Journal of Theoretical and Applied Information Technology
28th February 2018. Vol.96. No 4
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FUZZY MULTI-CRITERIA RANDOM SEED AND CUTOFF POINT APPROACH FOR CREDIT RISK ASSESSMENT

[1] BEULAH JEBA JAYA Y., [2] DR. J. JEBAMALAR TAMILSELVI

[1]Research Scholar, Bharathiar University, Coimbatore, Tamilnadu, India

[2]Professor, Department of Computer Applications, Jaya Engineering College, Chennai, Tamilnadu, India

E-mail: [1]beulahprince@yahoo.co.in, [2]jjebamalar@gmail.com

ABSTRACT

Data mining classification techniques have been studied extensively for credit risk assessment. Existing techniques by default uses 0.5 as the cutoff irrespective of datasets and classifiers to predict the binary outcomes, thus limiting their classification performance on imbalanced group sizes of datasets. This paper addresses two key problems with the existing techniques and talks about the advantages of using Multiple Criteria Decision Making (MCDM) technique on multiple evaluation criteria. The first key problem is applying default cutoff irrespective of datasets and classifiers. The second one is utilizing single criteria for evaluating classification performance and predicting cutoff point. This research work identifies the best cutoff point with respect to datasets and classifiers and integrates MCDM under fuzzy environment in all data mining stages of evaluation to take better decisions on multiple criteria, selection of initial random seed in the clustering phase for better cluster quality and Best Seed Clustering combined Classification (BSCC hybrid algorithm) with selected features to improve classification performance. The integration of these techniques gives a better hand to improve cluster quality and classification performance score with respect to datasets and classifiers because the cutoff point varies from dataset to dataset and classifiers to classifiers. Experimental outcomes from applied credit dataset of UCI machine learning repository found to be competitive and the proposed BSCC hybrid algorithm increases the performance score on obtained cutoff point over non-hybrid approach with default cutoff.

Keywords: *Credit Risk, Classification, Clustering, Fuzzy MCDM, Cutoff Point, Random Seed*

1. INTRODUCTION

Credit risk occurs when the debtor fails to repay the loan which brings significant losses to the financial institutions. It exists in all business that generates income and inappropriate evaluation leads to failure [1]. Intelligent credit analysis technique is essential for any financial institutions to classify the borrowers from good and bad credits. Data mining classification techniques are used widely for financial risk assessment [2]. These data mining techniques use classifiers for prediction with default cutoff point (0.5). For imbalanced group sizes of datasets, the 0.5 default probability cutoff point limits the classification performance scores by a significant extent. Most of the researchers used default cutoff, irrespective of datasets for identifying classification accuracy [3]. But this threshold limits the classification performance to a greater extent when the dataset is imbalanced. The dataset is imbalanced if the class representations are not evenly distributed [4]. In this situation, choosing the cutoff point is imperative for imbalance datasets. An instance that is greater than or equal to the cutoff point is considered as good credits and the instances that are lesser than the cutoff point is considered as bad credits. For cutoff point selection, ROC curve has been used to distinguish between two classes [5]. Hong proposed a quality criterion true rate to determine the optimal threshold for highly imbalanced datasets [6]. Cost sensitive un-weighted accuracy was taken as the quality criteria to ascertain the optimal cutoff [7]. Classifiers and clusters performance cannot be judged by a single criteria, because each single criterion performs well on different classifiers and clusters [3][8]. Since classifiers and clusters performance need to examine multiple criteria, it is modeled as a MCDM problem in literatures. It is noticed in existing literatures, the classifiers which are used for prediction use default cutoff point irrespective of datasets and classifiers i.e. the default cutoff point is used for all classifiers and datasets. In this case, the prediction accuracy is greatly limited for imbalanced datasets. Hence, the best cutoff point for classifiers with respect to datasets and classifiers need to be identified to improve the prediction process. Moreover, the performance of these data mining classifiers varies with different evaluation metrics. Therefore, the main motivation of this research work is to effectively make the classifier to choose the cutoff point with respect to datasets and classifiers using MCDM approach. In addition, the performance evaluation of the credit risk prediction at every stage is based on multiple criteria. Thus, this research work is intended to find the cutoff point for credit risk classification by integrating the following techniques (1) MCDM approach under fuzzy environment in all the data mining evaluation stages. (2) Selection of initial random seed in the clustering phase using an efficient K-Means algorithm. (3) BSCC hybrid algorithm is used to determine performance scores at arbitrary cutoff points for the classifiers such as Logistic Regression (LR), Naïve Bayes (NB) and Random Forest (RF). (4) Determination of the best cutoff point using MCDM-FAHP-TOPSIS method and performance evaluation is done at the obtained cutoff point.

In the first stage, optimal features are selected from the credit risk dataset using the Consistent Ranking Feature Selection (CRFS) method based on Fuzzy MCDM proposed in our previous work [9]. In the subsequent stage, the selected optimal features are given as input to BSCC hybrid algorithm which links Fuzzy Initial Random Seed KMeans (FIRS-KMeans) clustering and Fuzzy Multiple criteria decision making Cutoff point Classification (FMCC) approach for predicting the cutoff point. The FIRS-KMeans is an addition of traditional and general KMeans algorithm, in which the appropriate initial seed value are selected using MCDM approach under fuzzy environment and given as input to KMeans algorithm for clustering. This approach helps for better cluster quality and to improve the classification accuracy. Using the results of FIRS-KMeans, the FMCC is used to ascertain the performance scores for arbitrary cutoff points.

The paper is organized as follows: Section 2 concisely assesses the existing approaches in the literature, Section 3 talks about the techniques applied in the proposed work, Section 4 explains the environment to experiment the proposed work for cutoff point detection in credit risks datasets, Section 5 discusses the performance evaluation of credit risk assessment and Section 6 discusses the conclusion of the entire work.

2. LITERATURE REVIEW

In the past, several supervised and unsupervised learning techniques are proposed for credit risk assessment. One of the category of supervised learning algorithm is classification and algorithms such as linear discriminant analysis, logistic regression, decision trees, support vector machines and Artificial neural network are widely applied for credit risk prediction [10]. NB classifier is competitive with other classifiers and widely used in credit scoring domains [11]. LR classifier can be used for predicting the consequence of binary outcome [12]. Logistic regression model is used in credit scoring models with 0.5 default cutoff point [13]. RF classifier is an effective tool used for prediction and gives competitive results with other ensemble of classifiers [14].

Similarly, unsupervised clustering algorithms such as K-Means, Self-Organizing map are also widely used [15]. K-Means algorithm is widely used simple clustering technique which depends on initial seed value [16]. The initial seed value of K-means is determined using Taguchi method for better cluster quality [17].

Some researchers have shown by combining classification and clustering techniques, the prediction process shows better results [18][19]. Classification accuracy is improved when clustering and classification methods are combined and better than single classification approach [20][21].

