Applied Mathematics, Modeling and Computer Simulation
C. Chen (Ed.)

doi:10.3233/ATDE220092

An Analysis of the Authenticity of Financial Data of Listed Companies Based on Vector Machines

Mingmin GONG[a,1], Lei ZHENG[b], Jiayang HOU[a], Letao WANG[a] and Xing LU[a]
[a] *School of Information Engineering, Wuhan College, China*
[b] *Shen Zhen Institute of Technology, China*

Abstract. Driven by the benefits of listing at a high premium, financial fraud in listed companies has become a widespread problem worldwide. In the face of economic interests, even developed countries with relatively complete laws and policies cannot completely eliminate financial fraud. The financial fraud incidents of listed companies have seriously disrupted the normal order of my country's capital market. Therefore, the specific financial data of listed companies are analyzed to obtain the influence of financial indicators on the identification of financial fraud and to determine whether there is financial fraud in listed companies.

Keywords. Machine learning, index mathematical modeling, feature extraction.

1. Introduction

Nowadays, Chinese companies are gradually moving toward the world stage in response to the tide of development of the times. But at the same time facing the external financial crisis, trade barriers established by hegemony and weak internal control, all industries are facing financial risks [1]. Although the market economy is slowly improving and the government has stepped up its control efforts, many small and medium-sized enterprises are at stake and are deeply troubled by financial risks. Even some listed companies do not have a strong sense of risk management and control, and financial problems frequently appear. The strategies of many listed companies tend to be ahead of the capital, taking too much steps, and failing to adopt effective risk control measures to deal with them. In recent years, due to the weak financial risk control of listed companies in different industries, financial fraud cases and even corporate bankruptcies have occurred frequently. The financial data of some superior companies is falsified. The false content includes inflated deposits, falsified income, and buying and selling of stocks through related parties.

Driven by the benefits of listing at a high premium, financial fraud in listed companies has become a widespread problem worldwide. In the face of economic

[1] Corresponding Author, Mingmin GONG, School of Information Engineering, Wuhan College, China; Email: zhangyingyi@bucea.edu.cn.

interests, even developed countries with relatively complete laws and policies cannot completely eliminate financial fraud. The financial fraud incidents of listed companies have seriously disrupted the normal order of my country's capital market. Therefore, machine learning algorithms are used to analyze the specific financial data of listed companies to obtain indicators that affect the true and false financial data, and it seems that whether there is financial fraud in listed companies. Imminent. This article will focus on more than 20,000 financial data reports of listed companies for many years, and complete task one (find out the main characteristic factors that affect the judgment of whether the financial company's data is falsified) and task two (determine the falsified financial data of the manufacturing company in the sixth year) Situation) and task three (determine the fraudulent financial data of non-manufacturing companies in the sixth year).

2. Analyze the Problem

Facing the ever-changing economic environment, there are endless cases of fraudulent financial data of listed companies. Whether financial data is falsified is to start with micro data such as financial statement data and management information, and use statistical methods and data mining techniques to analyze the true financial data of the listed company. This article will study the falsification of the financial data of more than 20,000 listed companies in different industries. After fully analyzing the associated characteristic indicators, it will be discovered whether the financial data of the company is falsified. On the basis of reading a lot of domestic financial data fraud case analysis, using variance and T test to test the indicators with significant characteristics, and then using the principal component analysis method to determine the coefficients of financial indicators, excluding subjective factors, to predict the financial Provide a basis for calculating whether the data is falsified. Compared with other model prediction methods, support vector machine has good generalization ability for limited sample data and can obtain the optimal solution under the existing information conditions. Therefore, the algorithm model of support vector machine is selected for the analysis of financial data. Falsification provides technical support and verifies the validity of its prediction results. Finally, the construction model is applied to the judgment of whether the financial data of listed companies is falsified. [2]The results show that the model has a high accuracy in predicting whether the financial data is falsified, and the output of some listed company financial data is the result of falsification.

