International Journal of Database Theory and Application
Vol.10, No.2 (2017), pp.1-22
http://dx.doi.org/10.14257/ijdta.2017.10.2.01

Chinese Small Business Credit Scoring: Application of Multiple Hybrids Neural Network

Chi Guotai[1*], Mohammad Zoynul Abedin[2, 3] and Fahmida – E – Moula[4]

[1, 2, 4]Faculty of Management and Economics, Dalian University of Technology, Dalian 116024, China

[3]Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur – 5200, Bangladesh

[1]chigt@dlut.edu.cn; [2]abedin@mail.dlut.edu.cn; [3]fahmidamz@yahoo.com

Abstract

In recent years, hybrid models have proven to be a promising approach for the forecasting of credit status, therefore, the aim of this project is to examine the prediction performance of hybrid classifiers. Particularly, the combination of the feature engineering with popular neural network (NN) classifiers; an hybridization approach, is compared with hybrid classifier, NN classifiers, and three well-known baseline classifiers, i.e. stepwise discriminant analysis (SDA), stepwise logistic regression (SLR), and decision trees (DTs). Overall, we executed a 12+8+ (8×8) experimental design that resulted in 84 unique classification models; i.e., 12 baseline models, 8 NN models, and 64 hybrid models, a multiple hybrid; are examined over a large credit scoring dataset from a Chinese commercial bank. Besides, thirteen evaluation measures are used for the assessment task and this may be the first effort to link up multiple hybrid classifiers with multiple performance metrics for the evaluation of small business credit. The results reveal that the predictive and distinguish ability of the F ratio based SDA with multilayer perceptron based NN classifier (SDA_{FR}+MLP), a hybrid model, outperforms both of the one–dimensional scoring models (baseline model and NN model) and its hybrid counterparts.

Keywords: *Small business, credit scoring, neural networks, multiple hybrids*

1. Introduction

In China, Small businesses (SBs) have been accountable for 60% of the national GDP, 50% of tax revenues, 70% of foreign trade, and nearly 80% of urban employment. Moreover, Chinese SBs have experienced significant growth in terms of number and size. In 2009, China had registered 43 million SBs, together responsible for 66% of patent applications, 74% of technological innovation, and 82% of new products [1]. These results show that SBs play a significant role of China's unique brand of economic transition. Despite its important contributions to economic vitality, credit default continues to be an important issue affecting SBs.

With the marvelous growth of the credit industry and the diversified loan portfolios, credit scoring has gained more and more attention as the credit industry can then benefit from reducing the possible credit risks, improving cash flow, insuring credit collections and enhancing the better managerial decisions. The accuracy of credit scoring is critical to financial institutions' profitability. Even a fraction of improvement on the scoring accuracy of credit decision will produce a significant future savings for financial institutions. Therefore, the ultimate goal of credit scoring models is to assign credit applicants to either a 'good credit' group that is likely to repay financial obligation or a 'bad credit' group whose application will be denied because of its high possibility of

ISSN: 2005-4270 IJDTA

defaulting on the financial obligation. Consequently credit scoring problems lie in the domain of the more general and widely discussed classification and prediction agenda [2].

Aiming to satisfy the above mentioned needs, more and more attention has been paid to credit scoring, and resulting in many different useful techniques, known as the credit scoring models, have been developed by financial institutions and researchers in order to solve the problems involved during the evaluation process. However, when we look at the last two decades, ANN, a fashionable credit scoring techniques comes out as an important alternative, and draws attention from many researchers with its high prediction accuracy. ANNs are computer systems developed to mimic the operations of the human brain by mathematically modeling its neuro–physiological structure [3]. Unlike statistical techniques, ANN does not require any assumptions, can generalize, can correctly infer the unseen part of a population, and in research about credit scoring, for many years, authors supported the superiority of the ANN model over versatile statistical models and optimization methods such as multi–variate discriminant analysis (MDA), LR, and KNN analysis [4–11].

The potentiality of the classifiers depend on the details of the problem, the data structure, the characteristics used, the extent to which it is possible to segregate the classes by using those characteristics, and the objective of the classification. No single classification algorithm can produce the best results for all classification problems. Hence, there is a growing interest that existing applications of single classifier can be further improved by hybrid method. Moreover, the hybrid classifier has been demonstrated to be outperformed by a single classifier in having greater accuracies and smaller prediction errors when applied to the credit scoring data sets [12].

In addition to, the machine learning techniques require an effective feature representation for both training and knowledge acquisition. In general, the main objective of feature selection is to determine a subset of representative features, by discarding features with little or no detective information, as well as redundant features that are highly correlated. Moreover, different feature selection techniques give different results on the same dataset. In this respect, the challenge to construct an accurate NN classifier with optimal and the smallest possible number of features has been addressed.

The idea of integrating multiple NN classifiers is not new in literature. For example, West [10] investigates the credit evaluation accuracy of five neural networks models: MLP, mixture of experts (MOE), radial basis function (RBF), learning vector quantization (LVQ), and fuzzy adaptive resonance (FAR). These five models are tested on two real world data sets partitioned into training and independent testing sets using 10-fold cross validation. In terms of the results, the difference of performance among these five neural networks model is marginal. However, the author suggested that although the MLP is the most commonly used neural network model, the MOE and RBF neural networks should be considered for credit evaluation applications. In addition, Bensic *et al*. [13] and Susac *et al*. [14] investigated four NN architectures, backpropagation network (BPN), RBF, Probabilistic neural network (PNN), and LVQ on Croatian small business data set and they concluded that the PNN was the most successful model for small business credit scoring.

Boyacioglu *et al*. [15] also investigated four NN tools, MLP, competitive learning (CL), SOM, LVQ on Egyptian bankruptcy dataset and found that the MLP was the most successful models in predicting the financial failure of banks. As well, Hájek [16] explored four NN models, FFNN, RBF, PNN, and cascade correlation (CC) NN on US municipal credit dataset and concluded that the PNN showed the best results for both four-class and nine-class municipal credit rating problem. In the same way, from early period credit scoring and bankruptcy application of NN [9] to the current period of application [17] it can be said that one NN architecture was superior on another architecture, for example, MLP is on modular neural network (MNN) [5], PNN is on BPN [18], BPN is on self-organizing map (SOM) [19], PNN is on multi-layer feed-forward network (MLFFN) based on average correct classification rate (ACCR) but MLFFN is on PNN based on misclassification cost (MCC) [4], MLFFN is on PNN [20], RBF is on backpropagation multi–layer perceptron (BPMLP) [21], PNN is on MLFFN [22].

Although neural network are increasingly found to be powerful in many classification application, the performance is actually dependent on network model itself, especially on initial condition, network topologies and training algorithms, which may be one reason why the results of neural network for credit risk evaluation varies when compared with different architectures, with different traditional models and even with ensemble classifiers. To find the optimal neural network architecture is still a challenging issue.

In this study, in the light of the above experiences, we developed multiple hybrid NN classifiers to classify Chinese small business credit. Our particular interest involves designing feasible NN architecture from popularly used NN models in the literature. Therefore, the core objective of this project is to examine the prediction performance of hybrid classifiers by comparing single and advanced statistical as well as artificial intelligence techniques. Particularly, the combination of the feature engineering with popular NN classifiers; an hybridization approach, is compared with hybrid classifier, single NN classifiers, and three well–known baseline classifiers, *i.e*. SDA, SLR, and DTs. Overall, we executed a 12+8+(8×8) experimental design that resulted in 84 unique classification models; *i.e*., 12 baseline models, 8 NN models (2 of them, for the first time, namely, generalized feedforward network, GFFN, and Jordan/Elmen network, JEN; as per our best knowledge), and 64 hybrid models, a multiple hybrid; are examined over a large credit scoring dataset from a Chinese commercial bank. In addition, thirteen performance measures are used for the assessment task. The results reveal that the predictive and distinguish ability of the 'SDA_{FR}+MLP' hybrid model outperforms both of the one–dimensional scoring models (baseline model & NN model) and its hybrid counterparts. Consequently, this study recommends applying the 'SDA_{FR}+MLP' hybrid model, an optimal credit scoring model, in bank's credit strategies.

Therefore, the advantage of this specific application related to the existing work is four–fold. First, we compare different state of-the-art classifiers to each other with different feature sets, to obtain the model with the highest accuracy and efficiency. Second, multi-dimensional evaluation measures are used for the assessment task to more rigorously examine the robustness and stableness of these techniques. Third, in our experiments we found an interesting phenomenon that JEN based hybrid, a novel hybrid model for credit scoring, provide sensible credit scoring results and the policy maker can consider it as an alternative model for credit classification. Fourth, the findings of this study can allow us to identify the best NN prediction model, which can be regarded as the reliable baseline for future research in small business credit.

The rest of the paper is organized as follows: Section 2 describes a brief overview of neural network credit scoring models. Section 3 explains the experiments with the inclusion of data background, data preprocessing, feature selection, classification models, and performance metrics. Section 4 presents and discusses the empirical results. Section 5, the final section, concludes the paper with the potential future directions of this study.

2. Neural Network Credit Scoring Models

A wide variety of neural network models has been proposed in the literature for commercial applications including credit scoring. Though they share some common features, they differ in structure and details. Therefore, eight neural network architectures are investigated in this research: the most popular MLP network, RBF, LVQ, MNN, PNN, SOM, GFFN, and JEN. Credit scoring accuracy, however, is expected to vary with the neural network model choice [10]. As advised by Khashei *et al*. [3], that the single hidden layer network is sufficient to model any complex system; the designed networks have only one hidden layer. Besides, it is noted that the comparisons between various training procedures are not the focus of this study. However, the conceptual differences between these eight neural network models are highlighted next.

