Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9825081, 1 page
https://doi.org/10.1155/2023/9825081

Retraction

Retracted: The Application of the Data Mining Method Based on Ensemble Learning to the Research on Early Warning of Financial Crisis in Small- and Medium-Sized Enterprises

Wireless Communications and Mobile Computing

Received 29 August 2023; Accepted 29 August 2023; Published 30 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Guo and W. Wang, "The Application of the Data Mining Method Based on Ensemble Learning to the Research on Early Warning of Financial Crisis in Small- and Medium-Sized Enterprises," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 8443988, 11 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 8443988, 11 pages
https://doi.org/10.1155/2022/8443988

Research Article

The Application of the Data Mining Method Based on Ensemble Learning to the Research on Early Warning of Financial Crisis in Small- and Medium-Sized Enterprises

Jie Guo and Wenyan Wang

School of Economics and Management, Cangzhou Normal University, Cangzhou Hebei 061000, China

Correspondence should be addressed to Jie Guo; guojie0610@caztc.edu.cn

Received 17 May 2022; Revised 23 June 2022; Accepted 28 June 2022; Published 19 July 2022

Academic Editor: Kalidoss Rajakani

In the big data environment, the factors affecting the operation of small- and medium-sized enterprises are becoming more and more complex. In order to more accurately measure the financial crisis of small- and medium-sized enterprises, from the perspective of multisource information fusion, based on the traditional financial data reflecting the private information of enterprises, the public information data for measuring the macroeconomic and market environment are integrated and analyzed. Considering the multidimensional heterogeneity of multisource information data, BPNN, SVM, KNN, LOG, and MDA are introduced to build models, and ensemble learning is used to integrate the results of different early warning models to reduce the risk of inconsistent results from different models. The empirical results show that the early warning model integrating multisource information can improve the early warning accuracy. The research results show that the financial crisis early warning of small- and medium-sized enterprises should not only pay attention to financial indicators but also pay attention to the impact of corporate governance, industry, and the overall economic environment of the country; the accuracy of the model in this paper is significantly higher than that of the other models in terms of the accuracy of financial crisis early warning. It shows that the model in this paper can better predict the real financial status of small- and medium-sized enterprises.

1. Introduction

With the development of the market economy and global integration, the total number of small- and medium-sized enterprises in China has exceeded 99%. At the same time, they contribute 60% of GDP and 50% of tax revenue and provide 80% of employment opportunities. They have become the main force of market development. Therefore, it is very important to construct a scientific and reasonable financial crisis early warning model for small- and medium-sized enterprises in China [1]. However, SMEs have significant heterogeneity compared with large-scale enterprises such as manufacturing and real estate, which are currently more concerned by scholars. The following are some discussions on SMEs: poor economic strength and lack of adaptability to changes in the external environment [2]. With the increasingly fierce market competition, its characteristics of being easily affected by external macroeconomics and market environment have become more prominent. At the same time, the era of big data has changed the internal and external environment of enterprises. Collecting valuable data for effective analysis can provide more scientific support for management decisions [3]. Therefore, the analysis of the operating conditions of small- and medium-sized enterprises not only should rely on financial indicators reflecting the private information of enterprises but should also consider public information that can fully reflect the macroeconomic conditions and market

conditions. However, the existing research on financial crisis early warning of small- and medium-sized enterprises pays more attention to the private information of enterprises, ignoring the impact of external public information on enterprise operation, which is not suitable for the research on financial early warning of small- and medium-sized enterprises in the new environment [4]. Second, existing studies have shown that financial data reflecting private information of enterprises is lagging and gray, so financial data has limitations in crisis warning. Third, early warning research on small- and medium-sized enterprises needs to integrate multisource information, but multidimensional heterogeneous data brought by multisource information and complex nonlinear relationships in the data pose challenges to traditional statistical measurement methods; for example, multivariate discriminant analysis (MDA) and logistic regression (LOG) all require strict assumptions and cannot handle multidimensional heterogeneous data [5].

In recent years, with the development and maturity of computer science, artificial intelligence learning methods have gradually developed. Artificial intelligence methods not only avoid the limitations of traditional statistical methods with many assumptions but also can better deal with nonlinear complex relationships [6]. Among them, data mining methods such as backpropagation neural network (BPNN), *K*-nearest neighbor algorithm (KNN), Decision Tree (DT), and support vector machine (SVM) have achieved good results in financial early warning research. However, the disadvantage is that each method has different characteristics and applicable conditions (such as data structure), which will lead to inconsistent prediction results of different methods, which has certain limitations. Therefore, how to integrate the advantages of different data mining methods to build a structure capable of handling multisource heterogeneity is a field of interest [7].

In order to solve the above problems, based on the perspective of multisource information fusion, this paper introduces the integrated learning method to build a financial crisis early warning model for small- and medium-sized enterprises in China based on integrated learning. Data (macroeconomic information and market information) adopt five early warning algorithms that perform well in crisis early warning—BPNN, SVM, LOG, KNN, and MDA to build models, and introduce an integrated learning mechanism to analyze the results of different early warning models. Integrated processing combines the advantages of multiple algorithms to reduce the risk of inconsistent results from different early warning models. Empirical research on small- and medium-sized enterprises in China shows that the new early warning method can significantly improve the accuracy and stability of early warning [8].

