Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5625271, 11 pages
https://doi.org/10.1155/2020/5625271

Research Article

A Financial Distress Prediction Model Based on Sparse Algorithm and Support Vector Machine

Sen Zeng, Yaqin Li, Wanjun Yang, and Yanru Li

School of Economics & Management, Wuhan Polytechnic University, Wuhan 430023, China

Correspondence should be addressed to Yaqin Li; leeyaqin@whpu.edu.cn

Received 14 July 2020; Revised 30 September 2020; Accepted 10 November 2020; Published 29 November 2020

Academic Editor: Chen Chen

Classification learning is a very important issue in machine learning, which has been widely used in the field of financial distress warning. Some researches show that the prediction model framework based on sparse algorithm has better performance than the traditional model. In this paper, we explore the financial distress prediction based on grouping sparsity. Feature selection of sparse algorithm plays an important role in classification learning, because many redundant and irrelevant features will degrade performance. A good feature selection algorithm would reduce computational complexity and improve classification accuracy. In this study, we propose an algorithm for feature selection classification prediction based on feature attributes and data source grouping. The existing financial distress prediction model usually only uses the data from financial statement and ignores the timeliness of company sample in practice. Therefore, we propose a corporate financial distress prediction model that is better in line with the practice and combines the grouping sparse principal component analysis of financial data, corporate governance characteristics, and market transaction data with support vector machine. Experimental results show that this method can improve the prediction efficiency of financial distress with fewer characteristic variables.

1. Introduction

In recent years, machine learning algorithms have been widely used in the field of corporate financial distress prediction. However, most algorithms regard financial distress prediction as a simple dichotomy problem and often ignore the timeliness of financial distress outbreak in practice [1]. And the potential correlation between features and tags may not be considered.

A good financial distress prediction scheme must be realistic and efficient [2]. However, a large number of redundant and unrelated attributes would affect classification performance by increasing computing costs and the time required to learn and test the classifier. Feature selection, as an important technology in data mining and machine learning, has been widely used in classification models. Selecting features before applying classification method to the original dataset has several advantages, such as refining the data, reducing calculation cost, and improving classification accuracy. Therefore, we adopt a feature selection algorithm to improve the quality of financial distress prediction.

In the field of financial distress prediction, multiple feature selection methods are proposed, such as rough set method, LASSO method, wrapper, and filter [3–5]. However, most of these approaches fail to take into account the attributes and data sources of individual features and the different effects they may have on the tag. The information characteristics of a company can be grouped according to the analysis of financial statements and data sources, such as the growth ability, solvency, operating ability of financial statements and corporate governance characteristics, and market transaction data. These groupings reflect the correlation, redundancy, and complementarity among the features. Therefore, it can be effectively applied to subsequent feature selection methods.

In consideration of the above problems, this paper proposes an SVM prediction method based on sparse principal component analysis [6]. Consider that the company's data can be divided into several groups of variables

according to the growth ability, debt-paying ability, profitability, and so on. In this paper, sparse principal component analysis is used to screen the characteristic indexes of each group; then a new dataset is formed and substituted into support vector machine (SVM) for classification and prediction. This method is expected to screen out the input variables that are determinants of the results in order to improve the prediction of the financial distress model and put forward a new idea for the variable screening of financial distress prediction.

Our major contributions are summarized as follows: (1) when considering the natural grouping of corporate information features, reducing the redundant data of each feature group means fewer opportunities to make decisions based on noise, thus reducing overfitting; (2) less misleading data improves the stability and accuracy of modeling; (3) less data means faster algorithm training; (4) a complete variable selection path is generated, and it can be used to measure the information category and the relative importance of all variables; and (5) the multicollinearity problem is naturally overcome, and fewer variables make the model easier to explain.

This paper is organized as follows. Section 2 describes related work. Section 3 introduces the method of combining sparse principal component analysis with support vector machine. Section 4 analyzes the application process and classification results of sparse principal component analysis and support vector machine, that is, the dimensionality reduction method and the results of financial distress prediction. Finally, conclusions are drawn in Section 5.

2. Related Work

Artificial intelligence technology has been gradually introduced into the field of corporate financial distress prediction, and a considerable number of AI prediction models with higher prediction accuracy have been developed. Vapnik introduced support vector machines (SVM) technology into financial distress prediction and achieved good predictive results [7, 8]. SVM is an artificial intelligence method based on the principle of structural risk minimization. It seeks optimal compromise between the complexity of the model and the learning ability based on the limited sample information in order to obtain the best generalization ability. SVM shows its unique advantages in solving small sample size problem and nonlinear problem. It can achieve accurate prediction with a prediction accuracy comparable to that of ANN model [9, 10]. However, SVM is difficult to process large-scale training samples, which limits the scope of SVM [11]. In addition, which is similar to the ANN model, the SVM model also has the nature of "black box," which makes it difficult for the user to understand the results of the model [9, 12]. Shie et al. [13] introduced the particle swarm algorithm (particle swarm optimization (PSO)) into the SVM. An innovative PSO-SVM model is proposed, the model uses 54 banks as research objects to make predictions with an accuracy rate of 97%. However, in practice, in order to give full play to the advantages of SVM, scholars often mix SVM with other technologies instead of using SVM algorithm alone for research on financial distress prediction [3, 14, 15].

Since the 21st century, some scholars have proposed to combine certain technology with another technology to develop a compound model of financial distress prediction with higher performance. The compound model is characterized by high prediction accuracy and complexity. Based on the various compound models proposed by scholars, we divide them into integrated model and hybrid model based on the research of Chaudhuri and De [16].