2.1. Multilayer Perceptron

A feed forward MLP network consists of an input layer, an output layer, and a hidden layer between them. The credit scoring data used as inputs (x_i) at the input layer and the calculated sum were transmitted through the network, layer by layer, and a set of output value (y) were obtained. The connections between the input layer and the hidden layer contain weights (w_i), which are usually determined through training system. The hidden layer sums the weighted inputs and uses the activation function, f, to create the credit scoring output value. The activation function used in this research is the sigmoidal function. Mathematically, MLP can be represented as in Eq. (1),

$$y^{(t)} = f\,(y_i), \quad \text{with} \quad f = 1/(1+e^{-x}), \quad \text{and} \quad y_i = \sum_{i=1}^{n} W_i\, x_i + b \qquad (1)$$

where, $y^{(t)}$ is the final credit scoring output, f is the activation function, y_i is the hidden layer output, W_i represents the weight vector; x_i is the input vector (i = 1, 2, …, n); and b is the bias unit.

2.2. Radial Basis Function Network

The RBF network has a simple architecture with a single hidden layer. In this study the architecture of an RBFN is similar to a three–layer MLP. Though, there are differences between the MLP and RBFN. First, in a MLP, each node (*i.e.*, hidden node and output node) has the same transfer function, such as the sigmoidal function. In an RBFN, each hidden node has its own radial basis function, such as a Gaussian function (Eq. (2)) which is used in this study,

$$Y_i\,(X) = exp\,[-\{(x-c_i)^T\,(x-c_i)\}/2\sigma_i^{\,2}], \qquad i = 1, 2, K, L \qquad (2)$$

where Y_i is the credit scoring output of the ith node in hidden layer, x is the feature pattern, c_i is the weight vector for the ith node in hidden layer, *i.e*., the center of the Gaussian for node i; $\sigma_i^{\,2}$ is the normalization parameter (the measure of spread) for the ith node; and L is the number of nodes in the hidden layer. The outputs are in the range from zero to one so that the closer the input is to the center of the Gaussian, the larger the response of the node.

2.3. Learning Vector Quantization

The LVQ network is a simple three–layer supervised manner competitive network that produces a credit scoring decision by using the hidden layer neurons as a set of prototype vectors; a subset of these prototype vectors is assigned to each credit group [10].

$$\|x - W_i\| = min_d\,\{\|x - W_i\|\} \qquad (3)$$

In basic LVQ learning, Euclidean distance is used. The distance between the input vector, x, and the weight vector, W_i, is computed in Eq. (3) that produces a new credit applicant group, d, and the nearest credit units are declared to be the winner, that is a minimum Euclidean distance from the input vector, x.

2.4. Modular Neural Networks

Modular neural networks (MNNs) have a hierarchical organization, to cover several ANNs; its architecture basically consists of two main components: local experts and an integration unit. A gating network controls the competition in between of desired network output vector and activation output vector for each respective module network and learns to assign different regions of the credit scoring data space to different networks. Both the local experts and the gating network have full connections with the credit scoring input layer. The gating network has as many output nodes as there are local experts, and the outputs of the gating network, g_i, are interpreted by using the softmax activation function, in Eq. (4). Where u_i is the weighted sum of inputs associated with the gating network. The output values of the gating network are positive and sum up to one due to the special form of the softmax being non–linear. The final credit scoring output vector, $y^{(t)}$ of the MNN is a weighted sum of the output vectors of the local experts with their corresponding gating network outputs [8], shown in Eq. (5).

$$g_i = e^{u_i} / \sum_{k=1}^{n} e^{u_k} \text{, with } u_i = W_i^T x^{(t)} \quad (4)$$

$$y^{(t)} = \sum_{i=1}^{t} g_i y_i \quad (5)$$

To adjust the network's weights, the training of MNN is also done using the backpropagation of error, like the BPN. For a comprehensive account of MNN refer to Ref., Desai *et al*. [5].

2.5. Probabilistic Neural Network

An alternative NN architecture, the PNN [23] is non–linear, nonparametric pattern recognition modeling technique that consists of four layers: input, pattern, summation, and output. The credit scoring features from the input layer distribute the credit inputs to the pattern units, where the pattern layer generally uses the following function as in Eq. (6).

$$g(z_i) = exp\{(z_i - 1)/\sigma^2\} \quad (6)$$

Here, z_i is the dot product of the credit input and weight vectors, while the scale parameter σ^2 defines the width of the area of influence, and normally decreases as the sample size increases. When the credit scoring input is presented, the pattern layer computes distance from the input vector to the training input vectors, producing a vector whose elements indicate how close the input is to a training input. The summation layer has one neuron for each class. Each summation neuron dedicated to a single class sums the pattern layer neurons corresponding to numbers of that summation neuron's class. Activation of summation neuron n is the estimated density function of population n. The output neuron is a threshold discriminator that identifies which of its inputs from the summation units is the maximum.

2.6. Self-Organizing Maps

A SOM which was introduced by Kohonen [24] is a feed forward neural network consisting of input and output layers of neurons. The neurons from the output layer are usually ordered in a low–dimensional grid. Each unit in the input layer is connected to all neurons in the output layer. Weights are attached to each of these connections. This is similar to a weight vector, with the dimensionality of the input space, being associated with each output neuron. When a credit scoring training vector X is presented, the weight vector of each neuron c is compared with X. One commonly opts for the Euclidian distance between both vectors as the distance measure. The neuron that lies closest to X is called the 'winner' or the Best Matching Unit (BMU). The weight vector of the BMU and its neighbors in the grid are adapted with the following learning rule:

$$W_c = W_c + n(t)\, A_{winner,\,c}(t)\, (X - W_c) \quad (7)$$

In this expression $n(t)$ represents the learning rate that decreases during training. $A_{winner,\,c}(t)$ is the so–called neighborhood function that decreases when the distance in the grid between neuron c and the winner unit becomes larger.

2.7. Generalized Feedforward Networks

The GFFNs are the efficient implementation of the BP algorithm to train general feedforward NNs where the networks have no feedback and these networks are also the generalization of MLP such that connections can jump on forward over one or more layers. Typically, in GFFN, the neurons in one layer are not only connected to the neurons of the next layer, but also to all those neurons of all the forward layers. This type of network can usually be trained more quickly than non–generalized FNNs [25]. Neurons in each layer receive inputs only from the preceding layer, calculate their outputs and transmit the resulting signals to the next layer. Hyperbolic tangent was used as activation function instead of the sigmoidal function utilized in MLP network.

2.8. Jordan/Elman Network

A neural network with at least one feedback loop is called a recurrent neural network (RNN). The basic structure of two examples of RNN, are known as the Jordan's [26] network and the Elman's [27] network. The network mainly consists of four layers: input layer, hidden layer, context layer, and output layer. These networks feature a set of context units, whose activations are copied from either the outputs or the hidden units, respectively, and which are then feedforward into the hidden layer, supplementing the inputs. The context units give the networks a kind of decaying memory, which has proven sufficient for learning temporal structure over short distances, but not generally over long distances. The hidden layer activates the output layer and refreshes the context layer with the current state of the hidden layer. The back–propagation learning algorithm is commonly employed to train the weights.

3. Experiments

3.1. Data Background

In order to verify the feasibility and effectiveness of the proposed multiple hybrid neural network credit scoring model, a small business data set from a selected commercial bank in China is used in this study. The selected bank established in 1998 which has 95 branches in Dalian, Tianjin, Beijing, Shanghai and four other cities in China. There are totally 3111 small business customers in the dataset with 3040 good credit customers while the remaining 71 are bad credit customers. The large size of the sample is an important strength for the reliability of our findings.

3.2. Experimental Protocol

The proposed multiple hybrid neural network model is comprised of the following four fundamental building phases: (1) Pre-processing: data preprocessing is required to determine the relative importance variables and ensure data field consistency in credit scoring model building. (2) Feature selection: we combine filter approach (FA) with the embedded expert knowledge (EEK) to design the 8 different feature sets those are the most relevant features from a larger feature set. (3) Classification phase: under this phase, we built 3 different base classifiers (containing a total of 12 benchmark models), 8 different NN classifiers those act as a single classifiers, and 3 hybrid classifiers based on 8 different feature sets and 8 different NN classifiers (containing a total of 64 experimental models), which is known as multiple hybrid. In total, we have analyzed the performance of 84 classifiers for Chinese small business credit scoring applications. And (4) Performance evaluation phase: at the last phase, we evaluate the performance of the resulting models according to the 13 performance metrics. These four phases are described in detail in the following section along with the steps involved and the characteristics of each phase and the overall architecture of the introduced approach is described in Figure 1.

3.3. Data Preprocessing

In this study, we collect bank's internal data and the collected databases may contain features that are obsolete or redundant, missing values, outliers, and data in a form not suitable for classification models. Therefore, a pre–processing stage should be considered to enhance the credit scoring quality before feature selection and classification process and to increase the efficiency of the classification and prediction process. However, data preprocessing consists of (1) data cleaning, (2) data integration, (3) data transformation, and (4) data reduction. After a successful pre–processing of data, we construct a final data base with 3111small business objects and 81 variables (3111*81), where each small business customer in the dataset contains 48 financial, 27 non–financial, and 6 macroeconomic variables.

