Hindawi
Scientific Programming
Volume 2022, Article ID 3398545, 9 pages
https://doi.org/10.1155/2022/3398545

Research Article

Bank Financial Risk Prediction Model Based on Big Data

Hua Peng,[1,2] **Yicheng Lin**,[2] **and Mingzheng Wu**[2]

[1]*Wuyi University, Wuyishan 354300, China*
[2]*National Changhua University of Education, Changhua 50007, China*

Correspondence should be addressed to Yicheng Lin; yclin@wuyiu.edu.cn

Received 17 October 2021; Revised 10 December 2021; Accepted 16 December 2021; Published 26 February 2022

Academic Editor: Rahman Ali

Financial risk prediction is an important technique to systematically predict the unforeseeable risks in banking systems. The issues involving ill-timing and low accuracy in the current risks prediction methods necessitate an effective risk prediction method. Akin to the use of big data in various domains, the technology has a significant role in financial services and can be used to accurately and timely predict the possibilities of risks. In this paper, an effective hybrid method is proposed to aptly and effectively predict financial risks in the banking systems. The method utilizes the Lasso and linear regression algorithms via the big data features and framework technologies. By proper formalization of the bank financial risk problems, the risk data is obtained and processed. To filter the initial text features and preprocess the annual report text data, the information gain method is used. With the Bag-of-Words (BoW) and the word frequency reverse document frequency weighting method, the text features of financial risk prediction are extracted. The bank financial risk prediction model is constructed based on weighted fusion adaptive random subspace algorithm. The prediction results obtained are integrated so as to realize the bank financial risks in a seamless way. The experimental results show that the proposed method can effectively improve the prediction accuracy and consumes comparatively lesser time in risk prediction.

1. Introduction

As an important financial institution, banks have strong financial strength and diversified financial services. The safe operation of banks is of great significance to a country's economic security and healthy development [1]. On the surface, the bank is only an intermediary agency for money circulation, but in fact, the essence of the bank is to manage risks to obtain benefits. The focus of competition among peers is risk management ability, which can not only obtain high returns, but also reduce risks and be a mean to attract more customers. Financial risk prediction is the emerging research area to accurately and timely predict the risks involved in the banking. With the development of the global economy and the deepening of financial liberalization, possibility of breakout of financial crisis is higher. Moreover, finical data is becoming more vulnerable to destructiveness. Banks are high-risk industries; high-risk factors are always involved in the process of bank operation and management. The risk factors may in turn lead to financial crisis with a wide range of influence [2]. Especially under the background of increasingly complex financial ecological environment, the occurrence mechanism of financial crisis is more complex and destructive. Therefore, it is of great significance to study the bank financial risk prediction and establish an effective model to accurately predict the bank financial risk levels. This will help in preventing and controlling the occurrence of the financial crisis and/or reducing the losses caused by the financial crisis [3].

At present, scholars in related fields have studied financial risk prediction and achieved some theoretical results. Pawiak et al. [4] proposed a credit score prediction method based on learners' deep genetic hierarchy network. Credit scoring is an effective and key method used by banks and other financial institutions for risk management. It provides appropriate guidance for issuing loans and reduces the risk in the financial field. Using deep genetic learner level network to improve credit score risk prediction, combined with support vector machine, probabilistic neural network, and fuzzy system, credit scoring risk prediction is realized.

This method is effective, and the prediction performance of credit score data set is the best. Niu et al. [5] proposed a resampling integrated evaluation method of P2P loan credit risk based on data distribution. The class imbalance problem is solved by using the undersampling method based on the distribution of most class data. In order to improve the classification performance of the resampling integration model based on data distribution, the basic classifier with good comprehensive performance on the verification set is used for classification prediction to realize the resampling integrated evaluation of P2P loan credit risk. This method has good prediction performance. However, the above methods still have the problems of low prediction accuracy, long time, and poor effect.

To solve the above problems, a bank financial risk prediction method based on big data is proposed. Lasso and linear regression algorithms are studied by using big data characteristics and framework related technologies. By defining the formalization of bank financial risk problems, obtain and process bank financial risk data. Using word bag model and word frequency reverse document frequency weighting method, the text features of financial risk prediction are extracted. The adaptive fusion method is then utilized to fuse the financial risk characteristics. Based on the weighted fusion adaptive stochastic subspace algorithm, the bank financial risk prediction model is constructed to realize the bank financial risk prediction. This method can effectively improve the accuracy of risk prediction within a shorter risk prediction time span.

The rest of the paper is arranged into 4 sections. The technology of big data is elaborated in Section 2. Relevant theories about the bank financial risks are discussed in Section 3. The multisource heterogeneous data based financial risk prediction method is presented in Section 4. The last section, Section 5, is about conclusion and future work.

2. Big Data Technology

The buzz-word big data refers to the use of software utility to extract information from a large and complex data set through analyses and statistical measures. The technology of big data is to mine structured and/or unstructured data to obtain meaningful information and to generate machine learning models.