Although, the existing researches use a lot of classifiers for credit risk prediction, it applies 0.5 as default cutoff point for prediction irrespective of datasets and classifiers. When it comes to imbalanced datasets, the drawback with the existing models is that default cutoff point leads to lesser accuracy in risk prediction. Peng et al. determined accuracy of classifiers at default cutoff point irrespective of datasets and classifiers [3]. Some of the researchers applied only single criteria such as ROC curve to ascertain the cutoff for imbalanced datasets [22]. Soureshjani and Kimiagari evaluated best cutoff point in terms of ROC curve and minimization of overall error using logistic regression and neural network [23].

Kou et al. suggested that more than one single evaluation criteria are necessary for evaluating clustering algorithms since no algorithm performs better in all the evaluation criteria [8]. MCDM techniques have proved to be better for choosing the alternatives if multiple criteria are involved [3]. In recent times, MCDM techniques established greater attention from data mining classification; for example, TOPSIS one of the MCDM methods suggested for selecting the classification model on multiple performance metrics [24]. MCDM have found immense recognition in all areas of multi-criteria decision problems [25]. MCDM techniques are used to evaluate the important data mining classification algorithms [26].

Since cutoff point detection involves multiple performance measures, this problem can be solved using MCDM approaches. Since there is an uncertainty in global business markets, fuzzy techniques help to provide clear and reliable information. Many researchers have applied fuzzy techniques to select the alternatives from multiple criteria. For example, Fuzzy Analytic Hierarchy Process (FAHP) and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) MCDM methods are integrated to evaluate the financial parameters in a bank [27]. The FAHP method is one of the MCDM technique proposed by Van and Pedrycz for ranking the alternatives among multiple criteria [28]. FAHP method helps to make better decision among multiple criteria as it involves fuzzy set theory [29]. The TOPSIS method is one of the extensively used well-known MCDM technique and was first suggested by Hwang and Yoon [30][31]. Multiple criteria like Overall_accuracy, True Positive Rate (TPR), True Negative Rate (TNR), FMeasure and ROC-Area performance measures needed to detect the cutoff point are identified [3][32]. Overall accuracy and True Positive (TP) rate are commonly used performance metric for classification problems [33]. True Negative (TN) rate shows high probability of correct classification of negative cases [34]. F-measure is a popular performance measure for imbalanced dataset [35]. It is based on the measures of precision plus recall (True Positive) and Area under ROC is a useful graphing performance measure [36][37]. It is a compromise between true positive (recall) and false positive [38]. Additionally, RRSE is also added in this research work as a performance measure. Root Relative Squared Error (RRSE) is well known performance criteria for prediction [39]. Also a better predictor for forecasting [40].

Credit risk classification normally involves many stages to enhance the prediction process. One such stage is clustering which clusters the data using KMeans. Since KMeans algorithm initially selects the seed value randomly, it is necessary to select the best seed value for K-Means algorithm to improve the cluster quality. For an initial random seed selection in K-Means using FAHP method, cluster validity measures such as Adjusted Rand Index (ARI), Jaccard Coefficient (JC), Fowlkes-Mallows Index (FMI) and Sum of Squared Errors (SSE) are also identified for this research from the literatures [41].

Considering the reviewed literature which is described above, credit risk assessment need an approach to identify the best cutoff point with respect to datasets and classifiers. Also, the prediction process at every stage need to evaluate the performance based on multiple criteria. These research gaps are effectively fulfilled in this research work. From the reviewed literatures, it is also observed that, so far no attempts are made to use the effective MCDM techniques for determining initial random seed value in KMeans and cutoff point with respect to datasets and classifiers. Furthermore, all the existing credit risk prediction models applied MCDM technique in either of one stage (clustering or classification) whereas in the proposed method, fuzzy MCDM is applied in all the data mining evaluation stages i.e., feature selection, choosing best initial random seed for K-Means and cutoff point detection with respect to datasets and classifiers.

3. PROPOSED WORK

The data mining evaluation stages for Credit Risk Assessment (CRA) such as BSCC hybrid algorithm, selected performance evaluation criteria and MCDM methods used are briefly discussed in the following subsections. The system architecture diagram of FIRS-KMeans and FMCC for CRA is shown in Figure 1.

3.1 CRA through BSCC Hybrid Algorithm

Credit risk assessment helps the organization to identify whether to grant loans or not. In order to classify credit risks, a BSCC hybrid data mining approach is proposed which combines FIRS-KMeans clustering and FMCC. Classification results are greatly improved when hybrid data mining approach is used. The FIRS- KMeans select the initial random seed through fuzzy MCDM approaches such as FAHP method. With the selected random seed, the general KMeans algorithm is applied for optimal clustering.

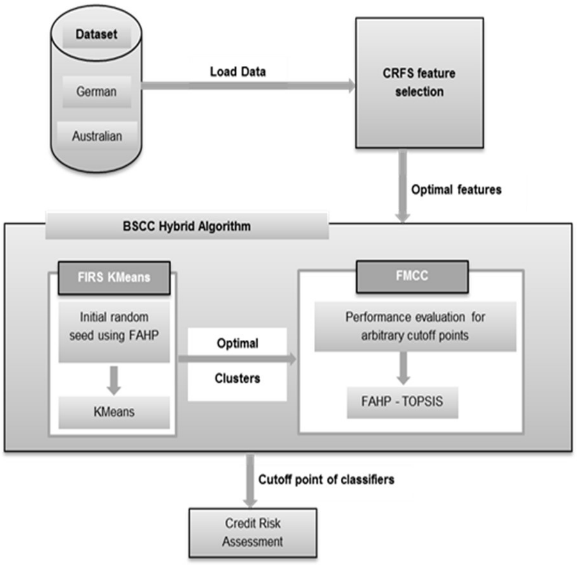

Figure 1: System Architecture Diagram of FIRS-KMeans and FMCC for CRA.

The main idea in KMeans is to ascertain the centroid coordinate or cluster center for each cluster. By selecting the appropriate initial random seed value, the general KMeans algorithm selects the better initial centroid coordinate which helps to improve better cluster quality and classification accuracy. Then, place the determined centroid coordinate isolated from each other. The subsequent step is to calculate the Euclidean distance of each data point based on cluster center and assemble the data points based on minimum Euclidean distance. Next, recalculate the centroid coordinate based on new grouping and repeat the previous steps until the data points do not move into groups. Then the clustered dataset is given as input to the FMCC classification to determine the performance scores for arbitrary cutoff points with respect to NB, LR and RF classifiers. NB data mining classifier is the most common machine learning algorithm which works on the concept of 'Bayes theorem'. It works as follows:

Given a training set 'ts' with set of attributes ts_1, ts_2, … ts_n, predefined class attribute PC= PC1, PC2, … PCj and new instance K for which classification need to be obtained. The most probable classification for the new instance K is determined using the following equation:

$$PC_{NB} = Max_{PC_j}\left\{Prob\ (PC_j) * \prod_{i=1}^{n} Prob\ (\, ts_i \mid PC_j\,)\right\}$$

where i ranges from 1 to 'n' attributes, j ranges from 1 to number of target class attributes Prob (ts_i / PC_j) is the individual probability ts_i on target class attributes PC_j.