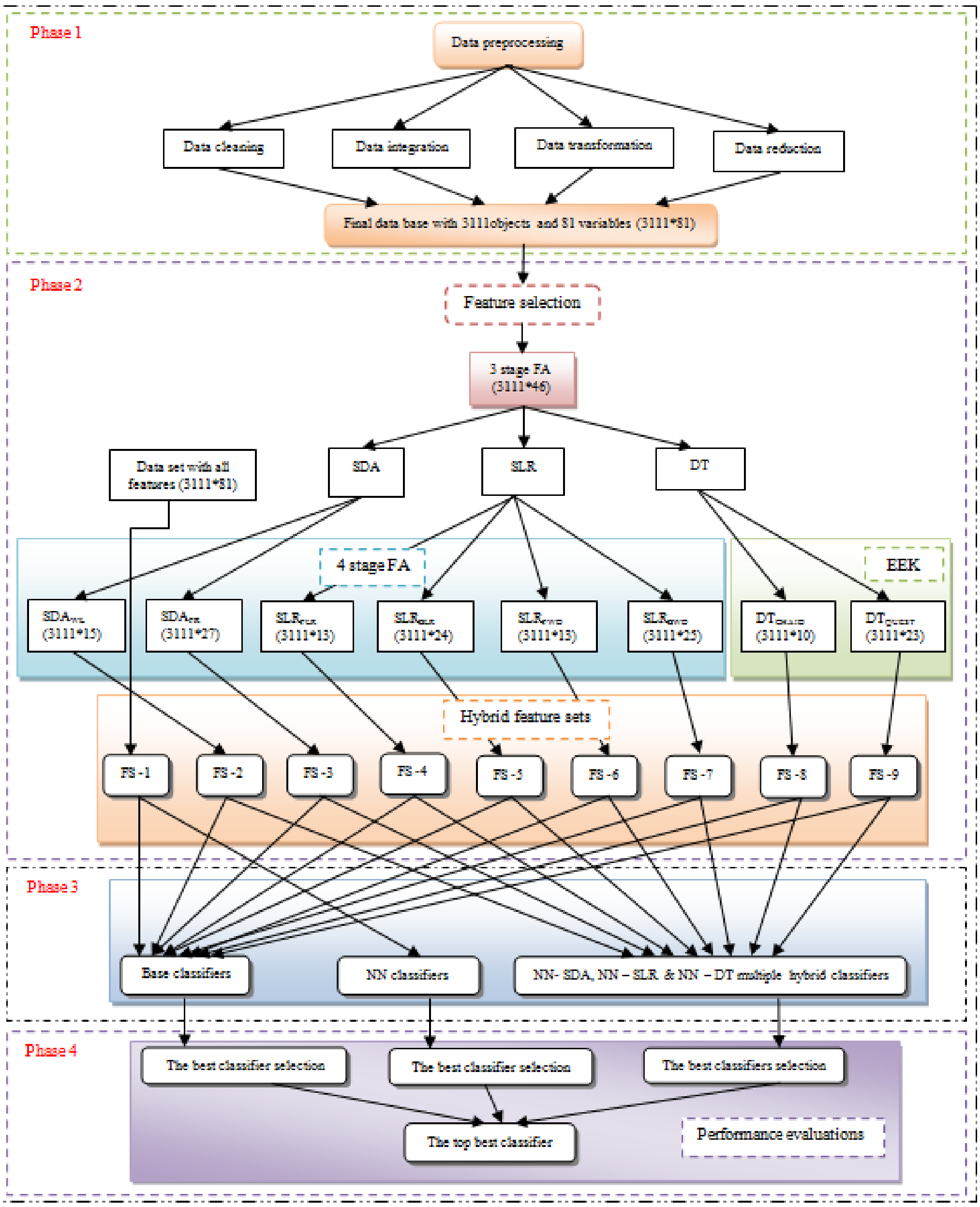

Figure 1. The Overall Experimental Architecture

3.4. Feature Selection

Kim *et al*. [28] suggested that the business problems are unstructured in nature; therefore combining machine–learning driven predictors with human–driven predictors may be a better approach. Once more, Oreski & Oreski [29] has reported that a simple combination of the best individual features selected by the single model does not necessarily lead to a good classification performance. Considering their observations, we propose an integrated approach to feature selection for the small business credit scoring that combine filter method with the embedded expert knowledge, a hybridization approach, to maximize the advantages of the aforementioned two types of models. Under phase 2 as shown in Figure 1, therefore, we design the 9 different feature sets through four steps those are the most relevant features from a larger feature set.

In the first step, after data preprocessing under phase 1 in Figure 1, a total of 81 features were selected and arranged them as features set 1 (FS–1).

In the second step, for the purpose of selecting only those features with both the greatest prediction capacity and the lowest correlation levels, we carried out correlation analysis, independent–samples *t*–test, one way ANOVA; a three stage hybrid filter approach. In this step 46 variables were selected.

In the third step, a stepwise analysis was used to reduce the dimensionality and to select final variables from the set of 46 variables. In stepwise analysis, we use stepwise discriminant analysis and stepwise logistic regression, four stage hybrids filter approach. As the stepwise discriminant method, we used Wilk's lambda and *F*–value as the criteria for determining the entry or removal of the input features. For the *F*–statistic, we set an entry value of 0.01 and an excluding value of 0.05 as the levels of significance. Following these two criterions, we selected 15 and 27 variables and formed them as feature set 2 (FS–2) and features set 3 (FS–3) respectively. In similar way, for the logistic regression (LR) we employed stepwise forward selection as well as backward selection method, where we set an entry value of 0.01 and an excluding value of 0.05 as the levels of significance for *z*–statistic. In forward selection, two different features sets; features set 4 (FS–4) & features set 6 (FS–6); were formed through likelihood ratio and Wald coefficient, each feature set contains 13 variables. Similar procedures were followed in backward selection and formed features set 5 (FS–5) & features set 7 (FS–7); containing 24 & 25 variables respectively.

In the fourth step, a hybrid model; 3 stage filter model with the decision tree embedded model that incorporate variable selection as part of the training process and are able to represent knowledge in a flexible and easy form; was applied to select the distinguish features from the set of 46 variables. Two popular DT algorithm, namely CHAID and QUEST, was applied and formed features set 8 (FS–8) & features set 9 (FS–9) respectively; 10 and 23 variables were contained under each feature set.

3.5. Classification

3.5.1. Base Classifier

The base classifier of discriminant analysis, logistic regression, CHAID and QUEST decision tree are too popular to be described here. The interested readers may refer to Prado *et al*. [30] for the detailed description and some previous results of these baseline classifiers.

3.5.2. Multiple Hybrid Classifiers

For the hybridization various strategies have been developed in order to overcome the deficiencies of single classification, to yield more accurate results, and to enforce diversity on the classifiers. For instance, Lasheras *et al*. [12] identified four basic approaches: (1) Hybrid Algorithms, (2) Ensemble Classifiers, (3) Clustering and

Classification, and (4) Feature Selectors. The last strategy constitutes the most commonly used methods. In this context, we consider two different representative approaches to hybrid paradigm to use the first technique for feature selection and the second for classification. More precisely, the first approach is to use parametric (SDA & SLR) & non–parametric (CHAID–DT & QUEST–DT) models to select the optimal features and the second is to use output, such as selected features, as an input to the neural models for classification. Therefore the hybrids that are developed in our study include SDA–NN, SLR–NN, and DT–NN, a multiple hybrid classification models which are a subcategory of hybrid classifiers.

Table 1. Performance Metrics and their Characteristics

Metrics	Characteristics
GC	GC is the number of good customers that are correctly predicted as good; *i.e.*, GC = $\{ G_g / TG \}*100$
BC	BC is the number of bad customers that are correctly predicted as bad; *i.e.*, BC = $\{ B_b / TB \}*100$
ACC	ACC is the proportion of correctly classified cases; *i.e.*, ACC = $\{(G_g + B_b)/ TN\}*100$
Type I error	Type I error is the number of good customers that are incorrectly predicted as bad; *i.e.*, Type I error = $\{G_b / TG\}*100$
Type II error	Type II error is the number of bad customers that are incorrectly predicted as good; *i.e.*, Type II error = $\{B_g/ TB\}*100$
AMC	AMC is the proportion of misclassified cases; *i.e.*, AMC = $\{(G_b + B_g)/ TN\}*100$
MSE	MSE is the average squared difference between the actual and estimated values; *i.e.*, MSE = $(1/N)\sum_{i=1}^{n}(\theta_i - P_i)^2$
RMSE	RMSE is the square root of the MSE; *i.e.*, RMSE = $\{(1/N)\sum_{i=1}^{n}(\theta_i - P_i)^2\}^{1/2}$
MAE	MAE is the mean absolute difference between the actual and estimated values; *i.e.*, MAE = $(1/N)\sum_{i=1}^{n}\lvert \theta_i - P_i \rvert$
EMCC	EMCC = $\{C_1* (G_b / TG)* (TG / TN)\} + C_2* (B_g / TB)* (TB / TN)$
AUC	AUC is the area under the ROC curve. The classifier achieving perfect accuracy corresponds to score of 1; while a score of 0.5 means that the classifier has no discriminative power.
AR	The AR taken from a ROC analysis provides an assessment of the discriminating power of the model; where, AR = 2*(AUC – 0.5). The AR lies between 0 (imperfect accuracy) and 1 (perfect accuracy).
Kappa	Ferri *et al*. [31] noted that Kappa = {P (A) – P (E)}/ {1 – P (E)}; where P (A) is the relative observed agreement among classifiers, *i.e.*, the proportion of correctly classified cases; hence, P (A) = ACC, and P (E) is the probability that agreement is due to chance or deviation. The Kappa value lies between 0 (imperfect accuracy) and 1 (perfect accuracy).