2.1. Big Data Concept. Big data refers to a data set that cannot be captured, managed, and processed by conventional software tools within a certain time range. It is a massive, high growth rate and diversified information asset that requires a new processing mode to have stronger decision-making power, insight and discovery power, and process optimization ability [6]. The big data industry takes data as the core. By collecting, storing, processing, analyzing, and applying the generated data and displaying it to users, the data processing efficiency is high and the cycle is short. The data processing technology contained in big data makes the bank financial risk prediction more scientific.

2.2. Big Data Features. Big data is not simply a huge amount of data but has its unique 4 V characteristics. The industry represented by IDC generally believes that big data has the characteristics of scale (Volume), diversity (Variety), high speed (Velocity), and value (Value). Big data 4 V characteristics are as Figure 1.

2.2.1. Large Data Scale. The huge order of magnitude is the basic attribute of big data. With the wide use and development of Internet technology, the number of Internet users is increasing rapidly. The acquisition and sharing of data information is becoming simple. At present, through a computer or a mobile phone, people can quickly and easily obtain a large amount of information. In addition, the sharing, clicking, browsing, and trading behaviors of network users on the Internet will produce a large amount of data. The quantity level of big data has jumped from TB level to the level of PB. The bank has the attribute of natural big data. Its huge financial transaction data is a natural data pool. The bank can easily understand the revenue and expenditure, deposits, and capital operation of customers.

2.2.2. Categories of Big Data. There are various types of big data and a wide range of sources. For the banking systems, the traditional enterprise financial database can no longer meet the needs of banks. In addition to the customer service, audio, network video, and online banking transaction records are retained by the banks. The bank can also obtain more data from website log data, enterprise ERP system, GPS global positioning system, e-commerce transaction records, government management department information platform, and other channels. Data types include not only traditional relational data types, but also unprocessed, semistructured, and unstructured information.

2.2.3. Fast Processing Speed. The faster frequency of data generation and update is also an important feature of big data. There is a saying about data processing in the era of big data, which is called the one-second law. Take online financial transactions as an example. On the trading platform, a large amount of financial transaction data, logistics, and transportation data are generated with every passing second. The data is generated and transmitted continuously; therefore larger storage and faster data processing tools are required.

2.2.4. Low Data Value Density. While the amount of data increases exponentially, the useful information hidden behind the data does not show the due growth proportion. Moreover, it is becoming increasingly difficult to obtain useful information. For banks, how to find useful information from a large amount of enterprise information is a problem. Because banks have strong financial strength, they can seek cooperation with professional data providers. At present, data providers represented by professional financial data service providers such as ninth power, IBM, and Intel provide banks with financial big data

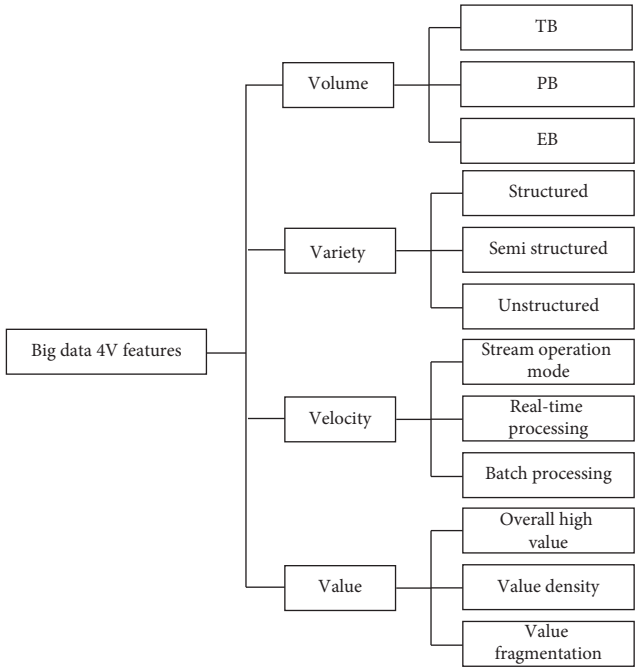

Figure 1: The big data 4 V features.

collection, analysis, and mining services to help banks mine data value.

2.3. Big Data Framework-Related Technologies. The frameworks of big data refer to the systematic expression of datasets so as to overcome the possible barriers in extracting information from data. The frameworks become necessary in such situations where the datasets are enormous and clumsy that meaning and/or information cannot be easily deduced from the data. Followings are some of the big data frameworks.

2.3.1. The HDFS File System. Hadoop distributed framework is the next mainstream big data processing framework, which is mainly used to process big data. The data level that Hadoop can handle is PB, which allows programs to perform distributed operations over thousands of nodes [7]. Hadoop has two core modules: (1) Hadoop Distributed File System (HDFS) and (2) the computing MapReduce framework. Among them, HDFS is a distributed file system that can be used on general hardware devices whereas MapReduce is used to realize distributed parallel computing. The principle structure of HDFS distributed file system is as Figure 2.

