

Received October 16, 2020, accepted October 28, 2020, date of publication November 2, 2020, date of current version November 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035128

Semi-Supervised Learning Classification Based on Generalized Additive Logistic Regression for Corporate Credit Anomaly Detection

SONG HAN 

Information Institute, Beijing Wuzi University, Beijing 101149, China

e-mail: hansong@bwu.edu.cn

This work was supported in part by the Beijing Social Science Fund Project under Grant 19YJC033, in part by the Beijing Education Commission Social Science Program Project under Grant SM201910037002, and in part by the Youth Top-Notch Talent Cultivation Plan of Beijing Municipal Colleges and Universities under Grant CIT&TCD201704060.

ABSTRACT Conventional corporate credit evaluation models are primarily based solely on financial variables in conjunction with supervised learning methods. However, the acquisition of the labeled sample information required by supervised learning methods is generally a costly and lengthy process, and is therefore difficult to obtain in practice, while the introduction of non-financial variables can be expected to provide greater diagnostic scope. The present study addresses these issues by proposing a semi-supervised generalized additive logistic regression model for detecting corporate credit anomalies based on a high proportion of unlabeled sample information that includes both financial and non-financial variables. The model not only can accommodate linear non-separable problems, but can also be trained using both labeled and unlabeled samples at the same time, while simultaneously realizing parameter estimation and variable selection. We also develop more precise definitions of corporate credit anomalies to increase the accuracy of corporate default risk assessments. The model is trained and tested using a dataset composed of actual financial and non-financial corporate data freely available on the Internet, and is demonstrated to provide better variable selection and credit anomaly prediction with better accuracy and robustness than other state-of-the-art models. The results reveal key financial variables correlated with corporate credit anomaly detection, and also verify that the non-financial variables significantly improve the corporate credit anomaly prediction accuracy of the model.


INDEX TERMS Corporate credit anomaly, semi-supervised learning, generalized additive logistic, risk transparency, surrounding risks.

I. INTRODUCTION

Credit is a generalized metric that appraises the likelihood that a debtor can repay the principal and interest on a loan as scheduled without default. A reliable appraisal of the risk of default for corporate entities is of particular importance for ensuring the financial health of banks, securities companies, funds, and other investors. Accordingly, corporate credit assessment has always been an important research topic in international academic and financial circles. While corporate credit ratings are published at regular intervals by professional rating agencies such as Standard & Poor's (S&P) and Moody's Investor Service, these institutions charge

high service fees, they do not assess all of the numerous small and medium-sized enterprises worldwide, and the sporadic nature of the ratings provided cannot reflect short-term issues that may affect a company's risk of default. Therefore, market participants require supplemental tools for evaluating corporate risk of default in real time for any corporate entity for which pertinent information can be obtained.

The evaluation of corporate credit is widely discussed in current literature. According to the existing research, corporate credit evaluation depends on corporate credit evaluation models, evaluation variables, and meaningful definitions of corporate credit anomaly. Conventional corporate credit evaluation models are primarily based on financial variables in conjunction with supervised learning methods. However, the

The associate editor coordinating the review of this manuscript and approving it for publication was Qingchao Jiang .

acquisition of the labeled sample information required by supervised learning methods is generally a costly and lengthy process, and is therefore difficult to obtain in practice. Moreover, the use of both financial and non-financial variables can be expected to increase the diagnostic scope of corporate credit evaluation models. However, while the limited availability of labeled sample information can be addressed by the use of a semi-supervised learning method, the introduction of non-financial variables will result in high-dimensional data that quickly erodes the training efficiency of semi-supervised learning methods. Finally, existing evaluation methods rely on corporate credit anomaly definitions that are subject to severe limitations.

The study addresses the above-discussed deficiencies in currently available tools applied by market participants for evaluating corporate risk of default in real time. First, we propose a semi-supervised generalized additive logistic regression (SSGALR) model for detecting corporate credit anomalies based on a high proportion of unlabeled sample information. Second, the model employs both financial risk variables and non-financial risk variables, which are obtained from publicly available information contained in the text of corporate financial reports and the risks of surrounding enterprises. The SSGALR model is able to make full use of unlabeled non-financial data samples to improve its learning performance by simultaneously conducting variable selection when processing high-dimensional data. We verify that the adopted non-financial variables improve the default risk prediction accuracy of the model. Third, we comprehensively define corporate credit anomalies according to a wide range of factors reflecting corporate dishonesty based on information publicly disclosed by market sources and other institutions. The proposed model is trained and tested using a dataset composed of actual corporate data, and is demonstrated to provide better variable selection and credit anomaly prediction with better robustness than other state-of-the-art models, including supervised semi-parametric logistic regression, supervised logistic regression, and random forest models.

The remainder of the article is organized as follows. A literature review is presented in Section II. The proposed SSGALR model and related algorithms are introduced in Section III. The model input variables, selection of empirical samples, and data collection are discussed in Section IV, and the model verification results are also presented. Section V concludes the study and presents future research directions.

II. RELATED STUDIES

A. CREDIT EVALUATION MODELS

Current mainstream corporate default risk assessment methods employ statistical and machine learning models based on historical information. The most frequently used models include neural network [1], [2], support vector

machine [3], [4], random forest [5], [6], AdaBoost [7], and logistic regression models. Of these, logistic regression models are most widely used because of their explanatory power, prediction accuracy, and relatively simple calculation methods [8]. Yan *et al.* [9] applied a logistic regression model to predict the loan default risk of listed companies, and the results indicated that the growth rate of capital maintenance, quick ratio, and asset turnover rate have a primary impact on the default risk of listed companies. However, logistic regression also faces the curse of dimensionality, which has become increasingly severe due to the reduced cost of data collection owing to advancements in computer technology. This was addressed by Fang *et al.* [10] by introducing lasso punishment into logistic regression for predicting loan default behavior, and demonstrated that the logistic lasso model not only offers faster calculations by selecting the most useful input variables from a high dimensional dataset, but can also conduct variable selection and coefficient estimation simultaneously. In addition, many logistic regression models are parametric, and therefore assume that the parameters of the model include all available information. However, little information is generally available regarding the form of a parametric regression model employed in risk assessment, and the presupposed model form can result in inaccurate risk predictions. This can be addressed through the adoption of a semi-parametric regression model that need not assume a model form in advance, while also retaining the flexibility of a non-parametric model [11]. This tactic was employed by Wang [12], who proposed a new type of default probability model based on the generalized additive model, and the model was demonstrated to provide high risk prediction accuracy. However, this model is not appropriate for use with high dimensional data because it cannot conduct variable selection. This was addressed by Zhang and Zhang [13], who incorporated group lasso punishment into a semi-parametric logistic regression model to obtain the generalized semi-parametric additive model for solving the credit scoring problem. The unique features of this model facilitates the simultaneous estimation of model parameters and the selection of input variables. This model serves as the basis of the semi-parametric additive model employed in the present work.

