Received 9 September 2020
Revised 14 December 2020
Accepted 21 February 2021

Prediction of credit risk with an ensemble model: a correlation-based classifier selection approach

Zhibin Xiong
School of Mathematical Sciences,
South China Normal University – Shipai Campus, Guangzhou, China, and
Jun Huang
Department of Management and Marketing, Angelo State University,
San Angelo, Texas, USA

Abstract

Purpose – Ensemble models that combine multiple base classifiers have been widely used to improve prediction performance in credit risk evaluation. However, an arbitrary selection of base classifiers is problematic. The purpose of this paper is to develop a framework for selecting base classifiers to improve the overall classification performance of an ensemble model.

Design/methodology/approach – In this study, selecting base classifiers is treated as a feature selection problem, where the output from a base classifier can be considered a feature. The proposed correlation-based classifier selection using the maximum information coefficient (MIC-CCS), a correlation-based classifier selection under the maximum information coefficient method, selects the features (classifiers) using nonlinear optimization programming, which seeks to optimize the relationship between the accuracy and diversity of base classifiers, based on MIC.

Findings – The empirical results show that ensemble models perform better than stand-alone ones, whereas the ensemble model based on MIC-CCS outperforms the ensemble models with unselected base classifiers and other ensemble models based on traditional forward and backward selection methods. Additionally, the classification performance of the ensemble model in which correlation is measured with MIC is better than that measured with the Pearson correlation coefficient.

Research limitations/implications – The study provides an alternate solution to effectively select base classifiers that are significantly different, so that they can provide complementary information and, as these selected classifiers have good predictive capabilities, the classification performance of the ensemble model is improved.

Originality/value – This paper introduces MIC to the correlation-based selection process to better capture nonlinear and nonfunctional relationships in a complex credit data structure and construct a novel nonlinear programming model for base classifiers selection that has not been used in other studies.

Keywords Risk analysis, Nonlinear programming, Modeling, Data analytics, Credit risk prediction, Ensemble model, Correlation-based classifier selection, Maximum information coefficient, Nonlinear optimization

Paper type Research paper

Journal of Modelling in Management
Vol. 17 No. 4, 2022
pp. 1078-1097

1746-5664
DOI 10.1108/JM2-09-2020-0235

This work is financially supported by the Ministry of Education of the People's Republic of China in Humanities and Social Sciences Planning Fund (No. 16YJA790053).

1. Introduction

Credit risk management has always been a challenging topic in the field of risk management. Assessing credit risk before issuing loans to individuals or firms has become a major task for banks and financial institutions. The failures of banks, mortgage lenders and insurers during the recent global financial crisis have revealed the importance of credit evaluation when making lending decisions. A widely adopted method for evaluating credit risk is that of using classification models. Specifically, several modeling techniques can be used to classify evaluation targets into predefined classes and identify applicants who pose a risk of default in the future.

There are two main types of techniques to build a classification model. Traditional techniques based on statistical models such as logistic regression (LR) and linear discriminant analysis (LDA), were first applied in credit risk evaluation(Altman, 1968; Martin, 1977). However, most statistical approaches usually impose assumptions that restrict model performance in analyzing credit data. Subsequently, techniques based on machine learning such as naïve Bayes (NB); K-nearest neighbors (KNN); support vector machine (SVM); neural networks (NN) and tree-based machine learning algorithms such as C5.0 decision tree (DT), random forest (RF) or gradient boosting machine (GBM), have been widely applied in credit risk modeling (Paleologo *et al.*, 2010; Verikas *et al.*, 2011; Geng *et al.*, 2015; Liang *et al.*, 2016; Maldonado *et al.*, 2017; Chang *et al.*, 2018; Fu *et al.*, 2020; Climent *et al.*, 2019; Carmona *et al.*, 2019).

However, the mixed results of the empirical studies on classifier performance reveal that using a single classifier to evaluate credit data may lack stability due to the complex nature, diversity and interrelationships of the credit data structure, with no stand-alone classifier being able to consistently outperform others across different credit data sets (Finlay, 2011; Tsai *et al.*, 2014; Abellán and Castellano, 2017). Therefore, multiple classifier systems (MCS) or ensemble models that combine multiple classifiers have gained popularity in recent years. As numerous classification algorithms can be used as base classifiers, the question is whether the selection of base classifiers in an ensemble model affects the accuracy of the prediction accuracy. An arbitrary selection of base classifiers may lead to either information loss due to lack of diversity or data redundancy when including too many similar classifiers, affecting the predictive power of an ensemble model.

In this study, we aim to improve the classification performance of the ensemble model by developing a correlation-based classifier selection using the maximum information coefficient (MIC-CCS) method. The concept of feature selection is applied to the selection of base classifiers, where the output of a base classifier is considered a feature. The correlation-based feature selection method finds out classifiers in such a way that the correlations between the outputs of each classifier are low and meanwhile the correlations between the outputs from each classifier and the actual value of the target variable (original class labels) are high. When measuring the correlation, the classic correlation measurements such as Pearson and Spearman coefficient fail to capture nonlinear or nonfunctional relationships and are sensitive to outliers.MIC proposed by Reshef *et al.* (2011), on the other hand, is based on mutual information which can capture linear, nonlinear and nonfunctional relationships between variables and far less sensitive to outliers. It has been confirmed to be an effective measure for detecting associations between variables in many fields (Sun *et al.*, 2018). However, very few studies in credit evaluation have applied MIC. Therefore, we introduce MIC to measure the correlations in this study due to the complex nature of the credit data structure. The MIC-CCS method then selects the features using a nonlinear programming model based on the correlation matrix, which is measured using MIC. Our proposed method, which balances the accuracy and diversity of the base classifiers, improves the performance

of the ensemble model by effectively selecting base classifiers that are significantly different and yet, have high levels of predictive accuracy.

The main contributions of our work include the following two points. First, many previous studies on credit risk evaluation using ensemble models focus on the strategies for combining classifiers; few have hitherto discussed the selection of the base classifiers. Our study provides an alternate solution to improve the classification performance of the ensemble model by effectively selecting base classifiers. Second, we introduce MIC to the correlation-based selection process to better capture nonlinear and nonfunctional relationships in a complex data structure and construct a novel nonlinear programming model for base classifiers selection that has not been used in other studies.

The rest of the paper is organized as follows. Section 2 briefly reviews several studies on credit risk evaluation using various classification models. In Section 3, we introduce the proposed classifier selection method. The experimental setup and data sets are described in Section 4. In Section 5, we present and discuss the results. Finally, conclusions and future research directions are provided in Section 6.

2. Related works

2.1 Comparison of stand-alone classifiers

Several studies have discussed assessment techniques for improving the performance of credit risk models such as a credit scoring system for consumer default risk prediction at the individual level and bankruptcy prediction and financial distress prediction at the firm level.

