IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China

WEI LI[1,2], SHUAI DING[1,2], YI CHEN[1,2] and SHANLIN YANG[1,2]

[1]School of Management, Hefei University of Technology, Anhui, Hefei 23009 China

[2]Key Laboratory of Process Optimization and Intelligent Decision-Making (Ministry of Education), Hefei University of Technology, Anhui, Hefei 23009 China

Corresponding author: S. Ding (dingshuai@hfut.edu.cn) and S.L. Yang (yangsl@hfut.edu.cn).

The work was supported in part by the Ministry of Science and Technology of China under Grant 2016YFC0503606, and in part by the National Natural Science Foundation of China under Grant 71571058, 71690235, 71501058, and 71490725.

ABSTRACT As a novel financing method, peer-to-peer (P2P) lending has drawn broad attention as it provides those financers who cannot participate in the traditional financial market with funds. In P2P lending marketplaces, one of the crucial challenges that P2P online lending platforms are facing is to accurately predict the default risk of each loan by tapping into default prediction models, thus effectively helping P2P online lending companies avoid credit risks. That traditional credit risk prediction models fail to meet the demand of P2P online lending companies for default risk prediction is because of the uneven distribution of credit data samples in the P2P lending marketplaces (i.e., the default sampled data are scarce). In this study, we designed a multi-round ensemble learning model based on heterogeneous ensemble frameworks to predict default risks. In this model, an extreme gradient boosting (XGBoost) is initially used for ensemble learning, and the XGBoost, deep neural network (DNN) and logistic regression are then regarded as heterogeneous individual learners to undergo a linear weighted fusion. To verify the designed default risk prediction model, real credit data from a famous P2P online lending marketplace in China were used in a test. The results of the test indicate that this model can effectively increase predictive accuracy compared with traditional machine learning models and ensemble learning models.

INDEX TERMS Ensemble learning, default prediction, imbalanced data, P2P lending.

I. INTRODUCTION

In China, P2P online lending has achieved development as an innovative financing model. That is due to its reliance on an end-to-end financing model made possible by the information technology and the Internet, which eliminates the involvement of traditional financial intermediaries. Such a financial model facilitates money to conveniently move from people who are rich in funds to the hands of those who are in need of them, thereby reducing operating costs of the market, improving the efficiency of funds allocation, and in the meantime, making the lending process more transparent [1]. Starting in 2007 when the first P2P online lending platform – PPDai - launched its online operations, P2P lending marketplaces in China have gained rapid development. By 2014, the number of platforms had exceeded 1000. As of the end of August 2017, the balance of loans in the online lending market totaled 1.12 trillion yuan and the number of operating platforms were as high as 2,065. The credit risks arising from the rapid growth of P2P online lending marketplaces should not be neglected. Presently, the balance of problematic loans is around 29.37 billion yuan, amounting to about 2.62% of the balance of loans in the industry as of the end of August 2017.

Given P2P online loans are mostly unsecured credit loans generated through online lending platforms rather than financial intermediaries, the assessment of default risks in these loans becomes increasingly prominent [2], [3]. In the meantime, despite being an innovative financing model, P2P lending can still inevitably acquire the risks commonly seen in traditional financial businesses. Thus, default risk prediction plays an important role in the quest of better facilitating P2P lending companies to identify credit risks and avoid losses. As can be found through analyses, the samples of credit data generated by Chinese P2P lending companies are characterized by high dimensions, uneven distribution of categories and large sample sizes. With respect to such high-dimensional, imbalanced and large-sized credit risk data,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2810864, IEEE Access

predicting default risks can be boiled down to a binary classification problem – particularly given the fact that the existence of imbalanced data has resulted in very few number of samples under the category of "default", making it a huge challenge for financial institutions, especially P2P lending companies, to accurately make predictions out of the high-dimensional and imbalanced credit data.

The P2P lending marketplace consists of innumerable individual lenders and borrowers through the Internet platforms. Although in real financial environment defaulting behaviors of loan applicants are rare cases compared to those who make regular repayments, defaults still take place from time to time [4]. In literature relating to prediction of default risks, many researchers have proposed a broad range of classification technologies, including statistical techniques such as linear discriminant analysis (LDA) and logistic regression, as well as nonparametric models like k-nearest neighbor and decision trees [5], [6]. In addition, machine learning techniques, such as the neural network and support vector machine [7]-[9], have also been used. However, the abovementioned models are not very effective in solving the classification of imbalanced data [10]. The later-proposed sampling methods, including the most common applied under-sampling and over-sampling methods, are used to change the original distribution of imbalanced data by eliminating samples of majority class and increasing those of minority class [11], [12]. Moreover, the Bagging and Boosting based ensemble methods are also widely used to deal with imbalanced data [13]-[15]. However, these methods adopt addition of minority class instances or deletion of majority class stances in order to change the original distribution of imbalanced data, in which some useful data may get eliminated.

Based on the analysis above, this paper designed a multi-round ensemble learning model based on heterogeneous ensemble frameworks to predict default risks. First, an ensemble learning of a data set which has undergone pre-processing and feature engineering was initially performed by using an extreme gradient boosting (XGBoost) model; then, parameter tuning in XGBoost for another training was conducted to obtain prediction result. In addition, the impact that each variable type in the credit data set exerts in the default risk prediction model can be known through the XGBoost model. Then, ensemble learning and linear weighted fusion of the XGBoost model and the logistic regression (LR) model, both treated as heterogeneous individual learners, were performed to obtain default prediction results from cleansed credit data set. Finally, another round of ensemble learning and linear weighted fusion of the XGBoost model, the deep neural network (DNN) model and the LR model, treated as heterogeneous individual learners, were performed, followed by the default prediction performed on data set to obtain predictive results. At last, predictive results obtained from each round of ensemble learning were compared and analyzed. The experimental results indicate that the prediction model proposed by this paper can effectively improve the accuracy of prediction.

Other sections of this paper are outlined as follows. In Section 2, relevant literature is reviewed. In Section 3, methods and models adopted in this paper are discussed, that is, three heterogeneous individual learners – XGBoost, DNN and LR - are employed to undergo ensemble learning and linear weighted fusion in a multi-round ensemble learning model based on heterogeneous ensemble frameworks. The experiment part is discussed in Section 4, which includes the introduction of the data sets, experiment settings such as data pre-processing and feature engineering, and comparison and analysis of predictive results of models. Finally, the summary of the paper is provided in Section 5.