The practical difficulty of acquiring the labeled sample information required by supervised learning methods limits the size of labeled datasets available for model training, and the credit status of most companies is unknown. This represents a serious liability because model training conducted with only a small number of labeled samples often leads to low generalization performance and also wastes a large volume of available unlabeled sample data. These issues can be addressed by the use semi-supervised learning, which relies on only a small number of labeled samples as guidance for selecting unlabeled sample data during model training [14]. This model training approach is employed in the present work.

The concept of semi-supervised learning was first proposed in 1992 [15], and can be traced back to the early development of self-training algorithms [16]. Semi-supervised learning is based on a combination of supervised and unsupervised learning in accordance with four general learning scenarios: clustering, classification, reduction, and regression. Of these, classification is the most widely used learning scenario, and representative classification algorithms can be grouped into four general categories: generative algorithms [17], semi-supervised graph algorithms [18], co-training algorithms [19], and transductive support vector machines [20]. Semi-supervised learning technology has advanced greatly in recent years, and has been widely applied in fields such as text classification [21], and facial recognition, image retrieval, and video segmentation [22]. Moreover, semi-supervised learning has been applied in the field of credit evaluation [23]–[26]. However, this machine learning method is subject to some key problems that remain to be solved. For example, increasing sample dimension inevitably results in an increasing volume of redundant variables, which do not contribute to the training process and can greatly detract from training efficiency. However, conventional semi-supervised learning methods generally seek to balance the use of labeled and unlabeled sample data, and therefore suffer from poor training performance when employing high-dimensional data. While efforts have been made to address the issue of data dimension reduction in semi-supervised learning [27], many of the proposed methods are based on black box operation and suffer from poor interpretability. The SSGALR model proposed in the present study accomplishes data dimension reduction by appropriately selecting unlabeled data samples during the training process.

B. CREDIT EVALUATION VARIABLES

Corporate credit evaluation methods have commonly relied on financial variables, such as solvency, developmental, and business classifications, cash flow, and profitability, and the significant effect of these variables on credit evaluation has been verified [28], [29]. However, efforts have been made to increase the diagnostic scope of corporate credit evaluation models by introducing a number of non-financial variables, such as customer satisfaction [30], national financial development [31], enterprise innovation classification [32], corporate governance [33], and corporate social responsibility [34]. These data have been demonstrated to be available to some extent from various public sources, such as annual financial reports, prospectuses, and audit opinions [35]. In addition, research has demonstrated that the probability of dishonest behavior in an enterprise increases if the owners, key personnel, or investors of the enterprise have exhibited dishonest behavior [36]. This factor has been denoted as the network effect. However, little research has focused on the use of publicly available non-financial information, or evaluation of the network effect in corporate credit evaluation. The present study addresses this issue by including risk information

contained in corporate financial reports and risk information associated with surrounding enterprises based on financial variables and other data.

C. DEFINITION OF CORPORATE CREDIT ANOMALIES

Corporate credit anomalies must be defined meaningfully to facilitate the classification of the corporate sample information used in constructing corporate credit evaluation models. However, different definitions of corporate credit anomaly are presently applied in corporate credit status research. The primary definitions presently applied are based on either special treatment (ST) classifications, which are assigned when a company has been operating at a loss for two consecutive years [37], [38], high default risk classifications [39], [40], or high default rate classifications [41]. However, these definitions are subject to severe limitations. Existing assignments of ST classification status apply only to a small number of corporate entities, while the number of small and medium-sized enterprises is very large. As such, ST status alone is not fully representative of the overall credit worthiness of corporate enterprises. With regard to the other two classifications, corporate loan default risk and default rate classifications are presently poorly disclosed in publicly available information. Therefore, this information is too unreliable to serve as a basis for defining useful corporate credit anomalies. The present work addresses this issue by defining corporate credit anomalies according to a wide range of factors reflecting corporate dishonesty based on information publicly disclosed by market sources and other institutions. This information includes warnings of financial loss or bankruptcy due to financial difficulties, disclosed violations in the provisions of the securities market, involvement in legal proceedings due to poor contract performance and business assets being frozen by court order, citations for abnormal business operations or serious violations of law and credibility, and unfulfilled performance of legal demands.

III. MODELS AND ALGORITHMS

A. MODEL SETTINGS

Suppose q attribute variables (categorical variables) with $D_i = (1, D_{i,1}^T, D_{i,2}^T \cdots, D_{i,q}^T)$ representing the i -th observation of the attribute variables, where $D_{i,g}^T \in R^{df_g}$ ($g = 1, 2, \cdots, q$) represents q groups of dummy variables generated by the q attribute variables, and df_g represents the degree of freedom of the g -th group variables. Suppose k continuous variables, with $Z_i = (Z_{i,1}, Z_{i,2}, \cdots, Z_{i,k})$ representing the i -th observation of the continuous variables. The response variable Y is a binary variable.

If the entire sample is divided into two parts comprising labeled sample $L = \{D_i, Z_i, Y_i\}_{i=1}^{n_L}$ of n_L samples and unlabeled sample $U = \{D_i, Z_i\}_{i=n_L+1}^{n_U}$ of n_U samples, where $n = n_L + n_U$, then the response variables, attribute variables, and continuous variables can be written respectively in matrix

form as

$$Y = \begin{pmatrix} Y_L \\ Y_U \end{pmatrix}, \quad D = \begin{pmatrix} D_L \\ D_U \end{pmatrix} = \begin{pmatrix} 1 & D_{1,1}^T & \cdots & D_{1,q}^T \\ 1 & D_{2,1}^T & \cdots & D_{2,q}^T \\ \vdots & \vdots & \ddots & \vdots \\ 1 & D_{n,1}^T & \cdots & D_{n,q}^T \end{pmatrix},$$