Although many different models (classifiers) can be used for credit risk modeling, existing studies have not yet identified a classifier that consistently outperforms the others (Finlay, 2011). In fact, different studies have different conclusions. As a prevalent data mining technique, NN has been extensively studied and compared with many other classifiers. Tam and Kiang (1990) predicted bank failures in Texas using an NN algorithm and compared its performance with that of the LDA, LR, KNN and DT algorithms. They found that NN has a better performance in predicting bankruptcy using data from the oil and gas industry in the U.S. Further, Lee *et al.* (2005) studied bankruptcy among 168 Korean companies using NN and compared it with LR and LDA. Their results showed that NN outperforms the other two models. In a recent study, Le and Viviani (2018) compared the performance of traditional statistical and machine learning techniques in predicting bank failure. The empirical results revealed that NN and KNN are better classifiers than LDA and LR. Hyewon and Zheng (2010) evaluated the financial distress of U.S. restaurants using LR and NN, showing that LR has a similar performance to NN in terms of prediction accuracy. Yang *et al.* (1999) compared the performances of NN and LDA, finding that LDA yields the best overall estimation results. Jing and Fang (2018) built models to predict bank failures in the U.S. using a logit model, NN and SVM. Their results indicate that the logit model outperforms the other two.

SVM has also been commonly used and compared with other techniques. For instance, Bellotti and Crook (2009) assessed the default risk on credit using different techniques, including SVM, LR and discriminant analysis. They found that SVM performed competitively and can be used to determine default risk. Shin *et al.* (2005) compared the performances of NN and SVM in predicting bankruptcy. Their results showed that SVM performs better than NN as the size of the training set decreases. Olson *et al.* (2012) found that DT performed better than NN and SVM in terms of prediction accuracy, transparency and transportability. Whiting *et al.* (2012) reported that RF provides better accuracy and interpretability. Based on 121 data sets, Fernández-Delgado *et al.* (2014) examined 179 classifiers, including SVM, NN, DT, RF, LDA, NB and LR. They found RF to be the best

classifier. Huang *et al.* (2017) evaluated the financial distress of Chinese firms with RF, NN, SVM, DT (C5.0), LDA and LR and concluded that RF and SVM perform best in predicting financial distress compared to the other methods.

2.2 Multiple classifier systems

There are multiple reasons why the studies on classifier selection yield different conclusions. For example, different studies use different samples and sampling methods, different time windows, include different variables in their models, etc. However, a key factor that causes these differences is the complex nature, diversity and interrelationships of the credit data structure (Chen *et al.*, 2016). Therefore, using a single classifier to evaluate credit data may lack stability, with empirical studies revealing that no single classifier is able to consistently outperform all others across different credit data sets. Samples misclassified by one classifier may be classified correctly by another. In other words, different classifiers may provide complementary classification information. Therefore, an increasing number of researchers are using MCS, also known as ensemble models, for credit risk modeling to improve classification performance. In MCS, each classifier is called a base or elementary classifier. Multiple classifiers are trained separately and aggregated in an ensemble for the final output.

Recent empirical evidence shows that ensemble models with multiple classifiers usually show better performance than models based on single classifiers for credit risk evaluation. Ramli *et al.* (2015) designed an early warning system for the currency crisis. They explored three ensemble models that combined SVM with KNN, LR with KNN and LAD Tree with KNN, whose outputs are further combined by linear regression as a meta-classifier. They found that these three ensemble models have a similar performance in terms of accuracy; however, LAD Tree-KNN has a higher area under the ROC curve (AUC) than the other two. However, all three ensemble models perform better than single classifiers. Liang *et al.* (2018) proposed a classifier ensemble approach to predict financial distress. Their method combined the outputs from multiple classifiers using a unanimous voting (UV) ensemble method in which the final prediction output is produced by choosing the class that all classifiers agree on, as opposed to the majority voting method. Their experimental results showed that UV outperforms the stand-alone classifiers and classifier ensembles based on other ensemble strategies including bagging, boosting, stacking and majority voting. Pławiak *et al.* (2019) created a 16-level cascade structure ensemble model to predict credit scoring. In their model, each level contained multiple classifiers based on the combination of two types of SVM classifiers and different hyperparameters, including different kernel functions, feature extraction methods and normalization types, which are optimized using a genetic algorithm. They reported that the accuracy rate derived from their ensemble model is much higher than that of any single machine-learning method. Shen (2019) proposed an ensemble model that combines multiple BPNNs (backpropagation neural networks) generated by the Ada Boost technique to classify credit data. They compared the performance of their method and other commonly used classifiers including LR, LDA, KNN, NB, SVM, classification tree (CT) and a stand-alone BPNN, finding that the proposed ensemble model outperforms the other stand-alone models. Additionally, a large body of literature has also reported the advantages of using ensemble models over single classification models (Zhou *et al.*, 2002; West and Dellana, 2009; Wang and Ma, 2012; Zhang *et al.*, 2019; Papouskova and Hajek, 2019; Montebruno *et al.*, 2020; du Jardin, 2019).

2.3 Strategies for combining multiple classifiers

Generally, the ensemble strategies that integrate multiple classifiers include voting, cascading and stacking. A voting strategy usually uses a single base learning algorithm and creates multiple classification models by varying the training data, using techniques such as bagging, boosting or random subspace, based on resembling, replication techniques or feature subspaces. These classification models are then combined in a voting manner, typically by weighted or unweighted voting (Papouskova and Hajek, 2019). A cascading strategy is characterized by the concatenation of multiple classifiers. That is, the first-level classifier's output is passed to the next level classifier and used as its input (Melville and Mooney, 2005). Under a stacking strategy, the outputs from multiple first-level classifiers (or base classifiers) are passed to the second-level classifier (or meta-classifier), where the outputs from the base classifiers are trained to combine. We chose the stacking strategy to combine heterogeneous classifiers. Hsu and Srivastava (2009) discussed the combination of heterogeneous classifiers from both theoretical and empirical aspects and reported that building ensemble models with heterogeneous classifiers increases diversity, and thus, performance, unlike using homogeneous classifiers. Xia *et al.* (2018) proposed a heterogeneous ensemble model that integrates the bagging algorithm under the stacking method. Their results also demonstrated the superiority of using heterogeneous classifiers. Nti *et al.* (2020) constructed 25 ensemble models using different ensemble strategies, namely, boosting, bagging and stacking to combine DT, SVM and NN. Their result revealed that heterogeneous ensemble classifiers under the stacking scheme outperform the other two combination techniques. In sum, a heterogeneous ensemble increases the complementarity and diversity of the base classifiers while stacking estimates and corrects the bias in the base classifier, thus improving prediction accuracy (Oza and Tumer, 2008; Papouskova and Hajek, 2019).

It is worth pointing out that when building ensemble models in the field of credit evaluation, some studies select base classifiers merely based on the literature. However, different studies apply different classifiers in different situations. Table 1 summarizes various base classifiers used by different studies. Some select base classifiers are largely based on the accuracy of individual classifiers, ignoring diversity, which may provide complementary information or, conversely, overemphasize diversity to the detriment of overall accuracy.

3. Methods

This paper attempts to improve the performance of the ensemble model by proposing a correlation-based classifier selection under the MIC-CCS method to select base classifiers whose outputs are then combined by the meta-classifier. Selecting base classifiers to build an ensemble model is treated as a feature selection problem, that is, the output – the probability estimations from a base classifier – can be considered a feature. The criteria used to build the classifier selection model are twofold. On one hand, the correlations between the outputs of each classifier should below so that the selected classifiers show significant differences and can, thus, provide complementary information. On the other hand, the correlations between the outputs from each classifier and the actual value of the target variable should be high, indicating these classifiers have a decent predictive capability for class labels. By doing so, the ensemble model built based on the selected base classifiers ensures not only the prediction accuracy of the base classifiers but also their diversity, thus providing complementary information. The detailed procedure is as follows.