II. RELATED LITERATURE

In the P2P lending marketplace, a personal credit-based model is largely adopted in microfinance lending, in which an individual's transaction detail lists and other credit records are used for qualification verification, while almost no mortgage or pledge is required to be provided by customers as in bank loan applications [16]. This could result in the presence of risks, which cannot be transferred through the means of pledging. In the P2P lending marketplace, the default risk of a borrower mainly stems from the occurrence of borrower's credit risk, which are mainly due to two reasons: first, insufficient willingness of performing contracts in credit receivers (i.e., the borrowers); second, inability to perform contracts despite such a willingness [2]. Thus, if the default risk of a borrower were to happen owing to a variety of changes arising from the life of a loan, this can eventually lead to financial losses of the creditor (i.e., the P2P lending platform). Therefore, many scholars have conducted plenty of studies, proposed a large number of models and approaches, and produced abundant research achievements in how to accurately predict credit risk levels of customers of P2P lending and avoid the occurrence of default risks [17]-[23].

As found in studies of classification problems such as prediction of default risks, the numbers of different classes of training samples could vary significantly, which is known as the "class-imbalance" problem [24]. The classification of imbalanced data has received growing attentions and become one of the hot research topics in the field of machine learning. In traditional machine learning studies, most classification algorithms assume that the prior probability of each class of samples is distributed evenly or the misclassification costs are identical. However, such an assumption does not necessarily hold true in the real world. In the case of imbalanced data, the information of minority class samples will be drowned in massive majority class samples, resulting in an exceedingly higher rate of classification error in the former than in the latter, as well as poorer generalization capacity of classifiers [25], [26]. Thus, the classification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2810864, IEEE Access

problem of imbalanced data constitutes one of the major difficulties in areas of data mining and machine learning. In the meantime, although the minority class samples generated from economic and financial fields, including those related to credit risk prediction and identification, credit card fraud detection, and market and business behavior analysis, etc., are rare, the accurate identification of minority classes is often of more value than that of majority classes, and people tend to pay more attention to the classification correctness of minority class samples [26], [27]. Thus, studying the classification problem of imbalanced data is of great significance. The performance metrics of imbalanced data classification is no longer determined by the overall identification rate of classifiers, instead, the classification accuracy in minority classes has received more attentions. Later, with advancement in studies of imbalanced data, classification algorithms based on imbalanced data have drawn ever greater attentions. Traditionally classical approaches to improve the overall performance of classifiers for imbalanced data mainly include resampling techniques at the data reconstruction level, and improvement of classical algorithms and proposition of new algorithms at the algorithm level.

The resampling technique refers to an approach in which the under-sampling method is adopted for majority class samples while the over-sampling method for minority class samples in a training sample set, in order to improve the extent to which the classes of training samples are evenly distributed, making it an effective approach to boost the performance of classifiers for imbalanced data [28]. Under the premise that minimum information loss is ensured and the majority of useful samples that are conducive to classification learning are retained, the core problems that need to be addressed in data reconstruction are: eliminating massive noisy information and significantly reduce the imbalance of data. Data resampling technique mainly consists of two basic methods: under-sampling and over-sampling [29]. Under-sampling technique, including approaches such as random under-sampling of majority class, neighborhood cleaning rule, condensed nearest neighbor rule, one-sided selection and Tomek Links, etc., is mainly used to remove majority class samples in a training set. In this process, some potentially useful majority class samples may also get removed, resulting in reduced performance of classifiers [30]. Over-sampling technique, including approaches such as random copy of original minority samples and SMOTE, is mainly used to increase the minority class samples in a training set [31]. When over-sampling technique is adopted to deal with the data with imbalanced distribution of classes, the simplest method is to copy minority class samples. However, excessive copy of original minority class samples would lead to the over-fit phenomena, and in the meantime, prolonged training time for classifiers as the new samples increase. On the other hand, the most typical over-sampling method is SMOTE, which can effectively reduce overfitting as a result of mechanical copy of minority class samples, thus improving the generalization capacity of classifiers [32].

Ensemble learning is a machining learning approach where multiple learners are trained to solve the classification problem. The central concept is to combine several “weak leaners” into a “strong learner”, thereby eventually boosting the performance of classifiers. Thus, ensemble learning has been attracting increasing attentions and more broadly use. LK Hansen and P Salamon pioneered an ensemble method for neural networks by which multiple neural networks can be trained and the training results can be combined to significantly improve the generalization capacity of learners in neural networks [33]. Through the approach of constructing functions, RE Schapire demonstrated that multiple weak learners can be boosted into a strong learner, creating the prototype of the Boosting algorithms. The Boosting algorithm proposed by RE Schapire runs into difficulty in solving actual problems as it requires prior knowledge of the lower bound of the learning algorithm’s generalization capacity, which is usually unavailable. Later, Y Freud and RE Schapire made further refinement by proposing the AdaBoost algorithm. Over recent years, improved algorithms, including those of AdaBoost.M1, AdaBoost.M2, AdaBoost.MR and AdaBoost.MH, have been proposed on the basis of the AdaBoost algorithm [33]-[36]. In the application of machine learning in classification of imbalanced data, traditional machine learning approach usually adopt the overall identification rate as the performance evaluation criteria for classifiers. That makes predictive results biased towards majority class samples, while minority class ones are often neglected or regarded as noise. However, it is minority class samples that often draw special attentions in real life. One of the characteristics of minority class samples themselves is the severe scarcity of sample data, thus it is very hard to generalize existing rules if just minority class samples are explored. L Breiman proposed a Bagging ensemble learning algorithm capable of effectively improve the generalization capacity of classifiers [37]. This algorithm no longer requires prior knowledge of the lower bound of generalization, thus is more broadly applicable [38].

When applying artificial intelligence and machine learning approach, unitarily adopting classification algorithms could result in poor effectiveness in dealing with complicated problems and the classification problem of imbalanced data [39]-[41]. When a unitary classification model is used to learn the training data, results obtained from a model with the highest training accuracy are often selected as the final results. However, the generalization capacity of such a learning model on test data sets is less than ideal. Moreover, despite that most ensemble methods rely on single base learners to produce homogenous ensembles, heterogenous ensembles which are combined by multiple heterogeneous individual learners are increasingly gaining attentions. As

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2810864, IEEE Access

found in studies of classification algorithms related in ensemble learning, using multiple individual classifiers in ensemble can effective solve overfitting of individual learning models, and produce superior predictive results and better generalization capacity in most imbalanced data sets [42]. Therefore, in this paper, XGBoost, DNN and LR models were selected as individual learners for a heterogeneous ensemble, and a default risk prediction was performed on credit data of P2P lending, achieving a fairly good predictive effect.