$$Z = \begin{pmatrix} Z_L \\ Z_U \end{pmatrix} = \begin{pmatrix} Z_{1,1}^T & \cdots & Z_{1,k}^T \\ Z_{2,1}^T & \cdots & Z_{2,k}^T \\ \vdots & \ddots & \vdots \\ Z_{n,1}^T & \cdots & Z_{n,k}^T \end{pmatrix}, \quad (1)$$

where the subscripts L and U respectively represent the labeled and unlabeled components of the response variables, attribute variables, and continuous variables. The proposed SSGALR model is defined as

$$\log \left(\frac{E(Y|D, Z)}{1 - E(Y|D, Z)} \right) = h(D) + f_1(Z_1) + f_2(Z_2) + \cdots + f_k(Z_k), \quad (2)$$

where $h(\cdot)$ is a linear function, which assumes that the attribute variables enter into the proposed model in linear form, and $f_j(\cdot)$ is an unknown smooth function, which is a function of infinite order continuously derivable in its field of definition. It is assumed that continuous variables enter the proposed model in the form of non-parametric smooth functions, where a spline is used to spread the estimation:

$$f_j(Z_j) = \sum_{t=1}^T \delta_{jt} \varphi_t(Z_j), \quad j = 1, 2, \dots, k. \quad (3)$$

Here, $\varphi_t(\cdot)$ is a B-spline base, T is the number of terms in the expanded base, and δ_{jt} is the coefficient corresponding to the t -th basis function of the j -th continuous variable. We adopt $T = 3$, and expand the continuous variable into a cubic spline form. Then, the j -th continuous variable of Z can be rewritten as the following combination of basis functions:

$$Z_j = (\vartheta_1(Z_j), \vartheta_2(Z_j), \vartheta_3(Z_j)) = \begin{pmatrix} W_{1j}^T \\ W_{2j}^T \\ \vdots \\ W_{nj}^T \end{pmatrix}, \quad (4)$$

where $W_{ij}^T \in R^3, i = 1, 2, \dots, n$. Therefore, Z can be rewritten as

$$Z = \begin{pmatrix} Z_L \\ Z_U \end{pmatrix} = \begin{pmatrix} W_{1,1}^T & \cdots & W_{1,k}^T \\ W_{2,1}^T & \cdots & W_{2,k}^T \\ \vdots & \ddots & \vdots \\ W_{n,1}^T & \cdots & W_{n,k}^T \end{pmatrix}, \quad (5)$$

where $W_{i,g}^T \in R^{df_g}, df_g = 3, i = 1, 2, \dots, n$ and $g = 1, 2, \dots, k$. Then, we combine the \mathbf{D} and \mathbf{Z} in matrix form

as follows.

$$X = \begin{pmatrix} X_L \\ X_U \end{pmatrix} = (D, Z) = \begin{pmatrix} D_L & Z_L \\ D_U & Z_U \end{pmatrix} = \begin{pmatrix} 1 & D_{1,1}^T & \cdots & D_{1,q}^T & W_{1,1}^T & \cdots & W_{1,k}^T \\ 1 & D_{2,1}^T & \cdots & D_{2,q}^T & W_{2,1}^T & \cdots & W_{2,k}^T \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & D_{n,1}^T & \cdots & D_{n,q}^T & W_{n,1}^T & \cdots & W_{n,k}^T \end{pmatrix} \quad (6)$$

Denoting the coefficients of the labeled component X_L as β , and the coefficients of the unlabeled component X_U as α yields $\beta = (\beta_0, \beta_1^T, \dots, \beta_{q+k}^T)^T \in R^{\sum_{g=1}^{q+k} df_g + 1}$, where $\beta_0 \in R^{df_g}, g = 1, 2, \dots, q+k$, and $\alpha = (\alpha_0, \alpha_1^T, \dots, \alpha_{q+k}^T)^T \in R^{\sum_{g=1}^{q+k} df_g + 1}$, where $\alpha_0 \in R^{df_g}, g = 1, 2, \dots, q+k$.

Based on the above formalizations, the objective function of the semi-supervised generalized additive logistic is

$$l(\alpha, \beta) = l_L(\beta) - \lambda \sum_{g=1}^{q+k} s(df_g) \|\beta_g\|_2 - P_U(\alpha, \beta, \gamma), \quad (7)$$

which consists of a supervised component $l_L(\beta)$, a second component that identifies groups of coefficients for implementing the group lasso penalty, and an unsupervised component $P_U(\alpha, \beta, \gamma)$. These components are defined as follows. The first component $l_L(\beta)$ is a logarithmic likelihood function, and is expressed as

$$l_L(\beta) = \sum_{i=1}^{n_L} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] = \sum_{i=1}^{n_L} [y_i \eta_i - \log(1 + \exp(\eta_i))], \quad (8)$$

where $p_i = \exp(\eta_i) / (1 + \exp(\eta_i))$ and $\eta_i = \beta_0 + \sum_{g=1}^{q+k} x_{i,g}^T \beta_g = \beta_0 + x_i^T \beta$.

The second component $\lambda \sum_{g=1}^{q+k} s(df_g) \|\beta_g\|_2$ regards the coefficient of a dummy variable corresponding to the same attribute variable as a group, and the coefficient of the basis function of the same continuous variable is also regarded as a group. The group lasso penalty is applied to the variable matrix to select or eliminate all variables in the same group simultaneously. Here, $s(df_g)$ is the weight coefficient. The third component $P_U(\alpha, \beta, \gamma)$ is formulated according to the correlation between the variable coefficients of unlabeled data and labeled data, and is therefore given as the sum of the squares of the linear differences between the variable coefficients of the unlabeled data and the labeled data as follows.

$$P_U(\alpha, \beta, \gamma) = \sum_{i=n_L+1}^n (x_i^T \alpha - x_i^T \beta)^2 + \gamma \sum_{g=1}^{q+k} s(df_g) \|\alpha_g\|_2 \quad (9)$$

The group lasso penalty is added to the variable coefficients α of the unlabeled data so that all variables in the same

group can be selected or eliminated simultaneously. Objective function (7) enables the simultaneous selection of variables for both the supervised and unsupervised components, and also allows different variables to be selected by the two components.

B. OPTIMIZATION ALGORITHM

The block coordinate descent method [42] is used to solve the SSGALR model. This method is an extension of the standard coordinate descent method that was developed for solving the group lasso problem. Similar to the coordinate descent method, the block coordinate descent method optimizes only one group of variables at each iteration, and the other groups are regarded as constants.