HDFS is a master-slave architecture. An HDFS cluster is composed of a *named* node and several data nodes. Usually the architecture consists of one node and one machine (data node). The machine manages the storage of the corresponding nodes. The named node is used to manage namespaces and tuning requests. The *data node* is mainly used for data storage. HDFS opens file namespaces to the public and allows user data to be stored as files. The

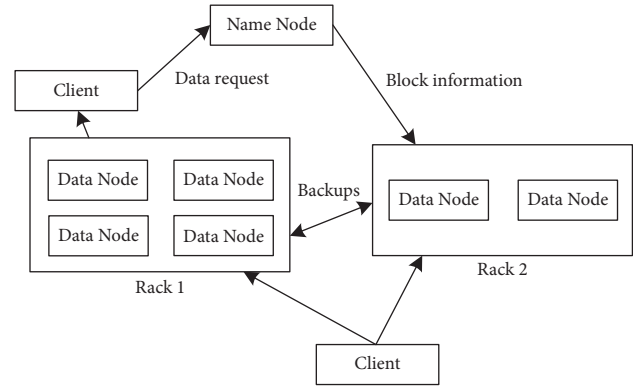

Figure 2: Principle structure of the HDFS file system.

data node also executes block creation, deletion, and block copy instructions from the name node.

2.3.2. Spark Distributed Computing Framework. The *spark* distributed computing architecture is currently the most popular big data computing framework. Compared with Hadoop's MapReduce framework, the *spark* is based on memory to do calculations, so the calculation performance is much better than MapReduce. The spark distributed computing framework is as Figure 3.

The main modules included in the spark framework are Spark-SQL data processing module, spark streaming data processing module, MLlib algorithm library module encapsulating mainstream machine learning algorithms, and the GraphX graph-based computing module [8]. Spark-SQL module is mainly used in data analysis, extraction, and index summary. The spark streaming is usually used for log analysis together with open source Kafka and Flume of Hadoop ecosystem. MLlib provides mainstream classification, clustering, and recommendation algorithms of machine learning, which is convenient for data science and technology to use spark for data mining.

2.4. Machine Learning and Statistics Related Algorithms. Machine learning algorithms are the dedicated programs that automatically learn from data and improve its performance with experience. Normal algorithms need program and data to produce output whereas the machine learning algorithm generates programs by taking output and data to operate without human intervention. Followings are the machine learning algorithms used in the domain of risk prediction.

2.4.1. Lasso Algorithm. In statistics and machine learning, the *Lasso* algorithm is a regression analysis method of simultaneous feature selection and regularization. The algorithm aims to improve the prediction accuracy and interpretability of statistical model [9]. By forcing the sum of absolute values of regression coefficients to be less than a fixed threshold, some regression coefficients are forced to become zero. The variables corresponding to these regression coefficients are effectively selected, so as to build a

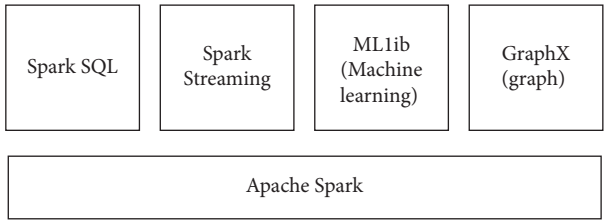

Figure 3: The spark distributed computing framework.

simpler model. The L_1 penalty term is added to the ordinary linear model. For ordinary linear regression, the Lasso estimate is

$$\begin{aligned} \widehat{\beta}_{\text{lasso}} &= \arg\min_{\beta \in R^d} Y - X\beta^2, \\ s.t. \sum_{j=1}^{d} \left|\beta_j\right| &\le t, t > 0. \end{aligned} \tag{1}$$

In formula (1), t and j correspond one-to-one, which is the adjustment coefficient.

It is equivalent to

$$\widehat{\beta}_{\text{lasso}} = \arg\min_{\beta \in R^d} Y - X\beta^2 + \lambda \sum_{j=1}^{d} \left|\beta_j\right|. \tag{2}$$

Order:

$$t_0 = \sum_{j=1}^{d} \left|\widehat{\beta}_j(\text{OLS})\right|. \tag{3}$$

In formula (3), OLS is estimated by the least square method. When $t < t_0$, when a part of the coefficient is compressed to a value of 0, the dimension of X is reduced to achieve the purpose of dimensionality reduction.

2.4.2. Linear Regression. The basic idea of linear regression method is to characterize the input data as a linear model and estimate and solve the parameters of the model by using the least square method under the principle of minimizing the mean square error [10]. Suppose the input data set is D where D has d features and m samples, and x_i is the i sample. At this time, the multiple linear regression model is described as follows:

$$\begin{aligned} X &= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \\ y &= (y_1, y_2, \ldots, y_m)^T, \\ f(x_i) &= w^T x_i + b_i, \\ (w^*, b^*) &= \arg\min_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2 \end{aligned} \tag{4}$$

When $\mathbf{X}^T\mathbf{X}$ full rank matrix or positive definite matrix, the weight parameter of the feature can be obtained as

$$\mathbf{w}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{5}$$

When $\mathbf{X}^T\mathbf{X}$ is not full of rank matrix or positive definite matrix, the optimal solution obtained by parameter estimation is not unique at this time, and the variance of the model can be reduced by adding regular constraints.