Note: N is the total number of the small business samples, θ_i is a binary indicator for the actual realization of the default variable (1 if default, 0 if no default) and P_i is the estimated probability of default. Low MSE, RMSE, and MAE values (0; perfect and 1; imperfect) indicate high confidence in the values predicted by the model.

3.6. Performance Metrics

We evaluate the performance of the resulting models according to the thirteen performance metrics listed in Table 1. For a two–class problem, most of these metrics can be easily derived from a 2*2 confusion matrix [32] as that given in Table 2, where each entry (i, j) contains the number of good/bad customers. The following performance measures are used and evaluated respectively: the percentage of good classification (GC), bad classification (BC), average correct classification (ACC), type I & type II errors, average misclassification (AMC), mean square error (MSE), root mean square error (RMSE), & mean absolute error (MAE) [31], expected misclassification cost (EMCC) [10], area under the ROC curve (AUC), cumulative accuracy ratio (AR) and Cohen's [33] Kappa statistics (Kappa). Though AUC and AR has produced identical ranking, both of them have used for future reference.

Table 2. The Confusion Matrix for Classification Problem

	Predicted observations		
	G	*b*	
Actual observations			
G	G_g	G_b	*TG*
B	B_g	B_b	*TB*
	Tg	*Tb*	*TN*

Notations: *G* = actual good; *g* = predicted good; *B* = actual bad; *b* = predicted bad; *Gg* = actual good predicted good; *Gb* = actual good predicted bad; *Bg* = actual bad predicted good; *Bb* = actual bad predicted bad; *TG* = total actual good observations; *TB* = total actual bad observations; *Tg* = total predicted good observations; *Tb* = total predicted bad observations; and *TN* = total number of observations in the small business dataset. Ref. Abdou [32].

4. Results and Discussion

In this study, we developed 84 different types of prediction models which fall into the three categories; base classifiers, single (NN) classifiers, and multiple hybrid classifiers. Our goal in this empirical evaluation is to show that hybrid methods, which compete quite well against base classifiers as well as single NN classifiers, are credible methods for credit scoring. For assessment purpose, the credit classifiers are ranked through the bold numbers indicating performances that significantly outperform the other classifiers based on the thirteen performance metrics listed in Table 2.

4.1. Base Classifiers

To verify the superiority of the proposed hybrid methods, models based on prior scholars as well as recognized feature selection algorithms are used as the point of references. Five different classification models are proposed under the famous DA model. Among of them, Altman's [34] Model and Beaver's [35] Model are based on the ratio set proposed by Altman [34] and Beaver [35], respectively; whereas three other DA models are SDA_{ALL}, SDA_{WL}, and SDA_{FR} based on the FS–1, FS–2, and FS–3, correspondingly. As shown in Table 3, among of five DA models, Altman's [34] Model presents the lowest performance, while the best performance is achieved by the SDA_{ALL}, based on all performance criterions. More precisely, the GC rate is 62.47% & 95.89%, BC rate is 66.20% & 78.87%, ACC rate is 62.55% & 95.50%, AUC value is 0.643 & 0.874, AR is 0.286 & 0.748, Kappa statistics is 0.033 & 0.426, type–I error rate is 37.53% & 4.11%, type–II error rate is 33.80% & 21.13%, AMC rate is 37.45% & 4.50%, MSE is 0.022 & 0.017, RMSE is 0.025 & 0.025, MAE is 0.001 & 0.001, and EMCC is 1.660 & 0.033; respectively.

Table 3. Performance Obtained by the Base Classifiers

Credit scoring models	GC (%)	BC (%)	ACC (%)	AUC	AR	Kappa	T – I Er. (%)	T – II Er. (%)	AMC (%)	MSE	RMSE	MAE	EMCC
Altman [34]	62.47	66.20	62.55	0.643	0.286	0.033	37.53	33.80	37.45	0.022	0.025	0.001	1.660
Beaver [35]	71.41	66.20	71.30	0.688	0.376	0.055	28.59	33.80	28.70	0.022	0.025	0.001	1.658
SDA_{WL}	91.02	73.24	90.61	0.821	0.642	0.234	8.98	26.76	9.39	0.020	0.025	0.001	1.310
SDA_{FR}	95.39	77.46	94.99	0.864	0.728	0.393	4.61	22.54	5.01	0.018	0.025	0.001	1.102
SDA_{ALL}	**95.89**	**78.87**	**95.50**	**0.874**	**0.748**	**0.426**	**4.11**	**21.13**	**4.50**	**0.017**	0.025	0.001	**0.033**
SLR_{FLR}	**99.74**	21.13	97.94	0.604	0.208	0.311	**0.26**	78.87	2.06	0.019	0.025	0.001	3.854
SLR_{BLR}	99.64	28.17	98.01	0.639	0.278	0.384	0.36	71.83	1.99	0.018	0.025	0.001	3.510
SLR_{FWD}	**99.74**	21.13	97.94	0.604	0.208	0.311	**0.26**	78.87	2.06	0.019	0.025	0.001	3.854
SLR_{BWD}	99.64	29.58	**98.04**	0.646	0.292	0.399	0.36	70.42	1.96	0.018	0.025	0.001	3.441
SLR_{ALL}	99.54	**67.61**	**98.81**	**0.836**	**0.672**	**0.716**	0.46	**32.39**	**1.19**	**0.011**	0.025	0.001	**1.583**
DT_{CHAID}	**100**	0.0	97.72	0.871	0.742	----	**0.00**	100	2.28	----	**0.018**	**0.000**	4.886
DT_{QUEST}	**100**	0.0	97.72	0.871	0.742	----	**0.00**	100	2.28	----	**0.018**	**0.000**	4.886

Again another five benchmark models; SLR_{FLR}, SLR_{BLR}, SLR_{FWD}, SLR_{BWD}, & SLR_{ALL}; are offered from another well-known classification model of LR. These five models are designed based on FS–4, FS–5, FS–6, FS–7, & FS–1, respectively. As can be observed in Table 4, among of five LR models, SLR_{ALL} model shows the best efficiency based on ACC rate, 98.81%; AUC value, 0.836; AR, 0.672; Kappa statistics, 0.716; type – II error rate, 32.39%; AMC, 1.19; MSE, 0.011; and EMCC, 1.583; the eight performance criterions. On the other hand, SLR_{FLR} & SLR_{FWD}, both of these models show the highest efficiency based on GC rate, 99.74% & 99.74%; BC rate, 21.13% & 21.13%; and type–I error rate, 0.26% & 0.26%; the three performance criterions and it is notable that these two models show the equal performance in case of these three criterions.

Furthermore, two different DT models, namely, DT_{CHAID} and DT_{QUEST} are proposed based on FS–8 and FS–9. Surprisingly, the two models provide the same results on all performance criterions. The distinguishing feature is that the two models show 100% GC rate, 0.871 AUC value, no type – I error, 0.018 RMSE and 0.000 MAE.

Focusing on above analytical results, we observe that the SLR_{ALL} model outperforms based on eight performance criterions; SDA_{ALL} model outperforms based on five performance criterions; DT_{CHAID} and DT_{QUEST} models outperform based on four performance criterions. Therefore, in line with West [10]; we find that the LR model outperforms SDA and DT in this type of classifier.

4.2. NN Classifiers

For verifying the applicability of hybrid models, we present the performance of eight popular NN classifiers as the benchmark those are designed based on original feature set, FS–1. The training parameters are arbitrarily specified in order to train the NN architectures as we mention earlier that the comparisons between various training procedures are not the focus of this study. Generally, the learning rate is set between 0.01 and 0.4, the momentum is set between 0.8 and 0.99 and the training lengths ranging from 1000 to 10,000 epochs [36].

Table 4. Performance Obtained by the NN Classifiers

NN Credit scoring models	GC (%)	BC (%)	ACC (%)	AUC	AR	Kappa	T – I Er. (%)	T – II Er. (%)	AMC (%)	MSE	RMSE	MAE	EMCC
MLP	99.77	**88.73**	**99.52**	**0.943**	**0.886**	**0.891**	0.23	**11.27**	**0.48**	**0.005**	**0.015**	0.001	**0.551**
RBF	**100**	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
LVQ	**100**	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
MNN	75.26	78.87	75.35	0.771	0.542	0.089	24.74	21.13	24.65	0.022	0.025	0.001	1.038
PNN	63.91	57.75	63.77	0.608	0.216	0.026	36.09	42.25	36.23	0.022	0.025	0.001	2.073
SOM	73.55	47.89	72.97	0.607	0.214	0.034	26.45	52.11	27.03	0.022	0.025	0.001	2.552
GFF	60.92	60.56	60.91	0.607	0.214	0.024	39.08	39.44	39.09	0.022	0.025	0.001	1.936
JEN	71.12	76.06	71.23	0.736	0.472	0.068	28.88	23.94	28.77	0.022	0.025	0.001	1.176

As the experimental results show in Table 4, we can observe that the predictive performance of MLP is the best among of eight NN classifiers. Generally MLP has the highest values of BC rate (88.73%), ACC rate (99.52%), AUC value (0.943), AR (0.886), Kappa statistics (0.891), type–II error rate (11.27%), AMC rate (0.48%), MSE (0.005), RMSE (0.015) and EMCC with a value of 0.551; ten performance criterions. Then, the RBF and LVQ are well performed on GC rate (100%) with no type–I error and no MAE; three performance criterions followed by MNN, SOM, JEN, PNN and GFF, the least performer classifier. The result is consistent with Boyacioglu *et al*. [15] who investigated four NN tools, MLP, competitive learning (CL), SOM, LVQ on Egyptian bankruptcy dataset and found that the MLP was the more consistent models in predicting the financial failure of banks.