In the first phase, each candidate base classifier participates in the J-fold cross-validation training. Given a set $D = (x_i, y_i), i = 1 \ldots k\}$, we randomly divide it into J subsets in similar sizes: D_1, D_2, $\cdots$, D_J. Let D_j and $D^{(-j)}$ (j = 1,2, $\cdots$, J) be jth testing and training sets, respectively. Assume n candidate classifiers and let $h_r^{(-j)}$ $(r = 1, 2, \cdots, n)$ be the rth

Studies	Base classifiers
Marqués *et al.* (2012)	1-nearest neighbor (1_NN), NB, LR, multilayer perceptron neural network (MLP), radial basis function neural network (RBF), SVM and C4.5
Xie *et al.* (2013)	LR, SVM and MLP
Chen *et al.* (2015)	LR, C4.5 DT, NB, KNN, RBF, MLP and SVM
Florez-Lopez and Ramon-Jeronimo (2015)	Chi-square Automatic Interaction Detector (ChAID), Assistant (based on ID3), C4.5 and Classification and Regression Trees (CART)
Ramli *et al.* (2015)	SVM, K-NN, LADTree and LR
Ala'raj and Abbod (2016)	SVM, NN, DT, RF and NB
Xiao *et al.* (2016)	DT, LR and SVM
Ekinci and Erdal (2017)	LR, J48 and voted perceptron
Abellán and Castellano (2017)	LR, MLP, SVM, C4.5 and creedal decision tree (CDT)
Liang *et al.* (2018)	SVM, MLP and CART
Zhang *et al.* (2018)	LR, SVM, NN, gradient boosting decision tree and RF
Zhang *et al.* (2019)	DT, KNN, NN, SVM, NB, quadratic discriminant analysis (QDA), LDA and parzen classifier
Papouskova and Hajek (2019)	LR, regression trees (RT), support vector regression (SVR) and MLP
Nti *et al.* (2020)	DT, SVM and NN

Table 1. Related studies of ensemble models using various base classifiers

classifier in $D^{(-j)}$ and $h_r(x_i)$ represents the corresponding output (probability estimations) for x_i in D_j using the classifier $h_r^{(-j)}$. The output of all candidate classifiers along with the corresponding original labels leads to a new training data set, $T = (h_1(x_i), h_2(x_i), \cdots, h_n(x_i), y_i), i = 1, 2, \cdots, k\}$. The correlation coefficient matrix based on data set T is then passed on to our proposed MIC-CCS model.

We introduce MIC to measure the correlation, k_{yi}, k_{ij}, between variables. MIC can detect the redundancy and relevance between two variables X and Y. Additionally, it captures not only linear correlations but also nonlinear and nonfunctional ones (Sun *et al.*, 2018). The rationale is that the relationship between two variables can be encapsulated by drawing a grid on the scatter plot of these two variables (Reshef *et al.*, 2011).

Given the observations of two continuous random variables in F: (X,Y), the x-y plane is divided into several smaller grids, so that all observations in data set F to fall into grid G. Specifically, the continuous variables in F are transformed into discrete variables to compute mutual information value. The mutual information $I(X, Y)$ is calculated according to equation (1):

$$\boldsymbol{I}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{x \in X} \sum_{y \in Y} \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{log} \frac{\boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y})}{\boldsymbol{p}(\boldsymbol{x}) \boldsymbol{p}(\boldsymbol{y})} \tag{1}$$

where $p(x,y)$ is the joint probability density of variable X and Y, $p(x)$ and $p(y)$ are the marginal probability density of variable X and Y.

MIC is defined as the highest normalized mutual information value between random variables X and Y:

$$\boldsymbol{MIC}(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{F}) = \underset{(\boldsymbol{x},\boldsymbol{y}):\boldsymbol{xy} < \boldsymbol{B}(\boldsymbol{n})}{\boldsymbol{max}} \frac{\boldsymbol{max}_{\boldsymbol{G} \in \boldsymbol{\Omega}(\boldsymbol{x},\boldsymbol{y})} \boldsymbol{I}(\boldsymbol{X}(\boldsymbol{G}), \ \boldsymbol{Y}(\boldsymbol{G}))}{\boldsymbol{log} \ \boldsymbol{min}(\boldsymbol{x}, \boldsymbol{y})}, \tag{2}$$

where $F = (x_i, y_i), i = 1...n\}$ is a set of ordered pairs (x, y), G represents an x-by-y grid of F, $\Omega(x, y)$ is a set of x-by-y grids and $I(X(G), Y(G))$ represents mutual information, where $X(G)$, $Y(G)$ are the discrete variables in $\Omega(x, y)$. B (n) is the maximal grid size that limits the sizes of the grids when searching over feasible partitions and is set to $n^{0.6}$, as suggested by empirical studies (Reshef *et al.*, 2011; Sun *et al.*, 2018). MIC takes values between 0 and 1, where 0 means statistical independence and 1 a completely noiseless relationship.

The MIC-CCS model is derived from the test theory (Ghiselli, 1964):

$$\boldsymbol{D_s} = \frac{\boldsymbol{r}\overline{\boldsymbol{k}}_{\boldsymbol{yi}}}{\sqrt{\boldsymbol{r} + \boldsymbol{r}(\boldsymbol{r}-1)\overline{\boldsymbol{k}}_{\boldsymbol{ii}}}}, \tag{3}$$

where D_s is the judgment value of the subset of classifiers M, r represents the number of selected classifiers M, $\bar{k}_{ii}$is the average sample correlation coefficient between feature variable i in M and $\bar{k}_{yi}$ represents the average sample correlation coefficient between target variable Y and feature variable i in M, where $i \in M, y \in Y$.

In equation (3), the numerator indicates the correlations between the features and target variable, which should be as high as possible. The denominator indicates the correlations between the feature variables, which are expected to be as low as possible. Hence, a maximized judgment value D_s indicates an optimal selection of base classifiers, taking into account the trade-off between the prediction accuracy and diversity of the base classifiers.

Assume a set of candidate classifier N, where N = 1,2,$\cdots$, n. M is the subset of N, where the number of classifiers in M is r, $r = 1, 2, \cdots, n$, r = 1,2,$\cdots$, n. The output from the i^{th} base classifier is h_i and y is the real value of the target variable. Let k_{yi} represent the MIC between h_i and y. k_{ij} represents the MIC between feature variables h_i and h_j (i = 1,2, $\cdots$, $n-1$; j = 2,$\cdots$, n). Obviously, the correlations between h_i and y should be high and those between h_i and h_j should be low, namely, $\sum k_{yi}$ should be as high as possible and $\sum k_{ij}$as low as possible. Additionally, we introduce decision variables u_i and v_{ij}:

$$\forall \boldsymbol{i} \in \boldsymbol{N},\ \boldsymbol{u_i} = \begin{cases} \boldsymbol{1}, & \boldsymbol{if}\ \ \boldsymbol{h_i} \in \boldsymbol{M}, \\ \boldsymbol{0}, & \boldsymbol{otherwise}, \end{cases} \tag{4}$$

$$\forall \boldsymbol{i}, \boldsymbol{j} \in \boldsymbol{N},\ \boldsymbol{and}\ \ \boldsymbol{i} < \boldsymbol{j}, \boldsymbol{v_{ij}} = \begin{cases} \boldsymbol{1}, & \boldsymbol{if}\ \ \boldsymbol{h_i} \in \boldsymbol{m},\ \ \boldsymbol{and}\ \ \boldsymbol{h_i} \in \boldsymbol{M}\ \ (\boldsymbol{u_i} = \boldsymbol{u_j} = \boldsymbol{1}), \\ \boldsymbol{0}, & \boldsymbol{otherwise}. \end{cases} \tag{5}$$

where $u_i = 1$ indicates that variable h_i is selected, while $v_{ij} = 1$ indicates variables h_i and h_j are both selected.

