**IAES International Journal of Artificial Intelligence (IJ-AI)**
Vol. 8, No. 2, June 2019, pp. 168~174
ISSN: 2252-8938, DOI: 10.11591/ijai.v8.i2.pp168-174 

# Prediction of bankruptcy using big data analytic based on fuzzy c-means algorithm

**Arup Guha, N. Veeranjaneyulu**
Vignan's Foundation for Science, Technology and Research, Vadlamudi, India

**Article Info**

***Article history:***

Received Feb 18, 2019
Revised Apr 14, 2019
Accepted May 16, 2019

***Keywords:***

Artificial neural network
Cluster-based sampling
Fuzzy c means clustering
Genetic algorithm
Machine learning
Under-sampling technique

***Corresponding Author:***

Arup Guha,
Vignan's foundation for Science,
Technology and Research,
Vadlamudi, India.
Email: guha.arup@gmail.com

**ABSTRACT**

This paper has suggested an optimization approach of the cluster-based sampling using Fuzzy c means algorithm to the classifier in order to select the most appropriate instances of bankruptcy. This method was examined with the help of a clustering method and GA based artificial neural network in order to solve the existing data imbalance issue. The objective of this paper is to optimize the selected design model of GA-ANN by using Fuzzy C means algorithm to predict corporate bankruptcies by considering different financial ratios of companies across several industries within the period from 1994 to 2014. Effectiveness of this method was proved by comparing its accuracy rate with the results of existing method. From the performance result the accuracy rate of this method was found to be 78.2% and misclassification rate to be 0.2178.



## 1. INTRODUCTION

Increasing amount of data has led to the evolution of data science and its application to solve complex classification issues in a set of data to take important managerial decisions [1]. This paper is based on an optimization approach in ANN (artificial neural network) using the concept of Fuzzy clustering which is a type of sampling technique for extracting appropriate information from the random data set [2]. Clustering is one of the effective form of data mining techniques that are widely used for performing descriptive learning technique in analytics for predicting the corporate bankruptcy [3]. This technique is based on the determination of similar groups with identical features among a huge random data set. This method is a popularly used sampling technique in the case of imbalance data within the set of random data because it is very difficult to identify patterns among data comprising of odd data values, either very high or very low [4]. The method of handling such imbalance set of data is important prior to model development because if the difference in data size is too large or too small, then the cases of bankruptcy are ignored in the analysis. The basic methods that were considered in this bankruptcy prediction were based on applying undersampling technique to the majority group and over-sampling to the minority class.

The paper is based on application of Genetic algorithm (GA) and its combination with Artificial Neural Network (ANN) i.e GA-ANN modelling technique [5]. Conducting classification tasks using unbalanced data usually deteriorates the classification performance. If the difference of the data size between the two categories is greater, most of the data is strongly classified as the majority class to decrease the overall misclassification [4]. Therefore, handling unbalanced data may be a crucial procedure in model development. This remains to be a major drawback in the classification/prediction techniques. The above drawback will definitely have an impact on classification performance. The performance metrics, like the AUROC, the AR, or the H-measure had no definite criteria to produce evidence for evaluating the excellence

of the model performance. Refining the data will add to the performance improvement of ANN as it can keep a check on the computation time and also reduce spending extra computing resources on training ANNs [5]. Optimizing the required data will aid in providing improved classification accuracy and thereby enhancing the results of prediction.

The proposed method is applied to the problem of bankruptcy predictions using the financial data that were collected in order to focus on the proportion of the small and medium-scale bankruptcy firms. The intention of the study is to clarify and investigate how a machine learning technique can be exploited within the field of economics. More specifically, aim of the research is to refine how the machine learning strategies could be harnessed to predict corporate bankruptcies. We intent to apply an approach for selecting the optimal training data set and found a proper connection weight to learn the ANN model where we can employ multi-modal GA using Fuzzy C means algorithm to find multiple solutions on the cut-off values of every cluster. This way by employing clustering and optimal selection approach the neural networks will be more improved because the feature selection method to identify the most effective features for the classifier will enhance the accuracy of their prediction of corporate bankruptcy [4]. The remaining section of the paper is organized in the following way. Section 2 describes the existing technique of sampling and how sampling technique has been used by many researchers over the period of time inorder to solve complex issues of imbalance data management. Section 3 mentioned the proposed Fuzzy cluster-based technique of solving bankruptcy problem in a more specific manner. Section 4 has presented the outcome of the proposed technique in the form of their implementation and experimental results. Section 5 briefs about the conclusion of the result considering the findings of the proposed algorithm with research gaps and limitations.

