Eng Int Syst (2022) 3: 211–216

Engineering
Intelligent Systems

Corporate Financial Fraud Identification and Crisis Forewarning Based on The Partial Least Squares Method

Yuying Li*

North China University of Science and Technology, Tangshan, Hebei 063210, China

Conducting research into financial fraud and predicting financial crises is an important research issue. In this paper, we describe the development of a model based on partial least squares (PLS) combined with a support vector machine (SVM). Components were extracted through PLS and were used as the input of SVM. Then, data were distinguished using SVM. An analysis was carried out on the collected samples. The results show that after component extraction by PLS, the model achieved an average accuracy of 83.61% for fraud identification, and the model also achieved good performance in relation to financial crisis forewarning, with an accuracy of 95%, 90.51% and 88.45% for Year T-1, T-2 and T-3, respectively. The results verify the reliability of the method and that the method can be applied in practice.

Keywords: partial least squares, financial fraud identification, financial crisis warning, support vector machine

1. INTRODUCTION

Due to the immaturity of the capital market and increasing economic pressure, enterprises are facing problems of financial fraud and financial crises. Therefore, to reduce economic losses, maintain market stability, and promote sustainable economic development, effective methods must be used to identify and warn enterprises about financial fraud and crises. An increasing number of algorithms and models have been successfully applied in related research (Kim et al., 2016; Xiong et al., 2021) established a financial fraud identification model based on back-propagation (BP) neural networks (BPNNs) and achieved an accuracy of 88.14%, a recall rate of 70.96% and an error rate of 7.19%. Qian et al. (2019) proposed a fuzzy cognitive graph-based method for financial crisis forewarning, analyzed the relationship between financial data, calculated the crisis value of the system, and verified the timeliness and efficiency of the method by validating the data of listed companies. Shang et al. (2021) mined financial indicators with big data technology, combined the fuzzy clustering method with association rules to determine the set of frequent fuzzy options, obtained fuzzy association rules, and validated the method on the relevant data of listed companies. Sun et al. (2021) constructed a financial forewarning model using BPNN for the financial crisis of mining companies and found that the method had a high prediction accuracy and the financial situation of Chinese A-share mining listed companies in 2018 was predicted to be overall good. Partial least squares (PLS) is a method of data analysis, which has more significant effects than traditional regression analysis when the number of variables is large and the amount of observation data is small. Therefore, based on PLS, a model combined with a support vector machine (SVM) was established to study the identification of corporate

*Email: tyu8ma@163.com

financial fraud and crisis forewarning in this study. An example analysis was carried out to verify the reliability of the method. This work makes several contributions to the smooth operation and sustainable development of the economy.

2. PARTIAL LEAST SQUARES-BASED MODELING

Regression analysis has a very wide application in data analysis (Ma, 2020) and PLS is a new regression analysis method (Hulland, 2015). PLS uses the technique of synthesis and screening of information in modeling and can simplify the data model while modeling, which has been very widely used in fields such as chemistry (Mahesh et al., 2015) and finance (Nitzl, 2016), and its theory and algorithm are developing.

Suppose there are p independent variables $(x_1, x_2, \cdots, x_p)$ and q dependent variables $(y_i, y_2, \cdots, y_q)$. According to the PLS method, component t_1 (the linear combination of $(x_1, x_2, \cdots, x_p)$) and component u_1 (the linear combination of $(y_i, y_2, \cdots, y_q)$) are extracted from X and Y. The two components meet: ? with as much variation as possible; and ? with the largest correlation degree.

It is assumed that the interpretability of an extracted component t_h for independent variable x_j is $r^2(x_j, t_h)$ and the interpretability for independent variables is $r^2(y_k, t_h)$. Then, the interpretability of component t_h for X and Y can be written as:

$$R(X; t_h) = \frac{1}{p}\sum_{j=1}^{p} r^2(x_j, t_h), \tag{1}$$

$$R(Y; t_h) = \frac{1}{q}\sum_{j=1}^{q} r^2(y_j, t_h). \tag{2}$$

Then, them cumulative explanatory power of the individual components for X and Y can be written as:

$$R(X; t_i, t_2, \cdots, t_m) = \sum_{h=1}^{m} R(X; t_h), \tag{3}$$

$$R(Y; t_i, t_2, \cdots, t_m) = \sum_{h=1}^{m} R(Y; t_h). \tag{4}$$

After extracting components t_1 and u_1, the regression of X on t_1 and the regression of Y on u_1 are carried out to verify the accuracy of the model. The algorithm is terminated if satisfactory accuracy is already achieved; otherwise, the second round of component extraction is required until satisfactory accuracy is achieved.

