An Ensemble Model of Outlier Detection with Random Tree Data Classification for Financial Credit Scoring Prediction System

V. Veeramanikandan, M. Jeyakarthic

Abstract: ***Recently, Financial Credit Scoring (FCS) becomes an essential process in the finance industry for assessing the creditworth of individual or financial firms. Several artificial intelligence (AI) models have been already presented for the classification of financial data. However, the credit as well as financial data generally comprises unwanted and repetitive features which lead to inefficient classification performance. To overcome this issue, in this paper, a new financial credit scoring (FCS) prediction model is developed by incorporating the process of outlier detection (OD) process (i.e. misclassified instance removal) prior to data classification. The presented FCS model involves two main phases namely misclassified instance removal using Naïve Bayes (NB) Tree and Random Tree (RT) based data classification. The presented NB-RT model is validated using the Benchmark German Credit dataset under different validation parameters. The extensive experiments exhibited that a maximum classification accuracy of 90.3% has been achieved by the proposed NB-RT model.***

Keywords: ***Classification; Credit Scoring; FCS; Naïve Bayes; Outliers.***

I. INTRODUCTION

Presently, the rapid increase in the financial crisis of firm in all parts of the globe leads to the development of the hot research topic of financial credit scoring (FCS) [1]. In any financial firm or organization, it is really hard to design a prediction model for forecasting the significant risks of the firm's financial condition in earlier. The FCSnormallydevelops a binary classification model that is resolved in a balancedmanner [2]. The output from the classifier model undergoes categorization into two kinds: representing failure condition of a firm and representing the non-failure condition of a firm. The contribution to the classifier technique is based on the statistical values attained from the financial details in present organizations. Presently, many classifier approaches have been presented by the use of diverse information for FCS. In general, the available FCS approaches undergo partitioning into statistical and artificial intelligence (AI) models. Recently, there is a much interest shown on the AI based FCS models.

Few of the AI approaches are decision tree (DT), random forest (RF), and so on. In addition, some hybridization of different models are also presented for FCS. A set of two new kernels by the use of soft computing approaches are presented for data classification [3]. The attained simulation outcome indicated that the presented method is an effective approach to determine the financial crisis. A new approach of utilizing ant colony optimization (ACO) and Nelder-Mead simplex model for training NN for FCS is presented [4]. A modified bacterial foraging model was introduced to carry out the training process of the wavelet neural network (WNN) for identifying the financial condition of banks [5]. An efficient model by the use of differential evolution (DE) and radial basis function (RBF) network called DERBF is developed for FCS [6]. The output indicated that the projected model is an efficient FCS technique on the applied banks dataset.

A DE model to train WNN for FCS in banks is developed in [7]. The experimental outcome on the employed 4 dataset indicated that the projected DEWNN technique was an efficient one over the compared models. An efficient approach known asprincipal component neural network (PCNN) to predict financial crisis in commercial banks is developed [8]. It showed that the projected PCNN technique is better than the other models on the identical FCS dataset. A kernel principal component neural network (KPCNN) model is presented in [9] where it was trained by the threshold accepting model. It was employed for predicting financial crisis in banks and the outcome indicated that the KPCNN attains considerable outcome over the compared models.

An efficient principal component analysis-WNN (PCATAWNN) technique that is also trained by the threshold accepting model is developed [10]. The simulation results reported that the projected technique is effectivefor FCS in banking domains. The simulation outcome reported that the projected technique shows maximum performance over other methods. In between the dissimilar models, ANN is the effective FCS technique due to its capability to take the nonlinearity relationship between diverse characteristics in real-time dataset [11-13]. But it is noted that the ANN learning models are based on gradient descent mechanisms that result in local optimal performance. Additionally, it is widely required that an appropriate quantity of network parameters should undergo tuning.