Objective function (7) includes two groups of coefficients α and β . However, we consider only a single iteration of β as an example here to describe the specific process of optimization by the block coordinate descent method. Fixing the value $\alpha = \tilde{\alpha}$, the objective function can be rewritten as follows.

$$l(\tilde{\alpha}, \beta) = \sum_{i=1}^{n_L} [y_i \eta_i - \log(1 + \exp(\eta_i))] - \lambda \sum_{g=1}^{q+k} \sqrt{df_g} \|\beta_g\|_2 - \sum_{i=n_L+1}^n (x_i^T \tilde{\alpha} - x_i^T \beta)^2 - \gamma \sum_{g=1}^{q+k} \sqrt{df_g} \|\tilde{\alpha}_g\|_2 \tag{10}$$

We first apply a second-order Taylor expansion to the supervised component $l_L(\beta)$ at $\tilde{\beta}$:

$$l_L(\beta) \approx l_L(\tilde{\beta}) + \frac{\partial l_L(\tilde{\beta})}{\partial \beta^T} (\beta - \tilde{\beta}) + \frac{1}{2} (\beta - \tilde{\beta})^T \frac{\partial^2 l_L(\tilde{\beta})}{\partial \beta \partial \beta^T} (\beta - \tilde{\beta}), \tag{11}$$

where

$$\frac{\partial l_L(\tilde{\beta})}{\partial \beta^T} = \sum_{i=1}^{n_L} (y_i - p_i(\tilde{\beta})) x_i^T, \frac{\partial^2 l_L(\tilde{\beta})}{\partial \beta \partial \beta^T} = - \sum_{i=1}^{n_L} x_i p_i(\tilde{\beta}) (1 - p_i(\tilde{\beta})) x_i^T,$$

and

$$p_i(\beta) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \exp(\beta_0 + x_i^T \beta) / (1 + \exp(\beta_0 + x_i^T \beta)).$$

Applying the Taylor expansion form of $l_L(\beta)$ along with the derivation of the g -th group coefficients in the k -th iteration yields the following:

$$\nabla_g l(\tilde{\alpha}, \beta) = \sum_{i=1}^{n_L} x_{i,g} (y_i - p_i(\tilde{\beta}))$$

$$- \sum_{i=1}^{n_L} x_{i,g} p_i(\tilde{\beta}) (1 - p_i(\tilde{\beta})) x_{i,g}^T (\beta_g - \tilde{\beta}_g) - \lambda \sqrt{df_g} \frac{\beta_g}{\|\tilde{\beta}_g\|_2} - 2 \sum_{i=n_L+1}^n x_{i,g} (x_i^T \beta - x_i^T \tilde{\alpha}) = 0, \tag{12}$$

where $\tilde{\beta}$ represents the result of the $(k - 1)$ -th iteration. This can be rewritten as follows:

$$\sum_{i=1}^{n_L} x_{i,g} (y_i - p_i(\tilde{\beta})) + \sum_{i=1}^{n_L} x_{i,g} p_i(\tilde{\beta}) (1 - p_i(\tilde{\beta})) x_{i,g}^T \tilde{\beta}_g - \sum_{i=1}^{n_L} x_{i,g} p_i(\tilde{\beta}) (1 - p_i(\tilde{\beta})) x_{i,g}^T \beta_g - \lambda \sqrt{df_g} \frac{\beta_g}{\|\tilde{\beta}_g\|_2} - 2 \sum_{i=n_L+1}^n x_{i,g} (x_i^T \beta_g - x_i^T \tilde{\alpha}) - 2 \sum_{i=n_L+1}^n x_{i,g} x_{i,g}^T \beta_g = 0, \tag{13}$$

where $\beta_{-g} = (\beta_1^T, \dots, \beta_{g-1}^T, 0^T, \beta_{g+1}^T, \dots, \beta_{q+k}^T)$. Simplification yields the following.

$$\left[\sum_{i=1}^{n_L} x_{i,g} p_i(\tilde{\beta}) (1 - p_i(\tilde{\beta})) x_{i,g}^T + 2 \sum_{i=n_L+1}^n x_{i,g} x_{i,g}^T + \frac{\lambda \sqrt{df_g}}{\|\tilde{\beta}_g\|_2} I_g \right] \beta_g = \sum_{i=1}^{n_L} x_{i,g} (y_i - p_i(\tilde{\beta})) + \sum_{i=1}^{n_L} x_{i,g} p_i(\tilde{\beta}) \times (1 - p_i(\tilde{\beta})) x_{i,g}^T \tilde{\beta}_g - 2 \sum_{i=n_L+1}^n x_{i,g} \times (x_i^T \beta_{-g} - x_i^T \tilde{\alpha}) \tag{14}$$

Here, the following definitions are applied.

$$a_g = \sum_{i=1}^{n_L} x_{i,g} p_i(\tilde{\beta}) (1 - p_i(\tilde{\beta})) x_{i,g}^T + 2 \sum_{i=n_L+1}^n x_{i,g} x_{i,g}^T + \frac{\lambda \sqrt{df_g}}{\|\tilde{\beta}_g\|_2} I_g, s_g = \sum_{i=1}^{n_L} x_{i,g} (y_i - p_i(\tilde{\beta})) + \sum_{i=1}^{n_L} x_{i,g} p_i(\tilde{\beta}) (1 - p_i(\tilde{\beta})) \times x_{i,g}^T \tilde{\beta}_g - 2 \sum_{i=n_L+1}^n x_{i,g} (x_i^T \beta_{-g} - x_i^T \tilde{\alpha}).$$

Finally, Equation (14) reduces to

$$a_g \beta_g = s_g, \tag{15}$$

which readily yields the result

$$\beta_g = a_g^{-1} s_g. \tag{16}$$

The coefficients of the $(q + k)$ -th group of β are iterated in turn. This same procedure is then applied to the coefficients α of the unlabeled data while holding the coefficients β fixed.

Algorithm 1 Optimization of Objective Function (7) Using the Block Coordinate Descent Method

Step 1: Expand the continuous variables in the explanatory variables by cubic spline, and initialize the β coefficient as $\beta^{(0)}$.

Step 2: Initialize the α coefficient as $\alpha^{(0)} = \beta^{(0)}$, and update a_g ($g = 1, 2, \dots, q + k$) in turn, repeating m times to obtain $\alpha^{(1)}$.

Step 3: Fix $\alpha = \alpha^{(1)}$, and update β_g ($g = 1, 2, \dots, q + k$) in turn, repeating m times to obtain $\beta^{(1)}$.