2. State of the Art

From the perspective of indicator selection, the existing research on corporate financial crisis early warning mainly focuses on the private financial information of companies. For example, Lv et al. (2004) found that profitability, asset-liability ratio, and company size have a significant impact on companies in financial crisis [9]. Ling et al. (2005) proved through experiments that indicators such as profitability, liquidity repayment ability, operational ability, and business development ability provide a lot of useful information for predicting financial distress. However, some studies have pointed out that financial indicators have the characteristics of hysteresis and grayness, and it is impossible to comprehensively judge the operation status of enterprises only by using financial indicators. Later, scholars at home and abroad began to pay attention to nonfinancial information. Lehman and Topol [10] found that in the early warning model with a long time window, indicators such as solvency, industry risk, and payment behavior are important factors for predicting bankruptcy, and combining financial and nonfinancial variables can make the results more accurate. Zhang et al. [11] comprehensively select corporate financial variables, nonfinancial variables, and macrovariables to establish a financial crisis early warning model. The empirical results show that the accuracy of the early warning model containing three sets of variables is better than that of only one or two sets of variables.

Judging from the selection of early warning methods, the methods for constructing early warning models for corporate financial crisis at home and abroad are mainly divided into two categories: one is statistical methods, including MDA and LOG. However, the application of statistical methods requires basic assumptions. When there are many variables, it is difficult to solve problems such as multicollinearity and autocorrelation, so there are limitations in practical applications. Most empirical research results show that logistic regression has higher accuracy and better stability in statistical measurement methods [12]. The second is artificial intelligence learning methods, including DT, rough set, genetic algorithm, and SVM [13]. At present, the models that perform well in corporate financial crisis early warning include DT, ANN, and SVM. However, due to the different characteristics of different models and the differences in applicable conditions, the empirical results are unstable and different and thus have certain limitations [14].

To sum up, the existing research on financial early warning of small- and medium-sized enterprises pays more attention to the private financial information of enterprises, ignoring the impact of external public information on enterprise operation [15]. In addition, compared with traditional statistical measurement methods, the existing data mining models show certain advantages in dealing with enterprise financial crisis early warning, but the differences in their applicable conditions are prone to the risk of inconsistent results. Therefore, based on the perspective of multisource information fusion, this paper comprehensively considers macroeconomic indicators and market indicators based on public information on the traditional financial indicators based on private information of enterprises [16]. In addition, in order to reduce the risk of inconsistent results of different data mining models, this paper introduces an ensemble learning mechanism to perform "secondary" fusion processing on the results of different models to improve the accuracy and stability of the final prediction results [17].

3. Methodology

3.1. Financial Crisis Early Warning Model Based on Ensemble Learning. This paper studies the financial crisis early warning of small- and medium-sized enterprises in China. Firstly, on the traditional financial indicators based on private information of enterprises, macroeconomic indicators and market indicators based on public information are added; secondly, in order to better integrate the advantages of different early warning models and improve the reliability of the prediction results, this paper introduces an integrated learning mechanism to integrate the results of different models to optimize the final decision results. The model design route is shown in Figure 1 [18].

Based on the perspective of multisource information fusion, this paper comprehensively selects multisource indicators that reflect corporate private information and public information. The indicators of corporate private information are mainly key corporate financial indicators, which are divided into five categories: solvency, operational capacity, profitability, structural stability, and business development capabilities; the indicators of corporate public information are divided into two categories: indicators reflecting the macroeconomic environment and indicators reflecting market conditions. In view of the reference to the existing research indicators and from the statistical point of view, the types of indicators with serious data missing are deleted. The indicators finally selected in this paper are shown in Table 1. There are 44 types in total, including 31 types of private information indicators and 13 types of public information indicators [19].

3.2. Financial Crisis Early Warning Model

(1) BP neural network (BPNN) [20]. A team of scientists headed by Rumelhart and McCelland proposed the BPNN algorithm in 1986, which is a multilayer feedforward network trained by an error backpropagation algorithm. When BPNN is used to establish financial crisis early warning, it first divides the set composed of multidimensional indicators and label outputs of enterprises (crisis enterprise DF and healthy enterprise NF) to obtain a training set and test set. The training set is used to train the network and establish the BPNN financial crisis prediction model; then, input the test set and output the prediction results of the test set enterprises

(2) Support vector machine (SVM). Vladimir Vapink (1995) proposed the SVM algorithm, which is a supervised learning model and related learning algorithms for analyzing data in classification and regression analysis. SVM transforms the linearly inseparable samples in the low-dimensional input space into a high-dimensional feature space by using the nonlinear mapping algorithm to make them linearly separable and uses the linear algorithm to linearly analyze the nonlinear characteristics of the samples. In the SVM financial early warning research, given the training set, based on the structural risk minimization principle of the SVM algorithm, a hyperplane in the space that can classify companies is found. Then, input the test set, and the output result is the possibility of financial crisis of the enterprises in the test set

(3) K-nearest neighbor algorithm (KNN). It was first proposed by Fix and Hodges (1952) and Cover and Har (1967). It is a multivariate discrimination method based on nonparameters. For a new input sample, if the k most adjacent samples found in the training set belong to a certain category (such as DF), the sample also belongs to this category. If the value of K is small, it means that good results can be achieved only when the samples to be predicted and the samples to be trained are close. If the value of K is large, it means that the approximation error of algorithm classification increases. At this time, samples far away from the input samples will also have an effect on the results

(4) Logistic regression (LOG). Martin (1977) first proposed the use of logistic regression models to predict corporate default probability and bankruptcy. The logistic regression model establishes a linearly classifiable model by the method of maximum likelihood estimation and realizes the classification prediction of the dataset of binary variables or multivalue variables. If the probability of a company's financial crisis is set as P (the probability of no financial crisis is $1-P, P \in (0,1)$), the logit transformation of P is used as the dependent variable, and the logistic regression equation is established:

$$\operatorname{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n, \quad (1)$$

$$\text{Available}: P = \frac{1}{\left(1 + e^{-(\beta_t + \beta_1 + \beta_2 + \cdots + \cdots + \beta_3)}\right)}. \quad (2)$$

Definition 1. Use $(X_1, y_1), (X_2, y_2), \cdots, (X_N, y_N)$ to represent training data and labels, where X_i is a vector, y_i is a label, and y_i is equal to +1 or -1.