4.3. Multiple Hybrid Classifiers

4.3.1. SDA_{WL} + NN Hybrid

From the results revealed in Panel A of Table 5, it can be said that the hybrids of SDA_{WL} +JEN>SDA_{WL}+RBF & SDA_{WL}+PNN>SDA_{WL}+GFF > SDA_{WL}+SOM> SDA_{WL}+MNN > SDA_{WL}+LVQ > SDA_{WL}+MLP from the view of all performance metrics. More specifically, SDA_{WL}–JEN hybrid classifier shows the best results in terms of BC rate (33.80%), ACC rate (98.46%), Kappa statistics (0.494), type–II error rate (66.20%), AMC rate (1.54%), MSE (0.015), and 3.235 EMCC value; the seven performance criterions. Then, an interesting result has given by the SDA_{WL}–RBF & SDA_{WL}–PNN; both of these hybrids produce the identical results and they are in to the front position in terms of five performance criterions, *e.g.*, GC rate (100%), AUC value (0.871), AR (0.742), followed by 0 type–I error rate and no MAE. It is notable that all hybrid classifiers in this category confirm almost identical results in terms of all criterions except BC rate and type–II error rate.

4.3.2. SDAFR + NN Hybrid

As the experimental results demonstrate in Panel B of Table 5, we can observe that the predictive performance of 'SDA_{FR}+MLP' hybrid classifier is the best among of eight NN based hybrids. Generally 'SDA_{FR}+MLP' hybrid has the highest values of BC rate (95.77%), ACC rate (99.87%), AUC value (0.979), AR (0.958), Kappa statistics (0.971), type–II error rate (4.23), AMC rate (0.13%), MSE (0.001), RMSE (0.015) and EMCC with a value of 0.207; ten performance criterions. Then, the RBF is in to the forefront on GC rate (100%) with no type–I error and no MAE; three performance criterions followed by SDA_{FR}+LVQ, SDA_{FR}+JEN, SDA_{FR}+GFF, SDA_{FR}+MNN, SDA_{FR}+SOM, and SDA_{FR}+PNN, the least performer hybrid classifier. unlike of previous hybrids, *i.e.*, SDA_{WL}–NN, here the results of performance techniques are deviated more in case of GC, BC, ACC, type–I error, type–II error, and AMC.

Table 5. Performance Obtained by the SDA + NN Hybrid Classifiers

NN Credit scoring models	GC (%)	BC (%)	ACC (%)	AUC	AR	Kappa	T – I Er. (%)	T – II Er. (%)	AMC (%)	MSE	RMSE	MAE	EMCC
Panel A: Performance obtained by the SDA_{WL} + NN hybrid classifiers													
MLP	99.97	5.63	97.81	0.528	0.056	0.103	0.03	94.37	2.19	0.021	0.025	0.001	4.611
RBF	**100**	0.0	97.72	**0.871**	**0.742**	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
LVQ	99.64	21.13	97.85	0.604	0.208	0.301	0.36	78.87	2.15	0.020	0.025	0.001	3.854
MNN	99.97	21.13	98.17	0.606	0.212	0.339	0.03	78.87	1.83	0.018	0.025	0.001	3.854
PNN	100	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
SOM	99.97	26.76	98.30	0.634	0.268	0.412	0.03	73.24	1.70	0.017	0.025	0.001	3.579
GFF	**100**	8.45	97.91	0.542	0.084	0.153	**0.0**	91.55	2.09	0.020	**0.004**	**0.000**	4.473
JEN	99.97	**33.80**	**98.46**	0.669	0.338	**0.494**	0.03	**66.20**	**1.54**	**0.015**	0.025	0.001	**3.235**
Panel B: Performance obtained by the SDA_{FR} + NN hybrid classifiers													
MLP	99.97	**95.77**	**99.87**	**0.979**	**0.958**	**0.971**	0.03	**4.23**	**0.13**	**0.001**	**0.015**	0.001	**0.207**
RBF	**100**	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
LVQ	99.97	11.23	97.94	0.556	0.112	0.196	0.03	88.73	2.06	0.020	0.025	0.001	4.335
MNN	68.09	70.42	68.15	0.693	0.386	0.051	31.91	29.58	31.85	0.022	0.025	0.001	1.453
PNN	60.07	61.97	60.11	0.610	0.220	0.024	39.93	38.03	39.89	0.022	0.025	0.001	1.867
SOM	67.93	61.97	67.79	0.650	0.300	0.040	32.07	38.03	32.21	0.022	0.025	0.001	1.865
GFF	70.20	67.61	70.14	0.689	0.378	0.053	29.80	32.39	29.86	0.022	0.025	0.001	1.589
JEN	72.70	74.65	72.74	0.737	0.474	0.072	27.30	25.35	27.26	0.022	0.025	0.001	1.245

4.3.3. SLR_{BLR} + NN Hybrid

It is evident from Panel A of Table 6 that that 'SLR_{BLR}+MLP' hybrid has secured the best performance in terms of almost all criterions except GC rate, type–I error rate and MAE. Closely following 'SLR_{BLR}+MLP' hybrid is the most superior with BC rate (92.96%), ACC rate (99.77%), AUC value (0.965), AR (0.930), Kappa statistics (0.948), type–II error rate (7.04%), AMC (0.23%), MSE (0.002), RMSE (0.015), and with EMCC of 0.344; the ten performance criterions. Like NN single classifiers, SLR_{BLR}+RBF and SLR_{BLR}+LVQ are comparable from 100% GC rate, following by 0 type–I error with 0 MAE.

4.3.4. SLR_{BWD} + NN Hybrid

Panel B of Table 6 summarizes the results of the SLR_{BWD}–NN hybrid classifiers. 'SLR_{BWD}+MLP' hybrid performing the best in respect of 9 out of 13 performance criterions are in bold. Closely observing 'SLR_{BWD} + MLP' performs the best in BC rate (95.77%), ACC rate (99.84%), AUC value (0.979), AR (0.958), Kappa statistics (0.964), type–II error rate (4.23%), AMC (0.16%), MSE (0.002), and EMCC (0.207). For the remaining four criterions, 'SLR_{BWD}+LVQ' performs the best, such as, GC rate (100%) with no type–I error & MAE but very insignificant RMSE (0.004). By the following of 'SLR_{BWD}+LVQ', 'SLR_{BWD}+RBF' also performs the best with respect to three criterions those are same to the 'SLR_{BWD}+LVQ' except RMSE. A careful examination of these results reveal that 'SLR_{BWD}+MLP' hybrid is the most competitive and 'SLR_{BWD}+LVQ' & 'SLR_{BWD}+RBF' hybrids are relatively competitive in this type of hybrid classifiers.

4.3.5. SLR_{FLR} + NN Hybrid

It is surprising to see in Panel C of Table 6 that in 'SLR_{FLR}–NN' hybrid classifiers; 'SLR_{FLR}+JEN' hybrid produce, on average, significantly better results in terms of almost all criterions. This can clearly be seen from 98.65% whole sample correct classification rate which is from 99.77% good group classification and 50.70% bad group classification; 0.625 Kappa statistics; 1.35% average misclassification following of 0.23% type I error & 49.30% type II error; 0.013 MSE; 0.001 MAE; with expected misclassification cost of 2.409. However, like other hybrid classifiers, 'SLR_{FLR}+MLP' hybrid is on the front line with producing the best result in terms of seven criterions; *e.g.*, BC rate (70.42%), ACC rate (99.07%), Kappa statistics (0.770), type–II error rate (29.58%), AMC (0.93%), MSE

(0.009), and EMCC (1.445). Once more, ‘SLR_{FLR}+RBF’ & ‘SLR_{FLR}+LVQ’ hybrids show the identical results in terms of all criterions and they are good position from five performance criterions; *e.g.*, 100% good classification with no type–I error, 0.871 AUC, 0.742 AR, and with no MAE. Following of three popular hybrids, the another popular hybrid, ‘SLR_{FLR}+PNN’, also shows the significantly better results with securing the best positions in four criterions; such as, 100% GC rate, no type–I error, no RMSE, and no MAE.

