Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4053916, 13 pages
https://doi.org/10.1155/2022/4053916

Research Article

Research on Audit Opinion Prediction of Listed Companies Based on Sparse Principal Component Analysis and Kernel Fuzzy Clustering Algorithm

Sen Zeng,[1] **Yanru Li**,[2] **and Yaqin Li**[1]

[1]*School of Management, Wuhan Polytechnic University, Wuhan 430023, China*
[2]*School of Accounting, Zhongnan University of Economics and Law, Wuhan 430073, China*

Correspondence should be addressed to Yanru Li; zenglee1993@163.com

Received 18 September 2021; Revised 18 October 2021; Accepted 29 January 2022; Published 7 March 2022

Academic Editor: Jiayi Ma

The prediction of audit opinions of listed companies plays a significant role in the security market risk prevention. By introducing machine learning methods, many innovations can be implemented to improve audit quality, lift audit efficiency, and cultivate the keen insight of auditors. However, in a realistic environment, category imbalance and critical feature selection exist in the prediction model of company audit opinions. This paper firstly combines batched sparse principal component analysis (BSPCA) with kernel fuzzy clustering algorithm (KFCM) and proposes a sparse-kernel fuzzy clustering undersampling method (S-KFCM) to deal with the imbalance of sample categories. This method adopts the kernel fuzzy clustering algorithm to down-sample the normal samples, and their features are extracted from abnormal sample sets based on the group sparse component method. The sparse normal sample set can maintain the original distribution space structure and highlight the classification boundary samples. Secondly, considering the company's characteristic attributes and data sources, 448 original variables are grouped, and then BSPCA is used for feature screening. Finally, the support vector machine (SVM) is adopted to complete the classification prediction. According to the empirical results, the SKFCM-SVM model has the highest prediction accuracy.

1. Introduction

The financial reports and audit reports regularly disclosed by listed companies form a vital basis for investment decisions to various stakeholders. The audit report refers to the written document of the audit opinion issued by the certified public accountant (CPA) on the financial statements of the audited entity based on performing the audit work in accordance with the provisions of the audit standards. For the audit of financial statements, it is to express an opinion on whether the financial statements have been prepared in accordance with the applicable accounting standards and whether the financial statements are fair in all material aspects, reflecting the financial position, operating results, and cash flow of the auditee. Due to the limitation of professional knowledge, time, and other conditions, it is difficult for users of financial statements to effectively review and accurately judge the authenticity and compliance of enterprise financial statements. As a third party independent of the audited entities and stakeholders, CPA issues related audit opinions regarding the assurance documents of the company's financial situation, operational results, cash flow, and other information to enhance the credibility of financial information of listed companies. Machine learning methods can be implemented to improve audit quality, lift audit efficiency, and cultivate auditors, which has become the industry consensus. Notably, by combining supervised and unsupervised learning technology, the prediction model of audit opinions can effectively solve the potential contradiction between audit efficiency and audit risks, making the audit work of listed companies more valuable. Meanwhile, it can vastly shorten data processing time, reduce simple duplicate labor, strengthen analysis and monitoring, and allow auditors to solve problems with their occupational judgments, thereby reducing audit risks and drawing more credit conclusions to ensure the quality of audit reports. In

addition, the machine learning methods are also applied to the early warning of audit risk, which can assist stakeholders to estimate the type of audit opinions issued by the registered accountants in accordance with the relevant data of listed companies, optimize the security market resource allocation, reduce capital market risk and maintain market economic order.

Nevertheless, two significant challenges hinder the practicality of artificial intelligence technology in the company's audit opinion prediction. First, most listed companies in the stock market belong to the "Standard Unqualified Opinion," Only a few companies have been issued in other audit opinions, which has a typical category-imbalanced problem [1–3]. Second, audit opinion prediction incurs the obsession with the "curse of data dimension". On the one hand, all data features are incorporated into prediction models, which result in excessively fitting and affecting prediction results [4, 5]. On the other hand, most audit opinion prediction models only adopt financial statements and fewer nonfinancial indicators [6–10]. Some empirical studies have revealed a significant relationship between abnormal nonfinancial indicators and auditing opinions [11, 12]. Moreover, the selection of features is based on experience or previous research, which makes the prediction model easily affected by human factors. Therefore, it is the key to applying machine learning to the audit opinion prediction of listed companies that how to perform data mining and analysis of the comprehensive information of the security market to extract the valuable information to determine whether the company has audit risk by establishing a prediction model with strong anti-interference ability.

Given this, our research combines supervised learning technology with unsupervised learning technology. First, to deal with the imbalance problem of the company's sample category, we combine the sparse principal component analysis with a fuzzy kernel-clustering algorithm and propose a novel undersampling method. Second, in terms of the feature screening, to retain the internal structure of the company's characteristics, we employ the sparse principal component analysis to select the characteristic data after grouping aiming to remove redundant information in listed company data, thereby selecting the optimal auditing risk prediction characteristics. Finally, the support vector machine is adopted to classify.

The main contributions of this study are as follows:

(1) For the category imbalance problem in the company audit opinion prediction, we combine the sparse principal component with kernel fuzzy clustering algorithm to propose a sparse kernel fuzzy clustering undersampling method. By comparing with the mainstream methods in processing the data imbalance problem, the testing performance of our proposed approach outperforms others.

(2) In the unified kernel function mapping space, the sparse fuzzy clustering and SVM are combined to integrate the undersampling and classification prediction, which improves the classification prediction effect and enhances the model's comprehensibility.

(3) To highlight the practicability of the prediction model in this paper, we take all listed manufacturing companies in China's A-share market from 2012 to 2019 as the research object to reflect the actual structure of the sample space of companies in the security market. In terms of sample features, we use the most comprehensive feature data in China Securities Market and Accounting Research (CSMAR) database. The corresponding algorithms determine the sample matching and feature screening in the experiment to avoid the influence of human factors.

The remainder of this paper is organized as follows. Section 2 reviews the related research literature. Section 3 introduces our model's specific process and core algorithms, including the BSPCA and kernel fuzzy clustering undersampling methods. Section 4 demonstrates the data, design, evaluation index, and results regarding this study. Section 5 concludes our research work and addresses future research directions.

2. Literature Review

Early scholars usually use statistical analysis methods (such as Logistic and Probabilistic models) to study audit risk early warning of companies. However, traditional research methods are limited by strict assumptions and have poor fault tolerance. With the wide application of artificial intelligence technology in corporate governance, an increasing number of scholars apply machine learning algorithms to predict company audit risk and financial fraud. Gaganis and Pasiouras et al. [13] applied a probabilistic neural network to predict audit opinions of listed companies. The listed companies in London Stock Exchange as the experimental objects found that their proposed model was superior to the traditional artificial neural network and logistic regression models. Perols [3] compared the application of six popular classifiers in the field of corporate financial statement fraud and found that logistic regression and SVM performed well relative to artificial neural networks, bagging, C4.5, and stacking. Fernandez-Gamez et al. [14] combined financial variables with corporate governance variables to form a feature set and used multilevel perceptron and probabilistic neural network to establish a prediction model of audit opinions. Heng-Shu [15] took the financial indicators of listed companies as variables and introduced Takagi-Sugeno fuzzy neural network to construct the prediction model of audit opinions. Salehi and Dehnavi [16] applied the grey model to predict audit reports and found that the Nash nonlinear grey Bernoulli model had the best prediction effect. Yao and pan et al. [17] adopted stepping-regression and principal component analysis (PCA) to reduce the dimension of company characteristics and used six machine learning methods to identify fraudulent activities in financial statements of Chinese listed companies. It is found that stepwise regression and SVM have the highest classification accuracy. Omidi and minetal [18] analyzed the financial statement fraud in China's stock market by combining supervised and unsupervised learning. First, the financial statement data were divided into three groups by the

cluster analysis. Then, multilayer feedforward neural network (MLFNN), probabilistic neural network (PNN), support vector machine (SVM), polynomial log-linear model (MLM), and discriminant analysis (DA) were used for classification and prediction, and the research found that fuzzy neural network had the best classification effect. Bao and ke et al. [19] used ensemble learning to predict accounting fraud of listed companies in the United States, and the input data were original accounting figures rather than financial ratios, which was proved to have a good prediction effect. Sánchez-Serrano and José Ramón et al. [20] taking a group of Spanish companies as research samples, compares the effects of several different neural networks in the prediction of audit opinions on the company's consolidated financial statements, and MLP obtains the best prediction effect, with an accuracy of more than 86%. Chyan-Long Jan [21] forecasts the CPA's going concern audit opinions of Listed Companies in Taiwan. The results show that the model's prediction accuracy is the highest in the case of important variables selected by CART and modeling by RNN.