Manuscript published on 30 September 2019
* Correspondence Author

V. Veeramanikandan*, Assistant Professor, Dept. Computer Science, T.K.Govt.Arts College, Vridhachalam, Email: klmvmani@gmail.com
M. Jeyakarthic, Assistant Director, Tamil virtual Academy, Chennai, Email:jeya_karthic@yahoo.com

At present, different methods have exhibited that the feed forward NN (FNNs) operates as a substitute method to classify data over the traditional statistical models [14]. Although different works have been made by exhibiting the goodness of the FNN in distinct models, few difficulties are presented in real time applications. Few of the difficulties are mentioned as follows: It is inflexible to recognize the proper FNN model that represents the characteristics of the problemssuch as network model, learning techniques and variables.

On the other hand, the nature of financial data contains unwanted, missing and repetitive instances. To resolve this issue, outliers or misclassified instances can be removed by the procedure of outlier detection (OD) technique. This OD model assists in the improvisation of the classification results. To overcome this issue, in this paper, a new financial credit scoring (FCS) prediction model is developed by incorporating the process of misclassified instance removal process (i.e. outlier detection) prior to data classification. The presented FCS model involves two main phases namely outlier detection (OD) and data classification. Naïve Bayes (NB) Tree technique is applied for removing the misclassified instances. Once the misclassified instances are eliminated, Random Tree (RT) technique is applied to classify the data. The presented NB-RT model is validated on the German Credit dataset under different validation parameters.

The upcoming parts are arranged as follows: Section 2 introduces the NB-RT approach. Section 3 performs experimental analysis and Section 4 provides conclusion.

II. PROPOSEDFCSMODEL

The process involved in the presented FCS model is shown in Fig. 1. Initially, the raw financial dataset is provided into the system and pre-processing takes place. This process converts the format of the dataset so that it is appropriate for further processing. Next, the sample selection process takes place. Furthermore, the OD process takes place by the use of NBTree. Once the outliers or misclassified instances are removed, the classification process is carried out utilizing RT model. Then, testing of data takes place and the performance measures are derived.

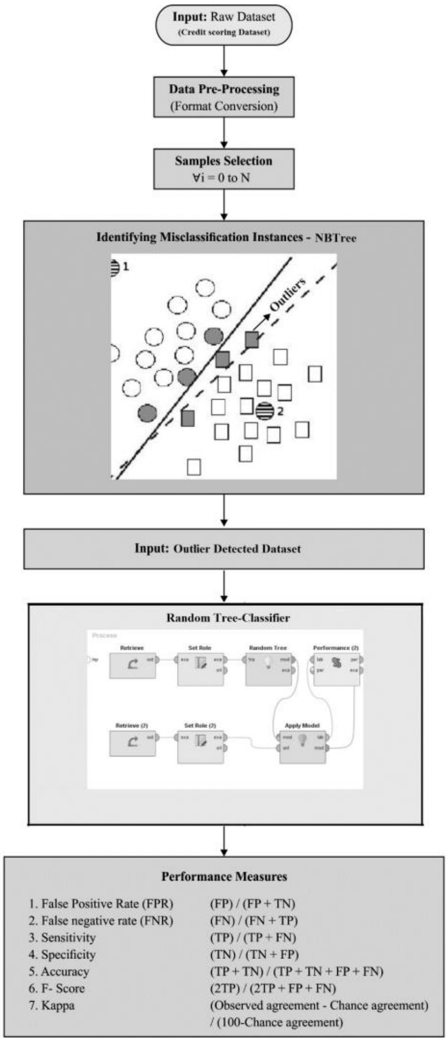

Input: Raw Dataset
(Credit scoring Dataset)
Data Pre-Processing
(Format Conversion)
Samples Selection
∀i = 0 to N
Identifying Misclassification Instances - NBTree
1
Outliers
2
Input: Outlier Detected Dataset
Random Tree-Classifier
Process
Retrieve
Set Role
Random Tree
Performance (2)
Retrieve (2)
Set Role (2)
Apply Model
Performance Measures
1. False Positive Rate (FPR) (FP) / (FP + TN)
2. False negative rate (FNR) (FN) / (FN + TP)
3. Sensitivity (TP) / (TP + FN)
4. Specificity (TN) / (TN + FP)
5. Accuracy (TP + TN) / (TP + TN + FP + FN)
6. F- Score (2TP) / (2TP + FP + FN)
7. Kappa (Observed agreement - Chance agreement) / (100-Chance agreement)