3. Solution

3.1. Task Analysis

Task 1: Analysis: When dealing with whether the financial data of listed companies in this different industry is falsified, the staff must first perform data cleaning according to certain scientifically based algorithms, and fill or delete some missing data items to avoid data noise. Cause problems such as low efficiency and high error rate. Therefore, for task one, not only need to classify the industry according to Annex 1, but also according to the imported data in Annex 2. After the data items are cleaned and

processed, use variance and T test to detect and find out the indicators with significant characteristics, and then principal component analysis , Research the correlation, remove the indicators that are too relevant, and what is left is the required characteristic indicators related to the falsification of financial data in various industries.[3]

Task 2: Analysis: According to the financial data of each listed company in the manufacturing industry given in Annex 2, determine the listed company whose financial data is falsified in the sixth year. Filter by the industry number given in Annex 1, and directly extract the data of all listed companies in the manufacturing industry to improve calculation efficiency. Therefore, for task two, it is necessary to compare the correct rate of support vector machine, decision tree and logistic regression model according to the eigenvalues obtained in task one, and compare the Gaussian kernel, polynomial and sigmoid of the optimal support vector machine model. Wait for the calculation results of different kernel functions, select the kernel function with the highest accuracy, and then import the data into the model to predict the listed companies that will be financially fraudulent in the sixth year.

Task 3: Analysis: According to the financial data of listed companies in other industries (except manufacturing) in Annex 2, identify the listed companies whose financial data is falsified in the sixth year. Therefore, to solve its task is to replace the data with the financial data of various industries outside the manufacturing industry on the basis of task two, and use the data model in task two that has been established to make the sixth-year financial data fraud prediction.

3.2. Feature Extraction

After importing the preprocessed data, delete the data for the sixth year, and divide the data into two categories according to whether the financial data of the listed company is falsified, so as to perform variance analysis and T test on the remaining features. Extract the feature values that have significant differences between the two types of data on whether the financial data is falsified, and keep these feature values as candidate indicators, and then calculate whether there is a correlation between the candidate indicators,[4] if they are related, perform principal component analysis, thereby Select uncorrelated characteristic indicators as the main indicators.

Definition: Analysis of variance, also known as "analysis of variance", is used to test the significance of the difference between the means of two or more samples.[5] Due to the influence of various factors, the data obtained from the study fluctuates. The reasons for the fluctuations can be divided into two categories, one is uncontrollable random factors, and the other is controllable factors that affect the results in the research [6].

The t test, also known as the student t test (Student's t test), is mainly used for a normal distribution with a small sample size (for example, $n < 30$) and an unknown population standard deviation σ. The t test is to use the t distribution theory to infer the probability of the difference, so as to compare whether the difference between the two averages is significant [7].

According to the different types of data design, there are two methods of analysis of variance as follows:

(1) To compare the mean of multiple samples in a group design, a completely randomized design of variance analysis, that is, a one-way analysis of variance, should be used;

(2) For the comparison of the mean values of multiple samples in the random block design, the analysis of variance of the compatibility group design should be used, that is, the two-factor analysis of variance.

The basic steps of the two types of analysis of variance are the same, but the decomposition method of variation is different. For the data of the group design, the total variation is divided into intra-group variation and inter-group variation (random error), namely: SS total = SS between groups + SS group For the data of the compatibility group design, the total variation includes not only the treatment group variation and random error, but also the compatibility group variation, namely: SS total = SS treatment + SS compatibility + SS error.

This project needs to perform principal component analysis on these variables after selecting certain indicators with significant differences. Because these variables often have a certain correlation, and many indicators have a strong correlation, it will not only increase the computational complexity, but also affect the analysis results of the model. Therefore, the idea of this article is to transform many variables into a few uncorrelated comprehensive variables through principal component analysis. The idea of the principal component analysis method used is to map n-dimensional features to m-dimensionality (m<n). This m-dimensionality is a brand-new orthogonal feature called principal component. This m-dimensional feature is reconstructed , Not simply subtracting the nm-dimensional features from the n-dimensional features, the core idea is to project the data along the maximum direction to make the data easier to distinguish.

In this paper, dimensionality reduction is a data set preprocessing technology. PCA extracts a smaller number of principal components to reduce the dimensionality of the data space. The extracted principal components must be able to contain the features of the original data to the maximum, and at the same time to the minimum Lose the amount of information of the original data, in order to discover the most important characteristics of the original data and the richer connotation behind the high latitude while reducing the dimensionality. The contribution rate of each principal component of PCA can be measured by its variance. The first principal component Comp.1 is the linear combination of all variables with the largest variance, followed by the second principal component Comp.2, and so on, n-dimensional variables have at most n principal components, generally extract the first p (p<n) components whose cumulative variance contribution rate is more than 80%. Determine the number of principal components that need to be retained and discard other principal components to achieve dimensionality reduction of the data, which can remove some redundant information and noise of the data, make the data simpler and more efficient, and improve the efficiency of other machine learning.