Therefore, the numerator in equation (3) can be written as:

$$\boldsymbol{r}\overline{\boldsymbol{k}}_{\boldsymbol{yi}} = \boldsymbol{r}\left(\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \frac{\boldsymbol{k_{yi}}\boldsymbol{u_i}}{\boldsymbol{r}}\right) = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \boldsymbol{k_{yi}}\boldsymbol{u_i}. \tag{6}$$

Similarly, the denominator in equation (3) can be written as:

$$\sqrt{r + r(r-1)\overline{k}_{ii}} = \sqrt{r + r(r-1)\left(\sum_{i=1,\ i<j}^{n-1} k_{ij}v_{ij} / \frac{r(r-1)}{2}\right)} = \sqrt{r + 2\sum_{i=1,i<j}^{n-1} k_{ij}v_{ij}}. \quad (7)$$

Equation (3) is finally rewritten as:

$$\textit{Maximize } D_s = \frac{\sum_{i=1}^{n} k_{yi}u_i}{\sqrt{r + 2\sum_{i=1,i<j}^{n-1} k_{ij}v_{ij}}}. \quad (8)$$

The objective of selecting a set of optimal base classifiers from all candidate classifiers is converted to a maximization problem in equation (8). The solution of this optimization problem returns a set of base classifiers, which are used to construct the ensemble model. Assume that m base classifiers are selected from n candidate classifiers. The selected base classifiers are denoted as $h_1^s, h_2^s, \cdots, h_m^s$ where $m < n$.

In the second phase, the same training data are applied to the base classifiers selected in the first phase. The probability estimates generated by these classifiers are then stacked and provided as input data to the meta-classifier, which is trained to optimally combine the estimations. Finally, the performance of the ensemble model is assessed with the testing data. The flowchart of the procedure is shown in Figure 1.

4. Experimental study

Australian, Taiwanese and German data sets are used to validate the proposed MIC-CCS method for credit risk assessment. These data sets are widely used in credit scoring studies and are available from the UCI Machine Learning Repository. The Australian data set includes 690 observations, 307 with good credit and 383 with bad credit. Each record contains a class label and 14 attributes, including six nominal and eight numerical attributes. The German data set contains a class label and 20 other features such as demographic and credit history information, for each applicant. The data set contains 700 observations for good credit applicants and 300 for bad credit applicants. The Taiwanese data set has 30,000 observations, 23,364 from creditworthy applicants and 6,636 from bad applicants. Each record has a class label and 23 features, including nine categorical attributes and 14 numerical attributes.

In this experimental study, we start with the eight most-used classification models as candidate base classifiers, namely, LR, LDA, KNN, NB, SVM, RF, light gradient boosting machine (LGB) and NN. However, more base classifiers can be added as need. For each of the three data sets, we randomly select training data and testing data with an 80–20 split. In the first phase of MIC-CCS, the models are fit to training data with 5-fold cross-validation. Python was used to conduct the experiments while our proposed nonlinear optimization model is solved with Lingo. Besides the ensemble model based on the proposed MIC-CCS method, we construct ensemble models based on three other classifier selection methods, namely, Pearson-CCS, BW-CCS and FW-CCS. Pearson-CCS uses the same proposed nonlinear optimization model as in this study. The only difference is that the correlation coefficient matrix is computed with the Pearson correlation coefficient. BW-CCS and FW-CCS use backward and forward selection methods to select base classifiers, respectively. The results from the ensemble model constructed with all candidate classifiers (ALL) are also presented and compared with those from other ensemble models in the following.

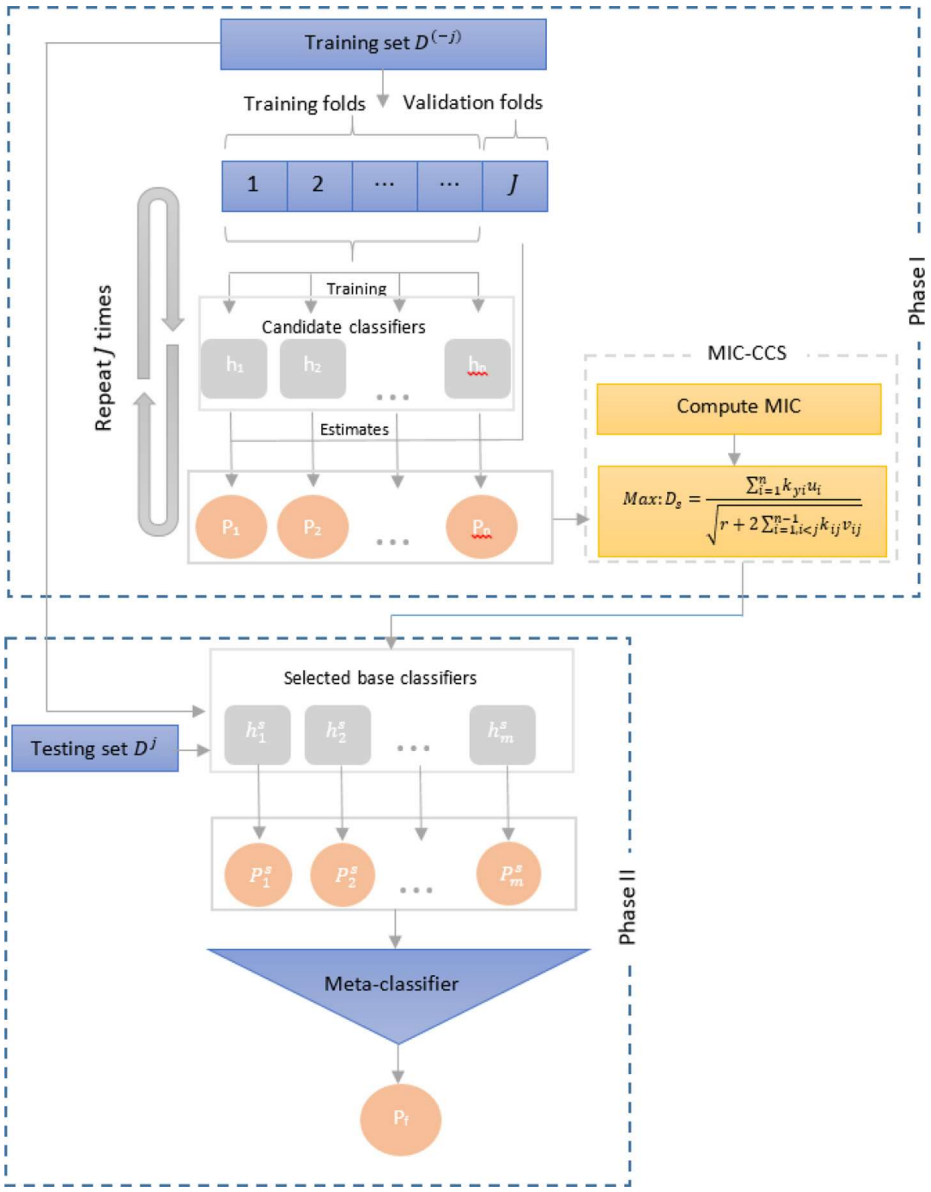

Figure 1.
Framework of our proposed method

5. Results

AUC is used to determine the performance of the classification model. It provides an aggregate measure of the performance of a model across all possible classification thresholds (Fawcett, 2006). Therefore, AUC is considered a more comprehensive measure for

binary classification problems, especially on imbalanced data set such as in a loan default problem where cases in the good class are significantly more than those of the bad class.