## 2. LITERATURE REVIEW

Previous studies in solving data imbalance problem were referred at the two approaches of data level and algorithm level. The various concepts, techniques and systems are discussed in this section based on the existing research in the current scenario.

### 2.1. Undersampling technique

Undersampling technique is referred to classification in terms of reduction in the number of instances to balance the dataset consisting of majority class and the minority class. This is an efficient model in the case of dealing with large amount of data. This technique is helpful since the training time of the dataset is reduced. However, this method possess disadvantages in the form of risk of distorting the original distribution of the majority class. Moreover, in this technique the potential useful data is discarded. It is crucial to have a relevant dataset to improve the classification performance of a model by sampling data with similar properties. Random under sampling reduces the dataset by removing a randomly sampled dataset from the majority class as the simplest method. However, partial data can also be used in data modeling because this huge amount of data is sufficient for analysis in the era of big data [4].

A cluster-based undersampling approach was performed where the approach has first conducted clustering of all instances of data and divided them into several clusters [6]. Next, it selects the potential relevant number of instances that is belonging to the majority class from each cluster on the basis of proportional instances majority class to the number of instances of the minority class within the cluster. Clustering, ensemble and undersampling methods were performed in one study to solve the class imbalance problem [7]. They first conducted clustering using instances of the majority class and then constructed multiple training datasets comprising of sampled instances of the majority class from each cluster, preserving instances of the minority class. The evolutionary sampling method based on GA has been deployed in order to selectively remove instances from the majority class [8,9]. However, previous studies on evolutionary sampling using GA have showed performance results of time-consuming tasks in exploring optimal or near optimal solutions, since instances of the majority class has become strings for GA searching. Thus, in this study a cluster-based sampling supported by GA is suggested in order to handle the in- efficiency problem of the previous existing evolutionary sampling method.

### 2.2. Clustering of non-bankruptcy firm data based on majority class

A cluster based boosting algorithm was performed in one study using the Instance Hardness Threshold and CBoost algorithm with a robust framework in order to predict bankruptcy effectively of the financial imbalance dataset [3]. This proposed framework is also verified by the KBD (Korean bankruptcy dataset) having a small balancing ratio in both the testing and training phases. The proposed model experiment results has achieved 86.8% in AUC i.e. the area under ROC curve. It has also outperformed other existing methods for bankruptcy prediction using imbalance set of data. Machine learning methods were applied to the dataset collected from the manufacturing companies in Korea, in order to know their future

state with the help of certain financial measures [10]. Using several machine learning method result showed an accuracy of more than 95%. However, this study has some limitation also in the form of dimensional issue.

### 2.3. Under sampling technique based on genetic algorithms (GA)

A re-sampling approach is performed in a study in order to solve the unbalanced data sets [5]. In this approach, both the oversampling and under sampling method are combined with the help of genetic algorithm (GA). The application of genetic algorithm is based on a set of determined criteria and the unbalance rate. This approach has been tested on literature as well as industrial datasets and a desired improvement on the classification performance has been observed [11].

An under sampling approach and GA-ANN model has been approached in a study to improve the existing traditional approach of classification, which were usually costly and slow [5]. The undersampling approach is based on K-means cluster distribution in order to solve the problems of imbalance set of data. This method is effective to enhance the rate of sampling and improved the final classification. At the same time, this method has lower time of processing. GA-ANN method used in their study uses the algorithm to optimize the bias and weight of the neural network and thereby resulted into better performance. To increasing the classification accuracy a new genetic algorithm was proposed based on over sampling in order to solve the class imbalance data sets [12]. It can create optimized minority class events to balance the training datasets. The experimental results on imbalanced datasets proved better performance over the previous sampling methods in terms of AUC and F-measure.

## 3. PROPOSED MODEL

The proposed cluster-based method is based on a clustering algorithm. In this study, the method is adopted using Fuzzy C clustering. Fuzzy c-means algorithm applies the concept of fuzzy logic where the objects of classifications are allowed for more than one cluster. This type of classification makes high clarity sense since all the clusters are well separated. In this technique, value are assigned to all the wights. Repetition is done until the centroid is computed for each of the cluster with the help of fuzzy partition. This concept is related with the development of k-means algorithm for the sensor network. Using the fuzzy c-means algorithm the partitioning of data is possible by the nodes into different measure-dependent set of groups [13]. The role of this algorithm is to classify the data into separate groups. Each of the separated groups are then used to find out the centroids and based on these, high priority and low priority values are determined for the bankruptcy and non bankruptcy data. The purpose of this newly proposed model is to determine the risk of bankruptcy within these predicted range of gathered data, considering 12 set of attributes. In our proposed implementation we are using enhanced GANN based multimodal GA based neural network.