Previous studies on the construction of audit risk early warning of listed companies have the following limitations:

First, there is little research on selecting the overall characteristics of listed companies. Existing studies have applied factor analysis, rough domain set, principal component analysis, and other techniques to filter or reduce the dimension of company characteristics, but most of their initial characteristics are determined according to previous studies. Considering that artificial intelligence technology can handle collinearity problems well, and the audit of listed companies is for annual financial statements, characteristics in the database should be considered as comprehensively as possible. All financial indicators should be included in the research scope.

Second, it is insufficient in research on the imbalance problem of audit risk prediction of listed companies. The existing research on the audit risk prediction of listed companies often ignores the category imbalance, resulting in companies with audit risk taking only a tiny part of the stock market. Most scholars artificially select control samples according to industry and asset size to construct balanced data sets. This method ignores the original sample structure of the stock market and reduces the practicability of the prediction model. Some scholars also use SMOTE oversampling technology to deal with imbalance problems. However, the new samples synthesized by this method may not provide too much helpful information.

Therefore, we should compare various techniques to deal with the imbalance problem and study the best method to deal with the audit risk samples of listed companies. This study uses a larger sample size and more comprehensive sample characteristics to analyze the prediction effect of the model to retain the real market environment faced by the listed companies to enhance the practicability of the model.

3. Model Description

We propose a hybrid classification model that combines Sparse principal component analysis (SPCA) [22], Kernel Fuzzy C-means algorithm (KFCM), K-Nearest Neighbor algorithm (KNN), and SVM. Figure 1 reflects the process of sample matching, feature screening, and classification prediction throughout the model, and Figure 2 shows the flow of the sparse-kernel fuzzy clustering undersampling method (S-KFCM) explicitly. As shown in Figure 1, the entire model is divided into three phases.

The first stage is a sample matching phase. In order to solve the problem of imbalanced problems in the prediction of listed companies' audit opinions, we apply the sparse-kernel fuzzy clustering undersampling method (S-KFCM) to choose the most similar and representative matching samples and build the balanced sample data set. As shown in Figure 2, we first divided the data set into several groups according to the year that the sample belongs to, and then conducted batched sparse principal component analysis (BSPCA) on the minority group (nonstandard audit opinion companies) in each group to obtain the feature set that best reflects the sample of the category of companies. Then the feature is used to cluster most of the samples in the same year, and the kernel fuzzy clustering algorithm is used to determine the clustering center of the majority group samples. Finally, the nearest neighbor algorithm is used to find the majority group samples closest to these clustering centers as the control samples of the minority group samples. The second stage is a feature screening phase. After obtaining the category-balanced data set, we also use BSPCA to perform feature screening, eliminate redundant information in raw data, and build the best feature dataset. The final stage is a classified prediction stage. We put the training data into SVM to train the model and then input the test sample for prediction, where SVM employs the same kernel with the nuclear fuzzy clustering algorithm in the first phase.

In the detection procedure, the key algorithms are BSPCA in stages 1 and 2 and the kernel fuzzy clustering undersampling method in Stages 1. Therefore, we provide a more explicit description of our proposed algorithm in the following two subsections.

3.1. Batched Sparse Principal Component Analysis (BSPCA). Sparse principal component analysis (SPCA) is proposed because principal component analysis can be transformed into a quadratic penalty regression problem. The objective function of SPCA is as follows:

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

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

By adding the LASSO penalty item to the above equation, sparse principal components can be obtained. Thus, the following optimization problem can be obtained:

$$(\widehat{\alpha},\widehat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} X_i - \alpha\beta^T X_i^2 + \lambda \sum_{j=1}^{k} \beta_j^2 + \sum_{j=1}^{k} \lambda_{1,j}\beta_{ji}, \tag{2}$$

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

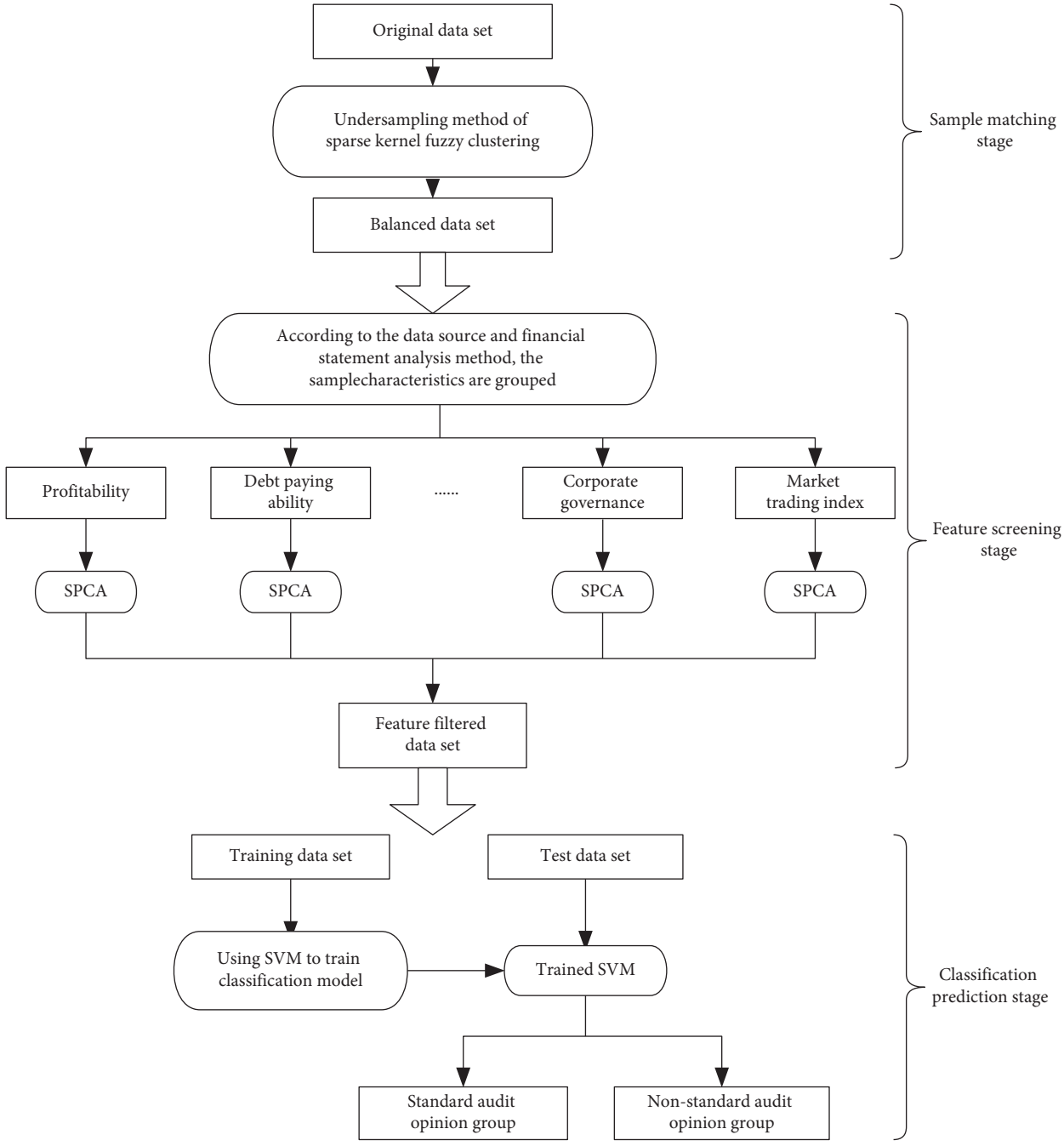

Figure 1: The procedure of audit opinion prediction with SKFCM-SVM.