The basic idea of the integrated model is to obtain a high-precision prediction model by combining several less accurate prediction techniques, which can give full play to the advantages of a single prediction technique and make up for its shortcomings. The main purpose of the proposed integrated model is still to improve the prediction accuracy of the model [17]. Most researches [18–21] also confirmed that the integrated model has higher predictive performance than the previous single-technology model and most hybrid models. Tsai et al. [22] combined the clustering technology with classifier integration technology. Clustering technology includes self-organizing mapping (SOM) and k-means clustering. Classifier technology includes logistic regression, multilayer perceptron (MLP) neural network, and decision tree. 21 different models were designed to predict financial distress. The study found that integration algorithm, composed of SOM and MLP classifier, offered the best prediction results. Li and Wang [23] proposed a support vector machine (SVM) integrated algorithm prediction model based on Choquet integral and used the Bagging algorithm to generate a new training set. Its prediction effect is better than that of a single SVM algorithm. Zieba et al. [24] used extremum gradient enhanced to improve the prediction ability of the decision tree. García et al. [25] compared and analyzed the performance of various integrated classifiers (bagging, AdaBoost, rotating forest, random forest, and random gradient enhancement). The empirical results showed that the overall performance of the model depended on the common type of positive samples.

One of the characteristics of the hybrid model is that it can get a more satisfactory prediction effect than a single-technology model. This is a two-stage modeling process. The technique used to select variables is called the basic technique, while the technique used to predict financial distress is called the mixed technique. Anandarajan et al. [26] and Pendharkar [27] combined the technology of genetic algorithm (GAs) with ANN and established GAs-ANN model by using the input variables of GAs selection neural network, which further improved the predictive performance of ANN. Ahn and Kim [28] proposed GAs-CBR model (Case-Based Reasoning (CBR)). They used GAs to select and optimize the enterprise's case required by the CBR model. Although this model requires more modeling time, it generates results with higher prediction accuracy. In addition, Yeh et al. [3] and Chuang [4] both adopted rough set (RS) technology to improve the quality of feature variables and reduce redundant attributes of the model. Liang and Tsai [5] compared and analyzed the influence of feature selection methods of multiple packers and filters on the prediction of

company financial distress and found that genetic algorithm (GA) and logistic model could achieve better prediction effect. Tian et al. [29] adopted LASSO method to screen model variables and found that accounting characteristics had stronger predictive ability than market characteristics. Huang et al. [30] also use LASSO technology to sort the information content of each financial indicator, so as to improve the interpretability of the financial distress model.

3. Grouping Sparse PCA-SVM Method

In the prediction of corporate financial distress, variables are divided into several groups according to market transactions, growth ability, solvency, profitability, and so on, and each group consists of several variables. At this point, the univariate selection method will ignore the information hidden in the variable grouping structure, which may reduce the performance of variable selection and may even misselect variables. There are more and more indicators reflecting the financial status of enterprises in reality, and many of them are noise variables. If all variables are included in the model indiscriminately, the accuracy of the model will be reduced. Therefore, variables should be selected in the modeling. The advantage of sparse principal component analysis lies in its consistency under the small disturbance of data change and its tendency to overcome multicollinearity naturally, and it can provide a complete variable selection path. Support vector machines (SVM) is a mainstream machine learning classification method at present. Due to its advantages in solving small samples and nonlinear problems and its good predictive performance, it has been widely applied in practice.

Therefore, combining the advantages of sparse principal component analysis and support vector machine, this paper proposes GSPCA-SVM method. Considering feature grouping, the effectiveness of sparse principal component analysis in identifying the most important feature indicators in each category is introduced, which enables us to build a better prediction model.

Figure 1 is the flowchart of target recognition combining SPCA and SVM. And the specific steps of the algorithm GSPCA-SVM are shown as follows. Firstly, according to the data sources and financial statement analysis methods, the characteristic indexes of listed companies are divided into several groups (such as solvency, profitability, and growth ability). Second, use sparse principal component analysis to screen the characteristics of each group of indexes. Third, combine the characteristic indexes screened by each group into a new dataset, and determine the training samples and test samples. Then, input the training samples with SVM method, obtain the coefficient and deviation of the discrimination function through learning, and construct the classification model. Finally, input the test samples to the classification model, then take the prediction processing, and finally calculate the accuracy.

3.1. GSPCA-SVM Algorithm. Given that X represents a standardized characteristic matrix of $n * m$, where n is the number of listed companies and m is the number of characteristics of listed companies, sparse principal component analysis is proposed on the basis that principal component analysis can be transformed into a quadratic penalty regression problem. That is, the solution of principal components is directly transformed into LASSO regression. Thereby, the solution of sparse principal components is effectively transformed into the variable selection problem of the linear model. On this basis, the penalty structure of the elastic net is introduced to obtain the sparse principal components.

The objective function of sparse principal component analysis is as follows:

$$(\widehat{\alpha}, t\widehat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \left\| X_i - \alpha\beta^T X_i \right\|^2 + \lambda\|\beta\|^2, \quad (1)$$

where X_i is the i^{th} row vector of X, $\lambda > 0$. When $\|\alpha\|^2 = 1$, $\widehat{\beta} \propto V_1$. In this way, regression knowledge is used to obtain the first principal component.

Assume that the matrix of the first k principal component orders α and β is, respectively, for any $\lambda > 0$, $(\widehat{\alpha}, t\widehat{\beta}) = \arg\min \sum_{i=1}^{n} \|X_i - \alpha\beta^T X_i\|^2 + \lambda\|\beta\|^2$, where $\alpha^T\alpha = I_k$, $\widehat{\beta}_i \propto^{\alpha,\beta} V_i$, $i = 1, 2, \ldots, k$. In this way, the original principal component analysis is transformed into a regression problem. By adding LASSO penalty item to the above equation, sparse principal components can be obtained. Thus, the following optimization problem can be obtained:

$$(\widehat{\alpha}, t\widehat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \left\| X_i - \alpha\beta^T X_i \right\|^2 + \lambda \sum_{j=1}^{k} \left\| \beta_j \right\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\| \beta_j \right\|_1, \quad (2)$$

where $\alpha^T\alpha = I_k$.

As stated above, the solution of sparse principal components can be transformed into a penalty regression problem. The general LASSO penalty regression problem can be solved by the least angle regression. Therefore, the calculation of sparse principal components can also be conveniently obtained by using the least angle regression algorithm.

Thus, the algorithm GSPCA-SVM is obtained as follows:

(1) Calculate the vectors corresponding to the first k principal components of the general principal components and let α start at $V[1, 2, \ldots, k]$.

