International Journal of Database Theory and Application
Vol.7, No.1 (2014), pp.71-82
http://dx.doi.org/10.14257/ijdta.2014.7.1.07

Identifying the Fraudulent Financial Information Based on Data Classification Method

Zhang Chen

Logistics Management Department, Hunan Communication Polytechnic
No. 635, South Shaoshan Road, Yuhua District, Changsha City,
Hunan Province, China

zhangchen28@163.com

Abstract

Finance fraud of companies is an international difficult problem with a long history. The finance fraud problem is concerned by lots of people. Some researchers make a lot of qualitative or quantitative researches and get some valuable conclusions. In this article , we mainly applies empirical research method, combined with normative research method. First of all, this paper reviews the relevant literatures of financial fraud detecting of listed companies, expounds existing research results from the aspects of motives, signs and detecting methods. We appraise these results are ordering to national conditions and characteristics, analyze the definition of financial fraud. We established a new method which is partial least squares (PLS) and support vector regression (SVR) to solve the above problem in finance. The PLS are able to reduce dimension effectively, acquire nonlinear factor matrix, and SVR has many advantages, such as high imitation degree, effective classification and strong robustness. The model which combines PLS and SVR has great recognition effect.

Keywords: *partial least squares, support vector regression, Finance fraud, accounting*

1. Introduction

The effect of company financial fraud include the following aspects: First, weakening the function of market resource allocation. To make the function of market resource allocation into full play, the most important premise is the demand of information is true, effective and open. However, the company's financial fraud that giving the false accounting information will mislead the market, and will lead the unreasonable allocation of resources. Secondly, misleading the investors. The investors will not make the investment blindly. Before making the investment, the investors will demand to know the operation status of the listing Corporation, and they will usually through the analysis of the financial statements of the company for this purpose. But the false accounting information will lead investors can not obtain the enterprise of real business. This will inevitably affect the decision-making and judgment of the investor, and will make investors suffer economic losses. Thirdly, damaging the related institutional interests. In order to obtain sufficient development funds, the company will loans to banks or other financial institutions. The company may be use the way to credit to purchase related products or services from their suppliers. In summary, the financial fraud is bound to affect the vital interests of the different institutions.

There are many research and analysis of scholars on the company's financial fraud problem. The preference [1] use the method of analytical review to analysis the accounting

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2014 SERSC

report, abnormal financial data and the financial indicators. The study confirmed the method of analytical review is effective for finding financial reporting fraud. The application of Logistic method to carry on the research of financial fraud recognition is more. Preference [2] uses the listing Corporation which is punished by CSRC because of the financial reporting fraud as a sample. From two aspects of ownership structure and board characteristics, he uses the Logistic regression analyses on the relationship between corporate governance and financial reporting fraud. The results show that, the fraud listing Corporation reflects the high proportion of corporate shares, the proportion of executive directors, the level of internal control, the size of the board of supervisors and lower proportion of shares in circulation. Preference [3] selects the proper financial indexes to establish the multivariate linear probability mode. His research shows that the identification of logistic model is better than the LMP model. Preference [4] use the analytical review methods and four kinds of statistical analysis methods (the method of single factor analysis, multivariate discriminate analysis, linear probability model and Logistic regression method) combined to establish the financial reporting fraud identification model. The study compare the recognition results of four kinds of recognition model, and find a effective financial method which is used to judge the enterprise credit status. Preference [5] put forward the data mining technology to improve the identification efficiency of false financial reports. They using data mining method on the establishment of rules and the corresponding algorithm, and put forward the solution and the corresponding procedures. Then, they use the RMTS model as an example of false financial reporting recognition system. Preference [6] using artificial neural network aided recognition system for listing Corporation that has the false financial statements. He pointed out that the artificial neural network is self-organizing, adaptive, self-learning, and the method is particularly suitable for processing massive in complex environment. Preference [7] argues that the change of financial data in the real financial statements has certain regularity. if the change is abnormal, there may be has the false composition data. It can be isolated by the data mining analysis function in the abnormal changes of recognition of financial report. Preference [8] using a neural network that is based on 36 listed companies financial index and the equity structure index as the modeling samples. After the sample training and learning, it has achieved a high correct rate. Preference [9] uses the data mining technology which can handle the massive information to identify the fraudulent financial reporting. They analyzed the necessity, feasibility, technical advantages to improve the efficiency and the effect of recognition. Preference [10] by using the LIBSVM software package, the application of support vector machine technology in financial report authenticity identification. The best recognition results for test set of sample are that the correct recognition rate reached seventy percent.