The PLS method can filter the relevant variables to improve the accuracy of subsequent fraud identification and crisis warnings, which can be considered a binary classification problem. In this paper, the SVM algorithm was selected for data discrimination.

SVM is a method based on statistical theory (Shi et al., 2015) and has outstanding applications in data prediction (Bui et al., 2016), classification (Ghimire and Lee, 2016) and fault diagnosis (Fadili and Boumhidi, 2018). For sample set $\{x_i, y_i\}$, the discriminant function in the D-dimensional space can be written as:

$$g(x) = wx + b. \tag{5}$$

To make the classification interval $\frac{2}{\|w\|}$ maximum, it should satisfy:

$$y_i(wx_i + b) \geq 1, i = 1, 2, \cdots, l. \tag{6}$$

It is solved by the Lagrange function, and the optimal classification function can be written as:

$$f(x) = sgn(wx + b) = sgn\left\{\sum_{i=1}^{l} a_i y_i K(x_i \cdot x) + b\right\}, \tag{7}$$

where $K(x_i \cdot x)$ refers to the kernel function. In this paper, the radial basis function (RBF) kernel function is used.

3. CORPORATE FINANCIAL FRAUD IDENTIFICATION AND CRISIS WARNING

3.1 Corporate Financial Fraud Identification and Feature Selection

Financial fraud refers to the act of making a financial gain (Li and Man, 2015) through deception or other means and causing others to suffer losses. It is generally realized through false financial reports (Craja et al., 2020) or the misappropriation of assets. The earliest financial fraud was reported in 1720. By bribing the government (Barnard, 2016), the South Sea Company proposed a plan to exchange stocks for national bonds, which led to the emergence of many "bubble companies" and caused market chaos. In 1720, Congress passed the Bubble Act to prevent other companies from competing with the South Sea Company for investors' capital, which led to a sharp drop in the stock price of the South Sea Company and damaged the government's integrity. Since the South Sea Company incident, cases of financial fraud have become common. For example, the Madoff scandal resulted in $21.2 billion in cash losses to investors (Hardy et al., 2020), Freddie Mac understated earnings of over $5 billion (Launder, 2006), and Luckin Coffee shares plunged 85% after it was revealed it fabricated a $2.2 billion deal. Current financial fraud includes fictitious transactions, no recording of related transactions, manipulation of accounting entries, fraudulent asset valuation, falsification of records or documents, manipulation of expected earnings and improper changes in accounting policies.

The characteristics of financial fraud can be divided into two, namely non-financial indicators i.e., related to corporate governance, and financial indicators, i.e., the data in financial statements. The fraudulent characteristic of the existing financial indicators is more reflected in the inconsistency between financial data and industry data. The research variables selected for this study are shown in Table 1.

Table 1 Research variables for financial fraud identification.

	Symbols	Name
Dependent variable	Y	Fraud or not
Independent variable	X_1	Turnover of account receivable
	X_2	Asset turnover ratio
	X_3	Inventory turnover ratio
	X_4	Current ratio
	X_5	Quick ratio
	X_6	Return on net assets
	X_7	Total assets growth rate
	X_8	Cash flow ratio
	X_9	Cash flow per share
	X_{10}	Board size
	X_{11}	Proportion of independent directors
	X_{12}	Ownership concentration
	X_{13}	Degree of affiliate transaction

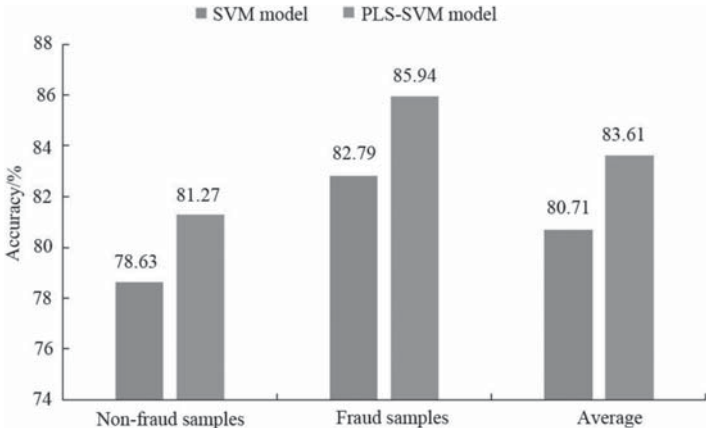

Figure 1 Performance comparison of different fraud identification models.