III. PROPOSED MODELS

A. FRAMEWORK OF DEFAULT PREDICTION MODEL BASED ON ENSEMBLE LEARNING

The heterogeneous individual learning devices used in the default predicting model proposed in this paper based on multiple-times ensemble learning are XGBoost, DNN, and LR. XGBoost, as an expansion of the gradient boosting system, demonstrates fast, efficient and configurable features, and it can improve the precision of the classification system. And meanwhile, XGBoost has been used to train models in Kaggle machine learning contests for many times and won championships. A typical deep learning model, DNN is a very deep neural network featuring a large amount of parameters and a high capacity, which means that it can perform more complicated learning tasks. And LR, as a classic classification method in statistical learning, is often used for modeling of credit risks. This is because LR produces not only less strict presumptions but also results between 0 and 1, which can be interpreted as probabilities of specific observations in a particular set of a certain group. Based on the above analysis, the above three heterogeneous individual learning devices were used to build a default risk prediction model for P2P lending through linear weighted fusion. The flow chart of the model's frame structure is shown in Fig. 1.

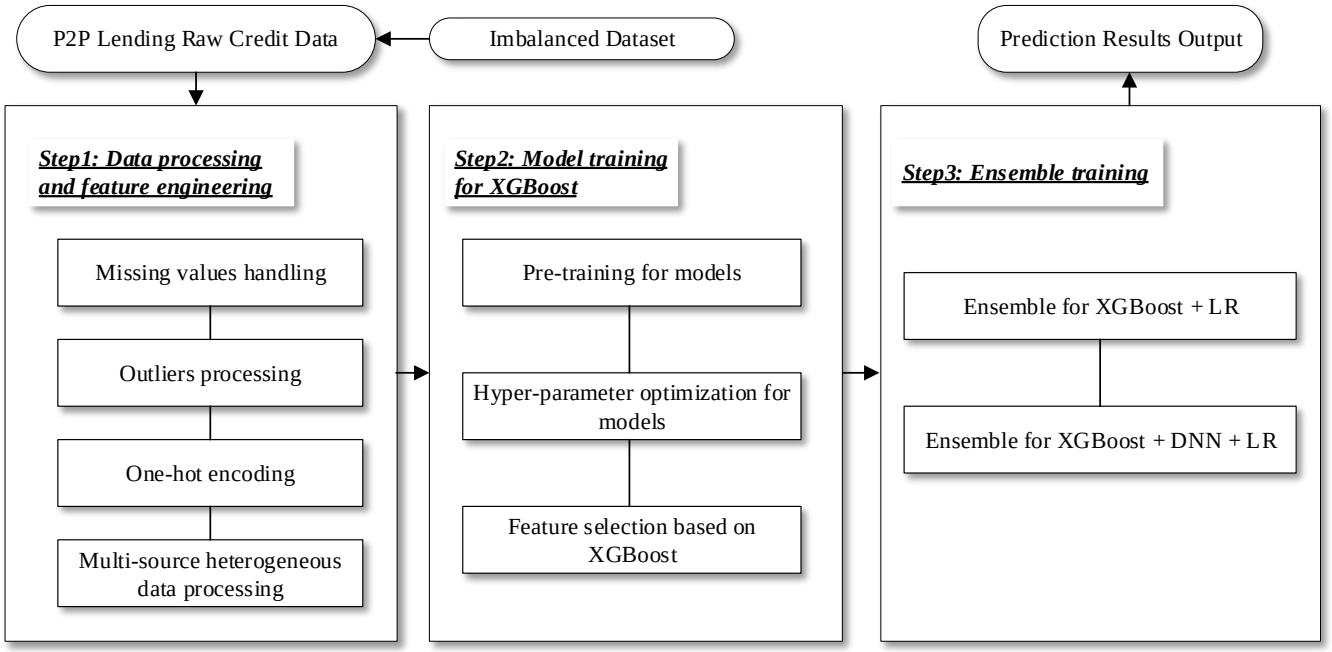

FIGURE 1. Framework of default prediction model based on ensemble learning

B. DATA ANALYSIS AND PROCESSING

At first, it' necessary to clean the credit dataset before building the default prediction model, so that the dataset can be used in the model. In data preprocessing and feature engineering, missing values and abnormal values of feature attributes are counted and processed, including deletion and padding. For categorized data, geospatial data and other unstructured multi-source data, the methods of sorting, one-hot encoding, etc., are used for processing. After adding and formally transforming the features of new datasheet built by consolidating the datasheet and the original features of the datasheet, the features are dynamically processed according to the cross-validation feedback results of the model, and the degree of importance of the variable data in the dataset is output by the XGBoost model.

C. MODEL TRAINING FOR XGBOOST

As a heterogeneous individual learning device, XGBoost (extreme gradient boosting) itself is an integrated model. It, an advanced gradient boosting (GB) system, is proposed by T Chen and C Guestrin and can be obtained freely as an open

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2810864, IEEE Access

source package [43]. XGBoost can optimize the objective function and its estimates. Fast, efficient and extendable, the system delivers excellent performance on many standard classification benchmarks. Different from the traditional gradient boosting decision tree (GBDT) which only uses the first-order derivative information, XGBoost conducts second-order Taylor series expansion on loss functions and adds regular terms in addition to the objective function to get optimal solutions overall so as to weigh the decrease of the objective function and the complexity of the model and thus avoid over-fitting.

Next, we used the XGBoost model to pre-train the data set. XGBoost has a large number of parameters, very difficult to produce excellent model performance, and thus hyperparameter optimization was required. Here, importance was placed on the iterative training of six parameters, namely, eta, gamma, max_depth, min_child_weight, colsample_bylevel, and lambda so as to find their optimal values. In addition, model-based feature selection was used to select features. This was because model learning and feature selection were performed simultaneously. And the decision tree based algorithm could output the significance of the features after the training of the model was completed. Therefore, we used XGBoost model to conduct feature selection, which produced high efficiency and precision.