Fig. 1. Working principle of ProposedModel

A. NBTree based outlier detection model

To find out the outlier data effectively, NBTree has been presented. NBTree is a hybridization method which makes use of naive Bayes classification on every leaf node of the constructed DT and has shown considerable classifier results with respect to detection rate. For constructing the NB tree, the classifier accuracy measure is employed rather than the information gain metric. For every attribute A_i, NBTree validates the classifier accuracy of the split based model on the attribute A_i and selects the attribute A_h with maximum classifier accuracy Acc_h for constructing a tree. When Acc_h is not considerably higher than the present node, then NB classification model is constructed for the present node, otherwise a split is generated to the attribute A_h.

NBTree iteratively constructs the tree till the number of instances is lesser than the preset leaf size 30 (default) or enhancement of Acc_h is lesser than 5% (by default). It can be shown that the NBTree selects the attributes in a greedy and recursive way with maximum classifier accuracy for constructing the tree. It forms the basis of why NBTree has the relatively maximum classifier rate and it incurs high computation complexity. For classifying a predefined test instance, NBTree arranges it down the tree from the root node to few leaf nodes and then employs the training instances which fall into the leaf nodes for constructing a NB classification model.

Finally, the NB classification model is employed for the removal of outliers.

B. RT based classification model

The RT operator operates such as DT operator with a single exception: for every split only an arbitrary set of attributes exists. The RT performs the learning of DT from both nominal and mathematical data. The DT is an efficient classifier model that can be simply understood. The RT operator operates in an identical way to Quinlan's C4.5. The subset size is mentioned using variable ***subset ratio***.

The way of representing the data as a tree holds the benefit of easier interpretation. The aim is to design a classifier model which identifies the label values depending upon diverse input features of the ExampleSet. Every internal node of tree indicates to one of the input features. The edge count of an internal node is equivalent to the number of probable values of the respective input attribute. Every leaf node indicates a value of the label provided to the values of the input attributes indicated by a route from root to the leaf. Next, pruning process takes place where the leaf nodes which are not included to the discriminative power of the DT are eliminated. It is carried out for enhancing the classification process of the unseen dataset. Pre-pruning is a kind of pruning which takes place parallel to the tree generation procedure. The post-pruning is carried out when the tree generation procedure is completed.

Differentiation

The RT operator operates with one exception that for every split an arbitrary subset of attribute is present.

Input

- **Training** set *(IOObject)*

This input port receives an ExampleSet which is obtained from the Retrieve operator in the linked Example Process. The output of other operators is also utilized as input.

Output

- **Model** *(Decision Tree)*

The RT is provided from this output port. This classifier method is employed to the unseen dataset to predict the model.

- **Example** set *(IOObject)*

The ExampleSet is provided as input with no modification to the output by this port. It is generally employed reusing the same ExampleSet in additional operators or viewing the ExampleSet in the Results Workspace.

Parameters

- **Criteria:** This variable chooses the criteria in which every attribute is chosen for splitting. It is one of the following values:
  - information_gain: The entropy of every attribute is determined. The attribute with least entropy is chosen for split. This technique has a bias towards the selection of attributes with larger values.
  - gain_ratio: It is an alternative of information gain which modifies it for every attribute.
  - gini_index: It is evaluation parameter of the impurity of an ExampleSet. Splitting on a selected attribute offers a decrease in the average gini index of the resultant subset.
  - accuracy: These variables are chosen for splitting the attributes which maximize the performance of the entire tree.

- **Minimal_gain**

The node gain is determined prior to the process of partitioning it. The node is divided when its gain is higher than the *minimal gain*. The maximum value of lower gain leads to less split and hence a smaller tree.

A higher value will be entirely eliminating the splitting and a tree with individual node is created.