LR classifier helps to discriminate between good and bad credits and works as follows:

Let 'Y' be the binary outcome variable indicating good or bad credit and 'prob' be the probability of Y equal to 1 given X i.e., prob = prob(y=1 | X). Let X=X1, X2.., Xn are the various predictor variables. The parameters of β0, β1. . . βn are chosen by maximizing the probability (Maximum Likelihood Estimation). Then the logistic regression model is generated using the logit transformation to predict probability and the equation is given below:

$$\text{logit}\ (\text{prob}(\text{Y} = 1 \mid \text{X}) = \log\left(\frac{\text{prob}}{1 - \text{prob}}\right) = \beta 0 + \beta 1 * \text{X1} + \ldots + \beta k * \text{Xn}$$

Log of odds can be translated to a probability using the following equation:

$$\text{prob}\ (\ \text{Y} = 1 \mid \text{X}) = \frac{\exp\text{onent}\big(\beta 0 + \beta 1 * \mathbf{X1} + \ldots + \beta k * \text{Xn}\big)}{1 + \text{exponent}\big(\beta 0 + \beta 1 * \mathbf{X1} + \ldots + \beta k * \text{Xn}\big)}$$

If prob > 0.5 then select good credit (class 1) otherwise bad credit (class 0).

RF data mining classifier is the well known machine learning algorithm which works well on classification techniques. It works as follows:

Two-third of the records in a dataset is selected by random replacement technique and designated as 'training set' to grow the tree. To branch or split a node in the tree, the attributes are selected at random (i.e., K) where K is the square root of the overall predictor attributes. For each generated tree, the rest of the records, i.e., one-third are used to calculate OutOfBag error percentage. Repeat this for all the generated trees to find the overall OutOfBag error percentage. For each tree, the RF classifier gives the number of votes for class attribute. Out of all trees, the algorithm chooses the classification with maximum votes.

3.2 Performance Evaluation Criteria

In this research work, the four commonly used cluster validity measures such as ARI, JC, FMI and SSE are identified as multiple criteria to select the appropriate initial random seed value for general KMeans clustering. The formula for calculating the cluster validity measures are given below:

Based on confusion matrix, the cluster validity measures ARI, JC, FMI and SSE are calculated as follows:

Let M=TP+FN, N=TP+FP, O=TP+TN+FP+FN, P= TP+FP+FN

1. ARI is a commonly applied cluster validation measure proposed by Hubert and Arabie to measure the agreement between two groups [42][43] and the equation is shown below:

$$\mathrm{ARI} = \frac{\mathrm{TP} - (\mathrm{M} * \mathrm{N})}{\mathrm{O}} \Big/ \frac{(\mathrm{M} + \mathrm{N})/2 - (\mathrm{M} * \mathrm{N})}{\mathrm{O}}$$

Higher ARI value, the higher is the quality of clustering.

2. JC is a simple and well known external cluster validity measure to compare how similar are the two groups [44] and it is defined as:

$$\mathrm{JC} = \frac{\mathrm{TP}}{\mathrm{P}}$$

Higher JC value results in higher similarity between the two groups.

3. FMI is a geometric mean of the measures of precision plus recall and can be used in flat clusterings [45] and it is defined as follows:

$$\mathrm{FMI} = \sqrt{(\mathrm{TP}/\mathrm{N}) * (\mathrm{TP}/\mathrm{M})}$$

Higher the FMI value, the greater is the similarity between the clusters.

4. SSE is another cluster validity measure which is the square of distance between individual point in a cluster Cj and the average of points in Cj and it is computed as follows [46]:

$$\mathrm{SSE} = \sum_{\mathrm{j=1}}^{\mathrm{K}} \sum_{\mathrm{x€Cj}} \Big(\mathrm{square}\Big(\mathrm{dist}\big(\mathrm{x(i)} - \mathrm{mean(j)}\big)\Big)\Big)$$

where K takes the value 2 (cluster of good and bad), x is the individual point in cluster Cj. Smaller SSE value results in better cluster quality.

The important performance measures needed in this research for credit risk classification are identified as overall_ accuracy, TPR, TNR, FMeasure, ROC-Area and RRSE which are described below:

Let TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

5. Overall_accuracy is the rate of correctly categorized good and bad credits and it is computed as:

$$\text{Overall accuracy} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$$

6. TPR is the rate of correctly categorized good credits and it is computed as:

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$$

7. TNR is the rate of correctly categorized bad credits and it is defined as:

$$\mathrm{TNR} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$$

8. FMeasure is is computed as:

$$\mathrm{FMeasure} = (\mathrm{Precision} * \mathrm{TP} * 2)/(\mathrm{Precision} + \mathrm{TP})$$

9. ROC-Area is a graph plotted as the rate of correctly categorized good credits in Y coordinates and the rate of incorrectly categorized good credits in X coordinates.

10. Root Relative Squared Error (RRSE) is added in this research work as a performance measure which is computed as follows:

$$\mathrm{RRSE} = \mathrm{SQRT}\left(\frac{\sum_{\mathrm{j=1}}^{\mathrm{n}} sqr\big(pred(j) - act(j)\big)}{\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{sqr}\big(\mathrm{act(j)} - \mathrm{mean(act)}\big)}\right)$$

where pred(j) is the predicted target values of the classifier and act(j) is the actual target values. Smaller the RRSE values better the performance classification.

3.3 MCDM Methods

The MCDM method FAHP combined TOPSIS for identifying the appropriate seed value and cutoff point for credit risk classification are discussed below:

3.3.1 MCDM-FAHP

In this research, the FAHP method is used to rank the alternatives (different initial random seed values say 5, 6, 7…) among multiple criteria such as ARI, JC, FMI, SSE cluster validity measures. The Triangular Fuzzy Number (TFN) is set up to convert the pairwise comparison matrices of multiple criteria with relative scores (formed by saaty scale of importance) [47]. The steps of FAHP method work as follows:

1. The TFN is set up with the triplet such as (r,s,t) where r is the lower bound, s is the average value, t is the upper bound and represents a

fuzzy number. Based on the parameter b, TFN is defined as (s-df, s, s+df) and the inverse TFN as (1/(s+df), s, 1/(s-df)) where df is the degree of fuzziness which is considered as 1 in this research. The pairwise comparison matrices formed for each criteria are converted to TFN to form a fuzzy comparison matrix.