Table 6. Performance Obtained by the SLR + NN Hybrid Classifiers

NN Credit scoring models	GC (%)	BC (%)	ACC (%)	AUC	AR	Kappa	T – I Er. (%)	T – II Er. (%)	AMC (%)	MSE	RMSE	MAE	EMCC
Panel A: Performance obtained by the SLR_{SLR} + NN hybrid classifiers													
MLP	99.93	**92.96**	**99.77**	**0.965**	**0.930**	**0.948**	0.07	**7.04**	**0.23**	**0.002**	**0.015**	0.001	**0.344**
RBF	**100**	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
LVQ	**100**	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
MNN	85.63	38.03	84.54	0.618	0.236	0.064	14.37	61.97	15.46	0.022	0.025	0.001	3.031
PNN	77.63	36.62	76.70	0.571	0.142	0.027	22.37	63.38	23.30	0.022	0.025	0.001	3.102
SOM	78.65	45.07	77.88	0.619	0.238	0.046	21.35	54.93	22.12	0.022	0.025	0.001	2.689
GFF	78.59	52.11	77.98	0.654	0.308	0.059	21.41	47.89	22.02	0.022	0.025	0.001	2.345
JEN	71.15	57.75	70.85	0.645	0.290	0.042	28.85	42.25	29.15	0.022	0.025	0.001	2.071
Panel B: Performance obtained by the SLR_{BWD} + NN hybrid classifiers													
MLP	99.93	**95.77**	**99.84**	**0.979**	**0.958**	**0.964**	0.07	**4.23**	**0.16**	**0.002**	0.015	0.001	**0.207**
RBF	**100**	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
LVQ	**100**	7.04	97.81	0.535	0.070	0.129	**0.0**	92.96	2.28	0.021	**0.004**	**0.000**	4.542
MNN	83.72	45.07	82.84	0.644	0.288	0.070	16.28	54.93	17.16	0.022	0.025	0.001	2.688
PNN	76.88	39.44	76.02	0.582	0.164	0.029	23.12	60.56	23.98	0.022	0.025	0.001	2.964
SOM	80.99	45.07	80.17	0.630	0.260	0.055	19.01	54.93	80.17	0.022	0.025	0.001	2.688
GFF	74.41	52.11	73.90	0.633	0.266	0.043	25.59	47.89	26.10	0.022	0.025	0.001	2.346
JEN	76.25	50.70	75.67	0.635	0.270	0.047	23.75	49.30	24.33	0.022	0.025	0.001	2.414
Panel C: Performance obtained by the SLR_{FLR} + NN hybrid classifiers													
MLP	99.74	**70.42**	**99.07**	0.851	0.702	**0.770**	0.26	**29.58**	**0.93**	**0.009**	0.025	0.001	**1.445**
RBF	**100**	0.0	97.72	**0.871**	**0.742**	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
LVQ	**100**	0.0	97.72	**0.871**	**0.742**	----	**0.0**	100	2.28	----	0.018	**0.000**	4.886
MNN	99.14	46.48	97.94	0.728	0.456	0.497	0.86	53.52	2.06	0.017	0.025	0.001	2.615
PNN	**100**	1.41	97.75	0.570	0.140	0.027	**0.0**	98.59	2.25	0.022	**0.000**	**0.000**	4.817
SOM	99.18	35.21	97.72	0.672	0.344	0.402	0.82	64.79	2.28	0.019	0.025	0.001	3.166
GFF	99.11	32.39	97.59	0.658	0.316	0.368	0.89	67.61	2.41	0.019	0.025	0.001	3.304
JEN	99.77	50.70	98.65	0.752	0.504	0.625	0.23	49.30	1.35	0.013	0.025	0.001	2.409
Panel D: Performance obtained by the SLR_{FWD} + NN hybrid classifiers													
MLP	99.84	**73.24**	**99.23**	0.865	0.730	**0.809**	0.16	**26.76**	**0.77**	**0.008**	0.025	0.001	**1.308**
RBF	**100**	0.0	97.72	**0.871**	**0.742**	----	**0.0**	100	2.28	----	**0.018**	**0.000**	4.886
LVQ	**100**	0.0	97.72	**0.871**	**0.742**	----	**0.0**	100	2.28	----	**0.018**	**0.000**	4.886
MNN	98.95	28.17	97.33	0.636	0.272	0.312	1.05	71.83	2.67	0.020	0.025	0.001	3.510
PNN	**100**	0.0	97.72	**0.871**	**0.742**	----	**0.0**	100	2.28	----	**0.018**	**0.000**	4.886
SOM	99.11	40.85	97.78	0.670	0.340	0.446	0.89	59.15	2.22	0.018	0.025	0.001	2.890
GFF	99.24	33.80	97.75	0.665	0.330	0.396	0.76	66.20	2.25	0.019	0.025	0.001	3.235
JEN	99.80	52.11	98.71	0.760	0.520	0.643	0.20	47.89	1.29	0.012	0.025	0.001	2.340

4.3.6. SLR_{FWD} + NN Hybrid

The results of SLR_{FWD}–NN hybrid classifiers are summarized in Panel D of Table 6 and showed that the BC rate (73.24%), ACC rate (99.23%), Kappa statistics (0.809), type–II error rate (26.76%), AMC rate (0.77%), MSE (0.008), and EMCC values (1.308); total of seven criterions; for the ‘SLR_{FWD}+MLP’ hybrid classifier are significantly better than those of other seven hybrid classifiers, and hence produce the best predictive results. Yet again, it is surprising that ‘SLR_{FWD}+RBF’, ‘SLR_{FWD}+LVQ’, and ‘SLR_{FWD}+PNN’ hybrids show the indistinguishable predictive performance and these hybrids are in good position in GC rate (100%) with 0 type I error, AUC value (0.871), AR (0.742), RMSE (0.018), and with 0 MAE; total of six criterions. Therefore, one can conclude that these three hybrids are relatively better in credit scoring application. Once more, the JEN based hybrid show the encouraging result which is supported by ACC rate (98.71%) with the

GC of 99.80%, Kappa statistics (0.643), AMC rate (1.29%), MSE (0.012), MAE (0.001), EMCC (2.340); consequently, the ‘SLR_{FWD}+JEN’ hybrid could be another competitive classification model for credit classification.

4.3.7. DT_{CHAID} + NN Hybrid

The DT_{CHAID}–NN hybrids results presented in Panel A of Table 7 reveal that the ‘DT_{CHAID}+MLP’ hybrid technique produces good results, with maximum values of ACC rate (98.52%), kappa statistics (0.616), AMC rate (1.48), MSE (0.014), and EMCC (2.271) among others. Alongside, ‘DT_{CHAID}+RBF’ & ‘DT_{CHAID}+LVQ’, two hybrids also illustrate the good extrapolative performance through 100% GC rate with 0 type I error, maximum AUC value (0.871) & AR (0.742), minimum RMSE (0.018) with 0 MAE. Yet again, in all respect, the very good average results are shown by the JEN based DT_{CHAID} hybrid.

4.3.8. DT_{QUEST} + NN Hybrid

Panel B of Table 7 lists the classification results of the DT_{QUEST}–NN hybrid models. Analytical results demonstrate that the ‘DT_{QUEST}+MLP’ hybrid outperforms the other hybrids in ACC rate (99.61%) with AMC rate (0.39%), AUC value (0.936), AR (0.872), Kappa statistics (0.910) and its MSE (0.004); six performance criterions. Like many times, ‘DT_{QUEST} +RBF’, & ‘DT_{QUEST}+LVQ’, the two hybrids has the best predictive results in GC rate (100%) with 0 type–I error, 0 MAE and minimum (0.018) RMSE; hence show the competitive position in credit scoring domain. Moreover, the ‘DT_{QUEST}+JEN’ hybrid model is superior in terms of BC rate (97.18%) with type–II error rate (2.82%), and the minimum EMCC (0.142). As a result, another time, it can be inferred that JEN based DT_{QUEST} hybrid could be an alternative for the classification application.

Table 7. Performance Obtained by the DT+ NN Hybrid Classifiers

NN Credit scoring models	GC (%)	BC (%)	ACC (%)	AUC	AR	Kappa	T – I Er. (%)	T – II Er. (%)	AMC (%)	MSE	RMSE	MAE	EMCC
Panel A: Performance obtained by the DT_{CHAID} + NN hybrid classifiers													
MLP	99.57	**53.52**	**98.52**	0.766	0.532	**0.616**	0.43	**46.48**	**1.48**	**0.014**	0.025	0.001	**2.271**
RBF	**100**	0.0	97.72	**0.871**	**0.742**	----	**0.0**	100	2.28	----	**0.018**	**0.000**	4.886
LVQ	**100**	0.0	97.72	**0.871**	**0.742**	----	**0.0**	100	2.28	----	**0.018**	**0.000**	4.886
MNN	98.68	50.70	97.59	0.747	0.494	0.477	1.32	49.30	2.41	0.017	0.025	0.001	2.409
PNN	99.90	0.0	97.62	0.500	0.000	-0.002	0.10	100	2.38	0.022	0.025	0.001	4.886
SOM	98.75	39.44	97.40	0.691	0.382	0.395	1.25	60.56	2.60	0.019	0.025	0.001	2.959
GFF	98.85	42.25	97.56	0.706	0.412	0.429	1.15	57.75	2.44	0.018	0.025	0.001	2.822
JEN	99.57	46.48	98.36	0.730	0.460	0.556	0.43	53.52	1.64	0.015	0.025	0.001	2.615
Panel B: Performance obtained by the DT_{QUEST} + NN hybrid classifiers													
MLP	99.90	87.32	**99.61**	**0.936**	**0.872**	**0.910**	0.10	12.68	**0.39**	**0.004**	0.025	0.001	0.620
RBF	**100**	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	**0.018**	**0.000**	4.886
LVQ	**100**	0.0	97.72	0.871	0.742	----	**0.0**	100	2.28	----	**0.018**	**0.000**	4.886
MNN	52.43	95.77	53.42	0.741	0.482	0.044	47.57	4.23	46.58	0.022	0.025	0.001	0.218
PNN	20.20	**97.18**	21.95	0.587	0.174	0.010	79.80	**2.82**	70.05	0.022	0.022	0.001	0.156
SOM	40.43	95.77	41.69	0.681	0.362	0.027	59.57	4.23	58.31	0.022	0.025	0.001	0.220
GFF	32.43	**97.18**	33.91	0.648	0.296	0.020	67.57	**2.82**	66.09	0.022	0.022	0.001	0.153
JEN	82.53	**97.18**	82.87	0.899	0.798	0.172	17.47	**2.82**	17.13	0.020	0.022	0.001	**0.142**

4.4. Comparisons of the Ten Best Credit Classifiers

Table 8 compares the ten best models in the ten types of credit scoring models in terms of thirteen performance matrixes. Regarding the comparative results, the ‘SDA_{FR}+MLP’ hybrid model performs the best for the whole comparisons (higher GC, higher BC, higher average accuracy, higher AUC value, higher AR, higher Kappa statistics, lower type–I error, lower type–II error, lower average misclassification, lower MSE, lower RMSE, lower MAE, and with a lower value of EMCC). Figures. 2–5 display the average performance indicators, *i.e.*, GC and BC rate, ACC rate, type I error and type II error, AMC results, for the ten best models. In particular, *F*-statistic of stepwise discriminant method used as the first component combined with MLP based neural networks as the second component, is the most superior to the other models. On the other hand, 'SDA_{FR}+MLP' hybrid model provides the lowest type I and type II errors (see Figure 4) which means that it can accurately control the credit risk and minimize the loss of good customers. Therefore, this hybrid model can be regarded as the optimal credit scoring model.