Financial fraud research in China is still in the stage of exploration, more and more scholars study is normative, and the empirical research is few. The main research methods are analytical review, single factor analysis method, linear probability model, and Logistic regression. The data mining technology is used less. With the stock market gradually improve and mature, empirical model results will be emerging. The application of neural network technology, support vector machine technology and other emerging data mining technology will be more widely.

We will discuss bout the basic model of SVR in Section 2 and The types and selection of Kernel function and the new method which is the PLS-SVR model will proposed in Section 3 and 4. In Section 5, the Simulation results of company financial fraud will be presented. Finally, conclusion is made in the last section.

 Copyright ⓒ 2014 SERSC

Considering the most classification methods and the characteristics of financial data, this paper is based on the classification method which can deal with the false financial report, and the flow figure is shown in below.

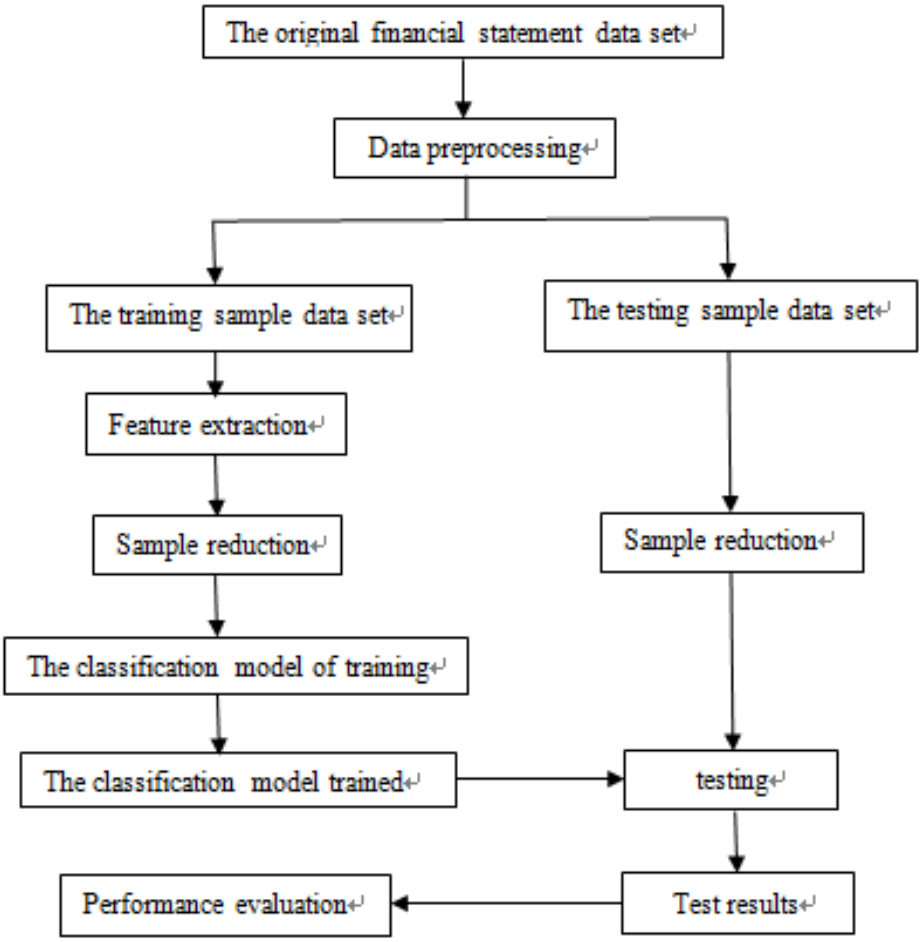

Figure 1. The flow figure of financial fraud identification which is based on data classification method

2. The Basic Model of SVR

$LS-SVR$ expands standard SVR by optimizing the square of relaxation factors and converting the constraints of inequality to equality, so the quadratic programming problem in traditional SVR becomes linear simultaneous equations, thus the calculating difficulty reduces a lot in company with the solution high efficiency and convergence speeding up.

The basic method of SVR :

Define $x \in R^n$ and $y \in R$, let R^n be the input space, by nonlinear transformation $\phi(\cdot)$, we let in the input space x map into a high dimensional characteristic space where we use the linear function to fit sample data while making sure the generalization.