D. ENSEMBLE TRAINING

At last, after pre-training and hyperparameter optimization of the model, once again we used ensemble learning to perform linear weighted fusion on the three heterogeneous individual learning devices, that is, XGBoost, DNN, and LR. DNN (deep neural network), a typical deep learning method, is a neural network with at least two hidden layers. The traditional multi-layer feedforward network with a single hidden layer already has a very strong learning ability. However, obviously it is more effective to add hidden layers than to add nerve cells on the hidden layer, which is because the addition of hidden layers increases not only the number of nerve cells that have an activation function but also the number of nested layers of the activation function [44]. Therefore, with DNN as a default risk predicting model, its input layers matched the feature space of the credit samples and its output layers contained dichotomous results, that is, default or normal. All layers were composed of nerve cells and work as elementary units of the DNN model. LR (logistic regression) was a typical classification method in machine learning, which was a statistics-based learning model with sound statistical basis and interpretability. However, due to the linear limitation of LR on variable relations, it was difficult to get the optimal precision. Therefore, its features could be given full play during modeling: (1) it reduced over-fitting by adding an L2 penalty function; (2) it served as a benchmark in rapid assessments on data cleansing effects and model phenotypes; (3) it made predictions by weighting and combining models of different structures to supplement the precision and robustness of the original model.

After that, XGBoost was combined with multiple models. This is because the starting point of XGBoost is that all variables are completely independent and approach to the real correlation through superposition of the dichotomous correlations of the decision tree. And the starting point of the neural network is that variables are full of complicated nonlinear correlations and approach to the real correlation through constant optimization of the network weight. Structures of the two models are highly complementary. In addition, LR is a simple, fast, and robust algorithm featuring high interpretability. At last, we used the weighted average method to combine models. Although it was to conduct linear complementation using partial independence of different models at the same level based on very limited knowledge about statistics and machine learning, at least it was robust as a method improving single-model results and avoiding increased over-fitting risks in a time and effort saving manner.

Therefore, the multiple-times ensemble learning model based on the heterogeneous integration framework designed in this paper was used to predict default risks, and it integrated XGBoost, DNN, and LR together as heterogeneous individual learning devices to conduct multiple times of ensemble learning as well as linear weighted fusion. First, XGBoost was used as an initial ensemble learning model to perform pre-training and then, the hyperparameters of the model were optimized to get an optimized predicting result. According to comparative analysis, the predicting result was better than that after training of DNN and LR models. Then, XGBoost model and LR model were further trained through linear weighted fusion. And finally, ensemble learning was conducted among XGBoost model, DNN model, and LR model. The model prediction precision after multiple times of ensemble learning was the highest, indicating that the model proposed in this paper delivered excellent performance.

IV. EXPERIMENTAL SETUP AND RESULTS ANALYSIS

A. DATA DESCRIPTION

The experimental data in this paper come from a well-known P2P lending enterprise in China. The credit dataset is the credit transaction data of the clients, including the credit default tag (dependent variable), modeling foundation and processing field (independent variable), original data of relevant user's network behavior, etc. The dataset is composed of three Comma-Separated Values (CSV) files, namely Master CSV, Log_Info CSV and Userupdate_Info CSV. Those files respectively record the borrowers' ID, characteristics, network behavior, education information, third party information, social network information, loan transaction time, default tag and other credit data information. There are 228 columns in Master CSV, 5 columns in Log_Info CSV and 4 columns in Userupdate_Info CSV, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2810864, IEEE Access

shown in Table 1. As the three CSV files are linked by the Idx field of each table, there are a combined total of 80,000 borrowers, after the P2P lending credit dataset is consolidated. Finally, in order to verify the validity and accuracy of the prediction model proposed in this paper, we use 75% of the borrow data in the credit dataset for training and the remaining 25% for testing. Additionally, in the test, the tenfold cross-validation is also used.

TABLE 1. Credit dataset feature field description

Name of Datasheet	Name of Datasheet Field Category	Datasheet Field Category Description
Master CSV	Idx	The unique key for each borrower, which can match Idx in two other files
	UserInfo_*	Borrower's feature field, totaling 24 columns
	WeblogInfo_*	Info network behavior field
	Education_Info*	Degree field
	ThirdParty_Info_Period*_*	Third-party credit-related data field
	SocialNetwork_*	Social network field
	LinstingInfo	Loan transaction time
	Target	Default tag
Log_Info	Idx	Unique key for every loan
	ListingInfo	Loan transaction time
	LogInfo1	Operation code
	LogInfo2	Operating category
	LogInfo3	Login time
Userupdate_Info	Idx	Unique key for every loan
	ListingInfo1	Loan transaction time
	UserupdateInfo1	Modification content
	UserupdateInfo2	Modification time

B. DATA PROCESSING AND FEATURE ENGINEERING

Before a credit dataset is used for default risk prediction, it's necessary to do the necessary data preprocessing and feature engineering. First of all, in this credit dataset, in addition to Master CSV, there are two CSV files, namely Log_Info and Userupdate_Info which are related to the historical records of a borrower and should be consolidated with the master file of Master CSV. The historical records are grouped by index, and various rows of information are summarized to each column, so that each index corresponds to the only line that leads to the main file. After the Master CSV and Log_Info CSV, Userupdate_Info CSV are consolidated with the field "Idx" as the only index, the file in the HDF5(Hierarchical Data Format version 5) data format is finally generated to facilitate the reading by the computer.

Second, in the field of credit risk identification and prediction, the perfection of the borrower's information will also affect its credit rating. Credit rating of the client with 100% perfect information can be more accurately predicted than that of the client with 50% perfect information. Based on this, we conducted a multi-dimensional analysis and processing of missing values. The number of missing values can be calculated by column (feature attribute), and the missing ratio of each column can be further obtained.

FIG. 2 displays the attributes with the missing value and the corresponding deletion ratios. As you can see from the figure, the missing value ratio for WeblogInfo_1 and WeblogInfo_3 is 97%. Thus, we remove two columns, as they basically contain no useful credit information. The missing value ratio for UserInfo_11, UserInfo_12, and UserInfo_13 which are categorical is 63%, and the missing values are populated with -1. "Whether to miss" is treated as another category here. Other numeric attributes with smaller missing value ratio are populated with the median value. In addition, the number of missing values can be used as a feature to measure the perfection of user credit information. In the control of credit risk, historical records of each loan, such as its start time and the total frequency of information updating, etc., are also important for measuring the behavior of borrowers. Furthermore, occurrence frequency of each type of sub-event is counted, so as to dig out more relevant factors that affect the credit of borrowers.