Step 4: Repeat steps 2 and 3 for k iterations until $(\alpha^{(k)} + \beta^{(k)} - (\alpha^{(k-1)} + \beta^{(k-1)}))$ is less than a preset threshold value, out of loop, and pass the convergent optimal solution to step 5.

Step 5: Screen coefficients by calculating the L_2 norm of each group of coefficients. If the L_2 norm of a group of coefficients is less than a preset threshold value, the values of the entire group of coefficients are set to 0.

The above-discussed steps are repeated until convergence is achieved. The processing flow of the optimization of objective function (7) is given as Algorithm 1.

IV. EMPIRICAL ANALYSIS

A. VARIABLES FOR CORPORATE CREDIT ANOMALY DETECTION

The financial and non-financial variables employed in the present study are listed in Table 1.

B. EMPIRICAL SAMPLE SELECTION AND DATA COLLECTION

The present study employs a dataset composed of actual corporate data for training and testing the proposed model. The processes employed for obtaining actual corporate data, and the transformation of that data into the inputs and target outputs of the model are described in the following subsections.

1) SAMPLE SELECTION

The data samples were selected from the listed companies in China. This listing included 3584 companies, as of 2018. Here, 322 labeled data samples reflecting anomalous credit were selected during the period of 2000 to 2018 according to the definition of corporate credit anomaly presented in Section II. The proportion of labeled anomalous credit samples to normal credit samples generally adopted is either 1:1, 1:2, or 1:5 [43], [44]. Therefore, we selected 644 labeled normal credit data samples over the same accounting period of 2000 to 2018 according to the ratio of 1:2. Therefore, the dataset includes 966 labeled samples and 2618 unlabeled samples. Training and testing datasets were randomly selected from the labeled and unlabeled samples separately, with 70% of the data being included in the training dataset and the remaining 30% in the testing dataset.

TABLE 1. List of selected financial and non-financial variables employed for corporate credit anomaly detection.

Variable	Description	Definition
Y	Corporate credit status	Binary variable, which is 1 if corporate credit is abnormal, and 0 otherwise
X_1	Current ratio	Current assets/current liabilities \times 100%
X_2	Quick ratio	Quick assets/current liabilities \times 100%
X_3	Equity ratio	Liabilities/owner's equity \times 100%
X_4	Cash flow to debt ratio	Net cash flow from operating activities/liabilities \times 100%
X_5	Asset liability rate	Total liabilities/total assets \times 100%
X_6	Equity multiplier	Total assets/owner's equity \times 100%
X_7	Inventory turnover rate	Main business cost/average inventory balance
X_8	Accounts receivable turnover rate	Total operating revenue/average balance of accounts receivable \times 100%
X_9	Current asset turnover rate	Total operating revenue/average total current assets \times 100%
X_{10}	Fixed assets turnover rate	Main business revenue/average balance of fixed assets \times 100%
X_{11}	Total assets turnover rate	Total operating revenue/average total assets \times 100%
X_{12}	Return on equity	Net profit/average balance of owner's equity \times 100%
X_{13}	Net profit on total assets	Net profit/average balance of total assets \times 100%
X_{14}	Net profit	Total profit minus income tax
X_{15}	Operating revenue growth rate	Increase in operating revenue in the current period/operating income in the previous period
X_{16}	Net profit growth rate	Net profit increase in the current period/net profit in the previous period
X_{17}	Total assets growth rate	Total assets increase in the current period/total assets in the previous period \times 100%
X_{18}	Net assets growth rate	Net assets increase in the current period/total net assets in the previous period \times 100%
X_{19}	Number of penalties	Number of instances of, e.g., administrative punishment, environmental protection punishment, and tax arrears
X_{20}	Credit of surrounding corporate entities	Binary variable, which is 1 if dishonesty and abnormal operations are reported in surrounding corporate entities, and is 0 otherwise
X_{21}	Number of surrounding corporate entities	Number of instances of, e.g., administrative punishment, environmental protection punishment, and tax arrears, in surrounding corporate entities
X_{22}	Risk transparency	$freq_r/len_t$, where $freq_r$ indicates the frequency of risk-related words in annual financial reports and len_t represents the total number of words in the text

2) DATA COLLECTION

The sample financial data were obtained from the Guo Tai'an database (www.gtarsc.com) and the Wind database (www.wind.com.cn). The historical penalty records and default risk data of surrounding corporate entities were obtained from the Tianyan credit reference agency (www.tianyancha.com).

The China Securities Regulatory Commission (CSRC) stipulates that a corporation must include a public disclosure in its annual financial report regarding its prospects for realizing its future business objectives and

Algorithm 2 Risk Transparency Calculation Method

Step 1: Obtain the text data source for risk analysis from the annual financial reports of the sample enterprises collected on the website www.cninfo.com.cn.

Step 2: Text segmentation and risk lexicon extraction. The text is segmented using the Python stuttering segmentation toolkit. In the absence of an established enterprise business risk lexicon, a business risk lexicon is constructed here based on the segmentation results. The constructed lexicon includes 439 negative words, such as pressure, cost, challenge, crisis, disadvantage, compression, severity, dependence, aging, and loss.

Step 3: Quantify the risk analysis of the text, and apply $freq_r/len_t$ to calculate the value of risk transparency.

development strategy. The analysis provided by management is intended to meet the information needs of investors by divulging the current development status and future operation direction of the enterprise. A quantitative analysis of these documents is applied in the present study to determine risk transparency as an index reflecting the level of corporate social responsibility consciousness. Because most of the accounting information and analysis disclosed in annual financial reports are not affected by writing style [45], we measure risk transparency according to number of identified negative words found in the documents. The specific steps are presented in Algorithm 2.

3) DATA PREPROCESSING

Generally, less than 20% of financial variable data, such as current ratio, quick ratio, and equity ratio, were absent from the dataset. These missing values were replaced using the mean imputation method. In addition, outlier data with values less than the 1% quantile and greater than the 99% quantile were respectively replaced with the 1% quantile and 99% quantile values. Finally, the values of the continuous variables were first standardized prior to expanding them with cubic splines.