We also compare the performance of the ensemble model derived from our classifier selection method based on MIC correlation with those based on forwarding selection (FW), backward selection (BW) and Pearson correlation. The MIC and Pearson correlation matrix for three credit data sets are shown in the Appendix. Table 2 shows selected classifiers under the different methods for the Australian credit data set. The results of the ensemble model with all eight base classifiers (ALL) and results from each stand-alone classifier are reported for comparison. Finally, this paper constructs five ensemble models, namely MIC-CCS, Pearson-CCS, BW-CCS, FW-CCS and ALL. The prediction results based on the test set from the Australian data set are shown in Table 3.

Based on Table 3, the ensemble models perform better than stand-alone models in the Australian case. The proposed model (MIC-CCS) provides the best performance among the ensemble models when assessing AUC. Among the stand-alone models, LDA performs better than the others. However, the value of 0.901 of AUC is 3.9% lower than the AUC value of 0.936 from the MIC-CCS model and lower than other ensemble models We also report the standard error and the p-value of AUC which tests the null hypothesis that the area under the curve equals 0.50. The p-values for all the models are statistically significant while MIC-CCS has the lowest standard error. This result further reflects that the ensemble model with the proposed method is robust and has better prediction performance.

Table 4 shows the selected classifiers under different methods for the two credit data sets. The number of selected classifiers in Tables 2 and 4 also show that our method can

Table 2. Selected classifiers under the different selection methods: Australian credit data

Classifier selection method	Selected classifiers	No. of selected classifiers
MIC-CCS	RF, LGB and NN	3
Pearson-CCS	LDA, KNN, SVM, RF, LGB and NN	6
Backward (BW)	LR, KNN, RF and LGB	4
Forward (FW)	LR, KNN, RF and LGB	4

Table 3. Estimation results based on testing set: Australian credit data

Type	Model	AUC	Standard error	p-value	95% confidence interval	
					Lower	Upper
Stand-alone classifier	LR	0.890	0.031	0.000	0.841	0.952
	LDA	0.901	0.029	0.000	0.845	0.958
	KNN	0.849	0.037	0.000	0.772	0.916
	NB	0.827	0.039	0.000	0.767	0.892
	SVM	0.888	0.030	0.000	0.840	0.949
	RF	0.896	0.030	0.000	0.842	0.955
	LGB	0.886	0.032	0.000	0.837	0.948
	NN	0.860	0.035	0.000	0.774	0.933
Ensemble models	MIC-CCS	0.936	0.019	0.000	0.902	0.976
	Pearson-CCS	0.934	0.021	0.000	0.890	0.972
	BW-meta	0.929	0.022	0.000	0.877	0.969
	FW-meta	0.929	0.022	0.000	0.877	0.969
	ALL-meta	0.924	0.023	0.000	0.875	0.963

generate ensemble models with smaller sizes compared with the other selection methods in this study. The Taiwanese and German credit data generate results similar to the ones above. The AUC in Tables 5 and 6 shows that the proposed MIC-CCS method provides the best classification performance compared with all other ensemble and stand-alone models.

As the meta-classifier can be any classifier, we tried different classifiers but only report the ones that provide the best results. For the Australian data, the meta-classifier is SVM, while for the Taiwanese and German, data the best meta-classifiers are NB and RF, respectively.

It is worth pointing out that different costs can be associated with different types of misclassification errors in the credit risk evaluation. Type I errors (FP), classifying bad applicants as good, usually causes higher losses than Type II errors (FN), classifying good applicants as bad, for a financial institution (Marqués *et al.*, 2012). Therefore, we also report F_β-score which is based on precision and recall. For a typical binary classification problem, these metrics can be derived from a confusion matrix given in Table 7.

Precision is the ratio of correctly predicted bad applicants (TP) to the total predicted bad applicants (TP + FP). It reflects the ability of a model to predict the bad applicants. The recall is the ratio of correctly predicted bad applicants (TP) to all bad applicants in the actual

Table 4.
Selected classifiers under the different selection methods: Taiwanese and german creditdata

	Classifier selection method	Selected classifiers	No. of selected classifiers
Taiwanese credit data	MIC-CCS	SVM and LGB	2
	Pearson-CCS	RF and LGB	2
	Backward (BW)	LR, LDA, KNN, NB, SVM, RF, LGB and NN	8
	Forward (FW)	LR, LDA, KNN, SVM, RF, LGB and NN	7
German credit data	MIC-CCS	SVM and LGB	2
	Pearson-CCS	SVM, RF and LGB	3
	Backward (BW)	LR, KNN, RF, LGB and NN	5
	Forward (FW)	LR, KNN, RF, LGB and NN	5

Table 5.
Estimation results based on the testing set: Taiwanese credit data

					95% confidence interval	
Type	Model	AUC	Standard error	*p*-value	Lower	Upper
Stand-alone classifier	LR	0.753	0.008	0.000	0.743	0.774
	LDA	0.755	0.008	0.000	0.742	0.773
	KNN	0.689	0.009	0.000	0.672	0.706
	NB	0.730	0.009	0.000	0.711	0.742
	SVM	0.711	0.009	0.000	0.702	0.733
	RF	0.721	0.008	0.000	0.703	0.736
	LGB	0.772	0.009	0.000	0.744	0.788
	NN	0.708	0.009	0.000	0.691	0.730
Ensemble models	MIC-CCS	0.774	0.008	0.000	0.749	0.790
	Pearson-CCS	0.744	0.008	0.000	0.738	0.766
	BW-meta	0.751	0.008	0.000	0.740	0.775
	FW-meta	0.751	0.009	0.000	0.739	0.777
	ALL-meta	0.751	0.008	0.000	0.740	0.775

Type	Model	AUC	Standard error	p-value	95% confidence interval Lower	Upper
Stand-alone classifier	LR	0.695	0.038	0.000	0.612	0.780
	LDA	0.672	0.039	0.000	0.603	0.778
	KNN	0.679	0.043	0.000	0.594	0.764
	NB	0.610	0.044	0.000	0.567	0.750
	SVM	0.732	0.037	0.000	0.657	0.807
	RF	0.691	0.043	0.000	0.584	0.775
	LGB	0.702	0.042	0.000	0.586	0.778
	NN	0.601	0.048	0.000	0.586	0.758
Ensemble models	MIC-CCS	0.772	0.036	0.000	0.699	0.845
	Pearson-CCS	0.756	0.037	0.000	0.681	0.825
	BW-meta	0.743	0.037	0.000	0.677	0.821
	FW-meta	0.743	0.037	0.000	0.677	0.821
	ALL-meta	0.760	0.037	0.000	0.689	0.829

Table 6. Estimation results based on the testing set: German credit data

class (TP + FN). It is a critical indicator as it indicates the ability of a model to identify the bad applicants, and thus avoid falsely accepting bad applicants.