(2) Given $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$, the following elastic net regression problem is solved: $\beta_j = \arg\min\ (\alpha_j - \beta)^T X^T X (\alpha_j - \beta)^{\beta} + \lambda\|\beta\|^2 + \lambda_{1,j}\|\beta\|_1$.

(3) For a given $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$, calculate the SVD of $X^T X\beta = \text{UDV}^T$; let $\alpha = \text{UV}^T$.

(4) Repeat the above two steps until convergence β.

(5) Standardize $\widehat{V_j} = \beta_j / \|\beta_j\|$, $j = 1, 2, \ldots, k$.

(6) Through the obtained feature vector matrix composed of the first k principal components after sparseness, the data matrix X can be feature selected and the feature dimension can be reduced. Therefore,

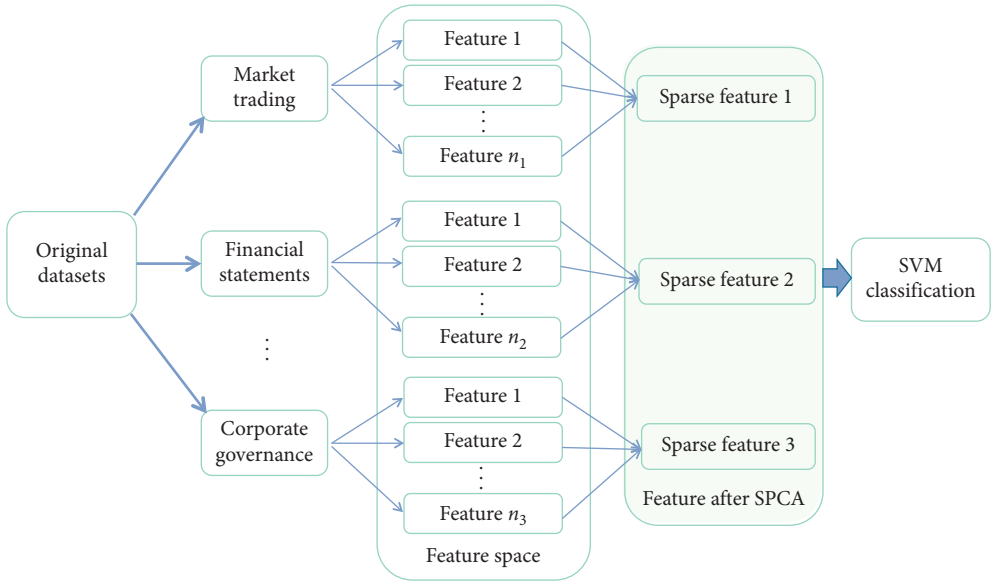

Figure 1: The flowchart of target recognition combining SPCA and SVM.

the dimensionality reduction data sample can be obtained through the formula $Y = X\alpha$.

(7) The dimensionality-reduced data samples are divided into training samples and test samples, the training samples are substituted into the SVM model for training, the coefficients a^* and deviation values b^* of the discriminant function are obtained, thereby constructing a stable classification model, and then the test samples can be classified and identified according to the discriminant function $f(x) = \operatorname{sgn} \sum_{i=1}^{n} a_i^* y_i K(x_i, x) + b^*$.

3.2. Time Complexity Analysis of GSPCA-SVM Algorithm. The time complexity of SVM is $O(N_1^3 + N_1^2 N_2 + d \cdot N_2 N_1)$. The time complexity of sparse principal component analysis is $O(N \cdot d^3)$. Therefore, the total complexity of this method is $O(N_1^3 + N_1^2 N_2 + d \cdot N_2 N_1 + N \cdot d^3)$, where N is the total number of samples, N_1 is the number of support vectors, N_2 is the training sample, and d is the dimension of the original sample. Furthermore, $d \gg N_1, N \gg N_1$. Thus, the total complexity of this method is about equal to $O(N \cdot d^3 + d \cdot N_1 N_2)$.

4. Application

4.1. Datasets. Filing for bankruptcy is often regarded as a sign that the company is in financial distress by many scholars. Considering the late delisting system in China's stock market, there is a lack of samples of listed companies that are already delisted. Therefore, this paper considers the Special Treatment of Chinese listed companies (i.e., ST) as a sign that the listed companies are in financial distress and selects the normal listed companies in the same industry during the same period as the control sample.

Considering the characteristics of the annual report disclosure system of listed companies in China, there is a lag that the company becomes ST in year T–1 which would be disclosed in the report which is actually being issued in year T. Moreover, it is insignificant to predict whether the company will be in financial difficulties in that year by using the data of the previous two years before the company was treated as ST, which will exaggerate the prediction accuracy of the model. Therefore, the research period is the previous three years (year T–3) before the companies become ST (year T) in order to reflect the predictive ability of the previous data and the trend of the company's financial status.

The sample in this paper is obtained from RESSET (financial research database) and CSMAR (financial index analysis database of Chinese listed companies). Related tests and model estimation are completed by MATLAB software. 188 listed companies that were specially treated (ST) in Shanghai and Shenzhen stock exchanges of China from 2015 to 2019 were selected as samples of companies in financial distress, while 188 non-ST companies with similar asset size that in the same industry and same period were selected for matching. In data selection, the ST year of the listed company is T, and the data of the first three years (T–3) of the above company (i.e., 2012–2016) are selected. Table 1 shows the sample distribution and financial years of the selected ST company and the control group of normal company.

4.2. Selection of Alternative Indicators. In the empirical research of financial distress prediction, researchers have not reached an agreement on the selection of indicators. This paper attempts to collect comprehensive data of financial characteristics, transaction characteristics, and corporate governance indicators of China's A-share listed companies. It contains 161 financial features, 10 market transaction features, and 8 corporate governance indicators.

According to the classification method of financial indexes of listed companies in RESSET financial research

Table 1: Distribution of sample companies.

Label year	Data year	Number of ST company samples	Number of normal company samples	Number of features
2015	2012	27	27	179
2016	2013	24	24	179
2017	2014	42	42	179
2018	2015	36	36	179
2019	2016	59	59	179

Table 2: Grouping of characteristics of the company.