V. RESULTS

A. MODEL PREDICTION PERFORMANCE

The corporate credit anomaly prediction performance of the proposed SSGALR algorithm was verified by comparisons with conventional logistic regression algorithms, including the supervised semi-parametric logistic regression (SSPLR) and supervised logistic regression (SLR) algorithms, in addition to extreme gradient boosting (XGBoost), which is a high-performance ensemble learning algorithm commonly employed in regression and classification applications. These algorithms were used to train the corporate credit anomaly detection models using the Python 3.6 programming language. The performance of the algorithms was validated in terms of the following recall (R), F1-score (F1), and accuracy rate (ACC) metrics defined according to the confusion matrix

TABLE 2. Confusion matrix.

		Predicted class	
		Positive	Negative
Actual class	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

entries listed in Table 2:

$$R = \frac{TP}{TP + FN}, \quad P = \frac{TP}{TP + FP},$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad F1 = \frac{2 \times P \times R}{P + R}. \quad (17)$$

The corporate credit anomaly prediction performance of the four algorithms are presented in Table 3 in terms of R, ACC, and F1 for the training and testing datasets. It can be seen from the results that the prediction performances of the four algorithms differ significantly, and that the SSGALR algorithm provides the best prediction performance of all algorithms considered in terms of all three performance metrics.

TABLE 3. Comparison of prediction performances for four algorithms in terms Recall (R), Accuracy (ACC), and F1 score (F1).

I. Algorithm	II. Training dataset			III. Testing dataset		
	R	F1	ACC	R	F1	ACC
SSGALR	0.85	0.84	0.87	0.82	0.80	0.85
SSPLR	0.79	0.80	0.82	0.76	0.75	0.79
SLR	0.76	0.77	0.79	0.75	0.76	0.78
XGBoost	0.83	0.82	0.85	0.80	0.79	0.82

The bootstrap method was applied to explore the performance of the four algorithms in terms of robustness. We resampled the data samples for 3584 companies 100 times, where 300 labeled samples and 811 unlabeled samples were resampled each time to ensure that the number of resampled samples was proportional to the number of corresponding samples employed originally. Box plots representing the classification errors of the SSGALR, SSPLR, and SLR algorithms over the 100 resampling operations are presented in Fig. 1. It can be seen that the average classification error of the SSGALR algorithm is less than the average classification errors of the SSPLR and SLR algorithms, and the fewer outliers indicates that the proposed algorithm is affected less by sample irregularities. This demonstrates that the proposed SSGALR algorithm is more robust than the two conventional supervised regression-based algorithms.

B. ANALYSIS OF KEY VARIABLES

Table 4 lists the values of the coefficients α and β obtained by the SSGALR algorithm for the list of variables given in Table 1. The results indicate that coefficient values are non-zero only for variables $X_3, X_4, X_6, X_{11}, X_{13}, X_{16}, X_{19}, X_{20}$, and X_{22} , indicating that these were the only variables considered that are correlated with corporate credit anomalies.

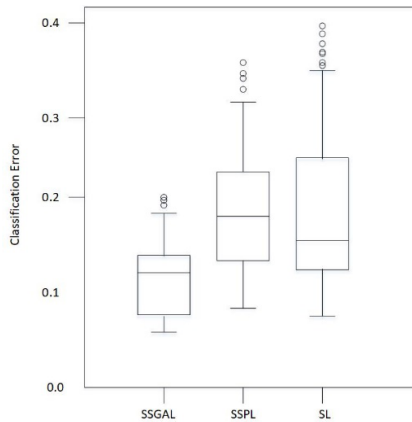


FIGURE 1. Box plots of the classification errors of the SSGALR, SSPLR, and SLR algorithms obtained after resampling the data 100 times.

TABLE 4. Prediction results OF the SSGALR algorithm.

Variable	Description	α	β
X_1	Current ratio	0	0
X_2	Quick ratio	0	0
X_3	Equity ratio	-30.17	-30.18
X_4	Cash flow to debt ratio	-8.13	-8.16
X_5	Asset liability rate	0	0
X_6	Equity multiplier	-1.43	-1.42
X_7	Inventory turnover rate	0	0
X_8	Accounts receivable turnover rate	0	0
X_9	Current asset turnover rate	0	0
X_{10}	Fixed assets turnover rate	0	0
X_{11}	Total assets turnover rate	-0.90	-0.92
X_{12}	Return on equity	0	0
X_{13}	Net profit on total assets	-27.25	-27.36
X_{14}	Net profit	0	0
X_{15}	Operating revenue growth rate	0	0
X_{16}	Net profit growth rate	-18.11	-18.13
X_{17}	Total assets growth rate	0	0
X_{18}	Net assets growth rate	0	0
X_{19}	Number of penalties	0.65	0.79
X_{20}	Credit of surrounding corporate entities	1.24	1.35
X_{21}	Number of penalties in surrounding corporate entities	0	0
X_{22}	Risk transparency	-3.25	-3.36

We note here as well that the absolute values of the coefficients α obtained for the unlabeled samples were slightly less than the absolute values of the coefficients β obtained for the labeled samples. This is because the group lasso penalty is added to the variable coefficients α in Equation (9), which decreases the absolute values of the coefficients. Of these, the absolute values of only the financial variables X_3 , X_4 , X_{13} , and X_{16} were relatively large, indicating that they are highly correlated with credit anomaly prediction. Here, X_3 (equity ratio) and X_4 (cash flow to debt ratio) reflect the solvency of an enterprise. The equity ratio is an important indicator of whether the financial structure of an enterprise is stable. An increasing equity ratio increases the likelihood of long-term solvency, such that the probability of anomalous corporate credit decreases. The cash flow to debt ratio reflects corporate solvency from two aspects: cash inflow and

cash outflow. An increasing cash flow to debt ratio increases the likelihood that an enterprise can repay its debts on schedule, such that the probability of anomalous corporate credit decreases. The variable X_{13} (net profit on total assets) is a measure of income from assets, and reflects the profitability of an enterprise. The variable X_{16} (net profit growth rate) reflects the developmental capacity and operating efficiency of an enterprise, which increase with increasing net profit growth rate. Accordingly, increasing X_{13} and X_{16} decrease the probability of anomalous corporate credit. From the perspective of non-financial variables, we note that X_{22} (risk transparency), X_{20} (credit of surrounding corporate entities), and X_{19} (number of penalties) reflect corporate credit anomalies to a lesser extent than the financial variables discussed above. Here, increasing risk transparency (X_{22}) represents an increasing level of corporate social responsibility, which decreases the probability of anomalous credit. An increasing risk of default in surrounding enterprises (X_{20}) is observed to increase the probability of anomalous credit. Finally, an increasing number of penalties levied against an enterprise (X_{19}) increases the probability of dishonesty, which increases the probability of anomalous credit.