3.2 Forewarning of Corporate Financial Crisis and Selection of Characteristics

A financial crisis refers to a situation where an enterprise's operation is unsustainable, going from business failure, to insolvency, to bankruptcy. The criterion for financial crisis as defined by Chinese scholars is whether the company is marked as ST or *ST by the Securities and Futures Commission. Forewarning of a financial crisis involves determining the possible risks to an enterprise using scientific methods to predict whether a financial crisis will occur.

Similar to financial fraud, the research variables of financial crisis forewarning also involve financial and non-financial indicators. The research variables selected in this paper based on information, such as business capacity and the cash flow of enterprises, are shown in Table 2.

4. EXAMPLE ANALYSIS

4.1 Validation of Financial Fraud Identification Model

The data for this study came from the China Stock Market & Accounting Research Database. Enterprises which were found guilty of fraudulent conduct between 2010 and 2019 as disclosed by the Shanghai Stock Market, Shenzhen Stock Market, the Securities and Futures Commission and other institutions between 2010 and 2019 were selected as the fraud samples. Enterprises with incomplete information were excluded. Finally, 114 fraud samples were obtained. Then, another 114 enterprises from the same industry with a similar asset size as the fraudulent enterprises but had no fraud accusations against them were selected as the non-fraud samples. The final grouping of the samples is shown in Table 3.

The performance of the model was compared in two cases, using PLS and inputting the data directly into SVM for identification and using PLS to extract the components before SVM identification and the results are shown in Figure 1.

Figure 1 shows that when using the raw data directly for identification, the SVM model achieved an accuracy of 78.63% for non-fraud samples and 82.79% for fraud samples, with an average accuracy of 80.71%; when using the PLS-extracted components as the input of SVM, the accuracy of the model for the non-fraud samples was 81.27%, which was an improvement of 2.64%; the accuracy of the model for the fraud samples was 85.94%, which was an improvement of 3.15%, and the average accuracy was 83.61%, which was an improvement of 2.9%. These results indicate that extracting components with PLS and recognizing them using SVM obtains better recognition results, which verifies the reliability of PLS in financial fraud identification.

Table 2 Identification of variables for financial crisis forewarning.

	Symbol	Name
Dependent variable	Y	Is there a financial crisis
Independent variable	X_1	Earnings per share
	X_2	Net assets per share
	X_3	Return on net assets
	X_4	Return on assets
	X_5	Current ratio
	X_6	Quick ratio
	X_7	Equity ratio
	X_8	Net profit growth rate
	X_9	Net assets growth rate
	X_{10}	Inventory turnover ratio
	X_{11}	Current asset turnover ratio
	X_{12}	Fixed asset turnover ratio
	X_{13}	Total assets turnover ratio
	X_{14}	Percentage of shares of the first majority shareholder
	X_{15}	Percentage of shares of top ten shareholders

Table 3 Grouping of enterprise samples.

	Training samples	Test samples
Fraud samples	70	44
Non-fraud samples	70	44

4.2 Validation of Financial Crisis Forewarning and Identification

The data for the study were obtained from the China Stock Market & Accounting Research Database. Forty enterprises which announced losses to be ST or *ST in 2018 and 2019 were selected as the financial crisis samples. Another forty enterprises from the same industry and with the same asset size as the selected *ST enterprises were selected as the non-fraud samples. The final sample grouping is shown in Table 4.

In financial crisis forewarning and identification, it is assumed that the year when the enterprise is announced to be ST is T, the previous year is T-1, the year before last is T-2, and three years ago is T-3. In this paper, the data from Year T-3, T-2, and T-1 were used to forewarn of financial risks in Year T, i.e., models were established respectively based on the data of Year T-3, T-2, and T-1. The performance of the models was tested.