As stated above, the solution of sparse principal components can be transformed into a penalty regression problem. The calculation of sparse principal components was obtained by using the least angle regression algorithm.

The steps of grouping SPCA are as follows:

(1) Collect and standardize the characteristic indicators of listed companies

(2) Divide the characteristic indexes of listed companies into several groups (such as solvency, profitability, growth ability, etc.) according to financial statement analysis methods and data source

(3) Use sparse principal component analysis to screen the characteristics of each group

(4) Combine the characteristics screened by each group into a new dataset

More details about the BSPCA can be referred to [23].

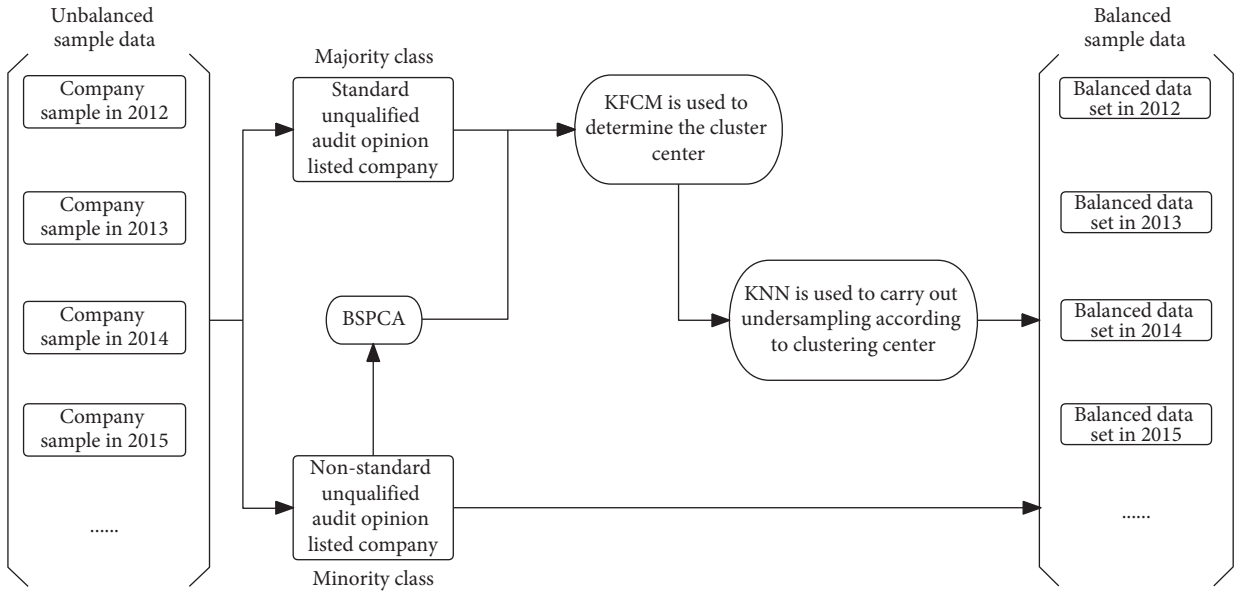

Figure 2: The procedure of sample matching with S-KFCM.

3.2. Undersampling Method of Kernel Fuzzy c-Means Clustering. The objective function of KFCM-K is,

$$Q = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\phi(x_k) - v_i^2, \quad s.t. \sum_{i=1}^{c} u_{ik} = 1, \quad k = 1, 2, \ldots, n. \tag{3}$$

Here, $\|\cdot\|$ is the Euclidean distance. u_{ik} the membership of data x_k belonging to the cluster i, represented by the prototype v_i, m is the fuzzification.

Given the Euclidean distance and optimizing Q concerning located in the kernel space such that $\nabla_{v_i} Q = 0$, we obtain [24–27]:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m}\phi(x_k)}{\sum_{k=1}^{n} u_{ik}^{m}}, \quad i = 1, 2, \ldots, c, \tag{4}$$

The prototype expression for the Gaussian kernel for $i = 1, 2, \ldots, c$ is then given as,

$$\widetilde{v}_i = \frac{\sum_{k=1}^{n} u_{ij}^{m} K(x_k, \widetilde{v}_i) x_i}{\sum_{k=1}^{n} u_{ij}^{m} K(x_k, \widetilde{v}_i)}. \tag{5}$$

The learning algorithm of KFCM-K iteratively updates u_{ik} as,

$$u_{ik} = \sum_{j=1}^{c} \left(\frac{\phi(x_k - v_i)}{\phi(x_k - v_j)} \right)^{-2/(m-1)}, \tag{6}$$

$$\phi(x_k - v_i)^2 = K(x_k, x_k) - 2\frac{\sum_{j=1}^{n} u_{ij}^{m} K(x_k, x_j)}{\sum_{j=1}^{n} u_{ij}^{m}} + \frac{\sum_{j=1}^{n}\sum_{t=1}^{n} u_{ij}^{m} u_{it}^{m} K(x_k, x_t)}{\left(\sum_{j=1}^{n} u_{ij}^{m}\right)^2}. \tag{7}$$

More details about the derivation of (6) and (7) can be referred to [28].

Use the KNN algorithm to calculate the Euclidean distance between the majority group samples and each cluster center by $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$.

Sort the distance values and select the K points with the smallest distance.

Our model is summarized as Algorithm 1.

4. Experiment Results and Discussion

In this section, we evaluate the performance of the proposed SKFCM-SVM method by using the real information data of manufacturing listed companies in China's A-share market. MATLAB2016b and Python3.8 as tools to obtain the calculation results.

4.1. Datasets. In accordance with the Guidance on Industry Classification of Listed Companies (2012) issued by the China Securities Regulatory Commission, this paper selects the manufacturing listed companies in China's A-share market from 2012 to 2019 as the research samples. According to the provisions of "China Registered Accountants Auditing Standards (CRAAS) No. 1504 - Communication of Key Audit Matters in Audit Reports", audit opinions in the audit of financial statements can be divided into standard unqualified opinion, unqualified opinion with emphasis, or other matters, qualified opinion, adverse opinion, and disclaimer of opinions. Considering that most listed companies are standard unqualified audit opinions, we define the remaining audit opinions as nonstandard audit opinions. The companies are divided into standard unqualified audit opinion groups and nonstandard audit opinion groups. The total number of samples in the experiment is 14,145, including 13,526 standard unqualified and 619 nonstandard audit opinions. Since the sample size in the model is large enough, the trained mixed model has high stability, practicability, and robustness, and the recognition result should be ideal.