- **Maximal_depth**

The tree depth is based on the size and nature of the ExampleSet. This variable is employed for restriction of the Tree size. The tree creation procedure is not continued in case the depth of the tree is equivalent to the Tree size. The tree creation procedure is not continued when the tree depth is equivalent to the *maximal depth*. When the value is assumed to be '-1', the *maximal depth* variable offers no limit on the tree depth, a tree of maximum depth is created. When the value is assumed to be '1', a Tree with an individual node is created.

- **Confidence**

This variable indicates the level of confidence employed for the pessimistic error computation of pruning.

- **Number_of_prepruning_alternatives**

Since pre-pruning executes parallel to the tree creation procedure, it might eliminate the splitting process at particular node. This variable modifies the number of alternate nodes managed to slit incase it is eliminated through the pre-pruning at a certain node.

- **Guess_subset_ratio:**

This variable indicates whether the subset ratio is guessed or not. When it is assumed to be true, *log(m) + 1* features are employed as subset, else, a ratio has to be represented by the *subset ratio* parameter.

- **Subset_ratio**

This variable indicates the subset ratio of arbitrarily selected attributes.

- **Use_local_random_seed**

This variable represents when a *local random seed* should be employed in an arbitrary way. By the use of identical value of the *local random seed*, it generates the identicalrandomization.

- **local_random_seed**

This variable represents the *local random seed*. This variable is only present when the use of *local random seed* variable is assumed to be true.

III. PERFORMANCE VALIDATION

For assessing the good characteristics of the presented NB-RT model, a series of experimental analysis is done using the benchmark German Credit dataset [15]. The information relevant to the dataset is given in Table 1. The details of the table indicate that the applied dataset holds a maximum of 1000 instances. A set of 20 attributes are present in the dataset under two classes namely bankrupt and non-bankrupt. Among the available instances, a total of 30% of instances comes under bankrupt class and the remaining 70% of instances comes under Non-Bankrupt. Next, Table 2 provides the description related to the features present in the dataset.

Since a total of 20 attributes exists in the dataset, it is represented as A1-A20 as mentioned in the table 2.

Table 1 Dataset Description

Parameter	Values
Dataset	German Credit
Source	University of California Irvine (UCI)
# of instances	1000
attributes	20
# of class	2
Bankrupt/Non-Bankrupt	300/700

Fig. 2 shows the distribution of the attributesunder the presence of instances under the applied German Credit dataset. The blue color in the figure indicates the non-bankrupt instances and the red color indicates the bankrupt instances. From this figure, it is easier to analyze the number of instances that falls under each attribute.

Table 3 displays the derived confusion matrix distinct classifier models namely MLP, RBF, LR and RT. The table values indicate that the presented NB-RT model classifies 553 instances under NB class and 62 instances under B class properly. In the same way, the RBF model classifies 609 instances under NB class and 134 instances under B class properly. Likewise, the LR model classifies 605 instances under NB class and 147 instances under B class properly. Similarly, the DT model classifies 591 instances under NB class and 121 instances under B class properly.

Table 2 German Credit Dataset Financial Features

Feature	Description	Feature	Description
A1	Credit amount	A11	Other instalment plans
A2	Status of existing checking account	A12	Personal status and sex
A3	Duration in months	A13	Foreign worker
A4	Age in years	A14	Other debtors/guarantors
A5	Credit history	A15	Instalment rate in percentage of disposable income
A6	Savings account/bonds	A16	Number of existing credits at this bank
A7	Purpose	A17	Job
A8	Property	A18	Telephone
A9	Present employment since	A19	Present residence since
A10	Housing	A20	Number of people being liable to provide maintenance