First, the PLS application was written in MATLAB 2017b to extract the components of the 15 indicators listed in Table 2. Two partial least squares components were extracted from the data of Year T-3, T-2, and T-1, respectively. The results are shown in Table 5.

Table 5 shows that the two components extracted from Year T-1 explain 97.64% of the dependent variable, the two components extracted from Year T-2 explain 93.25% of the dependent variable, and the two components extracted from Year T-3 explain 83.64% of the dependent variable, indicating that the extracted results are reasonable.

Then, the SVM model was written in MATLAB 2017b, and the test samples were brought into the trained model for testing. The results are shown in Table 6.

Table 6 shows that the prediction performance of the model is better for the financial crisis sample than the non-fraud sample. The average accuracy of the model is 95% for Year T-1, 90.51% for Year T-2, and 88.45% for Year T-3. In terms of the annual results, the closer the year in which the crisis arose, the higher the accuracy rate. Therefore, predicting the occurrence of a corporate financial crisis in the first three years can help corporate managers make early and reasonable predictions about the corporate situation so that they can take timely measures to avoid the crisis.

5. DISCUSSION

Financial fraud can affect investors' judgment of enterprises, disrupt the development of the national economy, and is detrimental to the stable operation of society. In times of economic recession, the possibility of financial fraud increases greatly (Bãnãrescu, 2015); therefore, the early identification of companies at risk of committing fraud can effectively deter companies which are intending to commit fraud. The manual method of identifying fraud is highly subjective and time-consuming. Identifying fraud using a simple and reliable model is very important in terms of stabilizing the capital market and promoting economic development, and it can also provide a certain basis for the government to formulate anti-fraud policies and effectively reduce the occurrence of financial fraud. A financial crisis is not only related to the survival of the enterprise, it also will have an impact on investors and other stakeholders (Huang et al., 2016) and affect the healthy development of the financial industry. Therefore, a forewarning of a financial crises can alert enterprise managers to deal with the crisis in a timely manner and provide a reliable

Table 4 Grouping of enterprise samples.

	Training samples	Test samples
Sample financial crisis	30	10
Non-fraud sample	30	10

Table 5 Component extraction results of PLS.

	Component	t_1	t_2
Year T-1	Explanatory capacity	0.8684	0.1108
	Cumulative explanatory capacity	0.8684	0.9764
Year T-2	Explanatory capacity	0.5632	0.4128
	Cumulative explanatory capacity	0.5632	0.9325
Year T-3	Explanatory capacity	0.7569	0.0641
	Cumulative explanatory capacity	0.7569	0.8364

Table 6 Prediction results of the SVM model.

	0 (non-fraud)	1 (financial crisis)	Average
Year T-1	90.12%	99.87%	95.00%
Year T-2	82.37%	98.64%	90.51%
Year T-3	79.64%	97.25%	88.45%

basis for the decision-making of investors, business partners, etc., which is beneficial to the long-term development of enterprises (Lin 2021).

In this research, a model is designed based on PLS combined with SVM for the identification of corporate financial fraud and financial crisis and data is collected to analyze these two models. The results of the financial fraud analysis show that when the SVM model was used for identification, the identification accuracy of the model was 78.63% and 82.79% for the non-fraud and fraud samples, respectively, with an average accuracy of 80.71%; when taking the components extracted by PLS as the input of the SVM model before identification, the accuracy of the model was 81.27% and 85.94% for the non-fraud and fraud samples, with an average accuracy of 93.61%. In conclusion, PLS is able to effectively improve the accuracy of fraud identification. The results of the financial crisis analysis suggest that the components extracted by PLS are able to fully explain the dependent variables. The results of the crisis warning show that the average accuracy of the model for Year T-1, T-2 and T-3 is 95%, 90.51% and 88.45%, respectively; the model had a prediction accuracy of 88.45% three years before the crisis. Thus, to better cope with a financial crisis, the financial situation of enterprises should be predicted as early as possible to ensure the healthy development of the enterprise.

Although several important results were obtained in this study of corporate financial fraud and financial crisis forewarning based on PLS, there are a few shortcomings. For example, the selection of research variables was somewhat subjective, and the performance of the model needs to be further improved, both of which will be addressed in future work.