Input: Given a set of n data points $X = \{x_i\}_{i=1}^{n}$, a basis Gaussian kernel function K s, the number of clusters c, the fuzzy index m, the termination criterion ζ and T, and the initialization partition matrix $U^{(0)} = \{u_{ij}\}_{i,j=1}^{c,n}$.
Output: The clustering prototypes V.
(1) **Procedure** KFCM_KNN (Data X, Number c, kernel functions K)
(2) The partition matrix $U^{(0)} = \{u_{ij}\}_{i,j=1}^{c,n}$ from FCM as initial membership matrix
(3) Calculate kernel matrix $K = \{K_{ij}\}_{i,j=1}^{c,n}$ by $K = \begin{bmatrix} \phi^T(x_1)\phi(x_1) & \cdots & \phi^T(x_1)\phi(x_n) \\ \vdots & \ddots & \vdots \\ \phi^T(x_n)\phi(x_1) & \cdots & \phi^T(x_n)\phi(x_n) \end{bmatrix}$.
(4) **Repeat**
(5) Calculate distances $D = \{d_{ij}^2\}_{i,j=1}^{c,n}$ by (7)
(6) Update partition matrix $U^{(t)} = \{u_{ij}\}_{i,j=1}^{c,n}$ by (6)
(7) Update clustering prototypes $V^{(t)}$ by (5)
(8) **Until** $|J^{(t+1)} - J^{(t)}| \leq \zeta$ or the number of iterations $t > T$.
(9) **Return** $U = \{u_{ij}\}_{i,j=1}^{c,n}$ and V.
(10) **End procedure**
(11) Find the sample closest to the clustering prototypes by the KNN algorithm.

Algorithm 1: Kernel Fuzzy C Mean with KNN (KFCM_KNN).

Table 1 reflects the annual sample distribution of nonstandard audit opinion companies and standard audit opinion companies in China's A-share manufacturing industry. The imbalance ratio (IR) is the number of companies in the standard audit opinion (majority category) divided by the number of companies in the nonstandard audit opinion (minority category). As shown in Table 1, with the gradually tightening supervision from China Securities Regulatory Commission (CSRC) on the capital market, the number of listed companies with nonstandard audit opinions continues to increase year by year nonbalance ratio (IR) gradually decreases. However, compared with the companies with standard unqualified audit opinions, the data set of manufacturing listed companies still has a severe category imbalance.

Considering that the prediction of audit opinions of listed companies has vital timeliness, the training and test samples are divided according to the year of data. We set four data sets, and their specific sample distributions are shown in Table 2. Since the data in the last year of each data set also has the problem of unbalanced sample categories, we carry out kernel fuzzy clustering analysis on majority category samples in the annual sample data from 2016 to 2019. The most representative majority class samples are found and matched with the minority class samples, and the balanced data sets are constructed as the test samples.

4.2. Selection of Alternative Indicators. At present, there is no specific economic theory to guide the selection of indicators to predict audit opinions. The fundamental reasons for various audit opinions are different, so it is difficult to fully describe a few simple ratio indicators. Therefore, through China Stock Market and Accounting Research Database, we download the financial indicators, corporate governance indicators, and market transaction data of listed manufacturing companies and attempt to obtain comprehensive information reflecting all aspects of the company. In order to retain the internal structure of the company's data information and apply it to batched sparse principal component analysis (BSPCA) introduced in Section 3.1, we use the feature grouping method of the CSMAR database for reference and divide all features into 15 groups. Among them, 390 indicators reflect the financial statement information of listed companies, which are divided into ten groups. There are 35 features of ratio structure, 27 features of solvency, 56 features of development ability, three features of risk level, eight features of dividend capacity, 67 features of operation ability, 88 features of index per share, seven features of financial disclosure index, 32 features of cash flow, and 67 features of profitability. In addition, 51 indicators reflect the governance ability of listed companies, divided into four groups, including five characteristics of ownership structure, ten characteristics of top ten shareholders, 28 indicators of relative value, and eight indicators of comprehensive governance information. There even are seven indicators that reflect the market trading information of listed companies.

4.3. Performance Evaluation Measures. Considering that the research in this paper belongs to the dichotomy problem when it comes to class imbalance or cost imbalance, the accuracy cannot sufficiently reflect the classification effect of the model. Therefore, we use the idea of a confusion matrix to measure the model performance. We set the nonstandard audit opinion as positive and standard unreserved audit opinion as negative. The correct prediction of the nonstandard audit opinion sample is true positive (TP), the false prediction of the nonstandard audit opinion sample is a false negative (FN), the correct prediction of the standard unqualified audit opinion sample is true negative (TN), and the false prediction of the standard unqualified audit opinion sample is false positive (FP). Specific evaluation indexes include accuracy, F1 value, G-mean, and Matthews correlation coefficient (MCC), and the specific calculation formulas are as follows:

Table 1: Company sample distribution.

Year	Number of companies	Number of companies with standard unqualified audit opinions	Number of companies with non-standard audit opinions	Imbalance ratio (IR)
2012	1447	1401	46	30.45
2013	1265	1219	46	26.5
2014	1343	1293	50	25.86
2015	1455	1401	54	25.94
2016	1894	1825	69	26.44
2017	2179	2108	71	29.69
2018	2242	2114	128	16.51
2019	2320	2165	155	13.96

Table 2: Sample distribution of the dataset.

Datasets	Year of the training sample	Number of training samples	Year of the test sample	Number of test samples
Dataset I	2012–2015	5510	2016	138
Dataset II	2012–2016	7404	2017	142
Dataset III	2012–2017	9583	2018	256
Dataset IV	2012–2018	11825	2019	310

$$\begin{aligned} \mathrm{CR} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \\ F_1 &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \\ G-\mathrm{mean} &= \sqrt{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\cdot\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}}, \\ \mathrm{MCC} &= \frac{\mathrm{TP}\cdot\mathrm{TN}+\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \end{aligned} \tag{8}$$

4.4. Comparison of Results and Discussion of Sample Matching Methods. In order to analyze the effect of the sparse kernel fuzzy clustering undersampling method we proposed to deal with the imbalance problem of sample categories, we will take the four data sets in Table 2 as experimental objects, introduce five popular sampling methods, and use the same classifier (SVM) for comparative study. The five sampling methods are random oversampling (RO), synthetic minority oversampling technique (SMOTE) [29, 30], adaptive synthetic sampling (ADASYN) [31], random undersampling (RU), and NearMiss [32]. The first three methods belong to oversampling, while the last two methods and the method in this paper (SKFCM) belong to undersampling.

Random oversampling (RO) randomly samples from the minority category samples and then adds the samples to the data set. SMOTE is to interpolate between the minority category samples to generate additional samples. ADASYN method also synthesizes new samples. The most significant feature of the ADASYN method is to adopt some mechanisms to automatically determine the number of synthesized samples to be generated for each minority sample, rather than synthesizing the same number of samples for each minority sample like SMOTE. Random undersampling (RU) is similar to random oversampling in that some samples are randomly selected from most samples. NearMiss is a prototype selection method that selects the most representative samples from the majority category samples for training, mainly aiming to alleviate information loss in random undersampling.

As shown in Table 3 and Figure 3, the S-KFCM method we proposed has achieved the best classification effect in four data sets, indicating that S-KFCM is more suitable for dealing with the matching problem of listed companies. In contrast, there is a specific defect in the sampling methods mentioned above for performing the comparative experiment. Random oversampling methods (RO), due to repeated sampling, often lead to severe extensions. Regarding the SMOTE method, if there are also a few samples around the selected minority sample, the new synthetic sample does not provide helpful information. If choosing a few samples around the majority category samples, this sample type may be noise, and the new synthetic sample will produce most of the surrounding samples to overlap, resulting in difficulties in classification. The ADASYN method and the NearMiss method are easily affected by the group point. The disadvantage of random undersampling (RU) is that the excluded samples may contain some critical information, leading to the learned model's poor effect.