3. Relevant Theories of Bank Financial Risk

Financial risk management is very important area in banking. Risk management in the realm of banking intends to systemically model possibilities of issues which in the long run may affect financial marketing and/or financial tweets.

3.1. Financial Risk Concept. The general definition of financial risk is the possibility of losses to financiers in the process of financial service transactions. It may also refer to forecasting whether the actual income is lower than the expected income, or the actual cost is higher than the expected cost [11]. From the perspective of the operation of financial institutions, this paper defines financial risk as banks are likely to suffer losses under the influence of various uncertain factors in the process of financial activities such as fund-raising and utilization. This shows that the actual income is lower than the operating cost.

3.2. Financial Risk Characteristics. The characteristics of financial risk are divided into five categories, including objectivity, uncertainty, latency, controllability, and periodicity. Details of the characteristics are given as follows.

3.2.1. Objectivity. Financial risk is accompanied by financial activities. As long as there are financial activities, there must be relevant risks. Moreover, with the continuous innovation of derivative financial instruments, it not only promotes financial development, but also brings new risks. Moreover, the occurrence of financial risks in a financial institution will inevitably affect its creditors and may further affect all aspects of economic operation.

3.2.2. Uncertainty. Financial institutions conduct business or decision-making activities in an uncertain environment; that is, the operating environment of financial business activities is constantly developing and changing, while it is difficult for the actors to accurately predict the future, and financial risks may arise at any time.

3.2.3. Latency. Financial risk is often manifested as the outbreak of financial crisis. In fact, financial activities may cover up some uncertain losses due to their own characteristics.

3.2.4. Controllability. Although uncertain changes in the economic situation may bring risks, the risks can be effectively controlled as long as targeted measures are taken.

3.2.5. Periodicity. For each financial institution, it operates in the established financial ecological environment, and the financial environment is affected by the whole economic environment. Therefore, when the periodic fluctuation of economy and the orderly change of monetary policy appear, it is easy to identify cyclical financial risks, which makes the monitoring of financial risks possible.

3.3. Financial Risk Classification. According to the scope of occurrence and influence of financial risk, this paper divides the risks into systematic financial risk and nonsystematic financial risk. Details of the risks are given in the following subsections.

3.3.1. Systemic Financial Risk. The systematic financial risk refers to the overall risk of the market including the impact of economic, political, social, and other environmental factors in the financial ecological environment on the whole market. Changes in external environmental factors may lead to financial crises in some banks and chain crises in the whole financial system. Therefore, only through a reasonable evaluation of the macroeconomic situation in a certain period of time can we identify the systemic financial risks faced by a country or region.

3.3.2. Nonsystematic Financial Risk. Nonsystematic risks refer to the possible loss caused by individual financial institutions in the financial industry. In the process of financial activities, these are the risks which are considered to be the decentralized risk. Nonsystematic financial risks can be reduced or even eliminated by improving bank management and asset allocation.

4. Bank Financial Risk Prediction Method Integrating Multisource Heterogeneous Data

This research work focuses on the bank financial risks intended to construct a multisource heterogeneous features set. The research proposes a banking financial risk prediction method that integrates multisource heterogeneous data.

4.1. Formal Definition of the Problem. In order to express the proposed method clearly, a formal definition should be made before introducing the specific method. Assuming that there are n samples in a given data set D, the data set is defined as $D = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}^T$, where $x_i \in R^n$ and the category label are $y_i \in \{-1, 1\}$. Suppose the number of features is p; then the feature space vector is $\mathbf{X} = (x_1^{(1)}, \ldots, x_{p_1}^{(1)}, \ldots, x_1^{(j)}, \ldots, x_{p_j}^{(1)}, \ldots, x_1^{(J)}, \ldots, x_{p_J}^{(J)})$, and J represents the number of different data sources. p_j is the number of features extracted from the j_{th} data source, $\mathbf{W} = (w_1, w_2, \ldots, w_p)^T \in R_+^p$ is the weight vector, and $|\cdot|$ represents the L_1 norm. For the linear regression model, the hypothesis is $y_i = \sum_{i=1}^{J} x_{ij}^T \beta_j + e_i$, where $\beta_j = (\beta_1^{(j)}, \ldots, \beta_{p_j}^{(j)}) \in R^{p_j}$ is the regression coefficient. Let the residual term e_i be an independent and identically distributed random variable and follow a normal distribution with a mean of 0 and a variance of σ^2. To this end, all feature vectors are normalized and centralized, that is, $\sum_{i=1}^{n} x_{ij} = 0$, $\|x_j\|^2 = 1$.

4.2. Data Acquisition and Processing. Bank financial risk prediction information can be divided into financial information and nonfinancial information. The information can generate quantitative financial characteristics and nonfinancial characteristics based on qualitative description. Among them, the financial features can be calculated and extracted by using the accounting information in the financial statements regularly issued by the bank. The nonfinancial features can be extracted by using the disclosure data in the form of financial reports, news, and other text related to the bank. Generally speaking, the information is regularly published on the network platform and is easy to obtain. The bank financial risk prediction data set collected and captured in this study will be described in detail in the next experimental design section. In addition, financial data can be transformed into structured data after simple processing, which can be directly used as the input of learning algorithm. The nonfinancial data in text form can be used for learning only after word segmentation, cleaning, filtering, and other natural language processing techniques.