Characteristics grouping	The company's information reflected by each indicator group	Number of features
Profitability	An index that reflects the profitability of an enterprise	40
Solvency	An indicator that reflects a firm's ability to repay long-term and short-term debt with its assets	17
Growth ability	An indicator that reflects the future development trend and speed of an enterprise	20
Operation ability	Indicators that reflect the operational capability of an enterprise	11
Cash flow	An indicator that reflects the amount of cash and cash equivalents in and out of an enterprise in a period	14
Dividend capacity	An indicator that reflects the distribution of the current year's earnings to shareholders	6
Capital structure	An index that reflects the relationship between the total capital of an enterprise and its composition and proportion	16
Earnings quality	An index that reflects the reliability of information related to the economic value of an enterprise expressed by accounting earnings	10
Index per share	According to the total number of shares discovered by the enterprise to measure all aspects of the enterprise information index	21
Dupont analysis index	An indicator that plays an integral role in the Dupont financial analysis system	6
Market transactions	An indicator that reflects the stock of a listed company traded in the secondary market	10
Corporate governance	An indicator that reflects the distribution of power within a company	8

database, 161 financial features are divided into 10 alternative index feature groups. There are 40 profitability indicators, 17 solvency indicators, 20 growth capacity indicators, 11 operating capacity indicators, 14 cash flow indicators, 6 dividend capacity indicators, 16 capital structure indicators, 10 earnings quality indicators, 21 per share indicators, and 6 Dupont analysis indicators. With the addition of market transaction characteristics group and corporate governance characteristics group, the distribution of relevant information and indicators reflected by each group is shown in Table 2.

4.3. Sparse Principal Component Analysis Results. According to the previous section, all characteristics of listed companies are divided into 12 groups according to data sources and financial statement analysis methods. Sparse principal component analysis was performed on each group of data, and the first principal component coefficient of each group was observed. If the first principal component coefficient of the characteristic variable was nonzero, it was selected. If the coefficient was zero, it was removed. Table 3 shows the selected characteristic indexes and the first principal component coefficient of each group. In order to compare whether grouping sparsity (GSPCA) better preserves the key information of the original dataset (OF) and eliminates redundant information, we conduct sparse principal component analysis for all 179 features (SPCA) and retain the characteristic indexes with nonzero coefficients of the top four principal components. The screening results are shown in Table 4.

Figure 2 visually reflects the distribution of all information features of listed companies and the results of screening features by sparse principal component analysis. From the grouping, the characteristic indicators (OF) mainly focus on profitability, solvency, growth ability, capital structure, and per share indicators. Through sparse principal component analysis of each group of features (GSPCA), a good dimension reduction effect can be achieved. 61 features can be selected from the original 179 indicators. Of the 40 indicators in the feature group reflecting the company's profitability, only 3 were selected by sparse principal component analysis, and only 3 of the 21 indicators in the feature group reflecting the index information per share were retained. Relatively, none of the indicators of the characteristic group reflecting the dividend ability of the company or the characteristic group of Dupont analysis are sparse, and 16 of the 20 indicators of the characteristic group reflecting the growth ability of the company are retained. In order to compare the effect of grouping sparsity in feature selection, this paper conducts sparse principal component analysis on all 179 indicators (SPCA). This method selects 12 characteristics from the original 179 indicators, and these 12 indicators mainly reflect the information of the enterprise's capital structure, earning quality, and corporate governance. As can be concluded from Figure 2, although this method

Table 3: Variables and first principal component coefficients after filtering of GSPCA.

Feature group	Selected features and first principal component coefficient	Number of features
Profitability	Gross profit margin **(0.57736),** gross profit margin of dynamic sales **(0.57734)**, and ratio of sales to cost **(−0.57734)**	3
Solvency	Liquidity ratios **(0.08438)**, quick ratio **(0.12058)**, conservative quick ratio **(0.12243)**, total shareholders' equity/liabilities **(0.21491)**, and operating net cash flow/net debt **(−0.95769)**	5
Growth ability	Growth rate of earnings **(−0.04697)**, diluted earnings per share growth rate **(−0.07908)**, revenue growth rate **(−0.00784)**, operating profit growth rate **(−0.01943)**, gross profit growth rate **(−0.00834)**, net profit growth rate **(−0.00734)**, growth rate of net profit attributable to the parent company **(−0.03645)**, growth rate of net profit (net deduction) attributable to the parent company **(−0.05498)**, three-year compound growth rate of net profit attributable to shareholders of the parent company **(−0.06823)**, net assets growth rate **(−0.03860)**, average increase of net profit attributable to the parent company in the past five years **(0.98782)**, growth rate of total assets **(−0.03625)**, sustainable growth rate **(−0.01687)**, net assets per share relative to the growth rate at the beginning of the year **(−0.01696)**, relative growth rate of shareholders' equity at the beginning of the year **(−0.03860)**, and relative growth rate of total assets at the beginning of the year **(−0.03625)**,	16
Operation ability	Turnover of liquid assets **(−0.09176)**, shareholder equity turnover rate **(−0.70697)**, and total assets turnover **(−0.70127)**	3
Cash flow	Net operating cash flow/net operating income **(−0.70477)**, dynamic net operating cash flow/net operating income **(−0.70477)**, cash component of net profit **(−0.00990)**, and cash recovery on total assets **(−0.08055)**	4
Dividend capacity	Balance of cash and cash equivalents per share **(0.22505)**, dividend per share **(0.14916)**, dividend guarantee multiple **(0.05615)**, cash dividend cover **(0.95628)**, dividend payout ratio **(0.01761)**, and retained earnings ratio **(0.09572)**	6
Capital structure	Long-term loans/total assets **(0.00822)**, bonds payable/total assets **(0.08242)**, current liabilities/total liabilities **(−0.70467)**, and noncurrent liabilities/total liabilities **(0.70467)**	4
Earnings quality	Net income/total profit from operating activities **(−0.29588)**, dynamic net income/total profit from operating activities **(−0.29588)**, and income tax/total profit **(−0.90825)**	3
Index per share	Net asset value per share **(0.09355)**, capital reserve per share **(0.67007)**, and reserve fund per share **(0.73637)**	3
Dupont analysis index	Equity multiplier **(0.904105)**, belonging to the parent company net profit of shareholders/net profit **(0.00105)**, net profit/total operating income **(−0.08675)**, net profit/total profit **(−0.01797)**, total profit/earnings before interest and tax **(−0.41034)**, and earnings before interest and tax/total operating income **(−0.07979)**	6
Market transactions	Annual turnover rate of total shares **(0.61204)**, annual turnover rate of tradable shares **(0.06291)**, average daily turnover rate of total shares **(0.78741)**, and average daily turnover rate of tradable shares **(0.03762)**	4
Corporate governance	Shareholding ratio of the largest shareholder **(0.94420)**, H-index of the largest shareholder **(0.24510)**, H-index of the top five shareholders **(0.16377)**, and H-index of top ten shareholders **(0.14691)**	4

achieves dimension reduction effect, it ignores many aspects of the company such as profitability, growth ability, operation ability, and market transaction. It may remove a lot of useful information related to a company's forecast of financial distress.