Definition 2. In the sample space of a linear model, the partitioning hyperplane can be represented by $\omega^T X + b = 0$, where ω and b are undetermined parameters, ω and X are vectors, and b is a constant.

Definition 3. A training set is linearly separable means that, for the training set $\{(X_i, y_i)\}_{i=1,2,\cdots,n}$, there exists (ω, b) such that for any $i = 1, 2, \cdots, N$, there is the following:

(a) If $y_i = +1$, then $\omega^T X + b \geq 0$

(b) If $y_i = -1$, then $\omega^T X + b < 0$

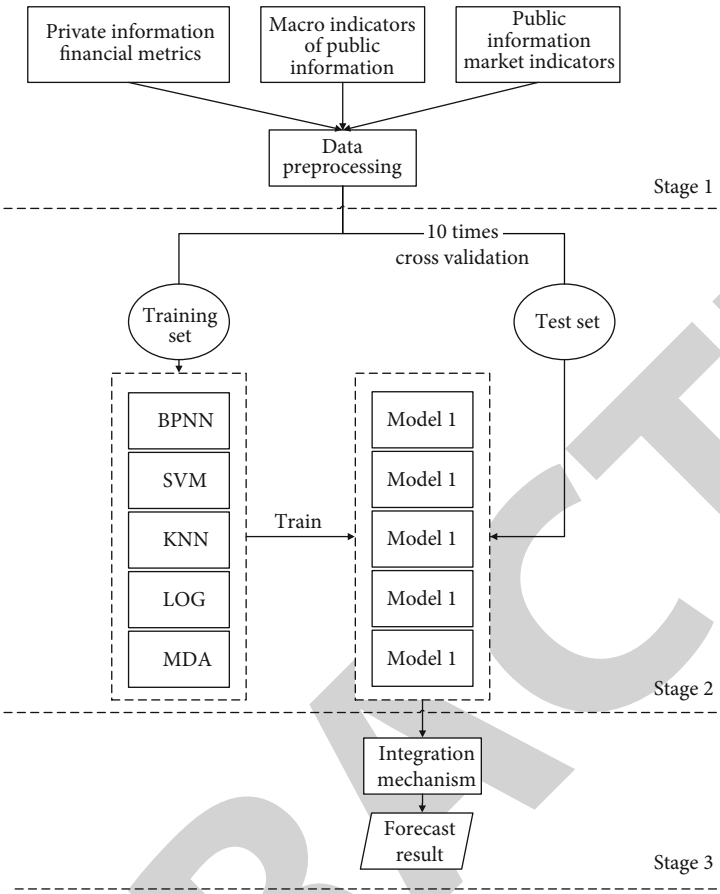

Figure 1: Model design route.

The above two formulas can be combined as

$$y_i\left[\omega^T X_i + b\right] \geq 0. \tag{3}$$

The optimization problem of support vector machine is as follows:

$$\begin{aligned} \text{minimize} &: \frac{1}{2}\|\omega\|^2, \\ \text{Constraints} &: y_i\left[\omega^T X_i + b\right] \geq 1, \end{aligned} \tag{4}$$

and distance from the vector X_0 to the hyperplane $\omega^T X + b = 0$ is $d = |\omega^T X_0 + b|/\|\omega\|$, of which

$$\|\omega\| = \sqrt{\omega_1^2 + \omega_2^2 + \cdots + \omega_m^2}.. \tag{5}$$

a can be used to scale $(\omega, b) \longrightarrow (a\omega, ab)$, and finally, there is $|\omega^T X_0 + b| = 1$ on the support vector X_0. At this time, the distance between the support vector and the hyperplane is $d = 1/\|\omega\|$. If you want to maximize the interval d, you only need to minimize $\|\omega\|$. For the convenience of solving, minimize $(1/2)\|\omega\|^2$.

Among them, x_i refers to each variable (multidimensional index), β_0 is a constant and has nothing to do with each variable x_i, and $\beta_1, \beta_2 \cdots \beta_n$ are regression coefficients, representing the contribution of each variable x_i to P. The probability P of a company's financial crisis calculated by the logistic regression model is determined and predicted by setting a threshold (this paper sets the threshold to 0.5; that is, if P is greater than or equal to 0.5, the company will have a financial crisis).

(5) Multiple linear discriminant analysis (MDA). MDA is mainly used in the classification problem of financial crisis early warning. The optimal linear decision model is used to classify new things. The general form of the discriminant function is

$$Z = \beta_0 + \sum_{i=1}^{n} \beta_i x_{it}. \tag{6}$$

When the linear relationship test (F test) of the model is significant, almost all regression coefficient βT-test was not significant. The sign of the regression coefficient is opposite to that expected. The tolerance of an independent variable is equal to 1 minus the judgment coefficient of the linear regression model obtained when the independent variable is the dependent variable and other independent variables are the predictive variables. The smaller the tolerance, the

Table 1: Multisource information indicators.