4.3. Feature Extraction of Financial Risk Prediction. Firstly, the collected annual report text data are preprocessed, and then unigrams, bigrams, and trigrams are extracted as text features by using word bag model and word frequency reverse document frequency (TF-IDF) weighting method. Because text features naturally face high-dimensional problems, high-dimensional text features may contain some redundant and irrelevant features [12]. Therefore, the information gain method is further used to filter the extracted initial text features, and the important features are retained to ensure the quality of the features. The calculation process of the information gain $IG(Y, F)$ is as follows:

$$IG(Y, F) = H(Y) - H(Y|F), \tag{6}$$

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y), \tag{7}$$

$$H\left(\frac{Y}{F}\right) = -\sum_{f \in F} p(f) \sum_{y \in Y} p\left(\frac{y}{f}\right) \log_2 p\left(\frac{y}{f}\right). \tag{8}$$

In formulas (6)–(8), $IG(Y, F)$ represents that when feature F is added, the information entropy of category Y decreases, $H(Y)$ represents the information entropy of category, $p(y)$ represents the probability of category y, and $H(Y|F)$ represents category under the condition of feature F. The information entropy of Y, A, represents the probability of $p(y|f)$ certain category distribution under a single feature condition. In the process of filtering text features, all unigrams, bigrams, and trigrams with an information gain

greater than 0.0025 are retained as important text features. In order to fully explore the role of different characteristics in bank financial risk prediction, the above characteristics are fully combined, and the combined characteristics are expressed as

$$F = F1 + F2 + F3. \tag{9}$$

In formula (9), $F1$ represents the set of extracted financial features, $F2$ represents the set of emotional features, and $F3$ represents the set of text features.

4.4. Construction of Financial Risk Prediction Model. Considering the demand for adaptive fusion of multisource data in bank financial risk prediction and comprehensively considering the advantages of the above random subspace method, adaptive Lasso method, and weighted fusion Lasso method for the prediction problem [13], this study proposes a financial risk prediction method based on weighted fusion adaptive random subspace. This method includes three main modules: firstly, the constructed adaptive fusion method is used to fuse the features, secondly, the base classifier is constructed, and finally, the learning results of the base classifier are integrated. The flow of financial risk prediction method based on weighted fusion adaptive random subspace is as Figure 4.

The goal of the financial risk prediction method based on weighted fusion adaptive random subspace in the first stage is to perform feature adaptive fusion to obtain the sampling weight $\mathbf{W} = (w_1, w_2, \ldots, w_p)^T \in R_+^p$ of the feature. To this end, first consider the classic Lasso model, which has the following form:

$$\beta^* = \arg\min_{\beta} \left\{ \frac{1}{2} \left\| y - \sum_{i=1}^{p} x_i \beta_i \right\|_2^2 + \lambda |\beta_i| \right\}. \tag{10}$$

In formula (10), λ represents the regular penalty parameter. After the weighted fusion adaptive estimation is performed on the features, a weight vector corresponding to each feature composed of regression coefficients will be obtained. Features with a weight of 0 will not be adopted. On the contrary, the greater the weight, the greater the probability of the feature being selected. When fusing multisource data, it is necessary to consider the impact of the relationship between different features on the prediction results. Therefore, the weighted fusion Lasso model is introduced on the basis of the Lasso model, and its form is as follows:

$$\beta^* = \arg\min_{\beta} \left\{ \frac{1}{2} \left\| y - \sum_{i=1}^{p} x_i \beta_i \right\|_2^2 + \lambda |\beta_i| + \frac{\lambda_2}{p} \sum_{i<j}^{p} a_{ij} \left(\beta_i - s_{ij} \beta_j \right)^2 \right\}. \tag{11}$$

In formula (11), $\lambda_2 / p \sum_{i<j}^{p} a_{ij} (\beta_i - s_{ij}\beta_j)^2$ is the penalty term, and $a_{ij} = \rho_{ij} / 1 - \rho_{ij}$, $s_{ij} = \operatorname{sgn}(\rho_{ij}) = \{+1, \rho_{ij} > 0 / -1, \rho_{ij} > 0$ and ρ_{ij} are the correlation coefficients between any two features x_i and x_j. Through the weighted fusion Lasso model, related features can be screened out or retained at the same time, which effectively solves the problem of multiple collinearities between features and improves the stability of the model. In order to be able to adaptively fuse different features, this research comprehensively considers Lasso, weighted fusion Lasso model and adaptive Lasso, and other methods and proposes a new regularized sparse model weighted fusion adaptive Lasso; its form is as follows:

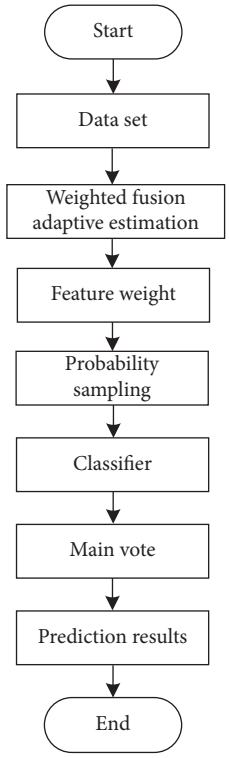

Figure 4: Flowchart of risk prediction method based on weighted fusion adaptive random subspace.