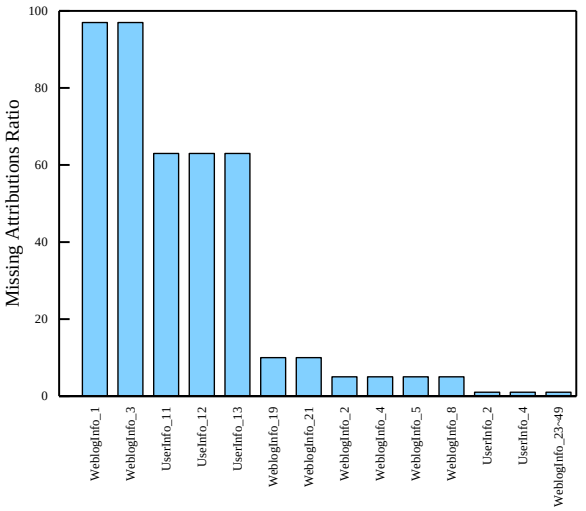

FIGURE 2. Ratio of missing attributions

Then, there are 190 numeric feature attributes in the original credit data. By calculating the standard deviation between attribute values, some of the features with small changes are removed. Table 2 lists 15 feature attributes whose standard deviation is close to 0, and which are removed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2810864, IEEE Access

Table 2. Removal of the standard deviation of numeric feature attributes

Attribute	SD	Attribute	SD	Attribute	SD
WeblogInfo_10	0.0707	WeblogInfo_41	0.0212	WeblogInfo_49	0.0071
WeblogInfo_23	0.0939	WeblogInfo_43	0.0372	WeblogInfo_52	0.0512
WeblogInfo_31	0.0828	WeblogInfo_44	0.0166	WeblogInfo_54	0.0946
WeblogInfo_32	0.0834	WeblogInfo_46	0.0290	WeblogInfo_55	0.0331
WeblogInfo_40	0.0666	WeblogInfo_47	0.0401	WeblogInfo_58	0.0609

The points which are inconsistent with the general behavior or features of other sample points in sample space are called outliers. Considering that the anomalous feature of outliers may be a multi-dimensional combination, a small number of outliers are removed by analyzing the number of missing values of sample attributes. Here, we adopt a simple and effective method to judge the outliers: training the original data with XGBoost model to get the importance of outputting feature with XGBoost model. On the basis of this, we choose the most important top 20 feature attributes (as shown in Fig. 3), and count the number of missing values of each sample in terms of 20 features. Then, take and remove the samples with missing values greater than 10 as outliers, which should be removed to facilitate the learning process.

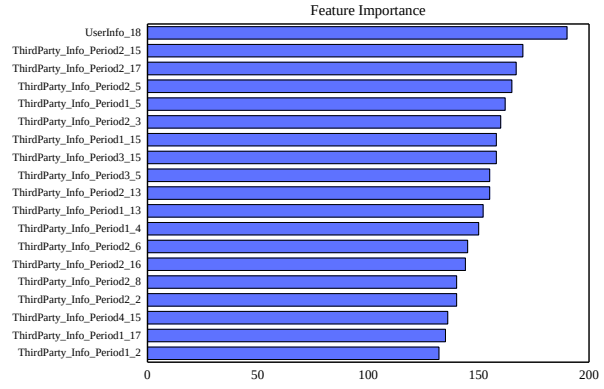

FIGURE 3. Feature importance

In addition, other processing including the conversion of upper and lower cases of characters, treatment of whitespaces and city names were performed. (1) with respect to upper and lower cases of characters such as the UserupdateInfo1 field in the Userupdate_Info sheet, their attribute values were extracted as per English characters which encompassed both upper and lower cases. For instance, the characters of “QQ” and “qQ" apparently should be assigned the same value. Thus, we uniformly converted all characters into lower case letters. (2) with respect to whitespaces such as the UserInfo_9 fields containing whitespace characters in the Master Sheet, such as “ZhongguoYidong” and “Zhongguo Yidong”, the same values were assigned to them after removing the whitespaces. (3) with respect to city names such as “Chongqing” and “Chongqing City” contained in UserInfo_8, the character “City” was removed since they actually refer to the same city.

Finally, the information on all geographical locations has undergone a categorical processing, where city names were replaced by longitudes and latitudes which had been collected from the sheet. In this way, categorical variables can be converted into numerical ones (for instance, “Beijing” can be replaced by the latitude and longitude value 39.92, 116.46, thereby obtaining two numerical characteristics on northern latitude and eastern longitude). As many algorithms (such as logistic regression and neural network models, etc.) can only deal with numerical characteristics, the categorical characteristics in the data set need to be encoded. We adopted a one-hot coding scheme to obtain 0-1 characteristics, thus resolving classifiers’ inability to deal with categorical characteristics.

C. HYPER-PARAMETERS OPTIMIZATION

Nonparametric classification algorithms used to predict default are almost rare. Thus, the hyperparameters must be optimized to boost the predicative power of algorithms. Firstly, hyperparameter optimization refers to the method of finding a group of parameters from possible value ranges of variables to help machine learning algorithms achieve satisfactory effects. Secondly, most machine learning algorithms require a vast amount of time in training, while some parameters need to be configured ahead of the training. Thus, these variables impose great impacts on training results. However, there is no such a group of variables that are suitable for all data sets, that means they should be specifically configured in particular applications. For instance, SVM involves hyperparameters like the kernel function and regular coefficient, decision tree involves tree building and pruning strategies, while neural network contains a large number of hyperparameters. In real-world application, manual parameter tuning is employed as a relatively traditional approach to obtaining experiential value or grid search; however, when the number of hyperparameters is too large (larger than 5, for instance), the above parameter tuning approach usually results in lesser effects than random searching.

Simply put, approaches like manual search, grid search and random search are all popular strategies applied in hyperparameter optimization. In this paper, a Python library – Hyperopt was selected as the parameter tuning tool. The open source library of Hyperopt has provided automatic optimization algorithms and software structure for hyperparameters, while Hyperopt has provided interfaces to transfer parametric spaces and evaluate functions. Presently, it supports three optimization algorithms including random

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2810864, IEEE Access

search, simulated annealing, and Tree-of-Parcen-Estimators (TPE). Nowadays, Hyperopt can already be used in deep neural networks, convolutional neural networks and Scikit-Learn open source machine learning library (a Python-based open source scientific computing package). TPE has been used for careful tuning and optimization of hyperparameters in XGBoost.