Level	Category	Indicator	
Private information	Solvency	Current ratio (X1)	Asset-liability ratio (X5)
		Equity multiplier (X2)	Current liabilities/total assets (X6)
		Quick ratio (X3)	Operating cash flow/current liabilities (X7)
		Earned interest multiple (X4)	
	Operational capability	Accounts receivable turnover ratio (X8)	Current asset turnover ratio (X11)
		Inventory turnover (X9)	Total asset turnover ratio (X12)
		Fixed asset turnover (X10)	Account payable turnover (X13)
	Profitability	ROE (X14)	Net interest rate on total assets (X18)
		Cost of sales rate (X15)	Return on current assets (X19)
		Net sales margin (X16)	Return on fixed assets (X20)
		Return on total assets (X17)	
	Structural stability	Fixed assets/total assets (X21)	Current liabilities/total liabilities (X23)
		Shareholders' equity/fixed assets (X22)	Current assets/total assets (X24)
	Business development capability	Net profit growth rate (X25)	Operating income growth rate (X27)
		Total asset growth rate (X26)	
	Capital expansion capability	Net assets per share (X28)	Net cash flow per share (X30)
		Earnings per share (X29)	Capital reserve per share (X31)
		Growth rate of total retail sales of social consumer goods (X32)	GDP growth rate (X37)
Public information (macroeconomics)		M1 growth rate (X33)	Growth rate of industrial added value (X38)
		M2 growth rate (X34)	Interest rate growth rate (X39)
		CPI growth rate (X35)	Interest rate (X40)
		PPI growth rate (X36)	Employment rate (X41)
Public information (market)		Total market cap/total liabilities (X42)	Yield increase (X44)
		Shares/total share capital (X43)	

more serious the multicollinearity. It is generally considered that there is severe multicollinearity when the tolerance is less than 0.1. The variance expansion factor is equal to the reciprocal of tolerance. Obviously, the larger the VIF, the more serious the multicollinearity. It is generally believed that there is severe multicollinearity when VIF is greater than 10.

3.3. An Ensemble Mechanism Based on Ensemble Learning. Ensemble learning is a machine learning algorithm that deals with classification problems. It establishes a set of independent models for the same problem for analysis and prediction and then fuses the output results of all independent models according to certain rules, thus obtaining an integrated model, with higher accuracy than that of a single model. At present, ensemble learning methods mainly include bagging, boosting, and so on. The majority voting (MV) method in bagging is the most widely used classical algorithm for ensemble learning. Therefore, in this paper, simple majority voting is used as a combination rule to construct an integrated mechanism, and an integrated analysis of the results of the above-mentioned different early warning models is carried out.

Table 2: Mixed matrix for a two-class early warning problem.

	Predicted as DF	Predicted as NF
Actual DF	TP	FN
Actual NF	FP	TN

TP (true positive): samples that are actually DF and predicted to be DF; FP (false positive): samples that are actually NF but predicted to be DF; FN (false negative): samples that are actually DF but predicted to be NF; TN (true negative): the sample that is actually NF and is predicted to be NF.

Denote the five early warning models in this paper as $F = \{f_1, f_2, f_3, f_4, f_5\}$. If the integrated mechanism of majority voting is used to predict the financial status of the i-th

Table 3: Prediction results of different models based on fused multisource information and based on fused multisource information.

Train-test ratio	Accuracy	Model					
		BPNN	SVM	KNN	LOG	MDA	Integrated learning
60 : 40	Acc-DF	0.8750	0.8125	0.6875	0.6250	0.8125	0.8750
	Acc-NF	0.8250	0.9250	0.9500	0.8250	0.9250	0.9250
	Acc	0.8393	0.8929	0.8750	0.7679	0.8929	0.9107
	F-measure	0.7638	0.8125	0.7627	0.6063	0.8125	0.8489
70 : 30	Acc-DF	0.8333	0.9167	0.8333	0.5000	0.8333	0.9167
	Acc-NF	0.8667	0.9000	0.8667	0.8333	0.9333	0.9333
	Acc	0.8571	0.9048	0.8571	0.7381	0.9048	0.9286
	F-measure	0.7715	0.8487	0.7715	0.5222	0.8333	0.8807

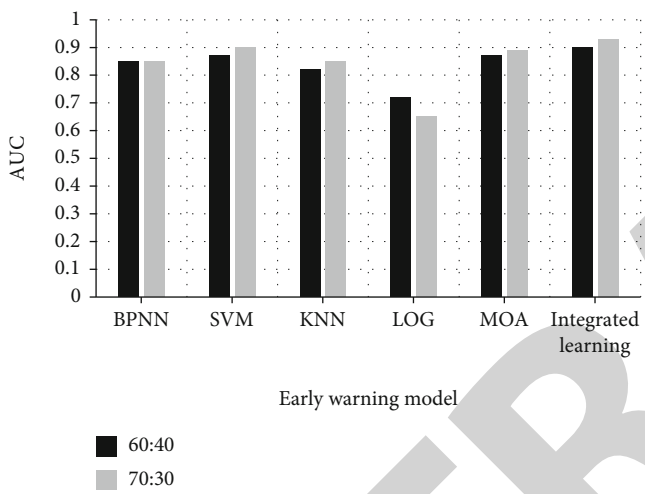

Figure 2: AUC values of different models based on fusion of multisource information.

enterprise Y_i, the calculation process is as follows:

$$F(Y_i) = \begin{cases} \text{NF}, P(F(Y_i) = NF) > P(F(Y_i) = \text{DF}), \\ \text{DF}, P(F(Y_i) = NF) \le P(F(Y_i) = \text{DF}), \end{cases} \quad (7)$$

$$P(F(Y) = \text{NF}) = \sum_{t:f_i(Y_i)=\text{NF}}^{5} \frac{f_t(Y_i)}{5}, \quad (8)$$

$$P(F(Y_i) = \text{DF}) = \sum_{t:f_i(Y_i=\text{NF})}^{5} \frac{f_t(Y_i)}{5}. \quad (9)$$

Among them, $f_t(Y_i)$ is the prediction result of the early warning model t on the enterprise Y_i, and $P(F(Y_i) = \text{DF})$ refers to the probability that the enterprise Y_i is predicted to be DF through the integration mechanism, and its value range is $[0, 1]$. Similarly, we can know that $P(F(Y_i) = \text{NF})$. $F(Y_i)$ is the prediction result of enterprise Y_i obtained by majority voting, and its value is NF or DF.

3.4. Model Evaluation Method. In order to measure the performance of the early warning model more accurately, this paper selects three categories of five indicators as the evaluation criteria for the model, which are described in detail as follows: for each test sample of the dataset, the model has four possible prediction results, as shown in Table 2 which is a mixture matrix of two-class early warning problems.