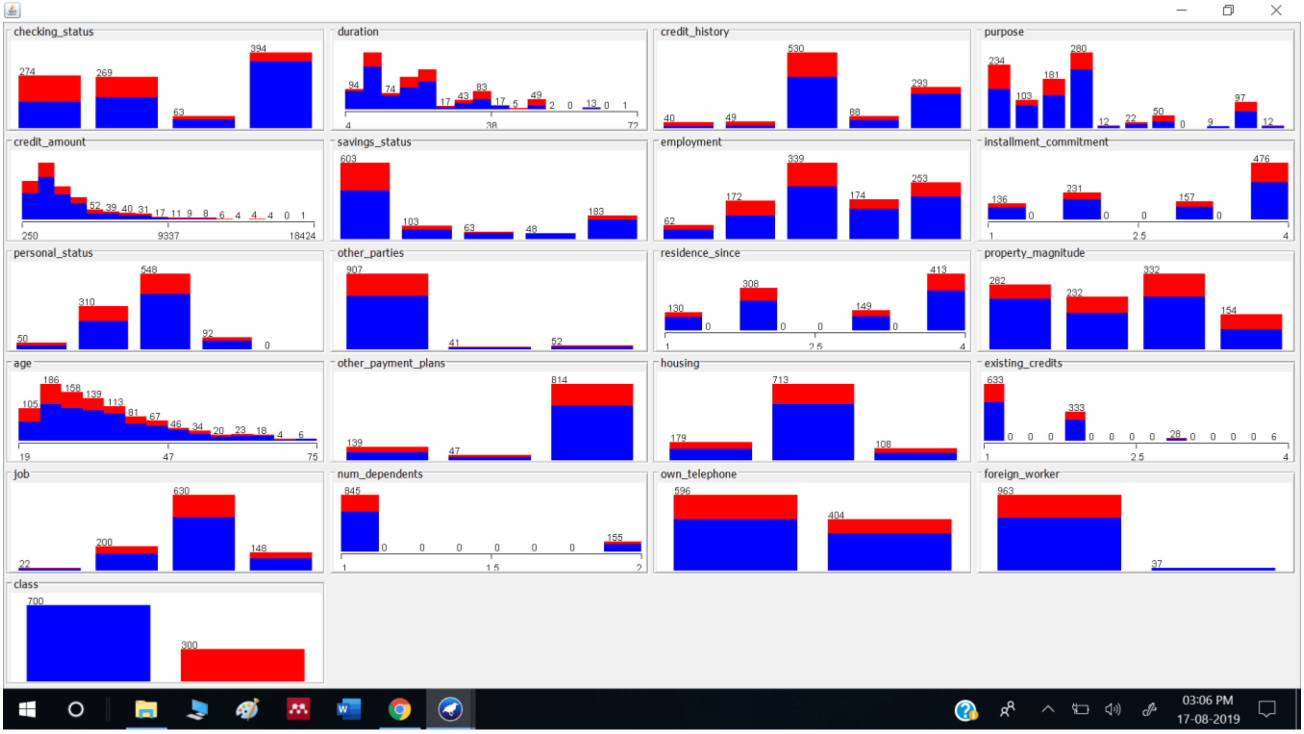
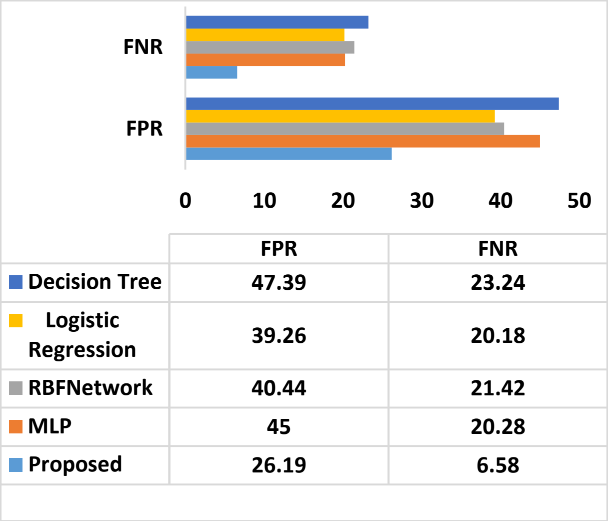


Fig. 2. Attributes Frequency Description of German Dataset

Table3Confusion Matrix of German Credit Bankruptcy Dataset using Various Classifiers