Prior to parameter tuning and optimization, an XGBoost model was used to predict default on credit data, obtaining the AUC value of 0.7850. After the employment of the TPE optimization package provided by Hyperopt, the results obtained are shown as in Fig. 4. As can be seen, after 30 iterations, the optimization has facilitated the XGBoost model to achieve an average prediction accuracy of 0.7869 in predicting defaults. A comparison found that the XGBoost model that had undergone hyperparameter optimization indeed resulted in improved prediction accuracy by 0.24% than before.

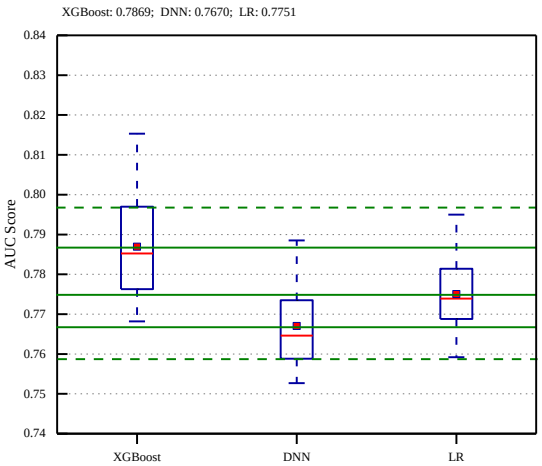

FIGURE. 4. The average AUC score for XGBoost, DNN, LR

D. PERFORMANCE METRICS

In this paper, the AUC score (the score of area under ROC curve) is used as the evaluation criteria for the default risk prediction model [40,41]. In the traditional two-category default prediction algorithms related to balanced data, the prediction accuracy is usually used as an evaluation indicator. However, for the two-category default prediction related to unbalanced data, the prediction accuracy cannot really reflect the performance of the classifier. Thus, if the accuracy acts as the evaluation indicator, the prediction model will predict the default clients as normal and cannot accurately identify the default clients, because the category tags are very unbalanced in the datasets. For example, in the training dataset in this dataset, normal users (i.e., non-default users) accounts for 92.7% of the totals. If a classifier predicts all users as normal, the classification accuracy of this classifier is up to 92.7%. However, in fact, such a classifier is of no real value, as it is unable to identify the default user. Our ultimate goal is to successfully identify the default clients in the credit dataset. Therefore, the precision evaluation criteria commonly used in two-category problems is not applicable to unbalanced datasets.

E. RESULT ANALYSIS AND DISCUSSION

After cleaning the credit dataset and performing feature engineering, we start to pre-train the model and get the unoptimized prediction results. First, the credit data is pre-trained using the XGBoost model to obtain the AUC score of 0.7850. Then, parameters of the XGBoost model are optimized by Hyperopt, a Python open source library. Here, Hyperopt provides an optimization interface that accepts an evaluation function and parameter space, calculates the loss function value for a point in the parameter space, and also requires specifying the distribution of parameters in the space. There are four important factors to Hyperopt, including specifying the function that needs to be minimized, the search space, trails database (optional), and the search algorithm (optional). 6 parameters of the XGBoost model, namely eta, gamma, max_depth, min_child_weight, colsample_bylevel and lambda are optimized with Hyperopt. After 30 iterations of hyperparameter optimization, the prediction result of XGBoost model is 0.7869, which is improved compared with the result not optimized, as shown in Table 3.

TABLE 3. Contrastive analysis for models

Model	XGBoost	DNN	LR	XGB+LR	XGB+DNN+LR
Category	GBDT	NN	LR	Ensemble	Ensemble
Basic Unit	Bipartite	Network	Linearity	-	-
AUC Score	0.7869	0.7670	0.7751	0.7882	0.7891
Interpretability	Normal	Poor	Good	Poor	Poor

From the analysis of Fig. 4 and Table 3, it can be seen that after optimization by hyperparameters, the AUC score of the prediction result is 0.7670 and 0.7751 respectively for DNN model and LR model which are both trained for 30 iterations as single individual learners. By comparing three different prediction results after hyperparameter optimization, it can be found that after a single ensemble learning, the prediction accuracy of the XGBoost model is higher than that of DNN and LR models, which illustrates that the single ensemble learning method can overtop the traditional machine learning models in predicting the credit risk default.

With model-based feature selection methods, we know the importance of each type of variable in the default risk prediction model. Here, based on the XGBoost model, the process of training XGBoost is the process of ranking the features by their importance. As shown in Fig. 5, although the number of ThirdParty variable is roughly the same as that of the UserInfo variable, the ThirdParty variable carries the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2810864, IEEE Access

highest weight and therefore plays the most significant role in the model. That is to say, the ThirdParty variable is the most influential factor. At the same time, it can be observed that the impact of the Education variable in the model is minimal. In other words, whether a person violates the contract or not is not completely determined by credit information such as educational background.

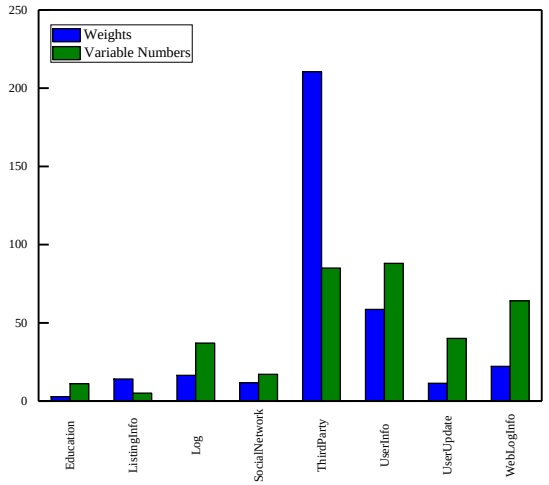

FIGURE 5. Weight distribution of variables using XGBoost

Subsequently, the ensemble learning method was continuously employed to jointly treat the XGBoost, DNN and LR models as heterogenous ensemble individual learners. Again, the ensemble learning method was employed to fuse the three models above by using linear weighting method. From Fig. 6 and Fig. 7, we can see that after the fusion of XGBoost and LR models, the optimal weight ratio of them in the linear fusion is 0.9:0.1, and the AUC value is 0.7882. After that, another round of linear weighting fusion across XGBoost, DNN and LR models were performed, which eventually produced a multi-ensemble learning model based on heterogeneous ensemble frameworks which is used to predict the default risks of P2P lending customers, where the optimal weight ratio of the linear fusion is 0.75:0.20:0.05, and the AUC value is 0.7891. A comparatively analysis and validation found that the multi-round ensemble learning model has resulted in superior performance in predicting default risks of P2P online customers.