Assuming that the numbers of FD and NM samples in the test set are Num1 and Num2, respectively, there are

$$\text{TP} + \text{FN} = \text{Num1}, \quad (10)$$

$$\text{FP} + \text{TN} = \text{Num2}. \quad (11)$$

(1) Accuracy. An accurate degree is the most common evaluation criterion in forecasting problems. This paper uses three indicators to measure the accuracy of the model. They are the prediction accuracy of crisis enterprises

$$(\text{Acc} - \text{DF}): \text{TP}(/\text{TP} + \text{FN}), \quad (12)$$

Forecast accuracy of noncrisis companies

$$(\text{Acc} - \text{NF}): \text{TN}\ (/\text{FP} + \text{TN}), \quad (13)$$

and overall accuracy

$$(\text{Acc}): (\text{TP} + \text{TN})\ (/\text{TP} + \text{FN} + \text{FP} + \text{TN}). \quad (14)$$

(2) *F*-measure method. *F*-measure, proposed by Powers (2011), is a statistic commonly used to evaluate the performance of early warning models. *F*-measure is a combination of recall and precision: *F*-measure = aunt recall × precision. Among them, precision is the proportion of DF (NF) companies that are actually predicted to be DF (NF) companies; recall is the ratio of the number of DF (NF) companies that can be correctly predicted to be all DF companies

Table 4: Comparison of integrated early warning models based on multisource indicators and financial indicators.

Train-test ratio	Accuracy	Integrated early warning model	
		Based on a multisource indicator model	Model based on financial indicators
60 : 40	Acc-DF	0.8750	0.8125
	Acc-NF	0.9250	0.9500
	Acc	0.9107	0.9107
	F-measure	0.8489	0.8391
70 : 30	Acc-DF	0.9167	0.8333
	Acc-NF	0.9333	0.9333
	Acc	0.9286	0.9048
	F-measure	0.8807	0.8333

When calculating the *F*-measure of DF,

$$\begin{cases} \text{Recall} = \dfrac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Precision} = \dfrac{\text{TP}}{\text{TP} + \text{FP}}. \end{cases} \tag{15}$$

When calculating the *F*-measure of NF,

$$\begin{cases} \text{Recall} = \dfrac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{Precision} = \dfrac{\text{TN}}{\text{TN} + \text{FN}}. \end{cases} \tag{16}$$

From the formula, if both recall and precision are relatively small, *F*-measure will be small; if recall (precision) is large and precision (recall) is small, *F*-measure will also be small; recall and precision are large only when the *F*-measure is larger. Therefore, the *F*-measure method can reasonably evaluate the prediction performance of the model for DF and NF samples.

The ROC curve can reflect the prediction performance of the model at different thresholds and has gradually become the mainstream method to measure the prediction performance of the model. The method uses Acc-DF and 1-Acc-NF as the horizontal and vertical coordinates, respectively, and each threshold corresponds to a point (1-Acc-NF, Acc-DF), constantly changing the threshold and converting the obtained (1-Acc- NF, Acc-DF) points that are connected, which is the ROC curve of the model on the test set. Since the ROC curve is a two-dimensional graph, it is not convenient for quantitative evaluation. Therefore, the area under the curve (AUC) is often used as the quantitative evaluation of the model performance by the ROC curve. The value range of AUC is [0, 1]. The closer the AUC is to 1, the better the performance of the early warning model.

4. Result Analysis and Discussion

4.1. Empirical Analysis of the Financial Crisis Early Warning Model

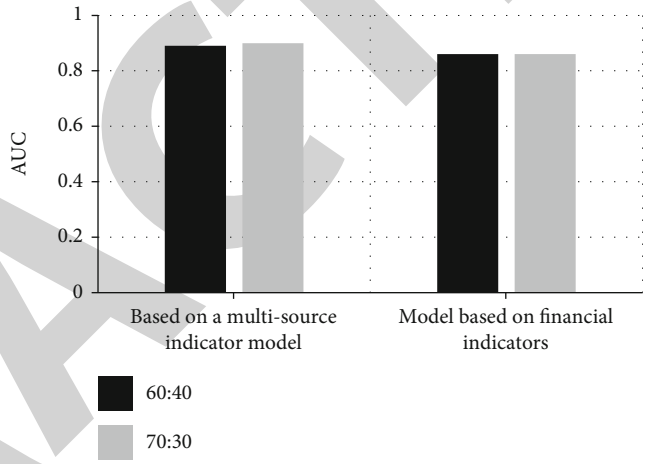

Figure 3: AUC comparison based on multisource indicators and financial indicator models.

(1) Selection of research samples and data sources

This paper studies the financial crisis early warning problem of small- and medium-sized board companies and takes the companies listed as "ST" as DF samples and the companies that have never been marked as "ST" as NF samples. The data of 791 small- and medium-sized board companies in China from 2005 to 2016 were collected (the data came from the Juling financial service platform), of which 40 companies were marked as ST and ST∗ from 2008 to 2016. Excluding the sample of companies with incomplete data, a sample of 721 companies was finally obtained. Considering that the impact of macroeconomic variables on corporate finance is related to the duration of macrovariables, this study selects data with a time window of 2 years for empirical analysis; that is, the company is marked as ST in year *T*; then, take the private data of *T*-2 years. Information and public information data are used for analysis

(2) Analysis of results

Since the total number of DF samples and NF samples is 40 and 681, respectively, and the sample ratio is about 1 : 17, it is a typical unbalanced dataset. In order to avoid the