In the characteristic space, the linear estimation function is defined as:

Copyright ⓒ 2014 SERSC

$$y = f(x,\omega) = \omega^T \phi(x) + b \tag{1}$$

Where ω is the weight and b is the skewness.

The aim function is:

$$\min_{\omega,b,\xi} J(\omega,b,\xi) = \frac{1}{2}\omega^T\omega + \frac{1}{2}C\sum_{i=1}^{N}\xi_i^2 \tag{2}$$

s.t.

$$y_i = \phi(x_i)\omega + b + \xi_i \ \ i = 1,\cdots,N \tag{3}$$

Where $\omega \in R^h$ is the weight vector and $\phi(\cdot)$ is non-linear mapping function, $\xi_i \in R^{N\times 1}$ is relaxation factor, $b \in R$ is the skewness while $C > 0$ is penalty factor.

Importing factors, $\alpha_i \in R^{N\times 1}$, we can easily get the function as:

$$L(\omega,b,\xi_i,\alpha_i) = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}C\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\alpha_i\left[\phi(x_i)\omega + b + \xi_i - y_i\right] \tag{4}$$

According to the KTT we get

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{N}\alpha_i\phi(x_i) = 0 \\ \dfrac{\partial L}{\partial b} = \sum_{i=1}^{N}\alpha_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = \alpha_i - C\xi_i = 0 \\ \dfrac{\partial L}{\partial \alpha_i} = \phi(x_i) + b + \xi_i - y_i = 0 \end{cases} \tag{5}$$

$$\begin{bmatrix} 0 & E^T \\ E & \phi\phi^T + C^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{6}$$

Where E is the matrix whose elements are all 1, I is a $N \times N$ identity matrix.

Inner product of regression in non-linear function can be replaced by kernel function satisfied *Mercer*. Let $\Omega_{ij} = \phi\phi^T$,

then

$$\Omega_{ij} = \phi(x_i)^T\phi(x_j) = K(x_i,x_j) \tag{7}$$

We then have the $LS-SVR$ regression function model

$$f(x) = \sum_{i=1}^{N}\alpha_i K(x_i,x_j) + b \tag{8}$$

Then, we will give the schematic diagram of the support vector classification.

 Copyright ⓒ 2014 SERSC

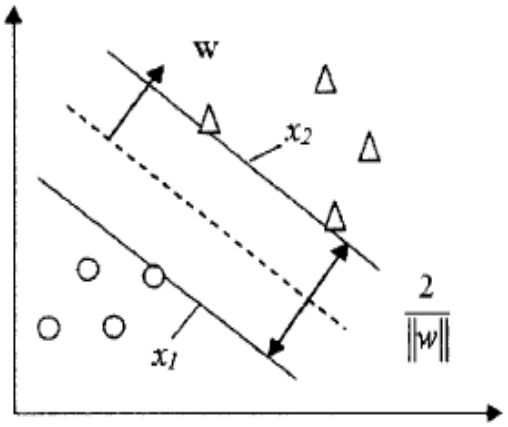

Figure 2. The schematic diagram of the support vector classification

3. The Types and Selection of Kernel Function

In general, the kernel function is commonly used as the linear kernel function, polynomial kernel function, the radial basis kernel function, and sigmoid kernel function. The functions are as follows:

(1) Linear kernel function

$$K(x, x_i) = x * x_i$$

(2) Polynomial kernel function

$$K(x, x_i) = [(x * x_i) + 1]^d$$

Where the d is the order of the polynomial

(3) Radial basis kernel function

$$K(x, x_i) = \exp(-\|x - x_i\|^2 / 2\sigma^2)$$

Where the σ is width of the kernel function

(4) Sigmoid kernel function

$$K(x, x_i) = \tanh(\gamma(x * x_i) + c)$$

Commonly used kernel function can be divided into two categories: one category is the global kernel function, the other is local kernel functions: The linear kernel function, polynomial kernel function, and Sigmoid kernel function is a global common kernel function

4. The PLS-SVR Model with Mixed Kernel Function

The PLS method has been used in this research to reduce the size of the input data. Next, we will introduce the partial least squares algorithm, and use mixed kernel function and partial least square method to establish the prediction model of mixed kernel partial least squares support vector regression.