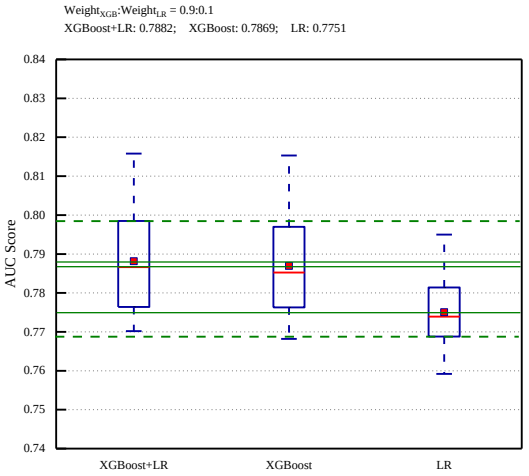

FIGURE 6. XGBoost + LR for prediction score

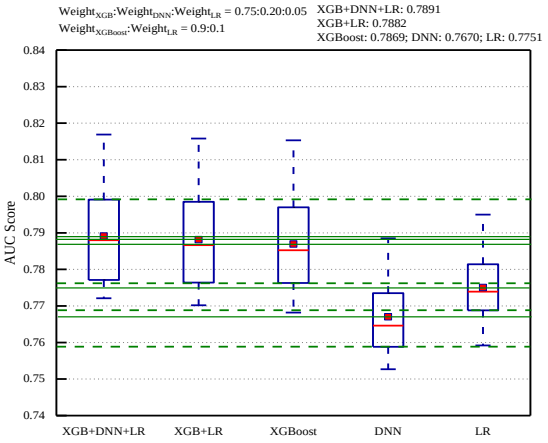

FIGURE 7. XGBoost + DNN + LR for prediction score

The XGBoost, DNN and LR models have been treated as individual learners for heterogeneous ensemble to undergo a linear weighting fusion. After hyperparameter optimization by using the Hyperopt, a Python-based open source library, the three models produced individual predicative results on default risks, as shown in Table 2. These results indicate that XGBoost, as an ensemble learning model itself, is of certain applicability to predicting default risks of P2P online borrowers. Following optimization of the DNN model, we have found that although prediction accuracy of an individual DNN is relative low, the performance of models can be further improved through the complementary effect of their structures. After parameter tuning, we obtained the optimal weight ratio for the combination of XGBoost, DNN and LR models being 0.75:0.20:0.05. The final predicative AUC value of such an ensemble model can reach 0.7891, as shown in Figure 5 and Table 2, indicating that the model has achieved an ideal predicative effect.

In conclusion, the linear weighted fusion model proposed in this paper, which uses heterogeneous individual learners for multiple integrations, is used to predict the client default in China's P2P lending marketplace and achieve good results. Our experiments verify that the repeated ensemble learning model has strong applicability in predicting the P2P credit risk, which is high-dimensional with unbalanced distributed categories and large samples. At present, China's lending market is developing rapidly with an increasing market capacity. Therefore, there is a strong demand for credit risk identification. Provision of an effective default risk forecasting model has an important practical demand for raising the industry risk management level.

V. CONCLUSION

In this paper, we proposed a multi-round ensemble learning model of heterogeneous ensemble frameworks comprised of XGBoost, DNN and LR individual learners, which was used to predict the default risks of P2P online borrowers in China. In the ensemble model, we firstly used XGBoost as the main modeling algorithm to the initial ensemble, producing better performances in terms of prediction accuracy, robustness, application range and running speed, etc. Then, complementary algorithms of DNN and LR were applied to boost the model's final prediction accuracy. The LR algorithm enjoyed high modelling speed and better performances in evaluating and interpreting the model. Additionally, the DNN algorithm excelled at exploiting complex non-linear relations, thus complementing the improvement of prediction accuracy of the ensemble model.

During data cleansing and pre-processing, a statistical treatment was performed on each characteristic attribute of the data set as per simple statistics (such as the median, variance, quartile, maximum and minimum values), followed by the cleansing of variables. The purpose of these procedures is to convert all variables into numerical variables for modelling under the premise that effective information has been retained to the highest extent. All categorical variables were converted into 0-1 variables by using the one-hot encoding scheme. By now, all variables have been converted into numerical variables. The majority of variables with missing or the same values were deleted in order to eliminate as many as possible the redundant variables containing very little effective information. Later, the median (or mean value) was used to fill in missing values, followed by the standardization of required data. In doing so, the cleansed data can be used for modelling.

Considering the poor performance of traditional classifiers in dealing with imbalanced data, as well as the difference of the case from traditional ensemble models, the XGBoost model was used in the initial ensemble, improving the default prediction accuracy of P2P online borrowers – the AUC value – to 0.7869 after experiencing hyperparameter optimization. Following the initial ensemble, the XGBoost, DNN and LR algorithms were used again for linear weighting fusion, eventually boosting the prediction accuracy – the AUC value – to 0.7891 and thus producing superior performance.

REFERENCES

[1] E. Lee and B. Lee, "Herding behavior in online P2P lending: An empirical investigation," Electron. Commer. *Res. Appl.*, vol. 11, no. 5, pp. 495-503, Sep.-Oct. 2012.

[2] Y. Guo et al., "Instance-based credit risk assessment for investment decisions in P2P lending," *Eur. J. Oper. Res.*, vol. 249, no. 2, pp. 417-426, Mar. 2016.

[3] C. Serrano-Cinca and B. Gutiérrez-Nieto, "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending," *Decis. Support Syst.*, vol. 89, pp. 113-122, Sep. 2016.

[4] G. Dorfleitner et al., "Description-text related soft information in peer-to-peer lending - Evidence from two leading European platforms," *J. Bank Financ.*, vol. 64, pp. 169-187, Mar. 2016.

[5] B. Baesens et al., "Benchmarking state-of-the-art classification algorithms for credit scoring," *J. Oper. Res. Soc.*, vol. 54, no. 6, pp. 627-635, Jun. 2003.

[6] D. West, "Neural network credit scoring models," *Comput. Oper. Res.*, vol. 27, no. 11, pp. 1131-1152, Sep. 2000.

[7] A. F. Atiya, "Bankruptcy prediction for credit risk using neural networks: A survey and new results," *IEEE Trans. Neural Netw.*, vol. 12, no. 4, pp. 929-935, Jul. 2001.