$$\beta^* = \arg\min_{\beta} \left\{ \frac{1}{2} \left\| y - \sum_{i=1}^{p} x_i \beta_i \right\|_2^2 + \lambda w_i^{(1)} |\beta_i| + \frac{\lambda_2}{p} \sum_{i<j}^{p} a_{ij} \left(\beta_i - s_{ij} \beta_j \right)^2 \right\}. \tag{12}$$

In formula (12), $w_i^{(1)} = 1/|(\beta_{ilasso} + 1/\sqrt{n})|$ is the adaptive weight. That is, before performing weighted fusion adaptive Lasso estimation, first perform Lasso estimation to obtain a set of regression coefficient vectors, and add its inverse as the adaptive weight of the feature to the weighted fusion adaptive Lasso. In this way, different features can be penalized according to their importance, and the model becomes an unbiased estimation, and a more accurate feature subset can be obtained [14].

Through the weighted fusion adaptive Lasso estimation, the adaptive feature weights $\mathbf{W} = (w_1, w_2, \ldots, w_p)^T \in R_+^p$ based on weighted fusion can be obtained. After using these weights to perform probability sampling on the features, the data subset $\{D_{\text{sub}}^1, D_{\text{sub}}^2, \ldots, D_{\text{sub}}^M\}$, $D_{\text{sub}}^i = \{(x_1^i, y_1^i), \ldots, (x_j^i, y_j^i), \ldots, (x_{p_i}^i, y_{p_i}^i)\}$ used for the training of the base classifier can be obtained. The sampling process is adjusted by the subspace ratio parameter r. The larger the r,

the higher the characteristic dimension of the sample subset.

In the second stage, the financial risk prediction method based on weighted fusion adaptive random subspace first determines the base classifier and then uses the data subset obtained in the first stage to train the base classifier. When the training samples are linearly separable, the representation of the hyperplane in the sample space is as follows:

$$\mathbf{w}^T\mathbf{x} + b = 0. \tag{13}$$

In formula (13), the normal vector $\mathbf{w}= [w_1, w_2, \ldots, w_d]$ and the displacement b, respectively, determine the direction of the hyperplane and its distance from the origin. At this time, the distance from any sample point $\mathbf{x}_i$ to the hyperplane is

$$r = \frac{\left|\mathbf{w}^T\mathbf{x}_i + b\right|}{\|\mathbf{w}\|}. \tag{14}$$

If the hyperplane $(\mathbf{w}, b)$ correctly classifies the sample $(\mathbf{w}_i, y_i) \in D$, there are

$$\begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq +1, y_i = +1 \\ \mathbf{w}^T\mathbf{x}_i + b \leq -1, y_i = -1 \end{cases}. \tag{15}$$

In formula (15), the sample points that can make the equation hold are support vectors. From a geometric point of view, the support vector is the sample points on the two classification boundaries $\mathbf{w}^T\mathbf{x}_i + b = 1$ and $\mathbf{w}^T\mathbf{x}_i + b = -1$. The classification boundary is only related to these support vectors. The sum of the distances from the support vector to the hyperplane is

$$\gamma = \frac{2}{\|\mathbf{w}\|}. \tag{16}$$

SVM can effectively deal with learning tasks with fewer samples, high feature dimensions, and nonlinear relationships between features [15]. Therefore, in the face of high-dimensional text data, this research chooses SVM as the base classifier of the financial risk prediction method based on weighted fusion adaptive random subspace.

The financial risk prediction method based on weighted fusion adaptive random subspace adopts the main voting strategy to synthesize the learning results of the base classifier in the third stage. Assuming that the category distribution is $\{c_1, c_2, \ldots, c_N\}$ and the output of the classifier h_i on the sample $\mathbf{x}$ is $\{h_i^1(\mathbf{x}), h_i^2(\mathbf{x}), ..., h_i^N(\mathbf{x})\}$, the main voting or majority voting method is expressed as follows:

$$H(\mathbf{x}) = \begin{cases} c_j, if \sum_{i=1}^{M} h_i^j(\mathbf{x}) > 0.5 \sum_{k=1}^{N}\sum_{i=1}^{M} h_i^k(\mathbf{x}), \\ \text{null, otherwise.} \end{cases} \tag{17}$$

According to formula (17), it can be seen that when a certain category label gets more than half of the votes, the main voting method uses it as the final output label. Corresponding to the main voting method is the relative majority voting method. The calculation process is as follows:

$$H(\mathbf{x}) = c_j^* = \arg\max_j \sum_{i=1}^{M} h_i^j(\mathbf{x}). \tag{18}$$

In the given formula (equation (18)), the category with the highest votes will be used as the final output category to obtain the final integrated prediction result. Through the above steps, the bank financial risk prediction is realized.