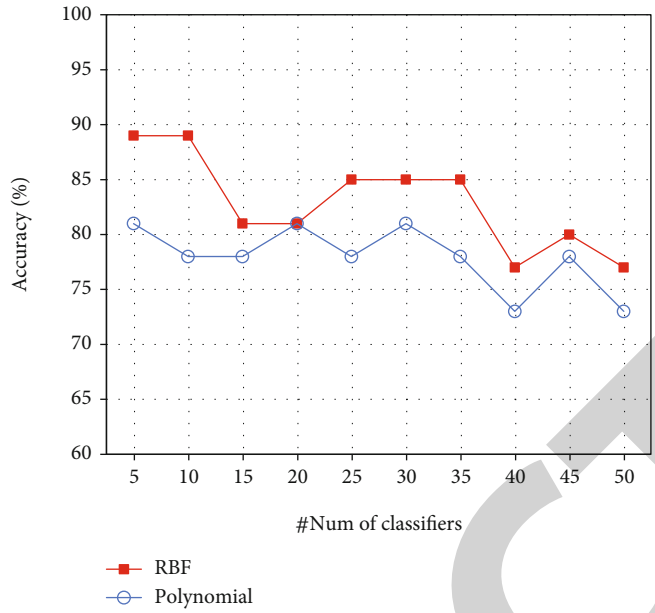

Figure 4: Prediction results after ensemble learning.

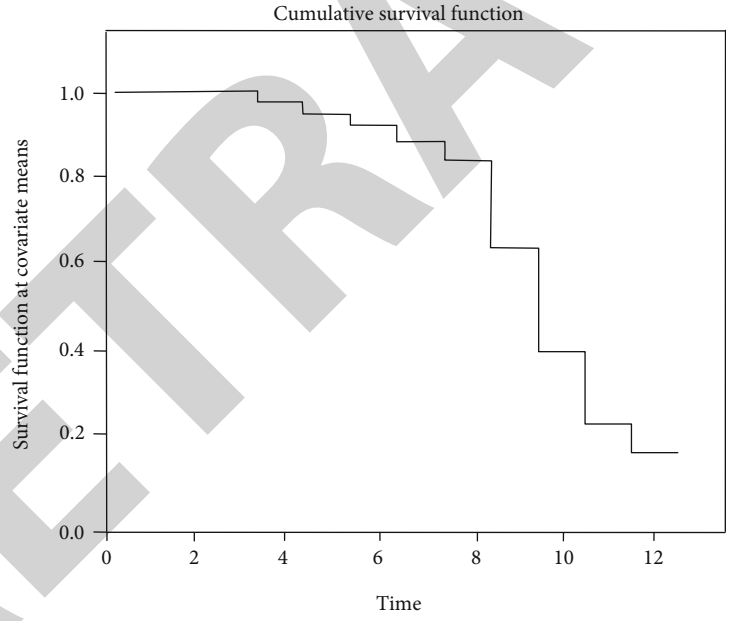

Figure 5: Survival rate graph.

influence of data imbalance on the accuracy of the model, this paper adopts the method of random simple undersampling, randomly selects 100 samples and 40 ST samples from the healthy group as the training and testing samples, and randomly divides the obtained sample data into training samples. With the training group and test group, the data is divided according to the training-test ratio (60 : 40, 70 : 30) (due to the small number of ST samples, this study will not explore the situation when the training-test ratio is 80 : 20 and 90 : 10). In this paper, 10 times of cross-validation is adopted, and the average result of 10 experiments is used as the final result to avoid overfitting and improve the robustness of the result.

In order to comprehensively evaluate the prediction performance of the new model, this part constructs two sets of controlled experiments: the first is to compare the prediction results of integrating private information and public information with the prediction results relying only on private information. The comparison results are shown in Table 3 and Figure 2. The second is to compare the prediction

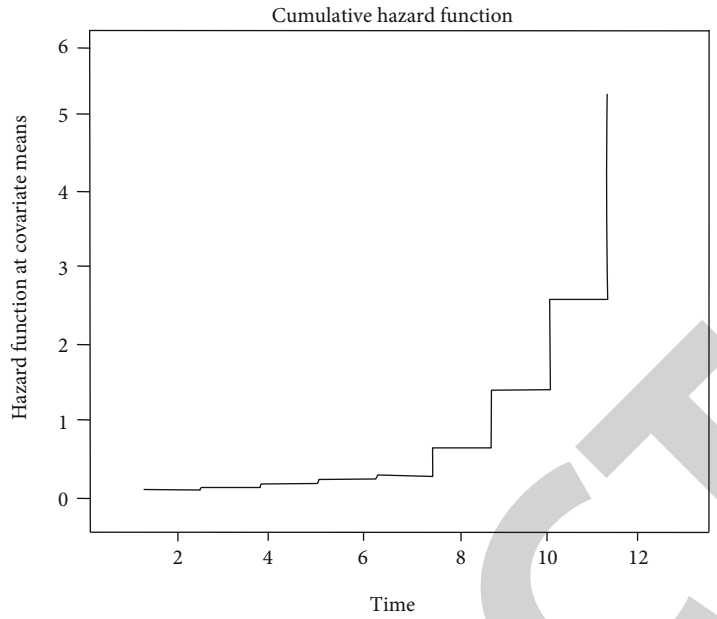

Figure 6: Risk rate curve.