Our proposed S-KFCM regards the minority category samples as the core object and uses batched sparse principal component analysis (BSPCA) to dig the vital feature combination of minority samples. Based on this feature set, the most representative samples in the majority category are found by the KFCM and KNN algorithms. Furthermore, all the underlying processes are completed according to the year of sample data, which fully considers the particular environment in the sample year to build the best sample balance dataset.

4.5. Comparison Results and Discussion of Characteristic Degradation Algorithms. In this section, we also use the relevant experiments to test the BSPCA method adopted in this paper for feature dimension reduction. We used three

Table 3: Classification results of different sample matching methods.

Dataset	Data set I				Data set II			
Method	Cr	F1	G	MCC	Cr	F1	G	MCC
RO	87.55	86.24	87.03	76.51	85.92	84.13	85.17	73.73
SMOTE	86.95	85.94	86.65	74.70	85.21	83.46	84.55	72.05
ADASYN	87.68	86.82	87.44	76.01	84.51	82.54	83.75	70.84
RU	89.85	89.19	89.63	80.32	87.68	86.36	87.14	76.80
NearMiss	77.54	80.25	76.30	57.29	77.46	79.22	77.00	55.73
S-KFCM	**94.20**	**94.03**	**94.16**	**88.55**	**88.03**	**87.02**	**87.69**	**76.99**
DATASET	Data set III				Data set IV			
Method	CR	F1	G	MCC	CR	F1	G	MCC
RO	81.25	78.56	80.19	63.60	85.23	83.54	84.61	71.98
SMOTE	81.25	78.76	80.40	64.29	83.87	81.88	83.15	69.43
ADASYN	80.86	78.22	79.95	63.61	83.55	81.59	82.87	68.67
RU	81.91	79.51	81.06	65.66	85.20	83.52	84.58	71.94
NearMiss	76.95	77.74	76.87	54.04	77.74	78.77	77.59	55.75
S-KFCM	**83.98**	**81.78**	**83.11**	**70.05**	**86.45**	**85.00**	**85.91**	**74.31**

The bold values are the maximum of the list.

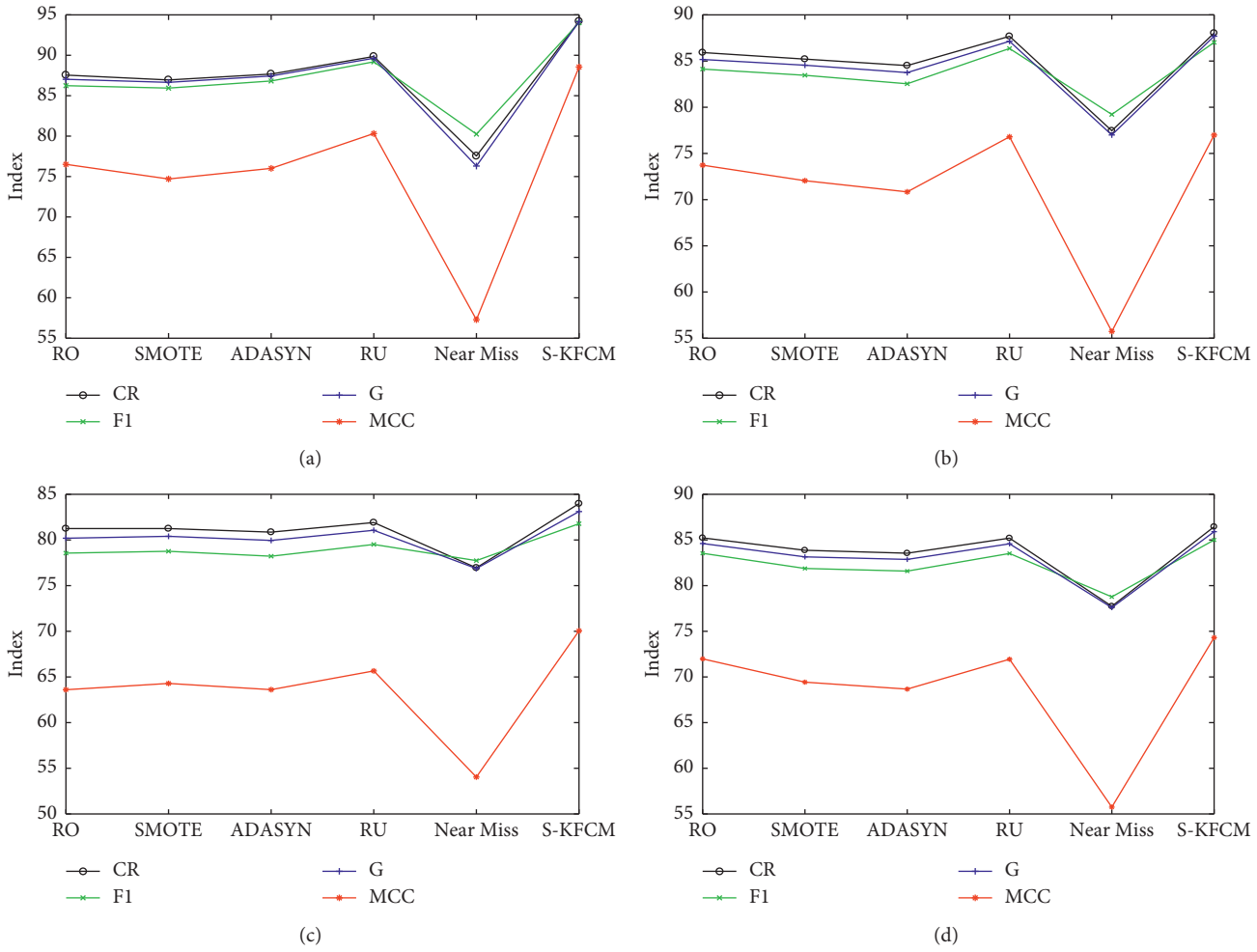

Figure 3: Index values of different sample matching methods under four data sets. (a) Data set I (b) Data set II (c) Data set III (d) Data set IV.

usual data dimensionality reduction methods for comparative analysis, including linear discriminant analysis, principal component analysis, and sparse principal component analysis. The nearest neighbor algorithm is the classifier in the linear discriminant analysis, and SVM is the classifier for the other dimensionality reduction methods. As shown from Table 4 and Figure 4, the classification effect of batched sparse principal component analysis (BSPCA) is the best in all data

Table 4: Classification results of different characteristic degradation algorithms.

Dataset	Data set I				Data set II			
Method	Cr	F1	G	MCC	Cr	F1	G	MCC
LDA	56.42	67.06	54.71	21.27	63.38	63.89	63.36	26.77
PCA	90.58	90.08	90.44	81.58	88.03	75.86	86.82	77.36
SPCA	93.48	93.23	93.41	87.19	87.32	86.36	87.04	75.40
BSPCA	**94.93**	**94.81**	**94.90**	**89.94**	**90.14**	**89.23**	**89.74**	**81.45**
DATASET	Data set III				Data set IV			
Method	CR	F1	G	MCC	CR	F1	G	MCC
LDA	51.17	49.80	51.10	2.350	51.94	48.44	51.49	3.910
PCA	83.20	81.39	82.63	67.71	85.16	83.69	84.68	71.50
SPCA	83.59	81.42	82.77	69.11	86.13	85.22	85.91	72.81
BSPCA	**83.98**	**82.10**	**83.32**	**69.53**	**87.10**	**85.82**	**86.63**	**75.43**

The bold values are the maximum of the list.