Experts	Proposed		MLP		RBFNetwork		LR		DT	
	NB	B	NB	B	NB	B	NB	B	NB	B
NB	553	22	574	126	609	91	605	95	591	109
B	39	62	146	154	166	134	153	147	179	121

Table 4 shows the classifier results analysis by various models on the applied identical dataset. Fig. 3 displays the analysis of classifier performance interms of FPR and FNR. Under the evaluation parameters such as FPR and FNR, the presented NB-RT model shows maximum classification by attaining lowest FPR and FNR values of 26.19 and 6.58 correspondingly. Furthermore, it is revealed that LR shows near optimal results obtained by the lower FPR and FNR values of 39.26 and 20.18 respectively. In addition, the RBF model exhibits that it shows poor performance over the LR and NB-RT model. However, it shows better classification over the MLP and DT by achieving the FPR and FNR values of 39.26 and 20.18 respectively. In the same way, the compared MLP shows poor classifier results with the higher FPR and FNR values of 45.00 and 20.28. Finally, the DT model exhibits worse classifier outcome with the maximum FPR and FNR values of 47.39 and 23.24 correspondingly.



	FPR	FNR
🟦 Decision Tree	47.39	23.24
🟨 Logistic Regression	39.26	20.18
⬜ RBFNetwork	40.44	21.42
🟧 MLP	45	20.28
🟦 Proposed	26.19	6.58

Fig. 3.Classifier results analysis with respect to FPR and FNR

Table 4 Various Classifier Results Analysis on German Credit Dataset

Classifier	FPR	FNR	Sensitivity	Specificity	Accuracy	F-score	Kappa
Proposed	26.19	6.58	93.41	73.80	90.97	94.77	61.85
MLP	45.00	20.28	79.72	55.00	72.80	80.84	33.98
RBFNetwork	40.44	21.42	78.58	59.55	74.30	82.58	34.10
LR	39.26	20.18	79.82	60.74	75.20	82.99	37.5
DT	47.39	23.24	76.75	52.61	71.20	80.41	26.93

Fig. 4 illustrates the classifier results interms of sensitivity and specificity. With respect to the performance measure of Sensitivity, the presented NB-RT model shows maximum classification by attaining highest Sensitivity value of 93.41 respectively. At the same time, it is noticed that the LR shows near optimal results obtained by the higher Sensitivity values of 79.82. In addition, the MLP model exhibits that it shows poor performance over the LR and NB-RT model. However, it shows better classification over the RBF and DT by achieving the Sensitivity value of 79.72. In the same way, the compared RBF model shows poor classifier results with the lower Sensitivity value of 78.58. At the end, the DT model exhibits worse classifier performance with the highest Sensitivity value of 76.75. These values clarified that the NB-RT model attained maximum classifier results interms of sensitivity.

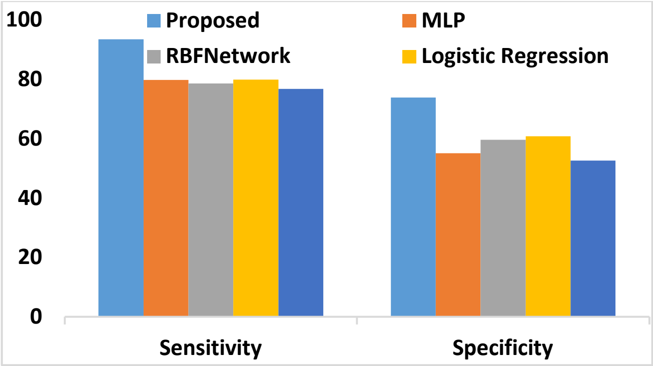


Fig. 4.Classifier results analysis interms of Sensitivity and Specificity

By means of specificity, the presented NB-RT model shows maximum classification by attaining highest Specificity value of 73.80 respectively. At the same time, it is noticed that the LR shows near optimal results obtained by the higher Specificity values of 60.74. In addition, the RBF model exhibits that it shows poor performance over the LR and NB-RT model. However, it shows better classification over the MLP and DT by achieving the Specificity value of 59.55. In the same way, the compared MLP model shows poor classifier results with the lower Specificity value of 55. At the end, the DT model exhibits worse classifier performance with the highest Specificity value of 52.61. These values clarify that the NB-RT model attained maximum classifier results interms of specificity.