- Constant capital or fixed assets.
- Current assets, inventory and receivables or short-term liabilities
- (Receivables * 365) / total assets
- (Net profit + depreciation) / total assets
- Total sales / total assets
- Short-term liabilities / total assets
- Working capital / total assets
- Working capital / sales
- (Current liabilities * 365) / cost of products sold
- (Current assets -inventory - receivables) / long-term liabilities
- (Inventory * 365) / sales
- Net profit/inventory

The step by step process of the proposed model is shown in the figure.
Figure: Proposed model of bankruptcy prediction
The process of the model comprises of the following process:
Step 1: In the first step of the model design, we have gathered the financial data of companies across several industries in India along with their different financial ratios within the period 1994 to 2014. Big Data related to bankruptcy is considered. These set of bankruptcy and non-bankruptcy data are being stored in merged data_10X.csv. The data is then preprocessed to clean noise data, null data and missing data and then stored in transformed_new data.csv by creating a specific path of preprocessed data. The Figure 1 shows clustered data along with their centroids, using Fuzzy c means clustering.

Figure 1. Fuzzy c-means clustering

Step 2: Data gathered is preprocessed using undergone fuzzy c means algorithm and followed by data filtering. With the help of this data, a 12*12 correlation matrix is formed considering each of the attributes. Then the matrix has been arranged considering their correlation heatmap. The Figure 2 shows the correlation heatmap. With this matrix, maximum priority can be determined of each attribute values with the help of correlation matrix.


Figure 2. Correlation matrices with heatmap

Step 3: These set of attribute clustered data is then analysed with the help of histogram in order to predict bankruptcy and no bankruptcy data, as shown in the figure. Hadoop map reduce algorithm has been applied to these preprocessed data.

Step 4: Bankruptcy and non-bankruptcy status of data is found with the first attribute i.e constant capital or fixed assets. Likewise we have proceeded with each attribute. The matrices were determined along with heat map that are classified colourwise with attributes range as shown in the Figure 3.

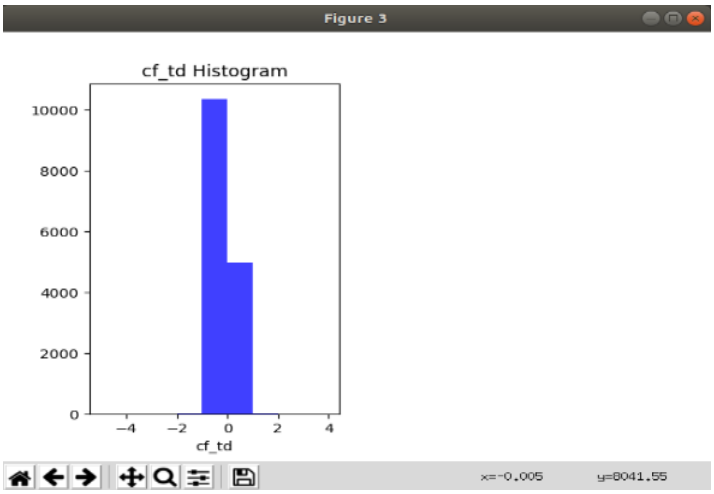


Figure 3. Histogram analysis of processed data

Agglomerative hierarchical Cluster technique has been employed in this case to improve the efficiency of the bankruptcy. After performing clustering on the extracted attributes, the *cluster feature vector* is applied to modify the classifiers for predicting bankruptcy from the data.

Step 5: The preprocessed data and the clustered data is stored into the transformed_new data.csv. The file is created automatically and renamed as data.csv, which is our main data. This main data is now separated into testing data and training data for the prediction of bankruptcy by considering them with the set of 12 attribute. The classification is done with classifier support vector machine, logistic regression and GA-ANN in order to compare.

Step 6: Before classification of the data is done, the classifier is trained in order to predict the exact bankruptcy. The prediction results for bankruptcy results are enhanced by employing multi modal GA based neural network. Correlation matrix will calculate the maximum values of attributes on the basis of mapping technique. After that, we need to give this data to the classifier, shown in the Figure 4 Correlation matrices of bankruptcy data and non bankruptcy data with the status of ID.