F_β-measure is a weighted harmonic approach between precision and recall shown in equation (9). It is useful when I Type and Type II errors have different costs. When the beta value is set to 1, it is known as F1-score which gives precision and recall the same weight. When beta is less than 1, it assigns more weight to precision and less to recall, otherwise more to recall and less to precision when the beta is greater than 1. As valid estimates of cost are not available, we set beta to 5 and 10 reflecting the higher cost of misclassifying bad applicants as good in the field of credit risk evaluation.

$$F_\beta = (1+\beta^2)\frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{\beta^2 Precision} + \boldsymbol{Recall}} \tag{9}$$

Table 8 Reports the ranking of the performance of models based on the results of F1-score and F_β score where beta equals 5 and 10 across the three data sets. Full F-score results for all models can be found in the Appendix. The proposed MIC-CCS method demonstrates its robustness as the models using this method are among the top three performing ones for all the metrics and data sets. The performances of other models present higher variance in different metrics or data sets. SVM, for example, ranks first in F5-score and F10-score for the German data set. However, it ranks 6th in F1-score for the same data set while ranges from 8th to 10th for Australia and Taiwan data sets in all F-scores.

In summary, the results of the three credit data sets indicate that ensemble models perform better than stand-alone ones, whereas the ensemble model based on the MIC-CCS method outperforms the other ensemble models based on different base-classifier selection

		Predicted/classified Positive (bad applicants)	Negative (good applicants)
Actual	Positive (bad applicants)	True positives (TP)	False negatives (FN)
	Negative (good applicants)	False positives (FP)	True negatives (TN)

Table 7. Confusion matrix

Table 8.
F-Score ranks based on the testing set: Australian, Taiwanese and German credit data

Type	Model	Australia credit data			Taiwan credit data			German credit data		
		F1-score Rank	F5-score Rank	F10-score Rank	F1-score Rank	F5-score Rank	F10-score Rank	F1-score Rank	F5-score Rank	F10-score Rank
Stand-alone classifier	LR	9	11	11	8	8	8	11	11	11
	LDA	10	8	8	7	6	6	10	8	8
	KNN	11	10	10	11	11	11	7	9	9
	NB	13	13	13	13	13	13	13	13	13
	SVM	8	9	9	10	10	10	6	1	1
	RF	7	7	7	5	3	2	8	7	7
	LGB	1	1	1	3	4	4	3	3	3
	NN	12	12	12	12	12	12	12	12	12
Ensemble models	MIC-CCS	1	1	1	2	1	1	2	3	3
	Pearson-CCS	1	1	1	4	5	5	9	10	10
	BW-meta	1	1	1	1	2	3	4	5	5
	FW-meta	1	1	1	8	8	8	4	5	5
	ALL-meta	1	1	1	6	6	6	1	2	2

methods. The results also show that our method is robust and can effectively select base classifiers that are significantly different, so that they can provide complementary information and yet these classifiers have good predictive capabilities, thus improving the classification performance of the ensemble model.

6. Conclusions

As the recent global financial crisis, the analysis of credit risk has become pivotal for banks and financial institutions. There exists considerable interest in improving the performance of credit risk evaluation models to better discriminate between applicants that pose a high default risk from creditworthy applicants, as a small improvement in classification performance can yield significant profit in the financial sector.

In this study, we propose a correlation-based classifier selection method (MIC-CCS) to improve the classification performance of the ensemble model under a stacking strategy. The proposed MIC-CCS model seeks to optimize the relationship between the accuracy and diversity of base classifiers. The final prediction of the ensemble model is improved by using the base classifiers selected by our nonlinear optimization model, allowing it to effectively select significantly different base classifiers that provide both complementary information and good prediction accuracy. We compare our approach with eight stand-alone classifiers and four other ensemble models using different base classifier selection methods. The experimental results suggest that the ensemble model based on our MIC-CCS yields the best performance. In addition, certain classifiers may be better than all classifiers for improving the performance of ensemble models. Further, MIC, as a nonparametric exploration statistic, can better measure the nonlinear relationship between variables with noisy data than a traditional correlation measurement such as the Pearson correlation coefficient can.

For future work, a dynamic classifier selection technique can be explored and compared with the static classifier selection method proposed in this study. The main idea of using dynamic selection techniques is to apply an unsupervised method to cluster similar samples and then, treat each cluster differently based on data particularities. By doing so, each cluster may select the best base classifiers for that group of instances, leading to a dynamic method of selecting classifiers.

References

Abellán, J. and Castellano, J.G. (2017), "A comparative study on base classifiers in ensemble methods for credit scoring", *Expert Systems with Applications*, Vol. 73, pp. 1-10.

Altman, E.I. (1968), "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The Journal of Finance*, Vol. 23 No. 4, pp. 589-609.

Bellotti, T. and Crook, J. (2009), "Support vector machines for credit scoring and discovery of significant features", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 3302-3308.

Carmona, P., Climent, F. and Momparler, A. (2019), "Predicting failure in the U.S. banking sector: an extreme gradient boosting approach", *International Review of Economics and Finance*, Vol. 61, pp. 304-323.

Chang, Y.C., Chang, K.H. and Wu, G.J. (2018), "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions", *Applied Soft Computing*, Vol. 73, pp. 914-920.

Chen, N., Ribeiro, B. and Chen, A. (2016), "Financial credit risk assessment: a recent review", *Artificial Intelligence Review*, Vol. 45 No. 1, pp. 1-23.

Climent, F., Momparler, A. and Carmona, P. (2019), "Anticipating bank distress in the eurozone: an extreme gradient boosting approach", *Journal of Business Research*, Vol. 101, pp. 885-896.

Du Jardin, P. (2019), "Forecasting bankruptcy using biclustering and neural network-based ensembles", *Annals of Operations Research*, Vol. 1.

Fawcett, T. (2006), "An introduction to ROC analysis", *Pattern Recognition Letters*, Vol. 27 No. 8, pp. 861-874.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014), "Do we need hundreds of classifiers to solve real world classification problems?", *Journal of Machine Learning Research*, Vol. 15, pp. 3133-3181.

Finlay, S. (2011), "Multiple classifier architectures and their application to credit risk assessment", *European Journal of Operational Research*, Vol. 210 No. 2, pp. 368-378.

Fu, X., Ouyang, T., Chen, J. and Luo, X. (2020), *Listening to the Investors: A Novel Framework for Online Lending Default Prediction Using Deep Learning Neural Networks. Information Processing and Management*, Vol. 57.

Geng, R., Bose, I. and Chen, X. (2015), "Prediction of financial distress: an empirical study of listed chinese companies using data mining", *European Journal of Operational Research*, Vol. 241 No. 1, pp. 236-247.