Fig. 5displays the investigation of the results interms of accuracy.Under the evaluation parameter of accuracy, the presented NB-RT model shows maximum classification by attaining highest Accuracy value of 90.97. On the other hand, it is noticed that the LR shows near optimal results obtained by the higher Accuracy values of 75.20. In addition, the RBF model exhibits that it shows poor performance over the LR and NB-RT model. However, it shows better classification over the MLP and DT by achieving the Accuracy value of 74.30. In the same way, the compared MLP model reveals poor classifier results with the lower Accuracy value of 72.80. At the end, the DT model exhibits worse classifier performance with the highest Accuracy value of 71.20. These values clarify that the NB-RT model attained maximum classifier results in terms of accuracy.

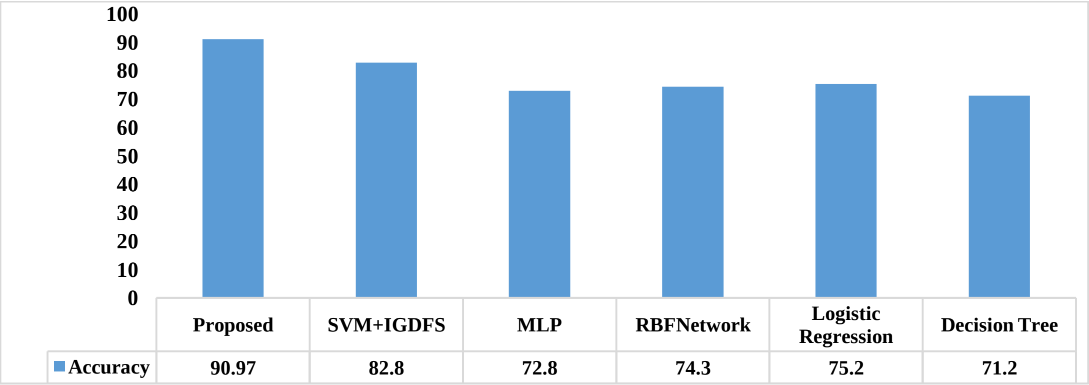


Fig. 5. Classifier results analysis interms of accuracy

Fig. 6 depicts the classifier results interms of F-score and kappa value.With respect to the evaluation parameter of F-score, the presented NB-RT model shows maximum classification by attaining highest F-score value of 94.77. In line with, it is noticed that the LR shows near optimal results obtained by the higher F-score values of 82.99. In addition, the RBF model exhibits that it shows poor performance over the LR and NB-RT model. However, it shows better classification over the MLP and DT by achieving the F-score value of 82.58. In the same way, the compared MLP model depicts poor classifier results with the lower F-score value of 80.84. At the end, the DT model exhibits worse classifier performance with the highest F-score value of 80.41. These values clarify that the NB-RT model attained maximum classifier results interms of F-score.

Under the evaluation parameter of kappa, the presented NB-RT model shows maximum classification by attaining highest Kappa value of 61.85. Next, it is noticed that the LR shows near optimal results obtained by the higher Kappa values of 37.5. In addition, the RBF model exhibits that it shows poor performance over the LR and NB-RT model. However, it shows better classification over the MLP and DT by achieving the Kappa value of 34.10. In the same way, the compared MLP model shows poor classifier results with the lower Kappa value of 33.98. At the end, the DT model exhibits worse classifier performance with the highest Kappa value of 26.93. These values clarify that the NB-RT model attained maximum classifier results interms of kappa.