2. Determine Fuzzy synthetic extent [48] by fuzzy addition operation of the total row for TFN and the inverse TFN triplets.

$$\mathrm{Fuzzyvector}_{\mathrm{TFN}} = \sum_{j=1}^{k} r_j \sum_{j=1}^{k} s_j \sum_{j=1}^{k} t_j$$

$$\mathrm{Fuzzyvector}_{\mathrm{InverseTFN}} = \frac{1}{\sum_{j=1}^{k} r_j}, \frac{1}{\sum_{j=1}^{k} s_j}, \frac{1}{\sum_{j=1}^{k} t_j}$$

$$\text{Fuzzy Synthetic extent} = \prod_{j=1}^{n} \mathrm{Fuzzyvector}_{\mathrm{TFN}} * \mathrm{Fuzzyvector}_{\mathrm{InverseTFN}}$$

where k and n represents the overall rows and columns in the matrix.

3. Calculate the set of weights for the criteria. Let K1=(r1,s1,t1) and K2=(r2,s2,t2) are two fuzzy numbers. the degree of possibility (d) is computed as:

$$d(k2 \geq k1) = \left\{ \begin{array}{c} 1 \text{ if } s_2 \geq s_1 \\ 0 \text{ if } r_1 \geq r_2 \\ \frac{(r_1 - t_2)}{(s_1 - t_2) - (s_1 - r_1)} \text{ otherwise} \end{array} \right\}$$

Weight vector is formed by choosing the minimum of (d) K≥Ki .

4. The obtained weights are normalized to convert to a non-fuzzy number.

5. Rank the alternatives based on the normalized weights.

3.3.2 MCDM-TOPSIS

TOPSIS method is used to rank the alternatives which are shortest to the best solution and farthest from the negative best solution. The method works as follows:

1. Construct a decision matrix with alternatives as rows and criteria as columns and X_{ij} is the score for the alternatives versus criteria.
   where i ranges from 1 to 'm' alternatives, j ranges from 1 to 'n' criteria.

2. Compute a Normalized_Decision_Matrix NDM_{ij}: $NDM_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}$
   where i ranges from 1 to 'm' alternatives, j ranges from 1 to 'n' criteria.

3. Construct weighted NDM $WNDM_{ij}$ by the following equation:
   $WNDM_{ij} = W_i * NDM_{ij}$
   where the weight of criteria W_i is determined by FAHP method.

4. Determine the positive best (PB*) and negative best(NB') solutions:
   PB*=Maximum($WNDM_{ij}$) and

   NB'= Minimum($WNDM_{ij}$).

5. Determine the separation measures from PB* (SP_i*)

$$SP_i^* = \sqrt{\sum_{j=1}^{n} (PB^* - WN_{ij})^2}$$

   where i ranges from 1 to 'm' alternatives, j ranges from 1 to 'n' criteria.

6. Determine the separation measures from NB' (SNi')

$$SN_i' = \sqrt{\sum_{j=1}^{n} (NB' - WN_{ij})^2}$$

   where i ranges from 1 to 'm' alternatives, j ranges from 1 to 'n' criteria.

7. Find closeness to the best solution (CB_i^*)

$$CB_i^* = \frac{SN_i'}{SP_i^* + SN_i'}$$

8. Greatest score of CB_i^* is considered as the best alternative for prediction.

4. EXPERIMENTAL SETUP

The proposed method for classification of credit risk is implemented using MATLAB 7.9 for initial random seed selection by FAHP method and cutoff point detection by FAHP-TOPSIS method. WEKA 3.7 [49] is used to cluster and classify the credit risks based on MATLAB output.

4.1 Dataset Used

UCI machine learning repository based two imbalanced datasets (German and Australian) of credit approval are used in this research [50]. The

German credit consists of 1000 records of 20 /1 features / class feature. 700 good credits and 300 bad credits. 700/300 which is highly imbalanced. The Australian credit approval consists of 690 records of 14 / 1 features / class feature with 307 good credits and 383 bad credits. 307/383 which is moderately imbalanced.

4.2 CRFS

An existing feature selection algorithm yields an inconsistent reduced set of different single evaluation criteria. In order to achieve consistency in determining the features with different feature selection algorithms, CRFS adopts three multiple assessment criteria such as distance, dependency and information measures instead of single evaluation criteria. In addition, multi-criteria FAHP ranking is applied to determine the optimal feature set which is part of our previous research work. Implemented step generates the following optimal features as an outcome for German credit dataset: checking-status, duration, credit-history, purpose, credit-amount, savings-status, employment, instalment_commitment, personal status, property-magnitude, age, housing, own-telephone. Similarly for the Australian credit approval dataset, nine features are selected as optimal features and implemented in the proposed approach.

4.3 BSCC Hybrid Algorithm

BSCC hybrid algorithm is applied to optimal features from previous step to determine the cutoff point with respect to dataset and classifiers for credit risk assessment. The BSCC hybrid algorithm combines the FIRS-KMeans clustering and FMCC.

4.3.1 FIRS-KMeans

FIRS-KMeans clustering determines the appropriate initial random seed which chooses better initial centroid coordinates for KMeans algorithm yielding better cluster quality by the effective MCDM approach under fuzzy environment such as FAHP. The weights for each cluster validity measure and overall weights are calculated using FAHP method. The results obtained from FIRS-KMeans for different arbitrary seed values for Australian credit approval dataset are depicted in Table 1.

Table 1. Ranks for FIRS-KMeans

Cluster Validity Measure with weights / Seed Values	ARI	JC	FHM	SSE	Overall Weight	Overall Rank
	0.168776	0.168776	0.168776	0.493671		
5	**0.586**	**0.635**	**0.777**	**1429.16**	**0.135**	**1**
6	0.644	0.657	0.794	1518.74	0.126	5
7	0.595	0.633	0.775	1436.98	0.070	6
8	0.586	0.634	0.776	1469.62	0.054	10
9	0.586	0.635	0.777	1429.16	0.135	2
10	0.595	0.633	0.775	1436.98	0.070	7
11	0.595	0.633	0.775	1436.98	0.070	8
12	0.586	0.635	0.777	1429.16	0.135	3
13	0.586	0.635	0.777	1429.16	0.135	4
14	0.595	0.633	0.775	1436.98	0.070	9

Highest rank with minimum seed value (i.e., 5) from Table 1 is selected as the best initial random seed for KMeans clustering, which shows improved classification performance when compared to default initial random seed (i.e.,10) in Australian credit approval dataset and the assessment results with LR classifier are shown in Table 2. But in the German credit dataset, the classification performance is better with default Random seed (i.e., 10) and the FIRS-KMeans algorithm also chooses the same.