Ghiselli, E.E. (1964), *Theory of Psychological Measurement*, McGraw-Hill, New York, NY.

Hsu, K.W. and Srivastava, J. (2009), "Diversity in combinations of heterogeneous classifiers", *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Aasia Conference*, Bangkok, Thailand, pp. 923-932.

Huang, J., Wang, H. and Kochenberger, G. (2017), "Distressed chinese firm prediction with discretized data", *Management Decision*, Vol. 55 No. 5, pp. 786-807.

Hyewon, Y. and Zheng, G. (2010), "Predict US restaurant firm failures: the artificial neural network model versus logistic regression model", *Tourism and Hospitality Research*, Vol. 10, pp. 171-187.

Jing, Z. and Fang, Y. (2018), "Predicting US bank failures: a comparison of logit and data mining models", *Journal of Forecasting*, Vol. 37 No. 2, pp. 235-256.

Le, H.H., Viviani, J.L. (2018), "Predicting bank failure: an improvement by implementing a machine-learning approach to classical financial ratios", *Research in International Business and Finance*, Vol. 44, pp. 16-25.

Lee, K., Booth, D. and Alam, P. (2005), "A comparison of supervised and unsupervised neural networks in predicting bankruptcy of korean firms", *Expert Systems with Applications*, Vol. 29 No. 1, pp. 1-16.

Liang, D., Lu, C.C., Tsai, C.F. and Shih, G.A. (2016), "Financial ratios and corporate governance indicators in bankruptcy prediction: a comprehensive study", *European Journal of Operational Research*, Vol. 252 No. 2, pp. 561-572.

Liang, D., Tsai, C.F., Dai, A.J. and Eberle, W. (2018), "A novel classifier ensemble approach for financial distress prediction", *Knowledge and Information Systems*, Vol. 54 No. 2, pp. 437-462.

Maldonado, S., Bravo, C., López, J. and Pérez, J. (2017), "Integrated framework for profit-based feature selection and SVM classification in credit scoring", *Decision Support Systems*, Vol. 104, pp. 113-121.

Marqués, A.I., García, V. and Sánchez, J.S. (2012), "Exploring the behaviour of base classifiers in credit scoring ensembles", *Expert Systems with Applications*, Vol. 39 No. 11, pp. 10244-10250.

Martin, D. (1977), "Early warning of bank failure: a logit regression approach", *Journal of Banking and Finance*, Vol. 1 No. 3, pp. 249-276.

Melville, P. and Mooney, R.J. (2005), "Creating diversity in ensembles using artificial data", *Information Fusion*, Vol. 6 No. 1, pp. 99-111.

Montebruno, P., Bennett, R.J., Smith, H. and Lieshout, C.V. (2020), "Machine learning classification of entrepreneurs in British historical census data", *Information Processing and Management*, Vol. 57.

Nti, I.K., Adekoya, A.F. and Weyori, B.A. (2020), "A comprehensive evaluation of ensemble learning for stock-market prediction", *Journal of Big Data*, Vol. 7 No. 1.

Olson, D.L., Delen, D. and Meng, Y. (2012), "Comparative analysis of data mining methods for bankruptcy prediction", *Decision Support Systems*, Vol. 52 No. 2, pp. 464-473.

Oza, N.C. and Tumer, K. (2008), "Classifier ensembles: select real-world applications", *Information Fusion*, Vol. 9 No. 1, pp. 4-20.

Paleologo, G., Elisseeff, A. and Antonini, G. (2010), "Subagging for credit scoring models", *European Journal of Operational Research*, Vol. 201 No. 2, pp. 490-499.

Papouskova, M. and Hajek, P. (2019), "Two-stage consumer credit risk modelling using heterogeneous ensemble learning", *Decision Support Systems*, Vol. 118, pp. 33-45.

Pławiak, P., Abdar, M. and Rajendra Acharya, U. (2019), "Application of new deep genetic cascade ensemble of SVM classifiers to predict the australian credit scoring", *Applied Soft Computing Journal*, Vol. 84.

Ramli, N.A., Ismail, M.T. and Wooi, H.C. (2015), "Measuring the accuracy of currency crisis prediction with combined classifiers in designing early warning system", *Machine Learning*, Vol. 101 Nos 1/3, pp. 85-103.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., Mcvean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M. and Sabeti, P.C. (2011), "Detecting novel associations in large data sets", *Science*, Vol. 334 No. 6062, pp. 1518-1524.

Shin, K.S. and Lee, T.S., Kim, H.J. (2005), "An application of support vector machines in bankruptcy prediction model", *Expert Systems with Applications*, Vol. 28 No. 1, pp. 127-135.

Sun, G., Li, J., Dai, J., Song, Z. and Lang, F. (2018), "Feature selection for IoT based on maximal information coefficient", *Future Generation Computer Systems*, Vol. 89, pp. 606-616.

Tam, K.Y. and Kiang, M. (1990), "Predicting bank failures: a neural network approach", *Applied Artificial Intelligence*, Vol. 4 No. 4, pp. 265-282.

Tsai, C.F. and Hsu, Y.F., Yen., D.C. (2014), "A comparative study of classifier ensembles for bankruptcy prediction", *Applied Soft Computing Journal*, Vol. 24, pp. 977-984.

Verikas, A., Gelzinis, A. and Bacauskiene, M. (2011), "Mining data with random forests: a survey and results of new tests", *Pattern Recognition*, Vol. 44 No. 2, pp. 330-349.

Wang, G., MA., J. (2012), "A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine", *Expert Systems with Applications*, Vol. 39 No. 5, pp. 5325-5331.

West, D. and Dellana, S. (2009), "Diversity of ability and cognitive style for group decision processes", *Information Sciences*, Vol. 179 No. 5, pp. 542-558.

Whiting, D.G., Hansen, J.V., Mcdonald, J.B., Albrecht, C. and Albrecht, W.S. (2012), "Machine learning methods for detecting patterns of management fraud", *Computational Intelligence*, Vol. 28 No. 4, pp. 505-527.

Xia, Y., Liu, C., DA, B. and Xie, F. (2018), "A novel heterogeneous ensemble credit scoring model based on bstacking approach", *Expert Systems with Applications*, Vol. 93, pp. 182-199.

Yang, Z.R., Platt, M.B. and Platt, H.D. (1999), "Probabilistic neural networks in bankruptcy prediction", *Journal of Business Research*, Vol. 44 No. 2, pp. 67-74.

Zhang, Z., Chen, Y., Li, J. and Luo, X. (2019), *A Distance-Based Weighting Framework for Boosting the Performance of Dynamic Ensemble Selection. Information Processing and Management*, Vol. 56, pp. 1300-1316.

Zhou, Z.H., Wu, J. and Tang, W. (2002), "Ensembling neural networks: many could be better than all", *Artificial Intelligence*, Vol. 137 Nos 1/2, pp. 239-263.