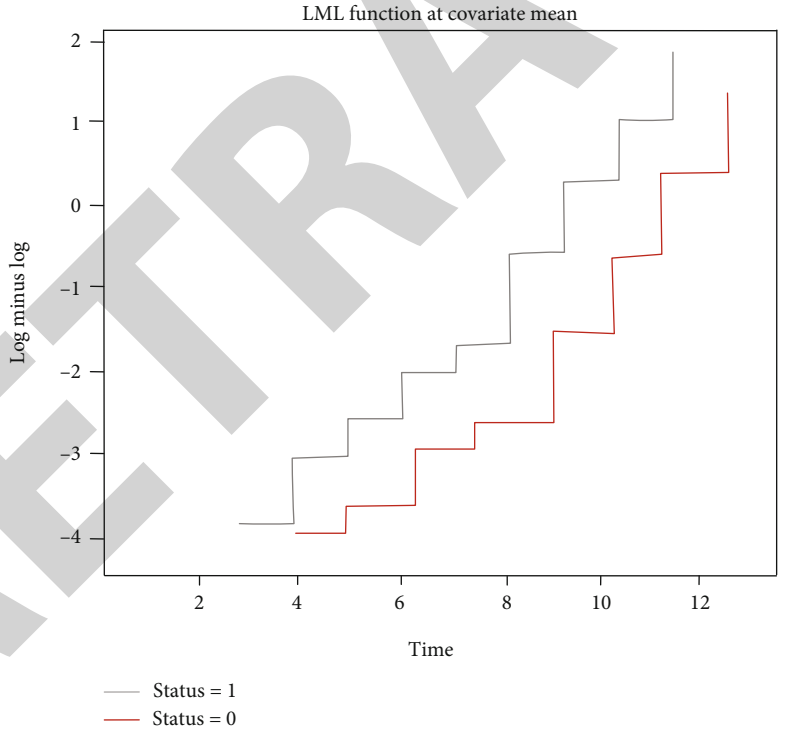

Figure 7: LML diagram of the two groups of samples.

results based on the ensemble learning model and the prediction results of a single data mining early warning model. The comparison results are shown in Table 4 and Figure 3.

The prediction results based on the ensemble learning model are compared with the prediction results of a single data mining early warning model. The comparison results are shown in Table 4 and Figure 3.

The following conclusions can be drawn from the experimental results:

(1) It can be seen from Table 4 and Figure 3 that the early warning model based on multisource indicators is obviously better than the early warning model based on financial indicators. For example, when the training-test ratio is 70 : 30, the prediction

accuracy of the multisource indicator model is higher than that of financial indicators. The model is high, where the Acc-DF of the two differs by 0.08%. There may be two reasons for this: first, SMEs are affected by their own management level, macroeconomic fluctuations, and market competition, and collecting multisource information can better reflect various factors affecting corporate finance; second, financial indicators lag behind. When used to predict financial crises, the results are less robust and less accurate

(2) As shown in Table 3 and Figure 2, the model prediction results based on ensemble learning are better than those of the single early warning model. Regardless of the training-test ratio of 60 : 40 or 70 : 30, the accuracy and stability of the ensemble learning prediction results are significantly better than those of the single early warning model. The main reasons for this may be the following two points: first, the single model shows instability in the multisource indicator early warning model. For example, when the train-test ratio is 60 : 40, the Acc-DF of BPNN is the highest, but the Acc-NF is lower, and other models are opposite to BPNN. When the training-test ratio is 70 : 30, the situation is just the opposite; second, ensemble learning can integrate the advantages of different early warning models and improve the prediction accuracy

(3) It can be seen from Table 3 that the performances presented by the five data mining early warning models are quite different. The prediction result of SVM is relatively the best, while the performance of LOG is the worst. The main reason is that SVM uses a nonlinear mapping algorithm to better handle nonlinear high-dimensional data, while LOG is easy to generate when dealing with complex nonlinear multidimensional data overfitting. The prediction accuracy of the model is that SVM and MDA are better than BPNN and KNN and BPNN and KNN are slightly better than LOG

4.2. Empirical Analysis of Integrated Learning. The figure is based on the prediction results of a single SVM model on the test sample data. This paper will use the AdaBoosting strategy for the RBF kernel function-based SVM single classifier and the polynomial kernel function-based SVM single classifier and classify at the same time. A series of experiments were carried out with a step size of 5, and the number of devices was gradually increased from 5 to 50. The above experimental process has shown the running results [10], and the number of equipment is closely related to the prediction accuracy. With the increase in the number of equipment, the prediction accuracy gradually decreases, and the prediction accuracy of the RBF method is higher than that of polynomial, and its statistical results are shown in Figure 4.

According to the survival time of the sample, the estimated value of the baseline survival rate function and the 8 covariates entered into the Cox regression model are, respectively, substituted into the final function expression of the Cox regression model, and the average value of each sample at the observation time t_i can be calculated as the survival function.

In addition, according to the change of the survival rate of the sample, the Cox regression model can also fit the survival rate curve and the hazard rate curve, as shown in Figures 5 and 6.

It can be seen from Figure 5 that after the survival time of 8 years, the survival rate of listed SMEs is relatively low. Similarly, it can be seen from Figure 6 that after 8 years of survival from SMEs, business risks also increase significantly. The average life span of small- and medium-sized enterprises is 8 years.

Since the application of the Cox proportional hazards model is based on the assumption of proportional hazards, that is, the relative hazard ratio is independent of time t and is a constant proportional value. In this paper, the sample data is divided into two groups according to the survival status = 1 and status = 0, and the quadratic logarithmic survival rate graph is made, respectively, namely, log minus log (LML). If the curves drawn using the two sets of data cross or the distance between the two curves varies greatly, it means that the hazard ratio changes with time, which does not have the premise of applying the Cox proportional hazards model; otherwise, the assumptions of the Cox proportional hazards model are met. Figure 7 shows the LML diagram of the two groups of samples.

5. Conclusion

Due to the characteristics of SMEs, only relying on financial indicators based on private information has been unable to obtain effective financial crisis early warning results. Based on the perspective of multisource information fusion, this paper integrates public information for measuring macroeconomic conditions and market conditions on the basis of traditional private information and establishes an indicator system that can more comprehensively reflect the financial status of SMEs. On this basis, five data mining models with better performance are selected for early warning analysis, and an ensemble learning method is introduced to perform "secondary" fusion processing on the prediction results of the five models, so as to obtain more effective and robust prediction results. The empirical analysis of the 12-year data of China's small- and medium-sized enterprises shows that the prediction results of the fusion of multisource information are significantly better than the prediction results of only using financial indicators, with higher accuracy. The empirical results prove the validity and reliability of the new perspective and method and also provide a more reliable method for financial early warning of small- and medium-sized enterprises.