Generally, the principal component analysis method mainly includes the following processes.

Step 1: Standardize the imported data, make it zero mean, and normalize it to eliminate the influence of different dimensions. The original data $Y = (Y_1, \ Y_2, \ Y_n)$ is a random variable in an n-dimensional space, where $Y_j = (Y_{1j}, Y_{2j}, \cdots, Y_{mj})\ T$, m is the sample size. Y is standardized to obtain n-dimensional random variables $X = (X_1, X_2, , X_n)$, $X_j = (X_{1j}, X_{2j}, \cdots, X_{nj})\ T$, where the calculation formula is:

$$X_{ij} = \frac{y_{ij} - \overline{Y}_J}{S_j} \tag{1}$$

Among them, y_{ij} $\overline{Y}_j$ S_j are the sample mean and standard deviation of respectively, and the calculation formula is:

$$\overline{Y}_J = \frac{1}{m}\sum_{i=1}^{m} y_{ij}, (j = 1,2,..., n) \tag{2}$$

$$S_j^2 = \frac{1}{m-1}\sum_{1}^{m} (y_{ij} - \overline{Y}_J)^2, (j = 1,2,..., n) \tag{3}$$

Step 2: Calculate the correlation coefficient matrix (or covariance matrix) of the indicator variables and solve the eigenvalues and eigenvectors. The correlation coefficient matrix of $=(r_{ij})n\times n$, the calculation formula of the correlation coefficient is as follows:

$$r_{ij} = \frac{\sum_{k=1}^{n} x_{ki} \bullet x_{kj}}{n-1}, (i, j = 1,2,..., n) \tag{4}$$

Solve the eigenvalues of the correlation coefficient matrix and sort them from largest to smallest $\lambda 1 \geq \lambda 2 \geq \cdots \geq \lambda n \geq 0$, the eigenvectors corresponding to the eigenvalues are a_1, a_2, $\cdots$, a_n, , where $a_j = (a_{ij}, a_{2j}, \cdots, a_{nj})$ Obtain n new index variables from the feature vector:

$$\begin{cases} Z_1 = a_{11}X_1 + a_{21}X_2 + ... + a_{n1}X_n \\ Z_2 = a_{12}X_1 + a_{22}X_2 + ... + a_{n2}X_n \\ ... \\ Zn = a_{1n}X_1 + a_{2n}X_2 + ... + a_{nn}X_n \end{cases} \tag{5}$$

Step 3: Calculate the variance contribution rate and cumulative contribution rate of the principal components, and select the first p 个（$p < n$）principal components based on the cumulative contribution rate of the variance. The variance contribution rate of the j-th principal component, and the cumulative contribution rate of the variance of the first p principal components:

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{n} \lambda_k} \tag{6}$$

According to the actual situation of the research problem, the first p principal components with a cumulative contribution rate of greater than 80% are usually selected to replace the original n index variables, so as to achieve data dimensionality reduction.

Step 4: Calculate the comprehensive score of the principal component according to the contribution rate of the principal component variance and the principal component expression.

$$w = \sum_{j=1}^{p} b_j z_j \tag{7}$$

Step 5: Combining the comprehensive scores of the principal components and the actual situation, do further statistical analysis [8].

4. Model Design

In order to solve the task of this project, this paper uses data preprocessing, feature value extraction, etc. to obtain the main indicators that mainly affect whether the financial data of listed companies is falsified.After data acquisition is completed, a certain algorithm is used to divide the training sample and the test sample, Import the constructed vector machine model. The model are shown in table 1.Comparing the support vector machine model, the logistic regression model and the decision tree model, it is not difficult to find that the optimized algorithm effect of the support vector machine model in these three models is much better than other algorithms. The following is the accuracy of the data training of the three models, figure1 shows:

Table 1. Accuracy of the three models for data training

Training set	Accuracy	Precision	Recall	F1	Auc
LR Training set	0.992	0.0	0.0	0.0	0.733
LRTraining set	0.992	0.0	0.0	0.0	0.769
SVC Training set	0.992	0.0	0.0	0.0	0.627
SVC Training set	0.992	0.0	0.0	0.0	0.776
RF Training set	1.0	1.0	1.0	1.0	1.0
RF Training set	0.992	0.0	0.0	0.0	0.559