Appendix

Table A1. Australian credit data: MIC correlation matrix

	LR	LDA	KNN	NB	SVM	RF	LGB	NN	Class label
LR	1	1	0.7378	0.7374	0.9812	0.7278	0.8402	0.8047	0.6889
LDA		1	0.7411	0.7406	0.9862	0.7139	0.8357	0.7789	0.7018
KNN			1	0.6136	0.8071	0.6729	0.6862	0.6097	0.6415
NB				1	0.6723	0.5378	0.6094	0.585	0.5625
SVM					1	0.7674	0.817	0.7281	0.7909
RF						1	0.9421	0.6538	0.9476
LGB							1	0.6793	0.9915
NN								1	0.6844
Class label									1

Table A2. Australian credit data: Pearson correlation matrix

	LR	LDA	KNN	NB	SVM	RF	LGB	NN	Class label
LR	1	0.974	0.887	0.66	0.96	0.868	0.835	0.897	0.796
LDA		1	0.884	0.57	0.979	0.859	0.832	0.855	0.796
KNN			1	0.595	0.904	0.869	0.846	0.803	0.82
NB				1	0.561	0.572	0.557	0.622	0.534
SVM					1	0.881	0.852	0.844	0.818
RF						1	0.971	0.817	0.96
LGB							1	0.805	0.995
NN								1	0.775
Class label									1

Table A3. Taiwanese credit data: MIC correlation matrix

	LR	LDA	KNN	NB	SVM	RF	LGB	NN	Class label
LR	1	0.8464	0.3047	0.5289	0.5052	0.2096	0.5224	0.2834	0.1853
LDA		1	0.3073	0.555	0.5199	0.2089	0.5229	0.2793	0.184
KNN			1	0.2795	0.3321	0.3131	0.326	0.1496	0.3056
NB				1	0.4732	0.1888	0.4536	0.2362	0.1651
SVM					1	0.232	0.481	0.2197	0.2235
RF						1	0.3326	0.1534	0.7249
LGB							1	0.3777	0.321
NN								1	0.1517
Class label									1

Table A4. Taiwanese credit data: Pearson correlation matrix

	LR	LDA	KNN	NB	SVM	RF	LGB	NN	Class label
LR	1	0.984	0.734	0.675	0.902	0.582	0.918	0.619	0.466
LDA		1	0.731	0.665	0.944	0.576	0.908	0.576	0.462
KNN			1	0.527	0.688	0.68	0.77	0.5	0.618
NB				1	0.636	0.452	0.63	0.575	0.373
SVM					1	0.543	0.841	0.449	0.442
RF						1	0.711	0.467	0.936
LGB							1	0.671	0.609
NN								1	0.387
Class label									1

Table A5. German credit data: MIC correlation matrix

	LR	LDA	KNN	NB	SVM	RF	LGB	NN	Class label
LR	1	0.883	0.41	0.512	0.668	0.3622	0.5708	0.502	0.3679
LDA		1	0.398	0.5375	0.6889	0.3477	0.5363	0.4563	0.3726
KNN			1	0.3461	0.5241	0.3309	0.4629	0.2554	0.3175
NB				1	0.4792	0.2942	0.3975	0.3233	0.2798
SVM					1	0.534	0.6623	0.3863	0.675
RF						1	0.8593	0.2701	0.8593
LGB							1	0.3843	0.8807
NN								1	0.2687
Class label									1

Table A6. German credit data: Pearson correlation matrix

	LR	LDA	KNN	NB	SVM	RF	LGB	NN	Class label
LR	1	0.972	0.658	0.715	0.865	0.621	0.647	0.681	0.564
LDA		1	0.642	0.713	0.869	0.619	0.651	0.641	0.574
KNN			1	0.532	0.757	0.648	0.661	0.502	0.61
NB				1	0.633	0.491	0.502	0.535	0.441
SVM					1	0.797	0.839	0.617	0.786
RF						1	0.947	0.474	0.941
LGB							1	0.494	0.987
NN								1	0.43
Class label									1

Table A7.
Prediction results based on testing set – Australia credit data

Type	Model	Precision	Recall	F1-score	F5-score	F10-score
Stand-alone classifier	LR	0.8772	0.8197	0.8475	0.82178	0.8202
	LDA	0.8154	0.8689	0.8413	0.8667	0.8683
	KNN	0.8095	0.8361	0.8226	0.8350	0.8358
	NB	0.7818	0.7049	0.7414	0.7076	0.7056
	SVM	0.8525	0.8525	0.8525	0.8525	0.8525
	RF	0.8689	0.8689	0.8689	0.8689	0.8689
	LGB	0.8852	0.8852	0.8852	0.8852	0.8852
	NN	0.8627	0.7213	0.7857	0.7259	0.7225
Ensemble models	MIC-CCS	0.8852	0.8852	0.8852	0.8852	0.8852
	Pearson-CCS	0.8852	0.8852	0.8852	0.8852	0.8852
	BW-meta	0.8852	0.8852	0.8852	0.8852	0.8852
	FW-meta	0.8852	0.8852	0.8852	0.8852	0.8852
	ALL-meta	0.8852	0.8852	0.8852	0.8852	0.8852

Table A8.
Prediction results based on testing set – Taiwan credit data

Type	Model	Precision	Recall	F1-score	F5-score	F10-score
Stand-alone classifier	LR	0.6496	0.37302	0.4739	0.3792	0.3746
	LDA	0.6501	0.3738	0.4746	0.3800	0.3754
	KNN	0.5372	0.3482	0.4225	0.3530	0.3494
	NB	0.6516	0.1959	0.3013	0.2013	0.1973
	SVM	0.6657	0.3587	0.4662	0.3652	0.3603
	RF	0.4714	0.4898	0.4804	0.4891	0.4896
	LGB	0.6115	0.4175	0.4962	0.4227	0.4188
	NN	0.5270	0.2570	0.3455	0.2622	0.2583
Ensemble models	MIC-CCS	0.4915	0.5019	0.4966	0.5015	0.5018
	Pearson-CCS	0.6453	0.3949	0.4900	0.4009	0.3964
	BW-meta	0.5597	0.4876	0.5212	0.4900	0.4882
	FW-meta	0.6496	0.3730	0.4739	0.3792	0.3746
	ALL-meta	0.6501	0.3738	0.4747	0.3800	0.3754

Table A9. Prediction results based on testing set – German credit data

Type	Model	Precision	Recall	F1-score	F5-score	F10-score
Stand-alone classifier	LR	0.4561	0.4333	0.4444	0.4341	0.4335
	LDA	0.4412	0.5000	0.4688	0.4975	0.4993
	KNN	0.5686	0.4833	0.5225	0.4861	0.4840
	NB	0.3594	0.3833	0.3710	0.3823	0.3830
	SVM	0.5139	0.6167	0.5606	0.6120	0.6155
	RF	0.5000	0.5167	0.5082	0.516	0.5165
	LGB	0.5862	0.5667	0.5763	0.5674	0.5669
	NN	0.3582	0.4000	0.3780	0.3982	0.3995
Ensemble models	MIC-CCS	0.5862	0.5667	0.5763	0.5674	0.5669
	Pearson-CCS	0.5490	0.4667	0.5045	0.4694	0.4674
	BW-meta	0.5763	0.5667	0.5714	0.5671	0.5668
	FW-meta	0.5763	0.5667	0.5714	0.5671	0.5668
	ALL-meta	0.5738	0.5833	0.5785	0.5823	0.5832

Corresponding author
Jun Huang can be contacted at: jun.huang@angelo.edu