Table 2. Comparison Results of FIRS-KMeans

Seed	Overall_ Accuracy	TPR	TNR	FMeasure	ROC-Area	RRSE
Default (10)	0.983	0.984	0.982	0.984	0.991	0.265
Best (5)	0.993	0.994	0.992	0.993	0.994	0.171

Using the optimal features and the best initial random seed, FIRS-KMeans algorithm divides the dataset into optimal clusters.

4.3.2 FMCC

With the best seed clustered dataset, FMCC approach is used for computing evaluation scores based on ten fold cross validation with respect to arbitrary choices of cutoff versus six key classifying evaluation measures such as overall_accuracy, TPR, TNR, FMeasure, ROC-Area and RRSE and three classifiers such as NB, LR and RF. Based on average_ranking, the evaluation scores are ranked to pick up the cutoff point by FAHP-TOPSIS approach. The results of the computation for this phase are tabulated in Table 3 for German credit dataset. Greatest FAHP-TOPSIS scores are considered as the cutoff point with respect to dataset and classifiers.

5. PERFORMANCE ANALYSIS

The proposed approach helps to ascertain the cutoff point with respect to datasets and classifiers.

From Table 3 some important conclusions can be drawn.

1. In previous studies, the cutoff point for classifiers were identified by single criteria such as ROC area or TP rate. But from Table 3, it has been observed that the ROC area gives the same scores for arbitrary choices of cutoff. Thus, only with ROC area, it is very difficult to judge the cutoff point for better classification performance. To overcome this, MCDM technique is applied to deal with multiple criteria for cutoff point detection.
2. To determine cutoff point, the balanced TP rate and TN rate scores help to achieve the best classification performance [51]. From Table 3, it is clear that the proposed approach determines cutoff point for balanced TPR and TNR (NB classifier, TPR=0.850, TNR=0.841 at cutoff point (0.3); RF: TPR=0.905 and TNR=0.930 at 0.4) when compared to other arbitrary choices of cutoff.
3. With respect to datasets and classifiers, some more comments can be drawn:
   - German credit dataset: For NB and LR, the cutoff point is identified at 0.3 whereas for RF, it is identified at 0.4 as shown in Figure 2. At this cutoff, the classification performance scores for all the criteria are high and balanced when compared to other choices of cutoff.
   - Australian credit approval dataset: For NB, the cutoff point is identified at 0.4 whereas for LR and RF, it is identified at default cutoff as shown in Figure 3.

From the above discussions, it is observed that the cutoff point varies with respect to dataset and classifier for better classification results. It is also noticed, it is difficult to evaluate the performance of the prediction process only with single criteria.

Table 3 Results for FMCC of three classifiers with FAHP-TOPSIS scores

Selected Features	Classifier	Arbitrary choices of cutoff	Overall_ Accuracy	TPR	TNR	FMeasure	ROC-Area	RRSE	FAHP-TOPSIS scores
14	Naïve Bayes (NB)	0.1	0.773	0.649	0.984	0.762	0.927	0.7569	0.44
		0.2	0.841	0.786	0.935	0.813	0.927	0.6938	0.59
		0.3	**0.847**	**0.851**	**0.841**	**0.803**	**0.927**	**0.6792**	**0.63**
		0.4	0.840	0.897	0.743	0.775	0.927	0.6854	0.60
		0.5	0.830	0.914	0.686	0.749	0.927	0.7014	0.56
		0.6	0.818	0.935	0.619	0.716	0.927	0.7218	0.52
		0.7	0.798	0.952	0.535	0.662	0.927	0.7445	0.48
		0.8	0.780	0.962	0.470	0.613	0.927	0.7683	0.44
		0.9	0.754	0.973	0.381	0.534	0.927	0.7966	0.41
		0.95	0.737	0.983	0.319	0.473	0.927	0.8157	0.40
14	Logistic Regression (LR)	0.1	0.987	0.989	0.984	0.982	0.998	0.2213	0.71
		0.2	0.986	0.989	0.981	0.981	0.998	0.2268	0.59
		0.3	**0.987**	**0.992**	**0.978**	**0.982**	**0.998**	**0.2291**	**0.82**
		0.4	0.987	0.992	0.978	0.982	0.998	0.2307	0.81
		0.5	0.986	0.992	0.976	0.981	0.998	0.2325	0.53
		0.6	0.986	0.992	0.976	0.981	0.998	0.2342	0.52
		0.7	0.986	0.992	0.976	0.981	0.998	0.2358	0.51
		0.8	0.986	0.992	0.976	0.981	0.998	0.2378	0.50
		0.9	0.985	0.992	0.973	0.980	0.998	0.2407	0.33
		0.95	0.984	0.992	0.970	0.978	0.998	0.2416	0.29
14	Random Forest (RF)	0.1	0.589	0.348	1.000	0.643	0.972	0.8751	0.41
		0.2	0.744	0.595	0.997	0.742	0.972	0.7637	0.49
		0.3	0.861	0.79	0.981	0.839	0.972	0.6758	0.59
		0.4	**0.914**	**0.905**	**0.930**	**0.889**	**0.972**	**0.6257**	**0.65**
		0.5	0.896	0.956	0.795	0.85	0.972	0.6200	0.64
		0.6	0.855	0.989	0.627	0.762	0.972	0.6507	0.58
		0.7	0.790	0.997	0.438	0.607	0.972	0.7022	0.52
		0.8	0.729	0.998	0.270	0.425	0.972	0.758	0.46
		0.9	0.671	1.000	0.111	0.200	0.972	0.8172	0.44
		0.95	0.649	1.000	0.051	0.098	0.972	0.8448	0.41

5.1 Comparitive Study

In this sub-section, the effectiveness of the proposed approach at obtained cutoff point is compared with the existing non-hybrid approach at default cutoff [3] for German credit and Australian credit approval dataset. Through this comparison, it is possible to show that best cutoff point is required for different classifiers and datasets instead of default cutoff for all classifiers and datasets. In addition, this comparison shows the improved performance prediction results at obtained cutoff point and depicted in Figure 2 and Figure 3. As illustrated in Table 3 for German credit, the cutoff point is identified at 0.3 for NB classifier, 0.3 for LR and 0.4 for RF. At this cutoff point, the performance scores are improved when compared to existing approach at default cutoff. It is observed from the comparison graph, there are significant improvements noticed with the following performance measures of proposed approach such as overall_accuracy of NB increased by 9.3%; LR 23.5% and RF 15% , TNR of NB 34.7%; LR 48.8% and RF 52.3%, F-Measure of NB 25.5%; LR 44% and RF 38.1%, ROC-Area of NB 14%; LR 21.3% and RF 18.1%, RRSE of NB 23.75%; LR 66.26% and RF 25.34% , TPR of LR 12.8% whereas TPR of NB and RF are high with existing non-hybrid approach for German credit dataset. But in the existing non-hybrid