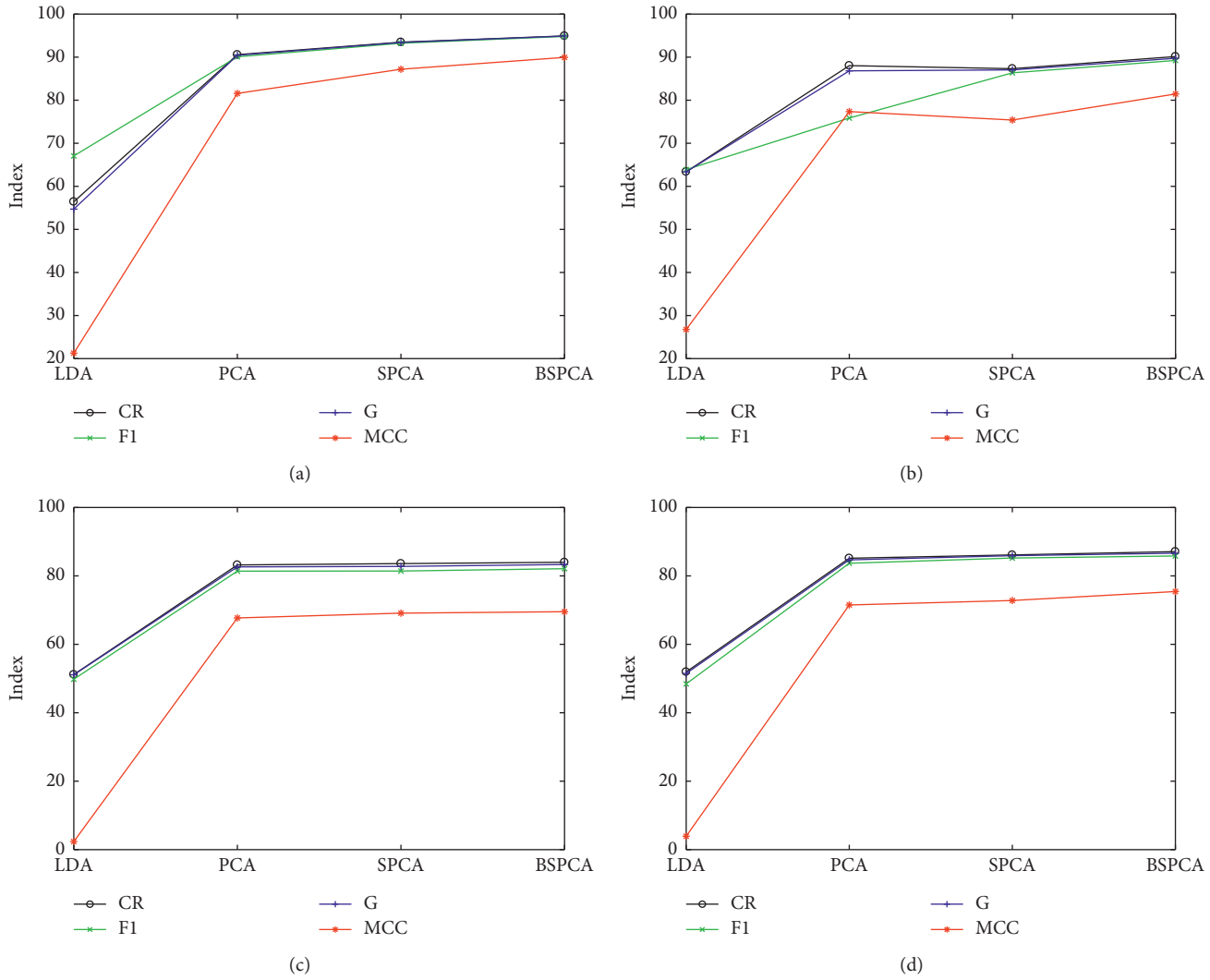

Figure 4: Index values of different characteristic degradation algorithms under four data sets. (a) Data set I (b) Data set II (c) Data set III (d) Data set IV.

sets. The cause is that the BSPCA method is different from the other three methods. BSPCA groups all features according to financial statement analysis methods and data sources and then uses sparse principal component analysis to screen each feature group. This method reduces the redundant data of each feature group, which means fewer opportunities to make decisions according to noise to reduce overfitting. It can be used to measure the information category and relative

Table 5: Classification results of different classifier algorithms.

Dataset	Data set I				Data set II			
Method	Cr	F1	G	MCC	Cr	F1	G	MCC
MLFNN	74.64	79.77	70.20	57.18	88.03	86.82	87.55	77.36
KNN	86.23	84.55	85.54	74.24	81.69	78.69	80.47	66.06
NB	50.72	66.34	20.55	3.880	85.93	84.13	85.17	73.74
C4.5	65.22	59.32	63.59	31.80	**90.85**	**90.91**	**90.84**	**81.70**
Bagging	80.43	79.39	80.27	61.19	79.58	80.27	79.50	59.30
SVM	**94.93**	**94.81**	**94.90**	**89.94**	90.14	89.23	89.74	81.45
DATASET	Data set III				Data set IV			
Method	CR	F1	G	MCC	CR	F1	G	MCC
MLFNN	79.30	74.64	77.14	62.99	79.03	80.24	78.79	58.51
KNN	78.52	75.77	77.69	58.55	76.45	72.45	75.06	55.28
NB	82.03	79.82	81.30	65.65	76.13	77.02	76.03	52.42
C4.5	82.03	**82.17**	81.25	64.07	68.06	58.58	64.10	40.64
Bagging	80.08	79.35	80.00	60.31	84.19	84.04	84.19	68.40
SVM	**83.98**	82.10	**83.32**	**69.53**	**87.10**	**85.82**	**86.63**	**75.43**

The bold values are the maximum of the list.