Table 8. Performance of the Ten Best Models

Credit scoring models	Classifier	GC (%)	BC (%)	ACC (%)	AUC	AR	Kappa	T – I Er (%)	T – II Er (%)	AMC (%)	MSE	RMSE	MAE	EMCC
SLR_{ALL}	Base classifier	99.54	67.61	98.81	0.836	0.672	0.716	0.46	32.39	1.19	0.011	0.025	0.001	1.583
MLP	NN classifier	99.77	88.73	99.52	0.943	0.886	0.891	0.23	11.27	0.48	0.005	0.015	0.001	0.551
JEN	SDA_{WL}+NN hybrid	99.97	33.80	98.46	0.669	0.338	0.494	0.03	66.20	1.54	0.015	0.025	0.001	3.235
MLP	SDA_{FR}+NN hybrid	**99.97**	**95.77**	**99.87**	**0.979**	**0.958**	**0.971**	**0.03**	**4.23**	**0.13**	**0.001**	**0.015**	**0.001**	**0.207**
MLP	SLR_{BLR}+NN hybrid	99.93	92.96	99.77	0.965	0.930	0.948	0.07	7.04	0.23	0.002	0.015	0.001	0.344
MLP	SLR_{BWD}+NN hybrid	99.93	95.77	99.84	0.979	0.958	0.964	0.07	4.23	0.16	0.002	0.015	0.001	0.207
MLP	SLR_{FLR}+NN hybrid	99.74	70.42	99.07	0.851	0.702	0.770	0.26	29.58	0.93	0.009	0.025	0.001	1.445
MLP	SLR_{FWD}+NN hybrid	99.84	73.24	99.23	0.865	0.730	0.809	0.16	26.76	0.77	0.008	0.025	0.001	1.308
MLP	DT_{CHAID}+NN hybrid	99.57	53.52	98.52	0.766	0.532	0.616	0.43	46.48	1.48	0.014	0.025	0.001	2.271
MLP	DT_{QUEST}+NN hybrid	99.90	87.32	99.61	0.936	0.872	0.910	0.10	12.68	0.39	0.004	0.025	0.001	0.620

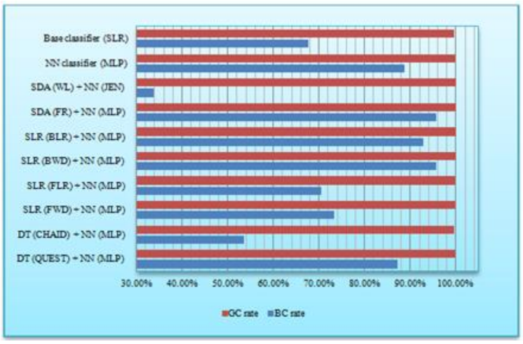

Figure 2. GC and BC Results of Ten Best Models

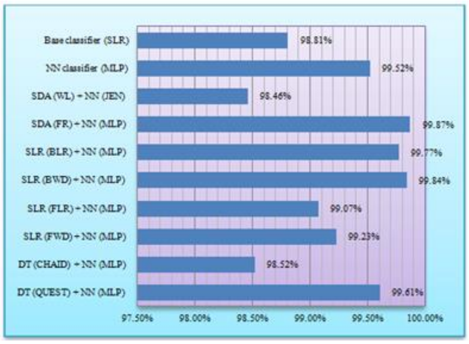

Figure 3. ACC Results of Ten Best Models

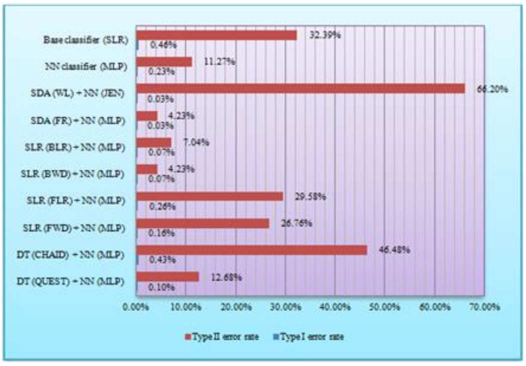

Figure 4. Type I Error and Type II Error of Ten Best Models

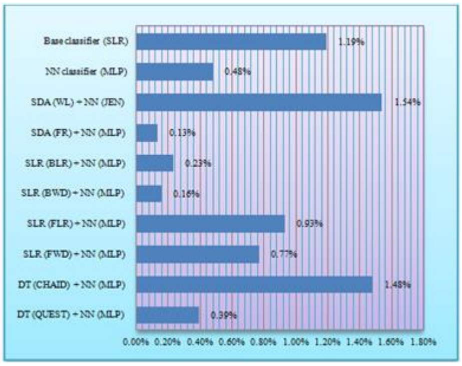

Figure 5. AMC Results of Ten Best Models

Moreover, on average, the difference in the performance between the 'SDA_{FR}+MLP' hybrid model and MLP based NN model is the least in comparison to the other seven models and their corresponding NN models. And this also proves that MLP based hybrid, *i.e.*, 'MLP+ SDA_{FR}' hybrid model is the most impressive technique for credit scoring. Although the 'MLP +SDA_{FR}' hybrid is the best model, we should note that, the difference in the performance criterions (Particularly in confusion matrix components) with its counterparts is the minimum. We believe that it arises from the independent variables. In the meanwhile, an improvement in accuracy of even a fraction of a percent translates into significant future savings and a central concern of these applications is the need to increase the scoring accuracy of the small business credit decision.

However, from Table 8, it is also evident that the SDA_{WL}–NN hybrid classifiers, DT_{CHAID}– NN hybrid classifiers, base classifiers, SLR_{FLR}–NN hybrid classifiers, and SLR_{FWD}–NN hybrid classifiers show limited power in correctly classifying bad observations and this produces the higher Type–II error rate (see Figures. 2 and 4) although it is more important to classify a bad applicant correctly than it is to classify a good applicant correctly. The reason may be that the application of these classifiers overfits the dataset especial to bad instances.

In our experiments we also found an interesting phenomenon that JEN based hybrid, a novel hybrid model for credit scoring, show the very good average results with securing the position in top ten models. This implies that the JEN based hybrid can provide sensible credit scoring results and the policy maker can consider it as an alternative model for credit classification.

Besides, these empirical results also indicate that SDA based *F* ratio method is the most optimal method to carry out feature selection for credit default prediction. While, SLR based BWD method comes secondly. SDA of WL is the third choice to find optimal feature subset for the credit scoring. And CHAID or QUEST of DT is not ideally applied. It also means that the filter approach is more applicable than the embedded approach to be used in the selection procedure for optimal feature subset to some extent. These results are consistent with Liang *et al*. [37], who compared filter and wrapper feature selection methods for credit scoring and bankruptcy application and found the filter based feature selection for credit scoring to be the optimal feature selection method.

Briefly it can be concluded that almost all hybrid models have a superior performance in terms of all performance metrics than their corresponding base and NN classifiers; however, the degree of improvement was dependent on the form of hybridization. Besides, hybrid models were more robust than their counterparts by reducing the variance of the single and NN classifiers. These characteristics of hybrids are very enviable for stake holders.

5. Conclusions

In this work, a new methodology for credit assessment has been developed by combining different classifier as multiple hybrid models, with the aim of obtaining better performance results than the single classifier. This can be viewed as a multiple hybrid approach that combines feature selection strategies with the classifiers for the construction of composite hybrids. In particular, SDA, SLR and DT have been taken as representatives of feature selection algorithms, an integrated approach that embeds expert knowledge with the filter method whereas the eight popular NN models have been used as classifiers.

Some interesting conclusions can be drawn from the experiments carried out. In general, almost all hybrid models have a superior performance but among of them the 'SDA_{FR} + MLP' hybrid have produced the best results in terms of all thirteen performance criterions, a noteworthy strength of this study, what may lead to significant cost savings in credit scoring applications for financial institutions. Since the construction of particular hybrid model is important, it seems that the jointly use of SDA with MLP in any order performs the best with comparing of other combinations. A final indication from the experiments is that using the DT algorithm with NN classifier is clearly worse than the SDA or SLR hybridization with NN. The above-mentioned research findings justify the presumptions that 'SDA_{FR} + MLP', the hybrid methods should be the best alternative in opposition to base classifiers and NN classifiers in conducting credit prediction tasks.