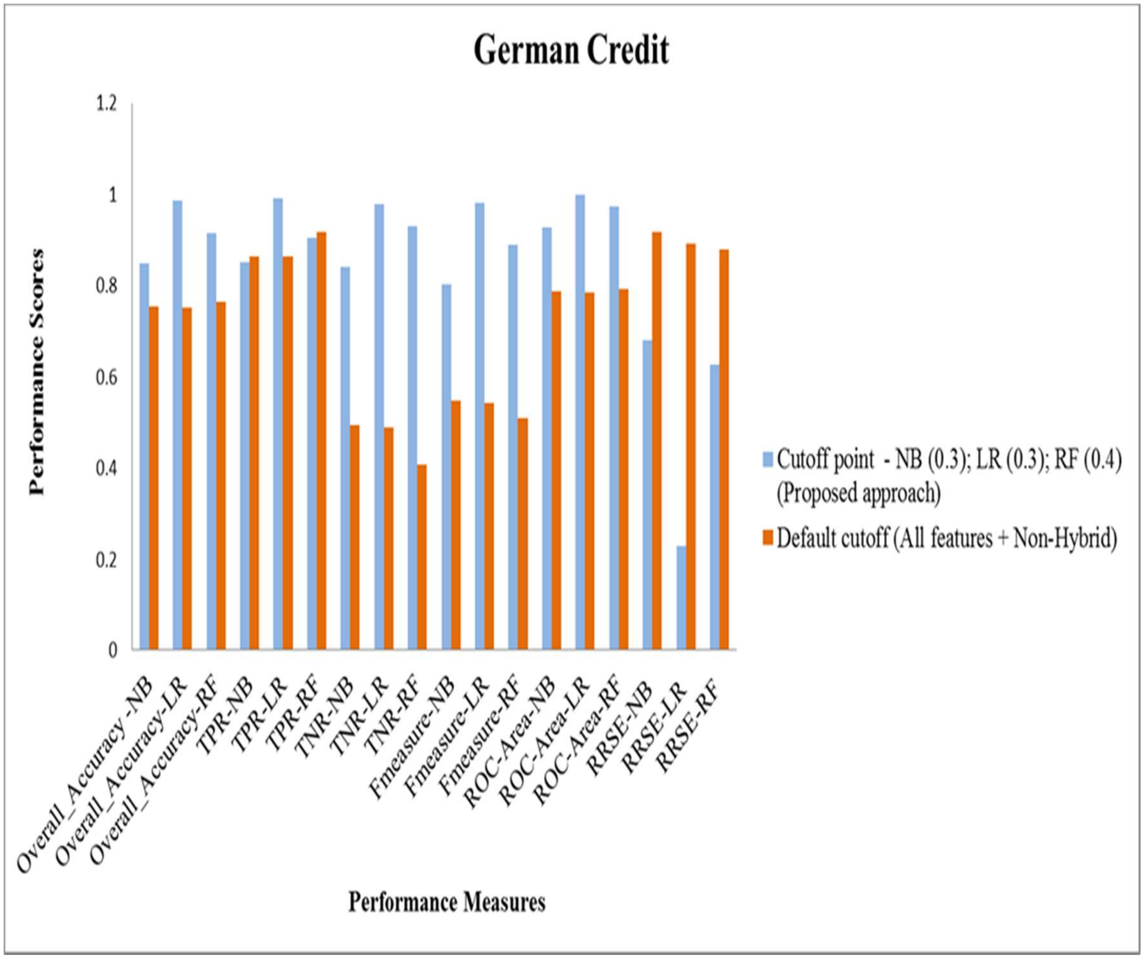

Figure 2: Comparison of Proposed approach at Cutoff Point Vs Existing Default Cutoff for German Credit Dataset

approach, TNR of NB and RF scores are not balanced with TPR. In Australian credit approval dataset, improvements are noticed with all the performance measures of proposed approach such as overall_accuracy of NB increased by 10.8%; LR 13.9% and RF 13.8% , TPR of NB 23.1%; LR 12.8% and RF 16.3%, TNR of NB 1.4%; LR 14.9% and RF 11.7%, F-Measure of NB 7.3%; LR 12.8% and RF 12.3%, ROC-Area of NB 6.3%; LR 8.8% and RF 7.8% , RRSE of NB 24.68%; LR 50.48% and RF 43.66%. Hence, it is easy to conclude that the proposed approach at obtained cutoff point outperforms with the existing default cutoff non-hybrid approach.These significant improvements with the credit risk datasets and three classifiers are possible due to the application of fuzzy MCDM approach in all the data mining evaluation stages which includes CRFS (optimal dataset), FIRS-KMeans (initial random seed selection for optimal clustering), FMCC (cutoff point selection for improved classification).