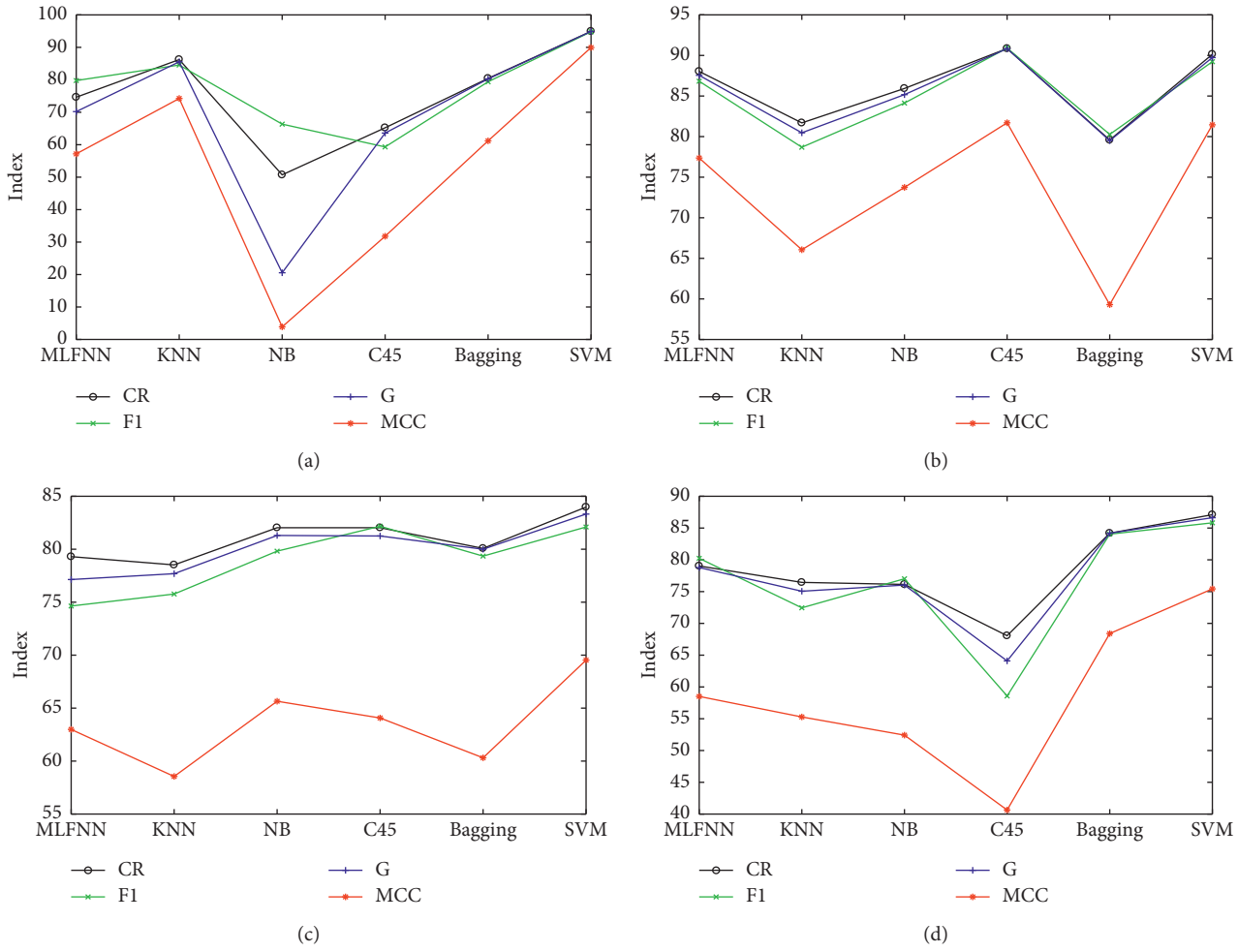

Figure 5: Index values of different classifier algorithms under four data sets. (a) Data set I (b) Data set II (c) Data set III (d) Data set IV.

importance of all variables. By grouping data features, it retains the internal structure of sample features to the maximum extent to better retain the adequate information in the original features in removing redundant information.

4.6. Comparison Results and Discussion of Model Classifier Algorithms. This section is the third part of the experiment. Under the condition that the model completes sample matching and data dimensionality reduction, we test the

influence of different classifiers on the prediction effect of audit opinions of listed companies. In this paper, multilayer feedforward neural network (MLFNN), K-Nearest Neighbor (KNN), Naive Bayesian classifier (NB), C4.5, Bagging and support vector machine (SVM) are introduced for comparative analysis.

Table 5 and Figure 5 show the prediction effects of different classifiers after using sparse kernel fuzzy clustering undersampling for sample matching and batched sparse principal component analysis for feature selection of four data sets. The best index results of each data set are shown in bold. As shown from Table 5, SVM achieves the best results in Dataset I and Dataset IV and falls behind the C4.5 algorithm only in the evaluation index of the Dataset II and the F1 value Dataset III. The difference between SVM and the other five classifier algorithms is that SVM uses kernel function in the classification process. Considering that the KFCM algorithm and SVM used in our model selected the same kernel function (Gaussian nuclear function), this indicates that the sample matching and classification prediction of our model is completed under the same kernel space, so SVM is suited to be the classifier for the company's audit opinion prediction model.

4.7. Significance Test. This section uses the Friedman test to determine significant differences between various methods in three partial experiments. We compared six sample matching algorithms, four feature drop-dimensional algorithms, and six classifiers. In Friedman's analysis, the card distribution is approximately Friedman's test statistics. We sorted the g mean in the above experiments and calculated the ranking of various methods in three partial experiments. The levels of various methods in three partial experiments are shown in Tables 6 to, 8.

$$* \chi^2_{0.05} = 11.071 < 17.57. \tag{9}$$

$$* \chi^2_{0.05} = 7.815 < 12. \tag{10}$$

$$* \chi^2_{0.05} = 11.071 > 7.14. \tag{11}$$

Under the null hypothesis, it would be no difference between all the methods, and therefore theoretically, R_j^2 should be equal. From the data in Tables 6–8, the value of the Friedman test statistics can be calculated as follows:

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

where n is the number of datasets and k is the number of methods, 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 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 freedom degrees.

Table 6: The ranks of the six sampling algorithms on four databases.

	RO	SMOTE	ADASYN	RU	NearMiss	S-KFCM
DATA SET I	4	5	3	2	6	1
DATA SET II	3	4	5	2	6	1
DATA SET III	4	3	5	2	6	1
DATA SET IV	2	4	5	3	6	1
Total rank	13	16	18	9	24	4
Ave rank	3.25	4	4.5	2.25	6	1

Table 7: The ranks of the four characteristic degradation algorithms on four databases.

	LDA-KNN	PCA-SVM	SPCA-SVM	BSPCA-SVM
DATA SET I	4	3	2	1
DATA SET II	4	3	2	1
DATA SET III	4	3	2	1
DATA SET IV	4	3	2	1
Total rank	16	12	8	4
Ave rank	4	3	2	1

Table 8: The ranks of the six algorithms on four databases.

	MLP	KNN	NB	C4.5	Bagging	SVM
DATA SET I	4	2	6	5	3	1
DATA SET II	3	5	4	1	6	2
DATA SET III	6	5	2	3	4	1
DATA SET IV	3	5	4	6	2	1
Total rank	16	17	16	15	15	5
Ave rank	4	4.25	4	3.75	3.75	1.25

According to Table 6 and Table 7, the first two parts of the experiment in this paper have passed the Friedman test, indicating significant differences between our sample matching method and feature screening method. According to Table 8, the third part of the experiment did not pass the Friedman test, indicating that our classifier methods only show specific differences.

5. Conclusion

Predicting audit opinions of listed companies is a research field with excellent application prospects. It can provide scientific and technical support for auditors to issue audit opinions, reduce audit risks and improve audit efficiency. However, the scarcity of non-standard audit opinion companies, the relative diversity of characteristic variables used in the existing literature, and the wide use of various classifiers pose challenges to the effective prediction of the model.

Given this, we combine batched sparse principal component analysis (BSPCA), kernel fuzzy clustering analysis and SVM, and put forward a whole set of models to deal with the prediction of audit opinions of listed companies in a real market environment. Our experiments are divided into three parts. In the first part, to show that the sparse kernel fuzzy clustering undersampling method can effectively deal with the imbalance problem of sample categories, we introduce random oversampling, SMOTE, ADASYN, random undersampling, and NearMiss for comparative experiments. It is proved that our method is most suitable for sample matching of audit opinion prediction of listed companies. In the second part, we studied the feature dimensionality reduction methods of listed companies, compared linear discriminant analysis, principal component analysis, sparse principal component analysis, batched sparse principal component analysis. Finally, we found that batched sparse principal component analysis has apparent advantages. In the third part, after determining the sample matching method and the feature dimension reduction method, we compare the influence of different classifiers on the prediction effect of audit opinions of listed companies, and the result shows that SVM has the best classification and prediction effect.