4.4. Determination of Training Samples and Test Samples and Selection of Kernel Function. In order to use a support vector machine (SVM) to build the company financial distress prediction model. The dataset division method of the company's financial distress proposed by Hsieh et al. [1] is referred to in this paper. Considering the timeliness of financial distress prediction, this paper takes the number of years the company has faced financial distress as the standard; two datasets were used for the model evaluation. The sample of 2015–2017 is taken as the training set, which contains 93 samples of financial distress companies and 93 samples of financial health companies. The sample of 2018 is used as the test set, including 36 samples of financially distressed companies and 36 samples of normal companies (Dataset I). Another set of data took the company samples from 2015 to 2018 as the training set, including 129 samples of financial distress and 129 samples of normal companies. Taking the sample of 2019 as the test set, there are 59 samples of financially distressed companies and 59 samples of normal companies (Dataset II). The two training sets are, respectively, used for model construction. Then the other two test datasets are used for model evaluation, respectively. With respect to the selection of kernel functions, there are specific kernel functions in specific application areas. The classification of corporate financial distress prediction is a general classification problem, so the radial basis kernel function commonly used in SVM is selected.

4.5. Evaluation Metrics. Considering that the prediction model of enterprise financial distress is a typical dichotomy problem, the commonly used evaluation indicators include accuracy, precision, recall, and F1 value. The larger these

Table 4: Variables and top 4 principal component coefficients after filtering of all features SPCA.

Selected features	First principal component	Second principal component	Third principal component	The fourth principal component
Asset-liability ratio	0	0.70571	0	0
Current assets/total assets	0	0	0	0.70408
Noncurrent assets/total assets	0	0	0	−0.70415
Ratio of fixed assets	0	0	0	−0.09175
Shareholders' equity/total invested capital	0	−0.04428	0	0
Interest-bearing debt/total capital invested	0	0.04428	0	0
Equity ratio	0	−0.70571	0	0
Net income/total profit from operating activities	0.29587	0	0	0
Dynamic net income/total profit from operating activities	0.29587	0	0	0
Income tax/total profit	0.90824	0	0	0
Shareholding ratio of the largest shareholder	0	0	−0.09218	0
Shareholding ratio of the top five shareholders	0	0	−0.72019	0
Shareholding ratio of the top ten shareholders	0	0	−0.68762	0

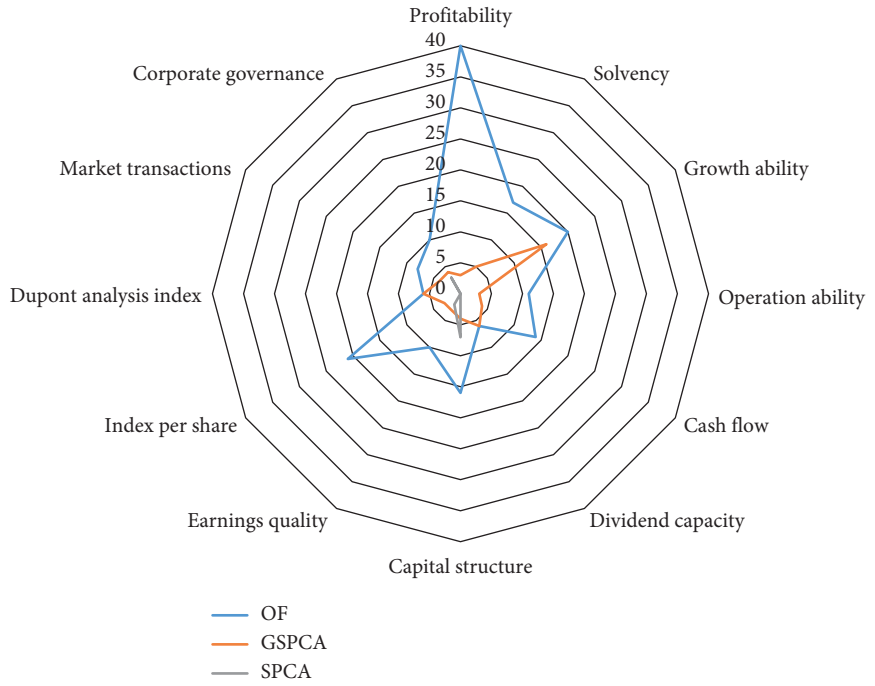

Figure 2: Comparison of information feature distribution and sparse principal component screening feature results of listed companies.

four indicators are, the better the prediction effect of the model is. In this paper, we consider financial distress as the positive label and financial stability as the negative label. There are 4 scenarios predicted by the model in the test dataset, and the occurrence times of each scenario are denoted as follows:

TP: classify enterprises in financial distress as financial distress

FN: classify enterprises in financial distress as financial health

FP: classify financially healthy enterprises as financial distress

TN: classify financially healthy enterprise as financial health

Accuracy is the proportion of the number of correct predictions to the total:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3}$$

Precision rate is the proportion of the number correctly predicted as a positive class to the number predicted as a positive class:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{4}$$

Recall rate is the proportion of the number of positive classes correctly predicted to the total number of true positive classes:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

F_1 is a weighting of accuracy and recall:

$$\frac{2}{F_1} = \frac{1}{\text{precision}} + \frac{1}{\text{recall}} \Rightarrow F_1 = \frac{2PR}{P+R} = \frac{2TR}{2TP + FP + FN}. \tag{6}$$