Moreover, in this study, we make some additional contributions to the credit scoring and financial distress literature. First, we note that the existing literature has compared the relative performance of various subsets of models. We execute a more comprehensive assessment of the performance of models through a large set of data with some new NN architectures, especially MNN, GFF, JEN. Second, we develop a number of additional appraisal metrics, which provide some new perspectives on the relative performance of the various approaches. Third, our study shows an extensive comparison among all models which can assist to take a feasible decision and produce a large financial and other benefits to organizations through credit approval, loan portfolio and security management.

Potential future directions of this study include (i) enhancing the information sources by including the data from different regions, particularly with more exploration of credit scoring data structures, (ii) extending the hybridization with different classifiers such as genetic algorithm, support vector machine, fuzzy neural network, *etc.*, and perhaps most importantly, (iii) the deployment of the system as a decision aid for administrators to assess its suitability and usability in real-world.

Acknowledgements

The research is supported by the support of the National Natural Science Foundation of China (71171031 and 71471027), National Social Science Foundation of China (16BTJ017), Social Science Foundation of Liaoning province of China (L16BJY016), Credit Risks Rating System and Loan Pricing Project of Small Enterprises for Bank of Dalian (2012–01) and Credit Risks Evaluation and Loan Pricing For Petty Loan Funded for the Head Office of Postal Savings Bank of China (2009–07). We thank the organizations mentioned above.

References

[1] P. Qiao, X. Ju and G. Fung, "Industry Association Networks, Innovations, and Firm Performance in Chinese Small and Medium-Sized Enterprises", China Economic Review, vol. 29, **(2014)**, pp. 213-228.
[2] M. Z. Abedin, C. Guotai and M. Bin, "Credit Default Prediction of Chinese Small Business: A Neural Network Methodology", European Journal of Economics, Finance and Administrative Sciences, vol. 77, **(2015)**, pp. 32-50.
[3] M. Khashei, A. Hamadani and M. Bijari, "A Novel Hybrid Classification Model of Artificial Neural Networks and Multiple Linear Regression Models", Expert Systems with Applications, doi:10.1016/j.eswa.2011.08.116, vol. 39, **(2012)**, pp. 2606-2620.
[4] H. Abdou, J. Pointon and A. E. Masry, "Neural nets versus conventional techniques in credit scoring in Egyptian Banking", Expert Systems with Applications, doi:10.1016/j.eswa.2007.08.030, vol. 35 (**2008**), pp. 1275-1292.
[5] V. Desai, J. Crook and Jr. Overstreet, "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment", European Journal of Operational Research, vol. 95, no. 1, **(1996)**, pp. 24-37.
[6] T. S. Lee and I. F. Chen, "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines", Expert Systems with Applications, doi:10.1016/j.eswa.2004.12.031, vol. 28, **(2005)**, pp. 743-752.
[7] T. S. Lee, C. C. Chiu, C. J. Lu and I. F. Chen, "Credit scoring using the hybrid neural discriminant technique", Expert Systems with Applications, vol. 23, **(2002)**, pp. 245-254.
[8] R. Malhotra and D. K. Malhotra, "Evaluating consumer loans using neural networks", Omega, doi:10.1016/S0305-0483(03)00016-1, vol. 31, **(2003)**, pp. 83-96.
[9] K. Y. Tam and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions", Management Science, vol. 38, **(1992)**, pp. 926-947.
[10] D. West, "Neural Network Credit Scoring Models", Computers & Operations Research, vol. 27, **(2000)**, pp. 1131-1152.
[11] G. Zhang, M. Y. Hu, B. E. Patuwo and D. C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis", European Journal of Operational Research, vol. 116, **(1999)**, pp. 16-32.
[12] F. Lasheras, J. Andrés, P. Lorca and F. Juez, "A Hybrid Device for the Solution of Sampling Bias Problems in the Forecasting of Firms' Bankruptcy", Expert Systems with Applications, doi:10.1016/j.eswa.2012.01.135, vol. 39, **(2012)**, pp. 7512-7523.
[13] M. Bensic, N. Sarlija and M. Susac, "Modeling Small Business Credit Scoring by using Logistic Regression, Neural Networks and Decision Trees", Intelligent Systems in Accounting, Finance and management, vol. 13, **(2005)**, pp. 133-150.
[14] M. Susac, N. Sarlija and M. Bensic, "Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models", 26th Int. Conf. Inf. Tec. Inter. (ITI), Cavtat, Croatia, **(2004)**, pp. 265-270.
[15] M. Boyacioglu, Y. Karaand and K. Baykan, "Predicting Bank Financial Failures using Neural Networks, Support Vector Machines and Multivariate Statistical Methods: A Comparative Analysis in the Sample of Savings Deposit Insurance Fund (SDIF) Transferred Banks in Turkey", Expert Systems with Applications, doi:10.1016/j.eswa.2008.01.003, vol. 36, **(2009)**, pp. 3355-3366.
[16] P. Hájek, "Municipal credit rating modelling by neural networks", Decision Support Systems, doi:10.1016/j.dss.2010.11.033, vol. 51, **(2011)**, pp. 108-118.
[17] C. Guotai, M. Z. Abedin and F. E. Moula, "Modeling Credit Approval Data with Neural networks: An Experimental Investigation and Optimization", Journal of Business Economics and Management, Under Review, **(2016)**.
[18] Z. R. Yang, M. B. Platt and H. D. Platt, "Probabilistic Neural Networks in Bankruptcy Prediction", Journal of Business Research, vol. 44, **(1999)**, pp. 67-74.
[19] K. Lee, D. Booth and P. Alam, "A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms", Expert Systems with Applications, doi:10.1016/j.eswa.2005.01.004, vol. 29, **(2005)**, pp. 1-16.

[20] H. A. Abdou, "An evaluation of alternative scoring models in private banking", Journal of Risk Finance, doi: 10.1108/15265940910924481, vol. 10, no. 1 **(2009)**, pp. 38-53.
[21] F. M. Tseng and Y. C. Hu, "Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks", Expert Systems with Applications, doi:10.1016/j.eswa.2009.07.081, vol. 37, **(2010)**, pp. 1846-1853.
[22] P. Ravisankar, V. Ravi, G. R. Rao and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques", Decision Support Systems, doi:10.1016/j.dss.2010.11.006, vol. 50, **(2011)**, pp. 491-500.
[23] D. Specht, "Probabilistic Neural Networks", Neural Networks, vol. 3, **(1990)**, pp. 109-118.
[24] T. Kohonen, "Self-Organizing Maps", Springer, Germany, **(1997)**.
[25] J. Principe, N. Euliano and W. Lefebvre, "Neural and Adaptive Systems: Fundamentals through Simulations", John Wiley and Sons, New York, **(1999)**.
[26] M. Jordan, "Attractor Dynamics and Parallelism in a Connectionist Sequential Machine", In: Proceedings of the 8th annual conference of the cognitive science society, **(1986)**, pp. 531-546.
[27] J. Elman, "Finding Structure in Time", Cognitive Science, vol. 14, **(1990)**, pp. 179-211.
[28] M. Kim, S. Min and I. Han, "An Evolutionary Approach to the Combination of Multiple Classifiers to Predict a Stock Price Index", Expert Systems with Applications, vol. 31, no. 2, (**2006**), pp. 241-247.
[29] S. Oreski and G. Oreski, "Genetic Algorithm-based Heuristic for Feature Selection in Credit Risk Assessment", Expert Systems with Applications, vol. 41, **(2014)**, pp. 2052-2064.
[30] J. Prado, V. Alcaˆntara, F. Carvalho, K. Vieira, L. Machado and D. Tonelli, "Multivariate Analysis of Credit Risk and Bankruptcy Research Data: A Bibliometric Study Involving Different Knowledge Fields (1968–2014)", Scientometrics, doi: 10.1007/s11192-015-1829-6. vol. 106, **(2016)**, pp. 1007-1029.
[31] C. Ferri, J. Orallo and R. Modroiu, "An Experimental Comparison of Performance Measures for Classification", Pattern Recognition Letters, vol. 30, **(2009)**, pp. 27-38.
[32] H. Abdou, "Genetic Programming for Credit Scoring: The Case of Egyptian Public Sector Banks", Expert Systems with Applications, vol. 36, **(2009)**, pp. 11402-11417.
[33] J. Cohen, "A Coefficient of Agreement for Nominal Scales", Educational and Psychological Measurement, vol. 20, no. 1, **(1960)**, pp. 37-46.
[34] E. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", Journal of Finance, vol. 13, **(1968)**, pp. 589-609.
[35] R. Beaver, "Financial Ratios as Predictors of Failure", Journal of Accounting Research, vol. 4, **(1966)**, pp. 71-111.
[36] S. Akkoc, "An Empirical Comparison of Conventional Techniques, Neural Networks and the Three Stage Hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) Model for Credit Scoring Analysis: The Case of Turkish Credit Card Data", European Journal of Operational Research, vol. 222, **(2012)**, pp. 168-178.
[37] D. Liang, C. Tsai and H. Wu, "The Effect of Feature Selection on Financial Distress Prediction", Knowledge Based Systems, vol. 73, **(2015)**, pp. 289-297.

Authors

Guotai Chi, (1955–) Male, PhD, Professor of Finance, Doctoral Adviser, Dalian University of Technology, Dalian 116024, China. His research interest includes financial risk management, asset – liability management.

Mohammad Zoynul Abedin, (1980–) Male, Investment theory doctor graduate student, Dalian University of Technology, Dalian 116024, China. His research interest includes financial risk management, credit analysis.

Fahmida–E–Moula, (1988–), Female, Investment theory doctor graduate student, Dalian University of Technology, Dalian 116024, China. His research interest includes financial risk management, credit analysis.