5. Experimental Analysis

To properly evaluate the proposed method experimentation was conducted based on real data obtained from the commercial banks. Details of the evaluation procedure along with the comparison of some state-of-the-art methods are presented in the following subsections.

5.1. Experimental Environment and Data. In order to verify the effectiveness of the banking financial risk prediction method based on big data, the experiment used the *spark cluster* as the experimental environment and adopted the operation mode of *spark on yarn*. In this study, 26 commercial banks were selected as experimental samples, and ST markers were used as a sign that banks were in financial risk, and 871 normal samples and 129 risk samples were obtained. From a feature point of view, the experimental data set consists of 39 financial features, 12 emotional features, and qualitative text features. For the extraction of sentiment words, the CNKI sentiment dictionary and the legal-related Sogou sentiment dictionary were used. The availed vocabularies contained various possible sentiments like the positive and negative sentiment, strong and weak modal sentiment, and the uncertain sentiment.

5.2. Risk Prediction and Evaluation Indicators. This article uses average accuracy rate, error rate, and prediction time as evaluation indicators. The average accuracy rate refers to the ratio of the number of correctly predicted samples to the total number of predicted samples. The greater the average accuracy rate, the higher the prediction accuracy. The calculation formula is

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}. \tag{19}$$

In the given formula (equation (19)), TP represents a true case, TN represents a true negative case, FP represents a false positive case, and FN represents a false negative case. The error rate refers to the ratio of the number of samples with prediction errors to the total number of samples. The smaller the error rate, the better the prediction effect. The calculation formula is

$$E = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} + \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}. \tag{20}$$

5.3. Comparison of the Accuracy of Bank Financial Risk Prediction. In order to verify the prediction accuracy of the proposed method, the methods of [4, 5] are compared with

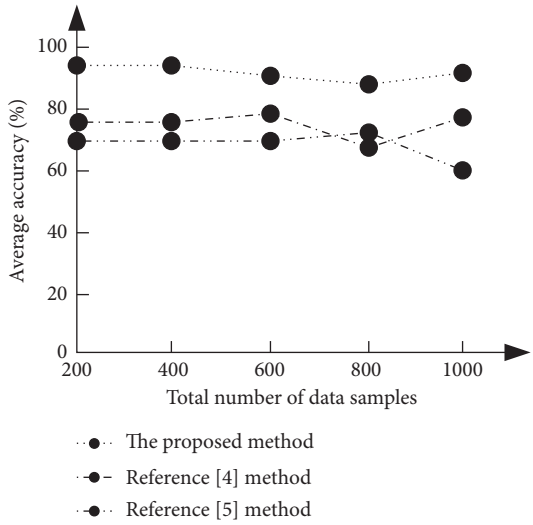

Figure 5: Comparison results of average accuracy results of different methods.

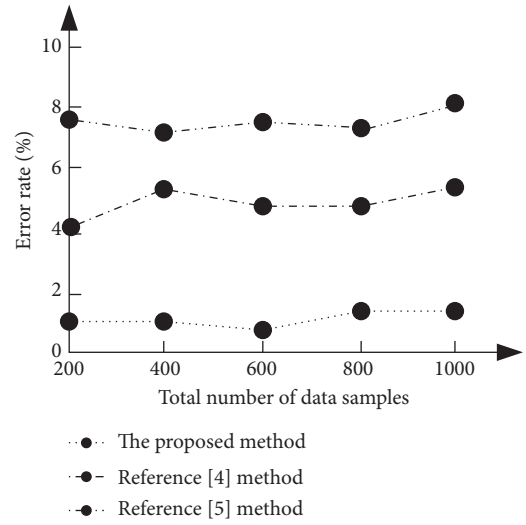

Figure 6: Comparative analysis of the bank financial risk prediction error rate.

the proposed method, respectively. The average accuracy of different methods is obtained and depicted in Figure 5.

It can be seen from Figure 5 that, under different total data samples, the average accuracy of the method in [4] is 75%, the average accuracy of the method in [5] is 73%, and the average accuracy of the proposed method is 92%. Therefore, compared with the methods of Pawiak et al. [4] and Niu et al. [5], the average accuracy of the proposed method is higher, and its bank financial risk prediction accuracy is higher.

5.4. Comparison of Bank Financial Risk Prediction Results. To further verify the prediction effect of the proposed method, the method is compared with that of the Pawiak et al. [4] and Niu et al. [5]. The comparison results about the bank financial risk prediction error rate of different methods are as Figure 6.

It is clear from Figure 6 that, under the total number of different data samples, the average error rate of bank financial risk prediction in [4] method is 4.4%. The average error rate of bank financial risk prediction in [5] method is 7.8%. The average error rate of the bank financial risk prediction by our proposed method is only 1.1%. It can be seen that, compared with the methods of Pawiak et al. [4] and Niu et al. [5], the average error rate of the bank financial risk prediction of the proposed method is smaller. Hence, the bank financial risk prediction of the proposed method is better.

5.5. Comparison of Bank Financial Risk Prediction Time. On this basis, the prediction time of the proposed method is verified. The methods of [4, 5] and the proposed method were compared in terms of risk prediction time. The comparison results of bank financial risk prediction time of different methods are shown in Table 1.