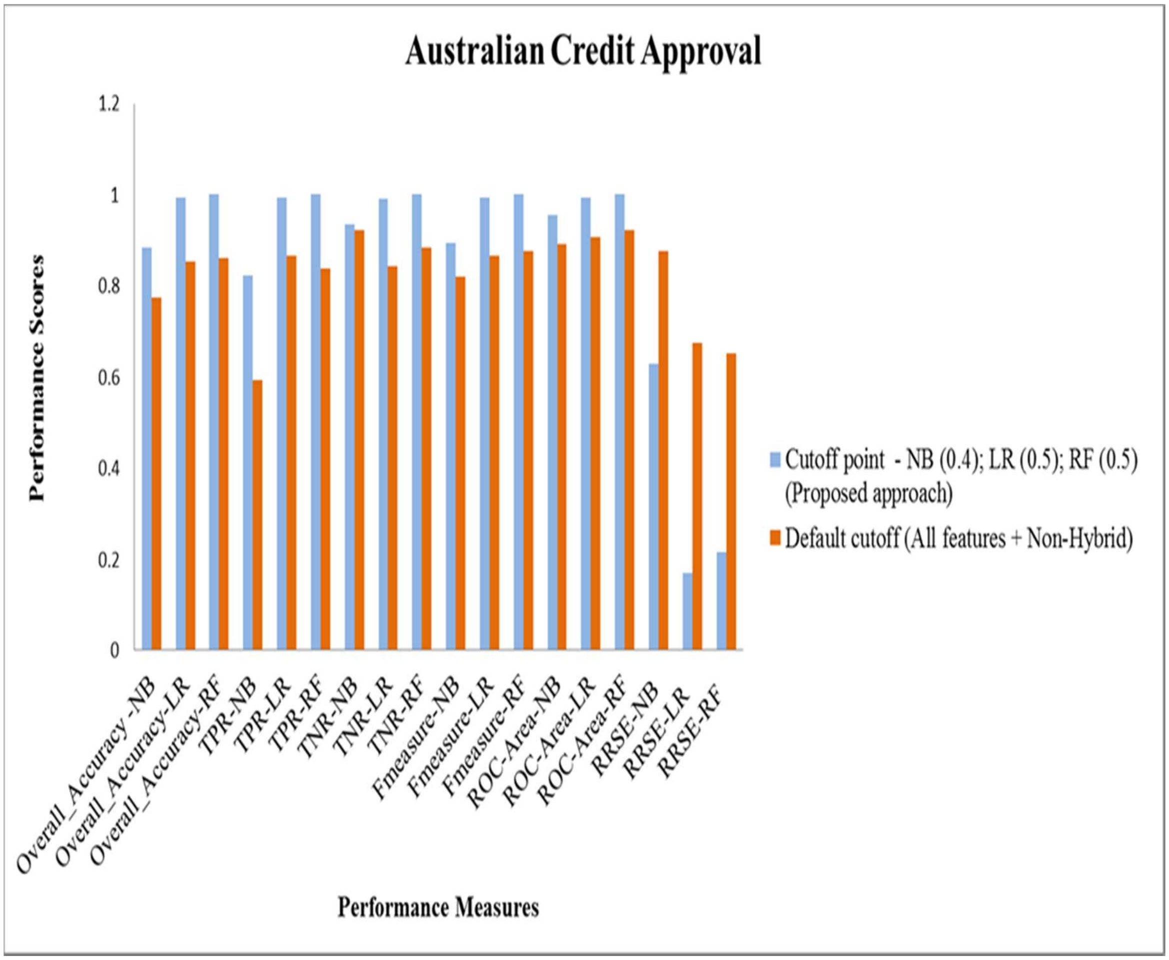

Figure 3: Comparison of Proposed approach at Cutoff Point Vs Existing Default Cutoff for Australian Credit Approval Dataset

6. CONCLUSION

Data mining classification techniques with default cutoff is used extensively for predicting classification performance irrespective of datasets and classifiers, thus limiting their improved classification performance. In this situation, choosing the cutoff point is very imperative for identifying the two classes (good and bad credits) accurately with respect to imbalanced datasets and classifiers. A new approach is proposed in this research work by integrating the following techniques to select the cutoff point for classifying the good and bad credits: fuzzy MCDM in all the data mining evaluation stages, appropriate initial random seed selection at clustering stage and BSCC hybrid algorithm is also integrated to improve the classification performance The experimental analysis of the proposed approach have shown promising classification results at obtained cutoff point.

To implement the proposed approach in credit risk prediction models, this research work aims to identify the cutoff point with respect to datasets and classifiers and determine appropriate techniques at every stage of data mining to improve the classification results at obtained cutoff point. Evaluation of classifiers at default cutoff for all classifiers and datasets and with non-hybrid approach significantly affects the classification results. Thus, the proposed approach is a best fit for credit risk assessment.With different classifiers and datasets, this approach identifies best cutoff point and obtain better prediction results. The limitation of the proposed approach is that it requires an integrated framework to effectively manage the intermediate process under one application. Another limitation is its subjective criteria weighting for MCDM evaluation. This leads to give more weightage for the specific criteria.

In the forthcoming work, the proposed work is to be extended to other data mining task such as association rule mining and ensemble approaches. In addition, integrated framework to manage all intermediate process under one application can be developed and suitable improvements to the criteria weighting method can also be identified.

REFERENCES:

[1] Kumar, M. N., and Rao, V. S. H. "A New Methodology for Estimating Internal Credit Risk and Bankruptcy Prediction under Basel II Regime", *Computational Economics*, Vol 46, 2015, pp. 83-102.

[2] Peng, Yi, Gang Kou, and Yong Shi., "Knowledge-rich data mining in financial risk detection", In *International Conference on Computational Science (ICCS 2009), Springer Berlin Heidelberg*, 2009, pp. 534-542.

[3] Peng, Y., Wang, G., Kou, G., and Shi, Y., "An empirical study of classification algorithm evaluation for financial risk prediction", *Applied Soft Computing*, Vol. 11, 2011, pp.2906-2915.

[4] Nitesh V. Chawla, "Data mining for imbalanced datasets: An overview", In *Data Mining and Knowledge Discovery Handbook*, 2010, pp. 875-886.

[5] Swets, J. A., "Measuring the accuracy of diagnostic systems", *Science*, Vol. 240, 1988, pp. 1285-1293.

[6] Hong, C. S., "Optimal threshold from ROC and CAP curves", *Communications in Statistics-Simulation and Computation*, Vol. 38, 2009, pp. 2060-2072.

[7] Calabrese, R., "Optimal cut-off for rare events and unbalanced misclassification costs", *Journal of Applied Statistics*, Vol. 41, 2014, pp. 1678-1693.

[8] Kou, G., Peng, Y., and Wang, G., "Evaluation of clustering algorithms for financial risk analysis using MCDM methods", *Information Sciences*, Vol. 275, 2014, pp. 1-12.

[9] Beulah Jeba Jaya Y., and Jebamalar Tamilselvi J., "Fuzzified MCDM Consistent Ranking Feature Selection with Hybrid Algorithm for Credit Risk Assessment", *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 11, 2015, pp. 1397-1403.

[10] Brown, I., and Mues, C., "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", *Expert Systems with Applications*, Vol. 39, 2012, pp. 3446-3453.

[11] Yao-huang, L. X. S. G., "Personal Credit Scoring Models on Naive Bayesian Classifier", *Computer Engineering and Applications*, Vol. 30, 2006, pp.058.

[12] Thomas P. Minka., "A comparison of numerical optimizers for logistic regression", *Technical report, Microsoft Research*, 2007.

[13] Bekhet, H. A., and Eletter, S. F. K., "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach", *Review of Development Finance*, Vol. 4, 2014, pp. 20-28.

[14] Breiman, L., "Random forests", *Machine learning*, Vol. 45, 2001, pp. 5-32.

[15] Gholamian, M., Jahanpour, S., and Sadatrasoul, S., "A new method for clustering in credit scoring problems", *Journal of mathematics and computer Science,* Vol. 6, 2013, pp. 97-106.

[16] Babu, G. Phanendra, and M. Narasimha Murty, "A near-optimal initial seed value selection in k-means means algorithm using a genetic algorithm*", Pattern Recognition Letters*, Vol. 14, 1993, pp. 763-769.

[17] Aparna K and Mydhili K Nair., "Selection of Initial Seed Values for K-Means Algorithm using Taguchi Method as an Optimization Technique", *International Journal of Engineering Research and Applications (IJERA)*, Vol. 4, 2014, pp. 214-217.

[18] Kotsiantis, S., Kanellopoulos, D., and Tampakas, V., "On implementing a financial decision support system", *International Journal of Computer Science and Network Security*, Vol. 6, 2006, pp. 103-112.

[19] Chen, W., Xiang, G., Liu, Y., and Wang, K., "Credit risk Evaluation by hybrid data mining technique", *Systems Engineering Procedia*, Vol. 3, 2012, pp.194-200.