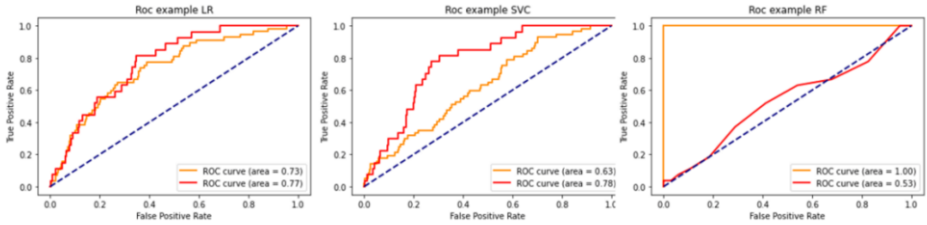

Figure 1. Comparison of Roc efficiency diagrams of the three models

In this paper, in order to make the comparison of the three models, in order to better improve the prediction accuracy and generalization ability of the model, the optimal support vector machine model is selected here to start the construction.

5. Data Selection and Indicators

In response to the problem in Task 1, this article uses the industry classification in Annex 1 and the financial data of related listed companies provided in Annex 2 to determine the data indicators related to the falsification of financial data. First, the data given in Annex 2 Carry out data preprocessing to improve the quality of the data, thereby helping to improve the accuracy and efficiency of data analysis. For each financial data-related indicator of a listed company, the data is read to obtain 363 characteristic factors for 22213 stocks in six years, of which the missing rate is greater than 80%, and the result of whether the financial data is fraudulent is not highly relevant After the indicator is directly deleted, there are 84 columns remaining, because the date does not have a particularly large causal relationship with whether the financial data of the listed company in the sixth year is falsified. The company did not falsify in the first five years, and it does not mean that the financial data of the listed company in the sixth year must not be falsified. Therefore, the impact of the 10 columns of related indicators on the date on whether the financial data of the listed company is falsified is excluded. Excluding the theoretically deleted part and standardizing the data, the remaining 74 columns of data can be calculated using certain algorithm principles, and indicators that have a significant impact on whether the financial data of listed companies are fraudulent or not are selected.

Use analysis of variance (the full name is one-way analysis of variance) to study the difference of FLAG for CASH_C_EQUIV, NOTES_RECEIV, AR, PREPAYMENT, OTH_RECEIV, etc., a total of 74 items. Because the analysis results of all indicators are too many, it is difficult to expand, so the analysis of variance is used. For the difference between X (fixed type) and Y (quantitative), take any 10 items shown in table 2 as an example for detailed analysis.

(1) Analyze whether there is significance between X and Y (p value is less than 0.05 or 0.01);

(2) If it is significant; describe the specific difference by comparing the average value;

(3) If it is not significant; it means that under different groups of X, there is no difference in Y;

(4) Summarize the analysis.

Table 2. ANOVA results of 10 random indicators

分析项	项	样本量	平均值	标准差	t	p
	0.0	10546	1118388981.94	4402981722.44		
CASH_C_EQUIV	1.0	84	633195357.65	818346245.31	1.010	0.313
	总计	10630	1114554902.50	4386355613.82		
	0.0	9404	370755067.90	1836001918.45		
NOTES_RECEIV	1.0	72	232778322.81	456576797.20	0.637	0.524
	总计	9476	369706700.91	1829478996.36		
	0.0	10390	747261669.12	2743924258.09		
AR	1.0	81	676003461.57	1267460908.42	0.234	0.815
	总计	10471	746710440.50	2735541235.10		
	0.0	10524	144061082.42	774581007.44		
PREPAYMENT	1.0	82	117518581.00	234465009.18	0.310	0.756
	总计	10606	143855869.80	771856135.51		

From the overall analysis results, different FLAG samples will not show significance for a total of 66 items such as CASH_C_EQUIV, NOTES_RECEIV, AR, PREPAYMENT, etc. ($p>0.05$), which means that different FLAG samples show consistency for all of these indicators, and there is no difference. In addition, FLAG samples are significant for a total of 18 items such as DILUTED_EPS, BASIC_EPS ($p<0.05$), which means that different FLAG samples are different for DILUTED_EPS, BASIC_EPS, etc. Comparing the analysis of a pair of oppositely significant indicators, figure 2 shows whether there are significant differences for comparison.