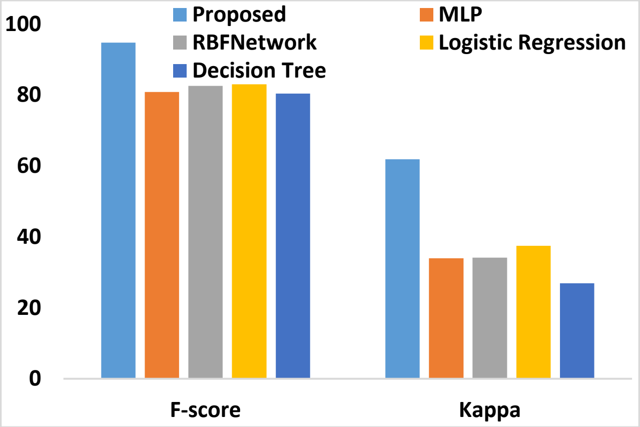


Fig. 6. Classifier results analysis interms of F-score and kappa value

From the above mentioned results and discussion, it is evident that the presented model shows extraordinary classifier results on the applied dataset over the compared methods.

IV. CONCLUSION

Presently, many classifier approaches have been presented by the use of diverse information for FCS. Although different works have been made by exhibiting the goodness of the FNN in distinct models, few difficulties are present in the designing and model applications. The nature of financial data contains unwanted, missing and repetitive instances. To resolve this issue, OD based classification model for FCS is developed. The presented FCS model involves two main phases namely NBTree based MIR model and RT model for data classification. The presented NB-RT model is tested against the German Credit dataset under different validation parameters. The extensive experiments exhibited that a maximum classification accuracy of 90.3% is achieved by the proposed NB-RT model.

REFERENCES

1. Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S.K., 2018. Financial crisis prediction model using ant colony optimization. International Journal of Information Management. https://doi.org/10.1016/j.ijinfomgt.2018.12.001
2. Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S.K., 2018. Intelligent hybrid model for financial crisis prediction using machine learning techniques. Information Systems and e-Business Management, pp.1-29.
3. Reddy, K.N., Ravi, V., 2013. Differential evolution trained kernel principal component WNN and kernel binary quantile regression: application to banking. Knowl. Based Syst. 39, 45–56.
4. Sharma, N., Arun, N., Ravi, V., 2013. An ant colony optimisation and Nelder-Mead simplex hybrid algorithm for training neural networks: an application to bankruptcy prediction in banks. Int. J. Inf. Decis. Sci. 5 (2), 188–203.
5. Paramjeet, Ravi, V., 2011. Bacterial foraging trained wavelet neural networks: application to bankruptcy prediction in banks. Int. J. Data Anal. Tech. Strateg. 3 (3), 261–280.
6. Naveen, N., et al., 2010. Differential evolution trained radial basis function network: application to bankruptcy prediction in banks. Int. J. BioInspir. Comput. 2 (3–4), 222–232.
7. Chauhan, N., Ravi, V., Chandra, D.K., 2009. Differential evolution trained wavelet neural networks: application to bankruptcy prediction in banks. Expert Syst. Appl. 36 (4), 7659–7665.
8. Ravi, V. and Pramodh, C., 2008. Threshold accepting trained principal component neural network and feature subset selection: Application to bankruptcy prediction in banks. Applied Soft Computing, 8(4), pp.1539-1548.
9. Ravisankar, P. and Ravi, V., 2009, December. Failure prediction of banks using threshold accepting trained kernel principal component neural network. In 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC) (pp. 7-12). IEEE.
10. Vasu, Madireddi, and Vadlamani Ravi. "Bankruptcy prediction in banks by principal component analysis threshold accepting trained wavelet neural network hybrid." In Proceedings of the International Conference on Data Mining (DMIN), p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2011.
11. Tsai, C.-F., Wu, J.-W., 2008. Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Syst. Appl. 34 (4), 2639–2649.
12. Atiya, A.F., 2001b. Bankruptcy prediction for credit risk using neural networks: a survey and new results. IEEE Trans. Neural Netw. 12 (4), 929–935.
13. Zhang, G., et al., 1999. Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. Eur. J. Oper. Res. 116 (1), 16–32.
14. Chauhan, N., Ravi, V. and Chandra, D.K., 2009. Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. Expert Systems with Applications, 36(4), pp.7659-7665.
15. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)