4.6. Comparison of Forecast Results of Classification Model. In order to analyze the advantages of this method in the process of feature selection and classification prediction, this paper also introduces principal component analysis, nuclear principal component analysis, linear discriminant analysis, and nuclear linear discriminant analysis for comparative study. In this section, seven corporate financial distress prediction schemes are formulated. The first scheme is to directly input all the original features of the company into the support vector machine for classification prediction (OF-SVM). The second scheme is to perform principal component analysis on all company information characteristics and then input the extracted principal components into support vector machine for classification (PCA-SVM). In the third scheme, linear discriminant analysis is applied to the characteristic data first, and then the nearest neighbor (KNN) is used for classification prediction (LDA-KNN). The fourth scheme is to conduct kernel principal component analysis for all company information features, in which radial basis kernel function is adopted, and then the extracted principal components are input into support vector machine for classification (KPA-SVM). The fifth scheme firstly adopts the kernel linear discriminant analysis, in which the radial basis kernel function is adopted, and then the nearest neighbor (KNN) is used for classification prediction (KDA-KNN). The sixth scheme uses sparse principal component analysis to screen variables for all corporate information characteristics and then inputs support vector machine (SPCA-SVM) for classification. The seventh scheme is the method proposed in this paper. First, all the company information features are grouped, and then each feature group is filtered by sparse principal component analysis in turn. Finally, the filtered features of each group were merged into a new database and input into support vector machine for classification learning (GSPCA-SVM).

Table 5 and Figure 3 show the accuracy, precision, recall, and F1 of each algorithm on the two datasets, respectively. The bold numbers in the table represent the best results for each dataset. In Dataset I, the GSPCA-SVM algorithm proposed by us is significantly superior to the other 6 algorithms in accuracy, precision, recall, and F1 value. In addition, in Dataset II, although GSPCA-SVM algorithm is not significantly superior to other algorithms in precision and recall, GSPCA-SVM algorithm is still superior to all other methods in the accuracy and F1 value of reflecting the overall prediction effect of the model. The precision value of GSPCA-SVM algorithm is second only to that OF-SVM algorithm and PCA-SVM algorithm, and the values are very close. And recall value of GSPCA-SVM algorithm is just after KPCA-SVM algorithm; this is because the kernel principal component analysis and support vector machine (SVM) method to forecast the effect of financial distress are not ideal (can be seen from the lower accuracy value). The method predicts more samples as financial distress companies, which caused the less amount of predicting financial distress companies as health companies (i.e., the smaller values of the FN), so the recall value is relatively greater.

Table 5: Classification results of sparse principal component SVM.

Dataset I				
	Accuracy	Precision	Recall	F1
OF-SVM	72.22	80.77	58.33	67.74
PCA-SVM	75.00	78.13	72.50	76.32
LDA-KNN	55.56	54.55	66.67	60.00
KPCA-SVM	48.61	49.15	80.56	61.05
KDA-KNN	58.33	57.89	61.11	59.46
SPCA-SVM	68.06	72.41	58.33	64.62
GSPCA-SVM	81.94	81.08	83.33	82.19
Dataset II				
	Accuracy	Precision	Recall	F1
OF-SVM	76.27	94.29	55.93	70.21
PCA-SVM	77.96	94.59	59.32	72.92
LDA-KNN	55.08	55.77	49.15	52.25
KPCA-SVM	48.30	49.00	83.05	61.64
KDA-KNN	55.93	56.14	54.24	55.17
SPCA-SVM	62.71	75.86	37.29	50.00
GSPCA-SVM	79.66	92.68	64.41	76.00

In addition, from the perspective of the practical application of the model and the cost of prediction error, regarding the ST company in financial distress mistakenly as a normal company would bring great losses to institutional and individual investors, while investors would not suffer loss from wrongly regarding normal company as ST company in financial distress. Therefore, the prediction results of positive samples should be focused on, namely, F1 value. Figure 3 shows that the F1 value of SVM classification results processed by the grouping sparse algorithm (GSPCA-SVM) is significantly higher than other methods. Thus, it can be seen that taking SVM classification method after dimension reduction by sparse principal component analysis has better performance in corporate financial distress prediction.

4.7. Significance Test. Two statistical tests were used in this experiment, the variance test and the Friedman test, to determine the significant differences between the various methods. We compared experimental data from two datasets on seven algorithms.

In variance analysis, we assume that there is no significant difference between various methods. According to the F1 index in the above experiment, the formula is used to calculate that the F value is 6.93 and P value is 0.0111. According to $n = 13, m = 6$, $F_{0.05}(m, n-m-1) = F_{0.05}(6,6) = 4.284$ can be obtained by looking up F test critical value table. The actual value of F is 6.93 which is greater than its value in the table, so it can be judged that there is a significant difference, and the P value is less than 0.05. Therefore, we can reject the null hypothesis, which indicates that there is a significant difference between our proposed method and the other six methods.

In Friedman analysis, the chi-square distribution is to approximate the Friedman test statistics. We calculated the ranks of seven methods by sorting the accuracy in the above