[8] Y. Yang, "Adaptive credit scoring with kernel learning methods," Eur. *J. Oper. Res.*, vol. 183, no. 3, pp. 1521-1536, Dec. 2007.

[9] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *J. Bank Financ.*, vol. 34, no. 11, pp. 2767-2787, Nov. 2010.

[10] S. Huda et al., "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis," *IEEE Access*, vol. 4, pp. 9145-9154, Nov. 2016.

[11] Z. H. Zhou and X. Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63-77, Dec. 2006.

[12] Y. Sun et al., "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358-3378, Jan. 2007.

[13] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. *Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263-1284, Sept. 2009.

[14] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowl. Inf. Syst.*, vol. 25, no. 1, pp. 1-20, 2010.

[15] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *IEEE Trans. Syst. Man Cybern. Paart A-Syst. Hum.*, vol. 41, no. 3, pp. 552-568, Nov. 2011.

[16] R. Emekter et al., "Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending," *Appl. Econ.*, vol. 47, no. 1, pp. 54-70, 2015.

[17] Y. Peng et al. "An empirical study of classification algorithm evaluation for financial risk prediction," *Appl. Soft. Comput.*, vol. 11, no. 2, pp. 2906-2915, Mar. 2011.

[18] T. C. Wu and M. F. Hsu, "Credit risk assessment and decision making by a fusion approach," *Knowledge-Based Syst.*, vol. 35, pp. 102-110, Nov. 2012.

[19] J. J. Liao et al., "An ensemble-based model for two-class imbalanced financial problem," *Econ. Model.*, vol. 37, pp. 175-183, Feb. 2014.

[20] Y. Wei et al., "Credit scoring with social network data," Mark. Sci., vol. 35, no. 2, pp. 234-258, 2015.

[21] S. Lessmann et al., "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124-136, Oct. 2015.

[22] P. Danenas, and G. Garsva, "Selection of support vector machines based classifiers for credit risk domain," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3194-3204, Apr. 2015.

[23] M. Ala'raj and M. F. Abbod, "Classifiers consensus system approach for credit scoring," *Knowledge-Based Syst.*, vol. 104, pp. 89-105, Jul. 2016.

[24] M. Ziba et al., "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy

in the lung cancer patients," *Appl. Soft. Comput.*, vol. 14, no. A, pp. 99-108, Jan. 2014.

[25] J. A. Sanz et al., "A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 4, pp. 973-990, Jul. 2014.

[26] W. C. Lin et al., "Clustering-based undersampling in class-imbalanced data," *Inf. Sci.*, vol. 409, pp. 17-26, Oct. 2017.

[27] A. Zakaryazad and E. Duman, "A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing," *Neurocomputing*, vol. 175, no. A, pp. 121-131, Jan. 2016.

[28] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," *Knowledge-Based Syst.*, vol. 41, no. 1, pp. 16-25, Mar. 2013.

[29] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Comput. Intell.*, vol. 20, no. 1, pp. 18-36, Jan. 2004.

[30] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 563-597, Jun. 2016.

[31] W. Lin et al., "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," *IEEE Access*, vol. 5, pp. 16568-16575, Sep. 2017.

[32] N. V. Chawla et al., "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, 2002.

[33] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993-1001, Oct. 1990.

[34] S. Jones, D. Johnstone, and R. Wilson, "An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes," *J. Bank Financ.*, vol. 56, pp. 72-85, Jul. 2015.

[35] S. Y. Kim and A. Upneja, "Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models," *Econ. Model.*, vol. 36, pp. 354-362, Jan. 2014.

[36] L. Zhou and K. K. Lai, "AdaBoost Models for Corporate Bankruptcy Prediction with Missing Data," *Comput. Econ.*, vol. 50, no. 1, pp. 69-94, Jun. 2017.

[37] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123-140, Aug. 1996.

[38] A. Ekinci, and H. İ. Erdal, "Forecasting Bank Failure: Base Learners, Ensembles and Hybrid Ensembles," *Comput. Econ.*, vol. 49, no. 4, pp. 677-686, Apr. 2017.

[39] J. Sun, M. Y. Jia, and H. Li, "AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9305-9312, Aug. 2011.

[40] J. Heo and J. Y. Yang, "AdaBoost based bankruptcy forecasting of Korean construction companies," *Appl. Soft. Comput.*, vol. 24, no. 24, pp. 494-499, Nov. 2014.

[41] H. Guo et al., "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification," *Eng. Appl. Artif. Intell.*, vol. 49, no. C, pp 176-193, Mar. 2016.

[42] E. Mayhua-López, V. Gómez-Verdejo, and A. R. Figueiras-Vidal, "A new boosting design of Support Vector Machine classifiers," *Inf. Fusion*, vol. 25, no. C, pp. 63-71, Sep. 2015.

[43] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, ACM, 2016, pp. 785-794.

[44] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May. 2015.

WEI LI received the M.S. degree in finance from Guizhou University of Finance and Economics, Guiyang, Guizhou, China, in 2014. He is currently pursuing the Ph.D. degree in business administration from Hefei University of Technology, Hefei, Anhui, China. His recent research interests include financial risk management and financial data analysis.

SHUAI DING received the M.S. and Ph.D. degrees from Hefei University of Technology, Hefei, China, in 2011 and 2008. From 2011 to 2013, he was a visiting scholar in the Department of Computer Science at University of Pittsburgh. He is currently an associate professor of management science and engineering at the Hefei University of Technology. He has published more than 20 articles in refereed journals such as IEEE Transactions on Fuzzy Systems, Decision Support Systems, International Journal of Production Research, and IEEE Systems Journal. His research interests include the cloud service recommendation, e-health and smart healthcare, social network modeling, and trust computing.

YI CHEN received the M.S. degree from Hefei University of Technology, Hefei, Anhui, China, in 2014, where she is currently pursuing the Ph.D. degree. Her recent research interests include healthcare analytics and medical decision-making, evolutionary game.

SHANLIN YANG is a member of the Chinese Academy of Engineering, and the leading Professor in management science and information system at School of Management, Hefei University of Technology. He is the director of academic board of Hefei University of Technology, and the director of National-Local Joint Engineering Research Center of "Intelligent Decision and Information System". He has own 2 second class prizes for State Scientific and Technological Progress Award, and 6 first class prizes for provincial and ministerial level science and technology award. He has published 5 academic works and more than 400 papers in important journals and international conferences worldwide. His research interests include information systems, social network, cloud computing, and artificial intelligence.