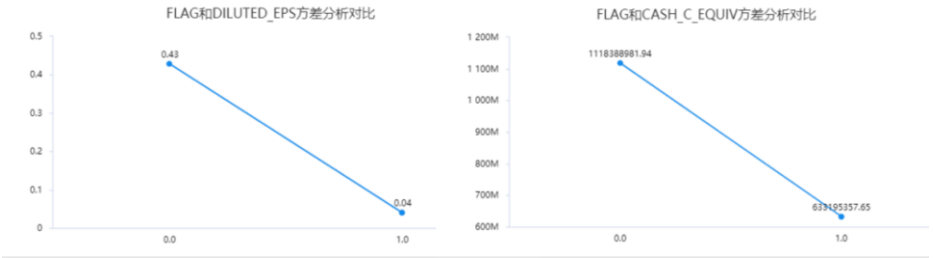

Figure 2. Comparison chart of whether there are significant differences

(1) FLAG shows a significant level of 0.01 for DILUTED_EPS (F=29.613, p=0.000), and the specific contrast difference shows that the average value of 0.0 (0.43) will be significantly higher than the average value of 1.0 (0.04);
(2) FLAG F is not significant for CASH_C_EQUIV (p>0.05).
Use t-test (the full name is independent sample t-test) to study the difference of FLAG to CASH_C_EQUIV, NOTES_RECEIV, AR, PREPAYMENT, OTH_RECEIV, etc. 74 items. Because the analysis results of all indicators are too many, it is difficult to describe, so the analysis of variance is used. For the difference between X (fixed class) and Y (quantitative), take any 10 items shown in table 3 as an example for detailed analysis.
The t-test studies the difference between X (determined category) and Y (quantitative):
(1) Analyze whether there is significance between X and Y (p value is less than 0.05 or 0.01);
(2) If it is significant; compare the average value and describe the specific difference;
(3) Summarize the analysis.

Table 3. Analysis results of 10 random indicators t-test
Analysis chart of random 10 indexes in the total number

分析项	S^2pooled(联合方差)	Cohen's d值
CASH_C_EQUIV	19240079976088072192.000	0.111
NOTES_RECEIV	3347203098608331776.000	0.075
AR	7483861660507940864.000	0.026
PREPAYMENT	5958126709196272 64.000	0.034
OTH_RECEIV	2350448051080611 52.000	0.003
INVENTORIES	1012088138918582 8864.000	0.117
OTH_CA	1974368026811378688.000	0.111
T_CA	1548033286051446 12864.000	0.105
FIXED_ASSETS	3317170512482898 7392.000	0.078
CIP	2963789607620398080.000	0.012

From the overall analysis results, different FLAG samples will not show significance for a total of 68 items such as CASH_C_EQUIV, NOTES_RECEIV, AR, PREPAYMENT, etc. (p>0.05), which means that different FLAG samples show consistency for all of these indicators, and There is no difference. In addition, FLAG samples are significant for C_PAID_FOR_DEBTS, C_PAID_TO_FOR_EMPL, DILUTED_EPS and other 16 items (p<0.05), which means that different FLAG samples are different for C_PAID_FOR_DEBTS, C_PAID_TO_FOR_EMPL, DILUTED_EPS, etc. Compare the analysis of a pair of oppositely significant indicators, and use figure 3 to show whether there is a significant indicator difference for comparison.

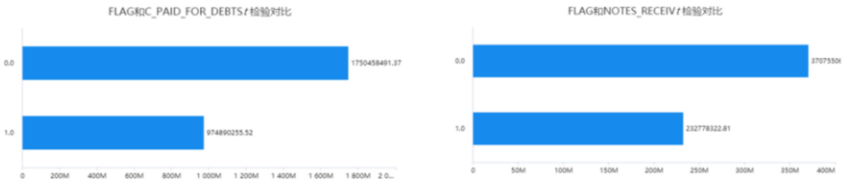

Figure 3. Comparison of whether there are significant differences in indicators

6. Experimental Results

For task 1, due to the existence of a large number of outliers in the original data given, this paper uses data depth refinement to characterize the data, make good use of data resources, and effectively apply big data preprocessing technology, feature engineering, and algorithms to this financial In the data analysis,[9] the correlation and causality between the things hidden behind the data are obtained, and the characteristic indicators related to the falsification of financial data in various industries are determined by discovering and comparing abnormal indicators. After analysis of variance, t test and principal component analysis, the final selected feature factors are "BASIC_EPS", "DILUTED_EPS", "C_PAID_TO_FOR_EMPL", "AP", "C_PAID_FOR_DEBTS", IAB", "FIXED_ASSETS", etc. Of the seven indicators.