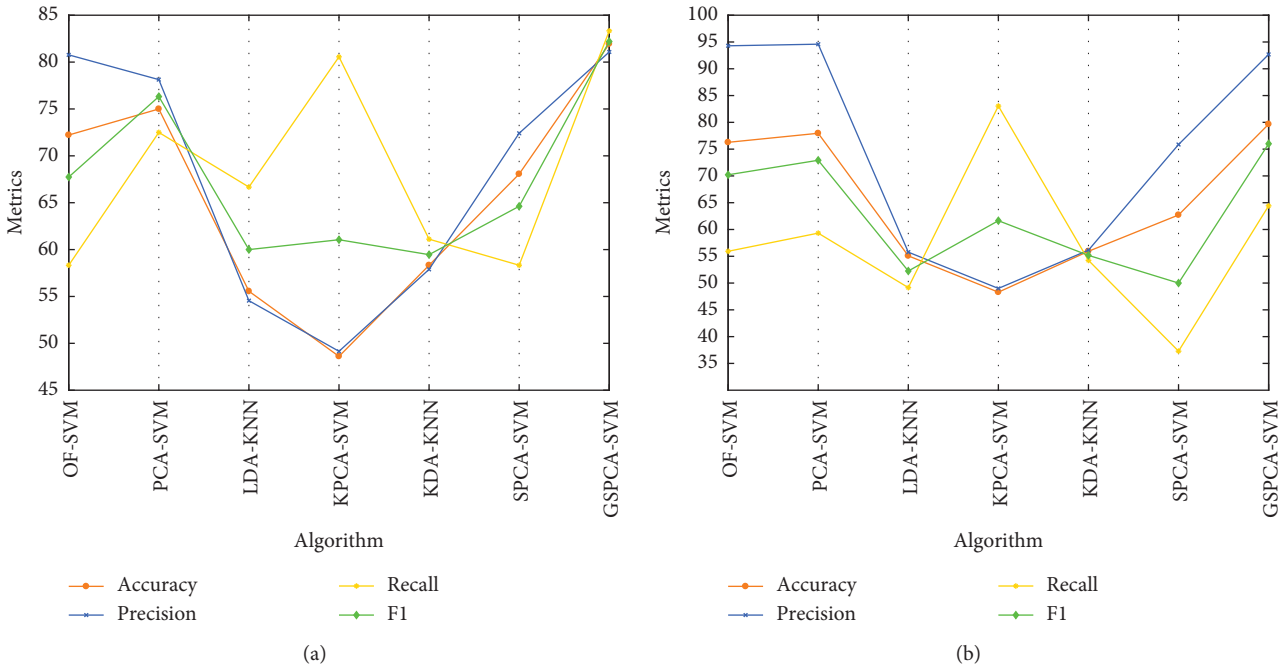

Figure 3: Index values of different algorithms under two datasets: (a) dataset I and (b) dataset II.

Table 6: The ranks of the seven algorithms on two databases.

	OF-SVM	LDA-KNN	PCA-SVM	KDA-KNN	KPCA-SVM	SPCA-SVM	GSPCA-SVM
Dataset 1	3	6	2	5	7	4	1
Dataset 2	3	6	2	5	7	4	1
Total rank	6	12	4	10	14	8	**2**
Ave rank	3	6	2	5	7	4	**1**

experiments. The ranks of the seven methods are shown in Table 6. Under the null hypothesis, there would be no difference between all the methods, and therefore theoretically R_j^2 should be equal. From the data in Table 6, the value of the Friedman test statistics can be calculated:

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) = 12, \tag{7}$$

where n is the number of datasets, k is the number of methods, and R_j^2 represents the sum of the ranks for all datasets under the k^{th} methods.

In statistical analysis, to reject the null hypothesis, the calculated value of χ_r^2 must be greater than or equal to the critical value of the chi-square distribution. In this set of experiments, we adopted the commonly used critical value of 0.05 degrees of freedom. By comparison, $\chi_{0.05}^2 = 12.592 > \chi_r^2$; it shows that there are some differences between these seven methods.

5. Conclusion

In this paper, an SVM model based on sparse principal component analysis (GSPCA-SVM) is proposed to deal with financial distress prediction. In the feature selection stage of the original dataset, we propose a method to group the features according to data sources and financial statement analysis. The purpose of this method is to investigate whether the predictive performance of the model can be improved by selecting fewer, relatively more important variables from each information feature category. Compared with other forecasting models, our method has a better forecasting effect because it combines the management method and machine learning method in the field of enterprise financial distress forecasting. Considering that the information feature of listed companies has the characteristics of natural grouping, applying the financial statement analysis method to the grouping of original datasets avoids the common dimension reduction method ignoring the information hidden in the variable grouping structure, which may reduce the forecasting effect of the model. In addition, the sparse algorithm selects fewer and more important variables from each information feature category to improve the prediction performance and explanatory ability of the model and further analyzes which feature categories of the company can provide more information for predicting financial distress.

From the perspective of management, the conclusion of GSPCA-SVM method proposed in this paper agrees with the management theory. In the original dataset, information

features are mainly concentrated in the feature group that reflects the corporate profitability, because the existence of an enterprise is that, in order to maximize the shareholders' wealth (or enterprise value), there will naturally be more indicators to reflect the profitability of the enterprise. However, in the dataset screened by the GSPCA-SVM model, more feature indicators are concentrated in the feature group reflecting the company's growth ability and dividend capacity. On the one hand, although there are a large number of profitability indicators, they generally have strong collinearity and do not have more information content. On the other hand, it also inspires the management and investors that, in order to avoid financial difficulties and investment failure, they should focus far more on the growth ability and dividend ability of the company. Financial distress prediction is the analysis of the company's future financial risk and development status. Among all the characteristic index categories, the growth ability of a company can be used to understand the future development potential of the enterprise, and it is the best reflection of the future development prospect, development trend, and development speed, including the changes of enterprise scale, profit, and owner's equity. Dividend capacity also plays an important role in financial distress prediction. According to the signal theory, a company sends a positive signal to the market that it is developing well by paying stable or more cash dividends to shareholders. However, for companies facing financial difficulties in the future, poor management and tight cash flow will prompt them to reduce cash dividends or even not pay cash dividends, which will be reflected in the dividend capacity index in advance. Therefore, the method in this paper can also provide a reference for the company management and investors to make correct decisions.

We will consider the following research directions in the future. First, we will consider the unbalanced problems in the realistic financial distress prediction samples of listed companies and put forward a more realistic and accurate prediction model. Secondly, we will explore related applications based on fuzzy theory in the field of corporate financial distress prediction, such as fuzzy clustering analysis [31] and fuzzy rough set model [32, 33]. Finally, we hope to combine the financial management theory with the machine learning algorithm to develop an effective method to predict the financial distress of companies.

Data Availability

The dataset and software code used to support this study's findings have not been made available because the data also forms part of an ongoing study. Requests for data, after the publication of the ongoing study, will be considered by the corresponding author Yaqin Li (leeyaqin@whpu.edu.cn).

