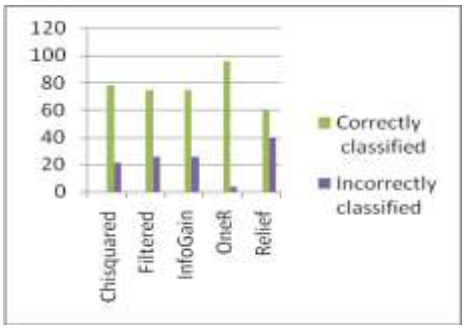


Figure 6 : Feature Selection Performance in weka

 ISSN: 2454-4248
53 – 58

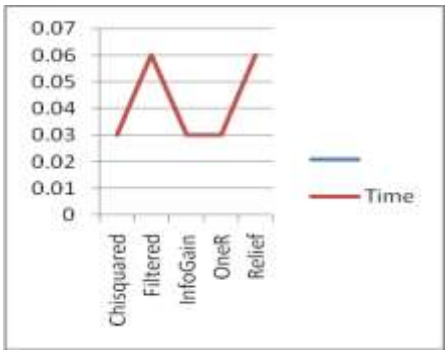


Figure 7: Time required for each FS

VI. Proposed Hybrid Model

The proposed hybrid model of classification is explained in figure 9. The credit dataset is preprocessed by the feature selection method OneRattributeEval followed by random forest classification process . The ranked attribute list obtained after OneRattributeEval feature selection method is listed in figure8.

Ranked attributes:	
71.7	3 credit_history
70.1	2 duration
70	20 foreign_worker
70	7 employment
70	8 installment_commitment
70	4 purpose
70	6 savings_status
70	19 own_telephone
70	9 personal_status
70	10 other_parties
70	11 residence_since
70	18 num_dependents
70	16 existing_credits
70	12 property_magnitude
70	15 housing
70	14 other_payment_plans
70	17 job
69.3	13 age
67.6	1 checking_status
66.1	5 credit_amount

Figure8: Ranked attribute list

The model classifies the credit data set using random forest after applying OneRattributeEval feature selection method leads to high accuracy. Also the model is generated using weka tool. The comparative studies stated in this paper proved that the proposed hybrid model can effectively classify the person into two categories such as who pay the loan promptly or not. The output prediction classifies the class label of the credit data set into two labels such as good or bad. The experiment done in R also proves that classification through Random Forest produce high accuracy than others for the loan repayment credibility prediction system.

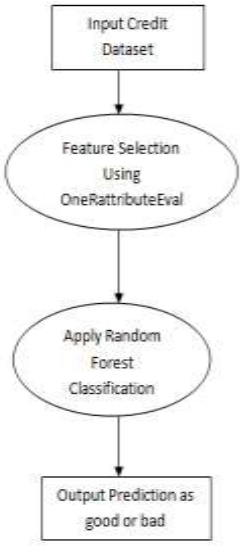


Figure 9 : Proposed Hybrid model for classification

VII. Conclusion

The experiment was conducted to predict the best feature selection method and best classifier performance. The experiment was implemented using the data mining tools such as weka and R programming. Random Forest classifier produces high accuracy in both weka and R under the credit data set. Several feature selection methods are evaluated and found that OneRAttributeEval will perform better in random forest classifier algorithm which was found to be the best classifier for classifying the persons who repay the loan amount promptly or not. The paper illustrates that the combination of Random Forest and OneRAttributeEval produce a good performance over credit dataset .

References

[1] Esra Mahsereci Karabuluta, Selma Ayşe Özelb, Turgay İbrikçib , "A comparative study on the effect of feature selection on classification accuracy" , Procedia Technology 1 (2012) 323 – 327

[2] L. Torgo, *Functions and data for "data mining with r" R package version 0.2.3*, 2012.

[3] Sudhamathy G. "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R" Vol 8 No 5 Oct-Nov 2016

[4] M. Dash and H. Liu, "Feature Selection for Classification,IIntelligent Data Analysis", vol. 1, nos. 1-4, pp. 131-156, 1997.

[5] K. Chitra, B.Subashini, "An Efficient Algorithm for Detecting Credit Card Frauds", Proceedings of State Level Seminar on Emerging Trends in Banking Industry, March 2013

[6] 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

[7] S. Doraisami, S. Golzari, "A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music ",ISMIR 2008 – Session 3a – Content-Based Retrieval, Categorization and Similarity 1

[8] K. Chitra, B.Subashini, Data Mining Techniques and its Applications in Banking Sector, International Journal of Emerging Technology and Advanced Engineering , ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8, August 2013

 ISSN: 2454-4248
53 – 58

[9] Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1 No. 4

[10] 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

[11] Nazeeh Ghatasheh , "Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study ",International Journal of Advanced Science and Technology Vol.72 (2014), pp.19 – 30

[12] http://www.listendata.com/2014/11/random-forest-with-r.html