This study simulates real scenes to show how these technologies are appropriately used to audit listed companies. Significantly, this paper combines Batched Sparse principal component analysis (BSPCA) with the kernel fuzzy clustering algorithm to down-sample normal samples (listed companies with standard audit opinions) to balance training samples. In this process, BSPCA processes the abnormal sample set to obtain the abnormal samples' salient features, which are used as the features of normal samples for fuzzy kernel clustering to achieve the purpose of downsampling. It can not only maintain the sample distribution space but also accurately extract the boundary samples. However, there are still shortcomings in this study. First, this study does not cover listed companies in other industries. In future research, all listed companies in China's A-share market should be collected to further test our method. Second, further research can be performed by adding more variables, such as the text information in the company's annual report and audit report, which seems to be an essential factor. Finally, a comparative empirical study can also be carried out when the data set is obtained from other stock markets. In addition, the methods and processes used in this paper can be extended to other fields (such as company financial distress prediction and bonds predictions) more than company audit prediction.

Data Availability

The data used to support the findings of this study are currently under embargo while the research findings are commercialized. Requests for data, 12 months after publication of this article, will be considered by the corresponding author.

Disclosure

Sen Zeng and Yanru Li are co-first authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Sen Zeng and Yanru Li contributed equally to this work.

Acknowledgments

This paper was supported by 2021 Philosophy and Social Science research project of Universities in Hubei Province, project name: "Research on Corporate Audit risk Prediction based on machine learning technology", the China University industry research Innovation Fund (2019ITA03044), the Graduate education innovation program of Zhongnan University of Economics and Law (201911135), and the Scientific Research Program of Wuhan Polytechnic University (2018J06).

References

[1] J. L. Perols, R. M. Bowen, C. Zimmermann, and B. Samba, "Finding needles in a haystack: using data analytics to improve fraud prediction," *The Accounting Review*, vol. 92, no. 2, pp. 221–245, 2016.

[2] T. B. Bell and J. V. Carcello, "A decision aid for assessing the likelihood of fraudulent financial reporting," *Auditing: A Journal of Practice & Theory*, vol. 19, no. 1, pp. 169–184, 2000.

[3] J. Perols, "Financial statement fraud detection: an analysis of statistical and machine learning algorithms," *Auditing: A Journal of Practice & Theory*, vol. 30, no. 2, pp. 19–50, 2011.

[4] D. G. Whiting, J. V. Hansen, J. B. McDonald, C. Albrecht, and W. S. Albrecht, "Machine learning methods for detecting patterns of management fraud," *Computational Intelligence*, vol. 28, no. 4, pp. 505–527, 2012.

[5] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, "MetaFraud: a meta-learning framework for detecting financial fraud," *MIS Quarterly*, vol. 36, no. 4, pp. 1293–1327, 2012.

[6] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," *Expert Systems with Applications*, vol. 32, no. 4, pp. 995–1003, 2007.

[7] P. Ravisankar, V. Ravi, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," *Decision Support Systems*, vol. 50, no. 2, pp. 491–500, 2011.

[8] I. Dutta, S. Dutta, and B. Raahemi, "Detecting financial restatements using data mining techniques," *Expert Systems with Applications*, vol. 90, pp. 374–393, 2017.

[9] C. W. Liu, Y. X. Chan, S. H. A. Kazmi, and H. Fu, "Financial fraud detection model: based on random forest," *International Journal of Economics and Finance*, vol. 7, pp. 178–188, 2015.

[10] S. Kotsiantis, E. Koumanakos, D. Tzelepis, and V. Tampakas, "Forecasting fraudulent financial statements using data mining," *International Journal of Computational Intelligence*, vol. 3, pp. 104–110, 2006.

[11] M. S. Beasley, "An empirical analysis of the relation between the board of director composition and financial statement fraud," *The Accounting Review*, vol. 71, pp. 443–465, 1996.

[12] Y. J. Kim, B. Baik, and S. Cho, "Detecting financial misstatements with fraud intention using multi-class cost-

sensitive learning," *Expert Systems with Applications*, vol. 62, pp. 32–43, 2016.
[13] C. Gaganis, F. Pasiouras, and M. Doumpos, "Probabilistic neural networks for the identification of qualified audit opinions," *Expert Systems with Applications*, vol. 32, no. 1, pp. 114–124, 2007.
[14] M. A. Fernández-Gámez, F. García-Lagos, and J. R. Sánchez-Serrano, "Integrating corporate governance and financial variables for the identification of qualified audit opinions with neural networks," *Neural Computing & Applications*, vol. 27, no. 5, pp. 1427–1444, 2016.
[15] H.-s. Tang, "Audit opinion of listed companies: a Takagi-Sugeno fuzzy neural network based study," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 20, no. 4, pp. 899–912, 2017.
[16] M. Salehi and N. Dehnavi, "Audit report forecast: an application of nonlinear grey Bernoulli model," *Grey Systems: Theory and Application*, vol. 8, no. 3, pp. 295–311, 2018.
[17] J. Yao, Y. Pan, S. Yang, Y. Chen, and Y. Li, "Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: a multi-analytic approach," *Sustainability*, vol. 11, no. 6, p. 1579, 2019.
[18] M. Piri, Q. Min, V. Moradinaftchali, and M. Omidi, "The efficacy of predictive methods in financial statement fraud," *Discrete Dynamics in Nature and Society*, vol. 2019, Article ID 4989140, 12 pages, 2019.
[19] Y. Bao, B. Ke, B. Li, Y. J. Yu, and J. Zhang, "Detecting accounting fraud in publicly traded U.S. Firms using a machine learning approach," *Journal of Accounting Research*, vol. 58, no. 1, pp. 199–235, 2020.
[20] J. R. Sánchez-Serrano, A. David, G. F. Lagos, and M. C. G. Angela, "Predicting audit opinion in consolidated financial statements with artificial neural networks," *Mathematics*, vol. 8, no. 8, 2020.
[21] C. L. Jan, "Using deep learning algorithms for CPAs' going concern prediction," *Information*, vol. 12, no. 2, 2021.
[22] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational & Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.
[23] S. Zeng, Y. Li, W. Yang, and Y. Li, "A financial distress prediction model based on sparse algorithm and support vector machine," *Mathematical Problems in Engineering*, vol. 2020, Article ID 5625271, 11 pages, 2020.
[24] J. H. Chiang and P. Y. Hao, "A new kernel-based fuzzy clustering approach: support vector clustering with cell growing," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp. 518–527, 2003.
[25] L. Zeyu, T. Shiwei, X. Jing, and J. Jun, "Modified FCM clustering based on kernel mapping," in *Proceedings of the Internat. Society for Optical Engineering*, pp. 241–245, CA, USA, August 2001.
[26] D. Zhang and S. Chen, "Fuzzy clustering using kernel method," . in *Proceedings of the International Conference on Control and Automation*, pp. 123–127, Washington D. C., USA, May 2002.
[27] S. Zhou and J. Q. Gan, "Mercer kernel, fuzzy C-means algorithm, and prototypes of clusters," *Lecture Notes in Computer Science*, vol. 3177, pp. 613–618, 2004.
[28] D. Graves and W. Pedrycz, "Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study," *Fuzzy Sets and Systems*, vol. 161, no. 4, pp. 522–543, 2010.
[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.
[30] H. Hui, W. Y. Wang, and B. H. Mao, "Borderline-Smote: A new over-sampling method in imbalanced data sets learning," in *Proceedings of the 2005 International Conference on Advances in Intelligent Computing*, Hefei, Anhui, China, August 2005.
[31] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Adaptive synthetic-nominal (ADASYN-N) and adaptive synthetic-KNN (ADASYN-KNN) for multiclass imbalance learning on laboratory test data," in *Proceedings of the 2018 4th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia, August 2018.
[32] I. Mani, "KNN approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of the Icml Workshop on Learning from Imbalanced Datasets*, Ottawa, Canada, August 2003.