[20] Zeng, H. J., Wang, X. H., Chen, Z., Lu, H., and Ma, W. Y., "Cbc: Clustering based text classification requiring minimal labeled data", In *Third IEEE International Conference on Data Mining (ICDM)*, 2003, pp. 443-450.

[21] Khanbabaei, M., and Alborzi, M., "The use of genetic algorithm, clustering and feature selection techniques in construction of decision tree models for credit scoring", *International Journal of Managing Information Technology*, Vol. 5, 2013, pp. 13-32.

[22] Hajian-Tilaki, K., "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation", *Caspian journal of internal medicine*, Vol. 4, 2013, pp. 627-635.

[23] Soureshjani, M. H., and Kimiagari, A. M., "Calculating the best cut off point using logistic regression and neural network on credit scoring problem-A case study of a commercial bank", *African Journal of Business Management*, Vol. 7, 2013, pp. 1414-1421.

[24] Kou, G., Peng, Y., and Lu, C., "MCDM approach to evaluating bank loan default models", *Technological and Economic Development of Economy*, Vol. 20, 2014, pp. 292-311.

[25] Jahanshahloo, G. R., Lotfi, F. H., and Izadikhah, M., "Extension of the TOPSIS method for decision-making problems with fuzzy data", *Applied Mathematics and Computation*, Vol. 181, 2006, pp. 1544-1551.

[26] Beulah Jeba Jaya Y., and Jebamalar Tamilselvi J., "Simplified MCDM Analytical Weighted Model for Ranking Classifiers in Financial Risk Datasets", In *proceedings of the IEEE International Conference on Intelligent Computing Applications (ICICA),* 2014, pp. 158-161.

[27] Mandic, K., Delibasic, B., Knezevic, S., and Benkovic, S., "Analysis of the financial parameters of Serbian banks through the application of the fuzzy AHP and TOPSIS methods", *Economic Modelling*, Vol. 43, 2014, pp. 30-37.

[28] Van Laarhoven, P. J. M., and Pedrycz, W., "A fuzzy extension of Saaty's priority theory", *Fuzzy sets and Systems*, Vol. 11, 1983, pp. 229-241.

[29] Tang, Y. C., and Lin, T. W., "Application of the fuzzy analytic hierarchy process to the lead-free equipment selection decision", *International Journal of Business and Systems Research*. Vol. 5, 2010, pp. 35-56.

[30] Aruldoss, M., Lakshmi, T. M., and Venkatesan, V. P., "A survey on multi criteria decision making methods and its applications", *American Journal of Information Systems*, Vol. 1, 2013, pp. 31-43.

[31] Hwang, Ching-Lai, Yoon, Kwangsun, "Multiple Attribute Decision Making Methods and Applications A State-of-the-Art Survey", *Lecture notes in Economics and Mathematical Systems, Springer*, 1981.

[32] Beulah Jeba Jaya Y. and Jebamalar Tamilselvi J., "Assessment of Fraud Pretentious Business Region Research Articles Using Data Mining Approaches", *International Journal on Computer Science and Engineering*, Vol. 5, 2013, pp. 653-659.

[33] Sherly, K. K., and Nedunchezhian, R., "BOAT adaptive credit card fraud detection system", In *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC),* 2010, pp. 1-7.

[34] Garanin, D. A., Lukashevich, N. S., and Salkutsan, S. V., "The Evaluation of Credit Scoring Models Parameters Using Roc Curve Analysis", World *Applied Sciences Journal*, Vol. 30, 2014, pp. 938-942.

[35] Ye Nan, Kian M. Chai, Wee S. Lee and Hai L. Chieu., "Optimizing F-Measure: A Tale of Two Approaches", In *Proceedings of the International Conference on Machine Learning*, 2012, pp. 289-296.

[36] Guo, X., Sun, H., Wang, L., Qu, Z., and Ding, W., "De-Word Classification Algorithm Based on the Electric Power of Large Data Library Retrieval", In *Computer Science and its Applications*, 2015, pp. 591-602.

[37] Abdelmoula, A. K., "Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks", *Accounting and Management Information Systems*, Vol. 14, 2015, pp. 79-106.

[38] Fawcett, T., "An introduction to ROC analysis", *Pattern recognition letters*, Vol. 27, 2006, pp. 861-874.

[39] Kotsiantis, S., "Credit risk analysis using a hybrid data mining model", International Journal of Intelligent Systems *Technologies and Applications*, Vol. 2, 2007, pp. 345-356.

[40] Krause-Traudes, M., Scheider, S., Rüping, S., and Mebner, H., "Spatial data mining for retail sales forecasting", In *11th AGILE International Conference on Geographic Information Science*, 2008.

[41] Peng, Y., Zhang, Y., Kou, G., Li, J., and Shi, Y., "Multicriteria decision making approach for cluster validation", *ProcediaComputer Science*, Vol. 9, 2012, pp. 1283-1291.

[42] Hubert, L., and Arabie, P., "Comparing partitions", *Journal of classification*, Vol. 2, 1985, pp. 193-218.

[43] Santos, J. M., and Embrechts, M., "On the use of the adjusted rand index as a metric for evaluating supervised classification", In *International Conference on Artificial Neural Networks*, 2009, pp. 175-184.

[44] Vendramin, L., Campello, R.J., and Hruschka, E.R., "Relative clustering validity criteria: A comparative overview", *Statistical Analysis and Data Mining*, Vol. 3, 2010, pp. 209–235.

[45] Silk Wagner and Dorothea Wagner, “Comparing clusterings - an overview”, *Technical Report. Informatics. Universit¨at Karlsruhe*, 2007.

[46] Lior Rokach and Oded Maimon , “Clustering Methods”, *Data Mining and Knowledge Discovery Handbook, Springer US*, 2005. pp. 321-352.

[47] Saaty, T.L., “The Analytic Hierarchy Process”, *McGraw Hill International*, 1980.

[48] Chang, D. Y., “Applications of the extent analysis method on fuzzy AHP”, *European journal of operational research*, Vol. 95, 1996, pp. 649-655.

[49] Witten, I.H., E. Frank, L. Trigg, M. Hall, G. Holmes and S.J. Cunningham, “Weka: Practical machine learning tools and techniques with java implementations”, In *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, 1999, pp. 192-196.

[50] Asuncion, A., Newman, D., “UCI machine learning repository”, *Available: http://mlearn.ics.uci.edu/ MLRepository.html. Irvine CA: University of California School of Information and Computer Science*, 2007.

[51] García, Vicente, Ramón Alberto Mollineda, and José Salvador Sánchez, “Index of balanced accuracy: A performance measure for skewed class distributions”, In *Iberian Conference on Pattern Recognition and Image Analysis*, 2009, pp. 441-448.