Figure 4. Correlation matrices of bankruptcy data and non bankruptcy data with the status of ID (0 and 1)

Step 7: At the end, the confusion matrices are calculated based on TP FP TN FN, to analyze the performance of GA-ANN classifier. The matrix will show the bankruptcy and non bankruptcy data prediction capacity of the classifier along with the misclassification rate.

TP means bankruptcy data was classified as bankruptcy and non bankruptcy data was classified as bankruptcy. FP means bankruptcy data was classified as non bankruptcy, FN means non bankruptcy data was classified as bankruptcy. TN means non bankruptcy data was classified as bankruptcy and bankruptcy data was classified as non bankruptcy. Once the proposed scheme is designed, the performance of the method will

be evaluated based on accuracy, precision, specificity and sensitivity, shown in the Figure 5.
Step 8: The final comparison has been done with the existing methods to know the effectiveness of the method.

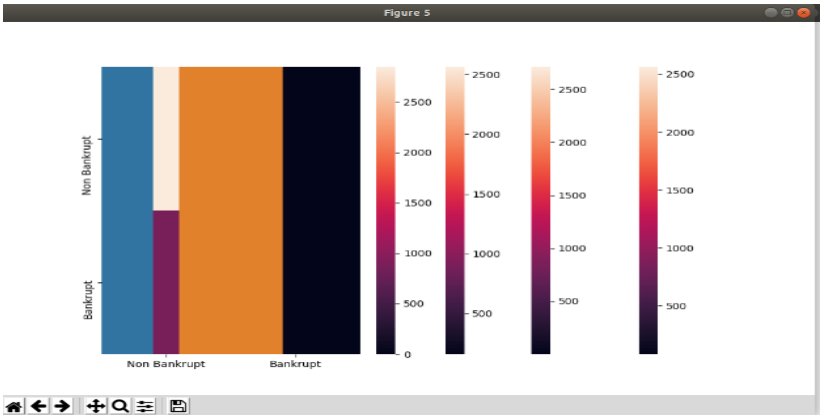


Figure 5. Performance evaluation matrix

## 4. EXPERIMENTS AND RESULTS

### 4.1. Research data and experiments

The dataset comprise of financial ratios of several small and medium scale companies from 1994-2014. The bankruptcy and non-bankruptcy data status are shown in the Figure 6. The number of non-audited companies are found comparatively higher that the total firms. The dataset was split into two subsets by considering 80% of the data for training dataset which is used to develop undersampling method for data class balancing and 20% for a validation dataset, which is arranged w.r.t the training data distribution. Two stage selection process of the input variable has been applied based on the previous method [1,3]. The chosen final variables were based on the *variant test* and these variable were used for the credit evaluation of the selected companies. The model is implemented using tools python 3.6 and Anaconda navigator.

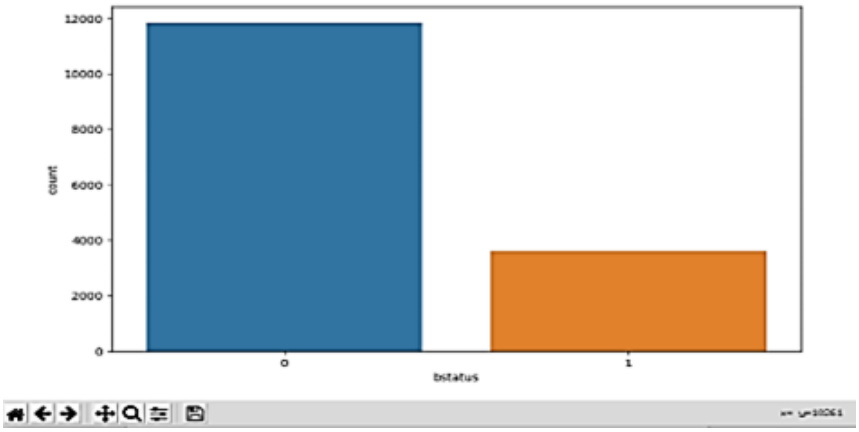


Figure 6. Status of bankruptcy and nonbankruptcy

### 4.2. Result and analysis

Effectiveness of the cluster-based GA-ANN undersampling method using Fuzzy C means algorithm applied to the classifier was being investigated for the bankruptcy prediction application. Here, we have set GA to search the cut-off for each cluster that represents the minimum distance of the clusters from the centroid. The optimization techniques are applied using GA-ANN, that has led to accurate prediction in this feature matrix. In the classification model, the applied classification algorithms used were Genetic Algorithm based Artificial Neural Networks, logistic Regression, Support Vector Machines and Decision

Trees to predict bankruptcy. Tested Genetic Algorithm based Artificial Neural Networks were found accuracy rate of 78.21% with comparison to existing method accuracy rate and showed misclassification rate 0.2178. Effectiveness of this method was proved by comparing its accuracy rate with the results of existing method. Thus, this method has proved effective in the handling of such imbalance dataset prior to model development, shown in the Figure 7.