TABLE 5. Prediction performances of the SSGALR algorithm with successively increasing levels of non-financial variables.

Model	F1	ACC
φ_1	0.76	0.78
φ_2	0.78	0.81
φ_3	0.80	0.83
φ_4	0.80	0.85

The effects of non-financial variables X_{22} , X_{20} , and X_{19} on the corporate credit anomaly prediction performance of the SSGALR algorithm were evaluated by comparing the performances obtained for different sets φ of variables including all significant financial variables $\varphi_1 = X_3, X_4, X_6, X_{11}, X_{13}, X_{16}$, and sets including the significant non-financial variables added in succession as $\varphi_2 = \varphi_1 + X_{22}$, $\varphi_3 = \varphi_2 + X_{20}$, and $\varphi_4 = \varphi_3 + X_{19}$. The prediction performances of the SSGALR algorithm obtained for the four variable sets are listed in Table 5 in terms of the ACC and F1 metrics. The results indicate that both the ACC and F1 values generally increase substantially with increasing non-financial information, although the addition of X_{19} is observed to have a negligible effect on the F1 value of the results. These results demonstrate that the non-financial variables associated with risk transparency, credit of surrounding enterprises, and number of penalties play significant roles in the detection of corporate credit anomalies.

VI. CONCLUSION

The present study addressed the generally costly and lengthy process associated with supervised learning methods and the loss of diagnostic scope associated with the sole use of financial variables in conventional corporate credit evaluation models by proposing a semi-supervised generalized additive

logistic regression (SSGALR) model for detecting corporate credit anomalies based on a high proportion of unlabeled sample information that includes both financial and non-financial variables. The model not only can accommodate linear non-separable problems, but can also be trained using both labeled and unlabeled samples at the same time, while simultaneously realizing parameter estimation and variable selection. We also developed more precise definitions of corporate credit anomalies to increase the accuracy of corporate default risk assessments. The model was trained and tested using a dataset composed of actual financial and non-financial corporate data freely available on the Internet, and its performance and robustness were demonstrated in comparison with SSPLR, SLR, and XGBoost algorithms. The results yielded several interesting findings.

The proposed SSGALR algorithm performed better than the other algorithms considered in terms of variable selection, corporate credit anomaly prediction, and model robustness. Increasingly high-dimension data requires that the variables be screened in the process of modeling, which further demonstrates the superiority of the proposed semi-supervised method.

The results demonstrated that the financial variables most correlated with corporate credit anomaly in the SSGALR model were the equity ratio, net profit on total assets, net profit growth rate, cash flow to debt ratio, equity multiplier, and total assets turnover rate, while the non-financial variables most correlated with corporate credit anomaly were risk transparency, credit of surrounding corporate entities, and the number of penalties.

The significant non-financial variables substantially improve the corporate credit anomaly prediction performance of the SSGALR model. These results verify that the number of negative words derived from the annual financial reports of companies and evaluation of surrounding risks are effective tools for evaluating the credit of corporate entities. This is particularly beneficial for evaluating the credit of small and medium-sized enterprises for which a full range of financial data may not be available.

Although this study has made contributions with both theoretical and practical significance in the field of corporate credit anomaly detection, topics remain to be explored. First, we considered only text associated with future default risk analysis in annual financial reports. In the future, we can mine other information in financial reports associated with credit evaluation. Second, we must expand the volume of corporate non-financial sample data employed in the model, particularly data associated with small and medium-sized enterprises, to further test the effect of non-financial variables in the SSGALR model.

REFERENCES

- [1] D. West, "Neural network credit scoring models," *Comput. Oper. Res.*, vol. 27, nos. 11–12, pp. 1131–1152, Sep. 2000.
- [2] H. Zhong, C. Miao, Z. Shen, and Y. Feng, "Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings," *Neurocomputing*, vol. 128, pp. 285–295, Mar. 2014.
- [3] C.-L. Huang, M.-C. Chen, and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert Syst. Appl.*, vol. 33, no. 4, pp. 847–856, Nov. 2007.
- [4] Y.-C. Lee, "Application of support vector machines to corporate credit rating prediction," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 67–74, Jul. 2007.
- [5] K. N. Fang, J. B. Wu, J. P. Zhu, and B. C. Xie, "Forecasting of card credit risk under asymmetric information based on nonparametric random forests," *Econ. Res. J.*, vol. 1, pp. 97–107, Dec. 2010.
- [6] J. Li and Q. Zhu, "Semi-supervised self-training method based on an optimum-path forest," *IEEE Access*, vol. 7, pp. 36388–36399, Mar. 2019.
- [7] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," *IEEE Access*, vol. 6, pp. 14277–14284, Feb. 2018.
- [8] X. H. Hu, W. Y. Ye, and B. Q. Miao, "Variable selection in credit risk models for Chinese listed companies," *J. Appl. Stat. Manage.*, vol. 31, no. 6, pp. 1117–1124, 2012.
- [9] B. Q. Yan, Y. Y. Zhao, and H. Zhang, "Research on logistic model of enterprise credit risk prediction based on financial data of listed companies," *J. Commun. Univ. China Sci. Technol.*, vol. 23, no. 4, pp. 36–47, 2016.
- [10] K. N. Fang, G. J. Zhang, and H. Y. Zhang, "Individual credit risk prediction method: Application of a lasso-logistic model," *J. Quantum Tech. Econ.*, vol. 2, pp. 125–136, 2014.
- [11] R. F. Engle, C. W. J. Granger, J. Rice, and A. Weiss, "Semiparametric estimates of the relation between weather and electricity sales," *J. Amer. Stat. Assoc.*, vol. 81, no. 394, pp. 310–320, Jun. 1986.
- [12] X. M. Wang, "Study on evaluation of default probability based on generalized additive models," *Syst. Eng. Theory Pract.*, vol. 28, no. 6, pp. 52–58, 2008.
- [13] J. Zhang and B. B. Zhang, "The application of generalized semi-parametric additive credit score model based on group-LASSO method," *J. Appl. Stat. Manage.*, vol. 35, no. 3, pp. 517–524, 2016.
- [14] S. Han and Q. H. Han, "A review of semi supervised learning," *Comp. Eng. Appl.*, vol. 56, no. 6, pp. 19–27, 2020.
- [15] C. J. Merz, D. C. St. Clair, and W. E. Bond, "SeMi-supervised adaptive resonance theory (SMART2)," in *Proc. IJCNN Int. Joint Conf. Neural Netw.*, Baltimore, MD, USA, 1992, pp. 851–856.
- [16] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 773–780, Jul. 1989.
- [17] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Mach. Learn.*, vol. 39, nos. 2–3, pp. 103–134, 2000.
- [18] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA: Morgan Kaufmann, 2001, pp. 19–26.
- [19] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory (COLT)*, New York, NY, USA, 1998, pp. 92–100.
- [20] X. Wang, J. Wen, S. Alam, Z. Jiang, and Y. Wu, "Semi-supervised learning combining transductive support vector machine with active learning," *Neurocomputing*, vol. 173, pp. 1288–1298, Jan. 2016.
- [21] W. Zhang, X. Tang, and T. Yoshida, "TESC: An approach to text classification using semi-supervised clustering," *Knowl.-Based Syst.*, vol. 75, pp. 152–160, Feb. 2015.
- [22] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.
- [23] Y. Zhang, C. G. Zhang, and X. H. Zhang, "Personal credit evaluation method based on improved graph semi-supervised learning," *Comput. Sci. Exploring*, vol. 6, no. 5, pp. 473–480, 2012.
- [24] J. Zhang, L. Li, and G. Zhu, "Research on credit prediction based on simulated annealing semi supervised learning," *J. China Univ. Sci. Technol.*, vol. 48, no. 6, pp. 447–457, 2018.
- [25] K. N. Fang and Z. L. Chen, "Credit scoring based on semi-supervised generalized additive logistic regression," *Syst. Eng. Theory Pract.*, vol. 40, no. 2, pp. 392–402, 2020.
- [26] F. Saitoh, "Predictive modeling of corporate credit ratings using a semi-supervised random forest regression," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage. (IEEM)*, Bali, Indonesia, Dec. 2016, pp. 429–433.
- [27] M. Zhao, Z. Zhang, and T. W. S. Chow, "Trace ratio criterion based generalized discriminative learning for semi-supervised dimensionality reduction," *Pattern Recognit.*, vol. 45, no. 4, pp. 1482–1499, Apr. 2012.