(1) When a single classifier solves a problem, there is often a phenomenon of weak classification.

However, the ensemble learning method just overcomes the shortcomings of weak classifiers. It transforms the combination of multiple weak classifiers into strong classifiers through the ensemble algorithm and obtains better results than weak classifiers

(2) The integrated learning method based on support vector machine shows many advantages in solving small-sample, nonlinear, and high-dimensional pattern recognition, and it adopts the principle of structural risk minimization, which overcomes the need to obtain a large number of BP neural networks. The training samples and the defect that only local optimal solutions can be obtained use the nonlinear kernel function to fully fit and learn the sample data, which can well solve the problems of "overlearning" and "underlearning," making the model easy to calculate, with speed and other advantages

Therefore, it is effective to introduce the SVM method into the model for constructing the financial crisis early warning model of listed companies and achieve a certain financial crisis early warning effect. At the same time, the following issues deserve further study:

(1) In terms of sample selection, since the number of ST companies in listed companies is far less than that of non-ST companies and because the traditional SVM algorithm (C-SVM) is not suitable for unbalanced samples, this paper selects two types of samples that match the number of samples. Therefore, it is necessary to further study the impact on the classifier when the two types of samples are unbalanced in order to ensure a more accurate prediction accuracy

(2) As an emerging method, support vector machine still has many unsolved or insufficiently solved problems, and there is still more room for development in application, which can be used more fully in commercial bank credit evaluation, loan classification, and risk management

Data Availability

The figures and tables used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The author would like to show sincere thanks to those techniques who have contributed to this research.

References

[1] L. I. Yujie, "Introduction to the special section on artificial intelligence and computer vision (VSI-aicv4)," *Computers and Electrical Engineering*, vol. 90, no. 3, p. 107055, 2021.

[2] S. Emanuele and R. Francesca, "Use of artificial intelligence in endoscopic training: is deskilling a real fear," *Gastroenterology*, vol. 160, no. 6, p. 2212, 2021.

[3] M. Dirk, "Availability bias and artificial intelligence," *Journal of the American Academy of Dermatology*, vol. 3, pp. 121–130, 2019.

[4] S. V. Patil, "Artificial intelligence in ophthalmology: is it just hype with no substance or the real McCoy," *Indian Journal of Ophthalmology*, vol. 67, no. 7, p. 1251, 2019.

[5] R. Bellazzi and A. Abu-Hanna, "Artificial intelligence in medicine AIME'07," *Artificial Intelligence in Medicine*, vol. 46, no. 1, pp. 1–3, 2009.

[6] C. Baraniuk, "Artificial intelligence is analysing heart scans in dozens of hospitals," *New Scientist*, vol. 244, no. 3258, p. 11, 2019.

[7] H. Jingsha and Y. Prasad, "Special issue on the application of artificial intelligence in advanced manufacturing," *International Journal of Computer Integrated Manufacturing*, vol. 33, no. 10-11, pp. 947-948, 2020.

[8] P. Leeson, "Deep learning your left from your right," *JACC: Cardiovascular Imaging*, vol. 13, no. 2, pp. 382–384, 2020.

[9] A. Onen, "Ahmet role of artificial intelligence in smart grids," *Electrical Engineering*, vol. 7, pp. 110–117, 2021.

[10] D. Lehman Constance and J. Topol Eric, "Readiness for mammography and artificial intelligence," *The Lancet*, vol. 398, no. 10314, p. 1867, 2021.

[11] J. Zhang, L. Tan, X. Tao, T. Pham, and B. Chen, "Relational intelligence recognition in online social networks – a survey," *Computer Science Review*, vol. 35, no. 2, p. 100221, 2020.

[12] O. Ziad, "Artificial intelligence, bias, and patients' perspectives," *The Lancet*, vol. 397, no. 10289, p. 2038, 2021.

[13] "Artificial intelligence in the age of cognitive endoscopy," *Gastrointestinal Endoscopy*, vol. 91, no. 6, pp. 1251-1252, 2020.

[14] P. Sharma, A. Pante, and A. Seth, "Artificial intelligence in endoscopy," *Gastrointestinal Endoscopy*, vol. 91, no. 4, pp. 925–931, 2020.

[15] G. Huguet, C. Schramm, E. Douard et al., "Measuring and estimating the effect size of rare non-recurrent deletions and duplications on general intelligence," *Biological Psychiatry*, vol. 87, no. 9, p. S196, 2020.

[16] "Your wish is my CMD," *Communications of the ACM*, vol. 63, no. 7, pp. 15-16, 2020.

[17] M. Atiquzzaman, J. Li, and W. Pedrycz, "Special issue on new advanced techniques in security of artificial intelligence," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 3, pp. 1255–1257, 2022.

[18] D. M. Maslove, P. W. G. Elbers, and G. Clermont, "Artificial intelligence in telemetry: what clinicians should know," *Intensive Care Medicine*, vol. 47, no. 2, pp. 150–153, 2021.

[19] A. Sarirete, Z. Balfagih, T. Brahimi, M. D. Lytras, and A. Visvizi, "Artificial intelligence and machine learning research: towards digital transformation at a global scale," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 7, pp. 3319–3321, 2022.

[20] W. Zhu, T. Zhang, Y. Wu, S. Li, and Z. Li, "Research on optimization of an enterprise financial risk early warning method based on the DS-RF model," *International Review of Financial Analysis*, vol. 81, p. 102140, 2022.