Disclosure

Sen Zeng and Yaqin Li are co-first authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Sen Zeng and Yaqin Li contributed equally to this work.

Acknowledgments

This paper was supported by the China University Industry-University-Research Innovation Fund, A New Generation of Information Technology Innovation Project 2019 (2019ITA03044), and the Scientific Research Program of Wuhan Polytechnic University (2018J06).

References

[1] T. Hsieh, H. Hsiao, and W. Yeh, *Mining Financial Distress Trend Data Using Penalty Guided Support Vector Machines Based on Hybrid of Particle Swarm Optimization and Artificial Bee Colony Algorithm*, Elsevier Science Publishers B. V., Amsterdam, The Netherland, 2012.

[2] J. L. Bellovary, D. E. Giacomino, and M. D. Akers, "A review of bankruptcy prediction studies: 1930 to present," *Journal of Financial Education*, vol. 33, pp. 1–42, 2007.

[3] C.-C. Yeh, D.-J. Chi, and M.-F. Hsu, "A hybrid approach of DEA, rough set and support vector machines for business failure prediction," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1535–1541, 2010.

[4] C.-L. Chuang, "Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction," *Information Sciences*, vol. 236, pp. 174–185, 2013.

[5] D. Liang and C. Tsai, "The effect of feature selection on financial distress prediction," *Knowledge-Based Systems*, vol. 73, pp. 289–297, 2015.

[6] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[7] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, Germany, 1995.

[8] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[9] K.-S. Shin, T. S. Lee, and H.-j. Kim, "An application of support vector machines in bankruptcy prediction model," *Expert Systems with Applications*, vol. 28, no. 1, pp. 127–135, 2005.

[10] J. Min and Y. Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," *Expert Systems with Applications*, vol. 28, no. 4, pp. 603–614, 2005.

[11] A. R. Sandin and M. Porporato, "Corporate bankruptcy prediction models applied to emerging economies: evidence from Argentina in the years 1991-1998," *International Journal of Commerce and Management*, vol. 17, no. 4, pp. 295–311, 2007.

[12] F.-M. Tseng and Y.-C. Hu, "Comparing four bankruptcy prediction models: logit, quadratic interval logit, neural and fuzzy neural networks," *Expert Systems with Applications*, vol. 37, no. 3, pp. 1846–1853, 2010.

[13] F. S. Shie, M.-Y. Chen, and Y.-S. Liu, "Prediction of corporate financial distress: an application of the America banking industry," *Neural Computing and Applications*, vol. 21, no. 7, pp. 1687–1696, 2012.

[14] W. F. Lin, Y. H. Hu, and C. F. Tsai, "Machine learning in financial crisis prediction: a survey,," *IEEE Transactions on Systems, Man and Cybernetics. Part C: Applications and Reviews.*vol. 42, no. 4, pp. 421–436, 2012.

[15] H. Li and J. Sun, "Predicting business failure using an RSF-based case-based reasoning ensemble forecasting method," *Journal of Forecasting*, vol. 32, no. 2, pp. 180–192, 2013.
[16] A. Chaudhuri and K. De, "Fuzzy support vector machine for bankruptcy prediction," *Applied Soft Computing*, vol. 11, no. 2, pp. 2472–2486, 2011.
[17] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.
[18] R. Ranawana and V. Palade, "Multi-classifier systems: review and a roadmap for developers," *International Journal of Hybrid Intelligent Systems*, vol. 3, no. 1, pp. 35–61, 2006.
[19] N. C. Oza and K. Tumer, "Classifier ensembles: select real-world applications," *Information Fusion*, vol. 9, no. 1, pp. 4–20, 2008.
[20] G. Wang, J. Ma, L. Huang, and K. Xu, "Two credit scoring models based on dual strategy ensemble trees," *Knowledge-Based Systems*, vol. 26, pp. 61–68, 2012.
[21] H. Xiao, Z. Xiao, and Y. Wang, "Ensemble classification based on supervised clustering for credit scoring," *Applied Soft Computing*, vol. 43, pp. 73–86, 2016.
[22] C.-F. Tsai, Y.-F. Hsu, and D. C. Yen, "A comparative study of classifier ensembles for bankruptcy prediction," *Applied Soft Computing*, vol. 24, pp. 977–984, 2014.
[23] X. Li and F. Wang, "Support vector machine ensemble based on choquet integral for financial distress prediction," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, 2015.
[24] M. Zieba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Systems with Applications*, vol. 58, 2016.
[25] V. García, A. I. Marqués, and J. S. Sánchez, "Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction," *Information Fusion*, vol. 47, pp. 88–101, 2019.
[26] M. Anandarajan, P. Lee, and A. Anandarajan, "Bankruptcy prediction of financially stressed firms: an examination of the predictive accuracy of artificial neural networks," *International Journal of Intelligent Systems in Accounting, Finance & Management*, vol. 10, no. 2, pp. 69–81, 2001.
[27] P. C. Pendharkar, "A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem," *Computers & Operations Research*, vol. 32, no. 10, pp. 2561–2582, 2005.
[28] H. Ahn and K.-j. Kim, "Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach," *Applied Soft Computing*, vol. 9, no. 2, pp. 599–607, 2009.
[29] S. Tian, Y. Yu, and M. Zhou, "Data sample selection issues for bankruptcy prediction," *Risk, Hazards & Crisis in Public Policy*, vol. 6, no. 1, pp. 91–116, 2015.
[30] J. Huang, H. Wang, and G. Kochenberger, "Distressed Chinese firm prediction with discretized data," *Management Decision*, vol. 55, 2017.
[31] S. Zeng, X. Wang, X. Duan, S. Zeng, Z. Xiao, and D. Feng, "Kernelized mahalanobis distance for fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, p. 1, 2020.
[32] K. Zhang, J. Zhan, and W.-Z. Wu, "On multi-criteria decision-making method based on a fuzzy rough set model with fuzzy α-neighborhoods," *IEEE Transactions on Fuzzy Systems*, vol. 99, 2020.
[33] J. Zhan, H. Jiang, and Y. Yao, "Three-way multi-attribute decision-making based on outranking relations," *IEEE Transactions on Fuzzy Systems*, vol. 99, 2020.