- [28] D. H. Downs, J. T. L. Ooi, W.-C. Wong, and S. E. Ong, "Related party transactions and firm value: Evidence from property markets in Hong Kong, Malaysia and Singapore," *J. Real Estate Finance Econ.*, vol. 52, no. 4, pp. 408–427, May 2016.
- [29] Z. Ntsalaze, G. Boako, and P. Alagidede, "The impact of sovereign credit ratings on corporate credit ratings in south africa," *Afr. J. Econ. Manage. Stud.*, vol. 8, no. 2, pp. 126–146, Jun. 2017.
- [30] E. W. Anderson and S. A. Mansi, "Does customer satisfaction matter to investors? Findings from the bond market," *J. Marketing Res.*, vol. 46, no. 5, pp. 703–714, Oct. 2009.
- [31] J. Couppey-Soubeyran and J. Héricourt, "The impact of financial development on the relationship between trade credit, bank credit, and firm characteristics: A study on firm-level data from six MENA countries," *Rev. Middle East Econ. Finance*, vol. 9, no. 2, Jan. 2013.
- [32] N. Fan and W. B. Zhu, "Theoretical selection and empirical analysis of credit evaluation indicators of small and medium-sized enterprises," *Sci. Res. Manage.*, vol. 24, no. 6, pp. 83–88, 2003.
- [33] X. Shu, "Research on the construction of credit rating index system of small and micro enterprises," *Finance Theory Pract.*, vol. 34, no. 5, pp. 105–108, 2015.
- [34] H. Ashbaugh-Skaife, D. W. Collins, and R. LaFond, "The effects of corporate governance on firms' credit ratings," *J. Account. Econ.*, vol. 42, nos. 1–2, pp. 203–243, Oct. 2006.
- [35] X. J. Jin, Y. Zhu, and X. L. Yang, "The influence of Internet media on the stock market—an empirical study of dongfang fortune net stock bar," *J. Commun. Res.*, vol. 20, no. 12, pp. 36–51, 2013.
- [36] T. Zhou, Y. L. Li, Q. Li, D. B. Chen, W. B. Xie, T. Wu, and T. Zeng, "Predicting corporate dishonesty by using network data," *Big Data Res.*, vol. 4, no. 5, pp. 41–49, 2018.
- [37] X. Y. Yang, Y. Y. Jiang, and Z. Z. Duan, "Applicability analysis and empirical test of KMV model in credit risk management of commercial banks in China," *Finance Theory Pract.*, vol. 37, no. 1, pp. 34–40, 2016.
- [38] J. Deng, T. Qin, and S. Huang, "Research on credit risk early warning of Chinese listed companies Based on logistic model," *Finance Theory Pract.*, no. 2, pp. 22–26, 2013.
- [39] J. X. Wang and L. Y. Yu, "Research on risk assessment model of commercial banks based on credit risk degree," *J. Ind. Eng. Eng. Manage.*, vol. 21, no. 4, pp. 85–90, 2007.
- [40] B. Zhang, Z. G. Zhou, W. Liu, and P. Jiao, "Uncertainty DE-KMV model for credit risk measurement of listed companies," *J. Syst. Eng.*, vol. 30, no. 2, pp. 165–173, 2015.
- [41] Z. P. Tang, W. H. Chen, and Y. P. Huang, "Measurement of credit risk of listed companies," *Statist. Decis.*, no. 24, pp. 174–179, 2016.
- [42] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, Feb. 2006.
- [43] J. Lai, M. Xiao, and Z. F. Zhou, "Modeling and empirical study on financial crisis prediction of listed companies in China," *Manage. Sci.*, no. 9, pp. 18–24, 2010.
- [44] Y. Yang, Y. M. Zhou, and Z. F. Zhou, "Enterprise credit risk assessment based on text big data," *Big Data Res.*, vol. 3, no. 1, pp. 44–50, 2017.
- [45] T. Loughran and B. McDonald, "Measuring readability in financial disclosures," *J. Finance*, vol. 69, no. 4, pp. 1643–1671, Jul. 2014.



SONG HAN received the B.E., M.E., and Ph.D. degrees in statistics from Beijing Forestry University, in 2003, 2006, and 2008, respectively. She is currently an Associate Professor with the Information Institute, Beijing Wuzi University. Her main research interests include data mining and machine learning.

• • •