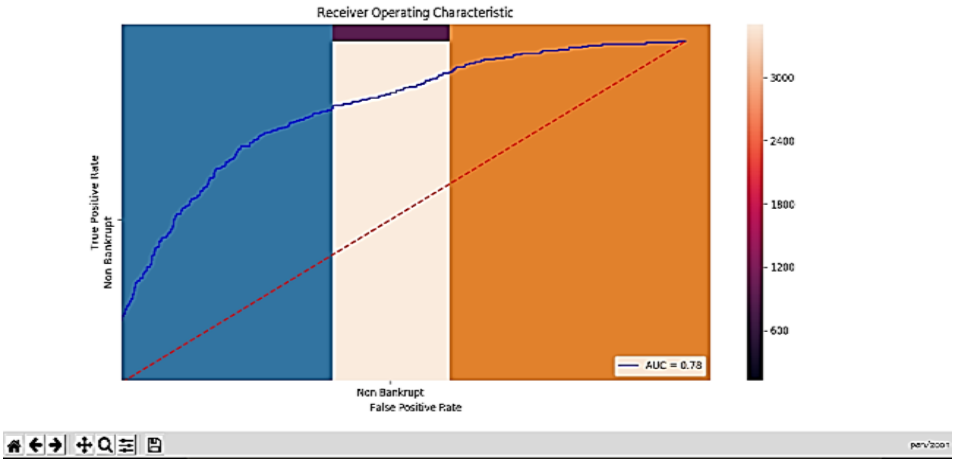


Figure 7. Comparison of model for accuracy rate

## 5. CONCLUSION

This study verified the effectiveness of the proposed approach of cluster-based under-sampling using Fuzzy C means algorithm in order to optimize GA-ANN for effective prediction of bankruptcy. In this the data is structured by classifying them using clustering technique and performing simultaneous optimization for the ANN model. This method has led to the effectiveness of the classifier and decreasing the data imbalance rate at the same time. The experimental result showed an accuracy of 78.2% as compared to the existing methods.

## REFERENCES

[1] Tambe, P. (2014). Big data investment, skills, and firm value. Management Science, 60 (6), 1452-1469.
[2] Kim, K. J., & Ahn, H. (2012). A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. Computers & Operations Research, 39 (8), 1800-1811
[3] Le, T., Le Son, H., Vo, M., Lee, M., & Baik, S. (2018). A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset. Symmetry, 10(7), 250.
[4] Kim, H. J., Jo, N. O., & Shin, K. S. (2016). Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. Expert Systems with Applications, 59, 226-234.
[5] Song, A., & Xu, Q. (2018). Imbalanced Data Classification Based on MBCDK-means Undersampling and GA-ANN. In International Conference on Artificial Neural Networks (pp. 349-358). Springer, Cham.
[6] Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36 (3), 5718-5727
[7] Kang, P., Cho, S., & MacLachlan, D. L. (2012). Improved response modeling based on clustering, under-sampling, and ensemble. Expert System with Applications, 39 (8), 6738-6753.
[8] Khoshgoftaar, T. M., Seliya, N., & Drown, D. J. (2010). Evolutionary data analysis for the class imbalance problem. Intelligent Data Analysis, 14 (1), 69-88
[9] García, S., & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary Computation, 17 (3), 275-306.
[10] Chow, J. C. (2018). Analysis of Financial Credit Risk Using Machine Learning. arXiv preprint arXiv:1802.05326.
[11] Vannucci, M., & Colla, V. (2017). Genetic Algorithms Based Resampling for the Classification of Unbalanced Datasets. In *International Conference on Intelligent Decision Technologies* (pp. 23-32). Springer, Cham.
[12] Dong, S., & Wu, Y. (2018, July). A genetic algorithm-based approach for class-imbalanced learning. In *Third International Workshop on Pattern Recognition* (Vol. 10828, p. 108281D). International Society for Optics and Photonics.
[13] Qin, J., Fu, W., Gao, H., & Zheng, W. X. (2017). Distributed $ k $-means algorithm and fuzzy $ c $-means algorithm for sensor networks based on multi agent consensus theory. *IEEE transactions on cybernetics*, 47(3), 772-783.