6. CONCLUSION

This paper described the design of a model which combined PLS with SVM for enterprise financial fraud identification and financial crisis forewarning and analyses the results, showing that that the PLS-SVM model achieved an accuracy of 81.27% in the identification of non-fraudulent enterprises, 85.94% for the identification of fraudulent enterprises, and the model had an accuracy of 95%, 90.51% and 88.45%, respectively, for financial crisis forewarning in Year T-1, T-2 and T-3. The results verify that the method is reliable. The method can be further applied in practice to contribute to the smooth operation of enterprises and the economy.

REFERENCES

1. Bãnãrescu, A. (2015). Detecting and Preventing Fraud with Data Analytics. *Procedia Economics and Finance*, *32*, 1827–1836.
2. Barnard, T.C. (2016). The South Sea Bubble and Ireland: Money, Banking and Investment, 1690–1721, by Patrick Walsh. *The English Historical Review*, (549), 463–465.
3. Bui, D.T., Tuan, T.A., Klempe, H., Pradhan, B. & Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, *13*(2), 361–378.
4. Craja, P., Kim, A. & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, *139*, 113421.
5. Fadili, Y. & Boumhidi, I. (2018). The adaptive fuzzy - support vector machine for fault detection and isolation in wind turbine. *Engineering Intelligent Systems*, *26*(1), 35–43.
6. Ghimire, D. & Lee, J. (2016). Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines. *Sensors*, *13*(6), 7714–7734.
7. Hardy, J., Bell, P. & Allan, D. (2020). A crime script analysis of the Madoff Investment Scheme. *Crime Prevention and Community Safety*, *22*(1), 68–97.
8. Huang, T.H., Leu, Y. & Pan, W.T. (2016). Constructing ZSCORE-based financial crisis warning models using fruit fly optimization algorithm and general regression neural network. *Kybernetes*, *45*(4), 650–665.

9. Hulland, J. (2015). Use of Partial Least Squares (PLS) in Strategic Management Research: A Review of Four Recent Studies. *Strategic Management Journal*, *20*(2), 195–204.
10. Kim, Y.J., Baik, B. & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, *62*(nov.15), 32–43.
11. Launder, W. (2006). Freddie Names New CFO Amid Accounting Overhaul. *American Banker*, *171*(199), 16–16.
12. Li, H. & Man, L.W. (2015). Financial Fraud Detection by using Grammar-based Multi-objective Genetic Programming with ensemble learning. *Evolutionary Computation*.
13. Lin, J. (2021). Design of enterprise financial early warning model based on complex embedded system. *Microprocessors and Microsystems*, *80*, 103532.
14. Ma, H. (2020). Enterprise Performance Regression Model Analysis Based On Management Accounting. *Engineering Intelligent Systems*, *28*(2), 163–167.
15. Mahesh, S., Jayas, D.S., Paliwal, J. & White, N.D.G. (2015). Comparison of Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR) Methods for Protein and Hardness Predictions using the Near-Infrared (NIR) Hyperspectral Images of Bulk Samples of Canadian Wheat. *Food and Bioprocess Technology*, *8*(1), 31–40.
16. Nitzl, C. (2016). The use of partial least squares structural equation modelling (PLS-SEM) in management accounting research: Directions for future theory development. *Journal of Accounting Literature*, 19–35.
17. Qian, W., Hui, F., Wang, X. & Ding, Q. (2019). Research on early warning and monitoring algorithm of financial crisis based on fuzzy cognitive map. *Cluster Computing*, *22*(2), 1–9.
18. Shang, H., Lu, D. & Zhou, Q. (2021). Early warning of enterprise finance risk of big data mining in internet of things based on fuzzy association rules. *Neural Computing and Applications*, *33*(9), 3901–3909.
19. Shi, J., Lee, W.J., Liu, Y., Yang, Y. & Wang, P. (2015). Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines. *IEEE Transactions on Industry Applications*, *48*(3), 1064–1069.
20. Sun, X. & Lei, Y. (2021). Research on financial early warning of mining listed companies based on BP neural network model. *Resources Policy*, *73*(2), 102223.
21. Xiong, T., Ma, Z., Li, Z. & Dai, J. (2021). The analysis of influence mechanism for internet financial fraud identification and user behavior based on machine learning approaches. *International Journal of System Assurance Engineering and Management*, 1–12.