According to the data in Table 1, as the total number of data samples increases, the bank financial risk prediction time of different methods increases. When the total number of data samples is 1000, the bank financial risk prediction time of the method of [4] is 22.9 s, the bank financial risk prediction time of the method of [5] is 31.5 s, and the bank financial risk prediction time of the proposed method is only 13.3 s. It can be seen that, compared with the method of [4] and the method of [5], the bank financial risk prediction time of the proposed method is shorter.

Table 1: Comparison results of bank financial risk prediction time with different methods.

Total number of data samples	The proposed method (s)	The method of [4] (s)	The method of [5] (s)
200	3.34	5.98	8.76
400	5.18	8.87	12.8
600	8.97	12.7	19.6
800	10.2	17.8	26.9
1000	13.3	22.9	31.5

6. Conclusion

The bank financial risk prediction method based on big data is proposed in this paper. The method intends to make the full use of big data technology. The bank financial risk prediction accuracy of the proposed method is high. Moreover, the method can effectively shorten the bank financial risk prediction time and has good risk prediction effect. However, in the process of bank financial risk prediction, due to the limitation of data acquisition channels, this study has not considered the prediction effect of other feasible and useful data sources. Therefore, in the next research, we have planned to further expand the multisource information and collect the bank financial risk data in real time. This will help to verify the effect of the bank financial risk prediction model. Besides, the model will be augmented to make the prediction results more accurate.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Flori, S. Giansante, C. Girardone, and F. Pammolli, "Banks' business strategies on the edge of distress," *Annals of Operations Research*, vol. 299, no. 1, pp. 481–530, 2021.

[2] M. Umar, X. Ji, N. Mirza, and B. Naqvi, "Carbon neutrality, bank lending, and credit risk: evidence from the eurozone," *Journal of Environmental Management*, vol. 296, p. 113156, 2021.

[3] C. Clab, A. Asr, and D. Teca, "Catastrophic expenditures in california trauma patients after the affordable care act: reduced financial risk and racial disparities - sciencedirect," *The American Journal of Surgery*, vol. 220, no. 3, pp. 511–517, 2020.

[4] P. Pawiak, M. Abdar, J. Pawiak, V. Makarenkov, and U. R. Acharya, "DGHNL: a new deep genetic hierarchical network of learners for prediction of credit scoring," *Information Sciences*, vol. 516, no. 2020, pp. 401–418, 2020.

[5] K. Niu, Z. Zhang, Y. Liu, and R. Li, "Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending," *Information Sciences*, vol. 536, pp. 120–134, 2020.

[6] A. Wibisono and D. Sarwinda, "Average restrain divider of evaluation value (ARDEV) in data stream algorithm for big data prediction," *Knowledge-Based Systems*, vol. 176, no. 15, pp. 29–39, 2019.

[7] M. T. Wu, G. Srivastava, M. Wei, U. Yun, and C. W. Lin, "Fuzzy high-utility pattern mining in parallel and distributed hadoop framework," *Information Sciences*, vol. 553, pp. 31–48, 2020.

[8] S. Kang, S. Lee, and J. Kim, "Distributed graph cube generation using Spark framework," *The Journal of Supercomputing*, vol. 76, no. 10, pp. 8118–8139, 2019.

[9] Y. Wen and Q. Lu, "Multikernel linear mixed model with adaptive lasso for complex phenotype prediction," *Statistics in Medicine*, vol. 39, no. 9, pp. 1311–1327, 2020.

[10] G. Goh and D. K. Dey, "Asymptotic properties of marginal least-square estimator for ultrahigh-dimensional linear regression models with correlated errors," *The American Statistician*, vol. 73, no. 1, pp. 4–9, 2019.

[11] A. L. Hamilton, G. W. Characklis, and P. M. Reed, "Managing financial risk trade-offs for hydropower generation using snowpack-based index contracts," *Water Resources Research*, vol. 56, no. 10, Article ID e2020WR027212, 2020.

[12] S. Salesi, G. Cosma, and M. Mavrovouniotis, "TAGA: tabu asexual genetic algorithm embedded in a filter/filter feature selection approach for high-dimensional data," *Information Sciences*, vol. 565, pp. 105–127, 2021.

[13] S.-B. Chen, Y.-M. Zhang, C. H. Q. Ding, J. Zhang, and B. Luo, "Extended adaptive Lasso for multi-class and multi-label feature selection," *Knowledge-Based Systems*, vol. 173, no. 1, pp. 28–36, 2019.

[14] N. Qiu, P. Gao, P. Wang, and Y. Tao, "Research on ACO-WNB classification algorithm based on improved information gain," *Computer Simulation*, vol. 36, no. 1, pp. 295–299, 2019.

[15] R. Touati, A. E. Oueslati, I. Messaoudi, and Z. Lachiri, "The Helitron family classification using SVM based on Fourier transform features applied on an unbalanced dataset," *Medical, & Biological Engineering & Computing*, vol. 57, no. 10, pp. 2289–2304, 2019.