It can reflect 90% of the data information of the original indicator, as shown in table 4.

Table 4. Feature name Feature value Feature interpretation

Feature name	Feature value	Feature interpretation
X1	BASIC_EPS	Basic earnings per share
X2	C_INF_FR_INVEST_A	Subtotal of cash inflows from investing activities
X3	DILUTED_EPS	Diluted earnings per share
X4	C_PAID_TO_FOR_EMPL	Cash paid to and for employees
X5	C_PAID_FOR_DEBTS	Cash paid for debt repayment
X6	FIXED_ASSETS	Assets and liabilities
X7	AP	accounts payable

The vector machine model is used to analyze whether the listed companies are fraudulent or not. The results show that the model selected in this paper performs well on the test set, and the accuracy rate obtained is 96.9%. Finally, in order to verify the superiority of the vector machine model, [10] this paper compares the experimental results of three models including logistic regression, decision tree and vector machine, and the results prove that the vector machine model established in this paper has the best effect.

For task two, in the selection of predictive models, this article chooses to divide the data into training set and test set, and model them on support vector machines, decision trees and logistic regression learning classification algorithms to evaluate the prediction effects of each model. Finally, debug and optimize the regularization parameter C and kernel parameters of the error term of the support vector machine model with the best prediction effect. The AUC score of the support vector machine model on the test set is 85.64%, which is higher than all the basic classifiers. It can be seen that the established The model is relatively stable, there is no serious overfitting, and the effect is good. And find out the prediction result of whether the company is fraudulent in the sixth year: 0.85% of listed companies are fraudulent, that is, 18 listed companies are fraudulent, and 2109 listed companies are not fraudulent.

For task three, add a non-manufacturing screening condition on the basis of task two, and use the established prediction model that does not have serious overfitting and has good effect to find listed companies in various industries except manufacturing Financial data falsification prediction results: 2.12% of listed companies have falsified, that is, 27 listed companies have falsified and 1244 listed companies have not falsified.

Acknowledgments

Fund Project: Supported by the Industry-University-Research Innovation Fund of the Science and Technology Development Center of the Ministry of Education (2018A02016)Wuhan University Scientific Research Innovation Platform-Private Enterprise Value Evaluation and Innovation Research Center.Introduction to the author: Gong Mingmin, female, born in 1977, master degree, associate professor, research interests in image processing, machine learning.

References

[1] Cong Ruonan, Shi Guoqing, Fu Yonghong, Shao Yixuan. Analysis and Countermeasures of Financial Fraud in Listed Companies-Taking Kang Dexin as an Example [J]. Chinese Township Enterprise Accounting, 2020(07): 79-80.
[2] Wei Jian, Zhao Hongtao, Jia Heping. Short-term traffic flow forecast based on improved LSTM model[J]. Science and Technology Innovation and Application, 2021, 11(12): 25-27.
[3] Huang Hengqiu. Python big data analysis and mining practice [M]. Beijing: People's Posts and Telecommunications Press. 2020: 104-107, 85-91.
[4] Yang Xingli. Overview of performance metrics for classification learning algorithms [J/OL]. Computer Science: 1-17 [2021-05-07].
[5] Lu Libiao, Ni Zhaomin, Du Bin. The efficacy analysis of DKI and DWI in the diagnosis of benign and malignant breast diseases[J]. Zhejiang Traumatic Surgery, 2021, 26(02): 366-368.
[6] Katirci R. and Aktas H. and Zontul M.. The prediction of the ZnNi thickness and Ni% of ZnNi alloy electroplating using a machine learning method[J]. Transactions(3) of the IMF, 2021, 162 -168.
[7] Mainak Bandyopadhyay and Minakhi Rout and Suresh Chandra Satapathy. Machine Learning Applications for Urban Computing[M].
[8] Manjurul Islam MM and Prosvirin Alexander E. and Kim Jong-Myon. Data-driven prognostic scheme for rolling-element bearings using a new health index and variants of least-square support machine and machinery [J]. , 2021, 160.
[9] Ma Yu and Sclavounos Paul D.. Support Vector Machinery Model of the Nonlinear Hydrodynamics of Fixed Cylinders[J]. J. Offshore Mech. Arc. Eng, 2021, 143(5).
[10] Zhan Shexia et al. Evaluation of source water quality and the influencing factors: A case study of Macau[J]. Physics and Chemistry of the Earth, 2021, 123.