Copyright ⓒ 2014 SERSC

4.1 PLS Regression Model

PLS is a reasonably new Multivariate statistical method of data analysis, which is a method for constructing predictive models when the explanatory variables are many and highly collinear. Its main focus is to extract the potential components, which uses data of multiple dependent variables and independent variables for analyzing and modeling.

First, let E_0^T and F_0^T are separately the transposed matrix of E_0 and F_0, then we can obtain the eigenvector w_1 associated with the largest eigen value of the matrix $E_0^T F_0 F_0^T E_0$, the component t_1 is:

$$w_1 = \frac{E_0^T F_0}{\left\| E_0^T F_0 \right\|}; t_1 = E_0 w_1; p_1 = \frac{E_0^T t_1}{\left\| t_1 \right\|^2}; E_1 = E_0 - t_1 p_1^T \tag{9}$$

In the same way, we can obtain the eigenvector w_h associated with the largest eigen value of the matrix $E_0^T F_0 F_0^T E_0$, the component t_h is:

$$\begin{cases} w_h = \dfrac{E_{h-1}^T F_{h-1}}{\left\| E_{h-1}^T F_{h-1} \right\|}; \\ t_h = E_{h-1} w_h; \\ p_h = \dfrac{E_{h-1}^T t_h}{\left\| t_h \right\|^2}; \\ E_h = E_{h-1} - t_h p_h^T \end{cases} \tag{10}$$

If the rank of $X_{n \times p}$ is A, we can use cross validation method for identifying, then,

$$\begin{cases} E_0 = t_1 p_1' + ... + t_A p_A' \\ F_0 = t_1 r_1' + ... + t_A r_A' + F_A \end{cases} \tag{11}$$

Where $r_1', ..., r_A'$ is a row vector of regression coefficient, F_A is error matrix. Concerning the least square regression equation of F_A is

$$\hat{F}_0 = t_1 r_1 + t_2 r_2 + ... + t_h r_h \tag{12}$$

Because $t_1, t_2 ..., t_A$ can express as the linear combination of $E_{01}, E_{02} ..., E_{0A}$, Hence, according to the property of PLS regression：

$$t_i = E_{i-1} W_i = E_0 W_i^* \qquad (i = 1, 2, ..., h) \tag{13}$$

Where

$$W_i^* = \prod_{k=1}^{i-1} (I - W_k P_k^T) W_i$$

 Copyright ⓒ 2014 SERSC

Then equation (13) substitute into equation (12),

$$\begin{aligned}\hat{F}_0 &= r_1 E_0 W_1^* + r_2 E_0 W_2^* + ... + r_h E_0 W_h^* \\ &= E_0 (r_1 W_1^* + r_2 W_2^* + ... + r_h W_h^*)\end{aligned} \tag{14}$$

Let $y^* = F_0$, $x_i^* = E_{0i}$, $\alpha_i = \sum_{k=1}^{h} r_k W_{ki}^* (i = 1, 2, ..., m)$

. Then, equation （14） can be revert to the regression equation of standardized variable as follows

$$\hat{y}^* = \alpha_1 x_1^* + \alpha_2 x_2^* + ... + \alpha_m x_m^* \tag{15}$$

Equation (15) can be written down raw variable y, and its PLS regression equation of estimated value $\hat{y}$ is obtained.

4.2 PLS-SVR Model

The processing of $PLS-SVR$ is divided into following steps:

Step1：PLS for feature extraction of the raw data

From compute the equation (9) and equation (10) we can contain the vector t_i, p_i and w_i.They respectively constitute the score matrix $T_{train} = [t_1, ..., t_h]$, load matrix $p = [p_1, ..., p_h]$ and correlation coefficient matrix $w = [w_1, ..., w_h]$ of training samples.

Step2：$LS-SVR$ Modeling

After h dimensions have been extracted. Which can use T_{train}, y_{train} train the $LS-SVR$ model, it contain $Lagrange$ multiplier and bias term b of the optimal parameter. On this basis, the following equation can be written,

$$\begin{bmatrix} 0 & E^T \\ E & \phi\phi^T + C^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y_{train} \end{bmatrix} \tag{16}$$

Then using equation (16) can result in coefficient b and α.

Step3：PLS-SVR modeling

Calculate the prediction value of test sample data is

$$y_{predict}(t) = \sum_{i=1}^{N} \alpha_i K_{mix}(x_i, x_j) + b \tag{17}$$

The flow figure as shown in Figure 3.

Copyright ⓒ 2014 SERSC

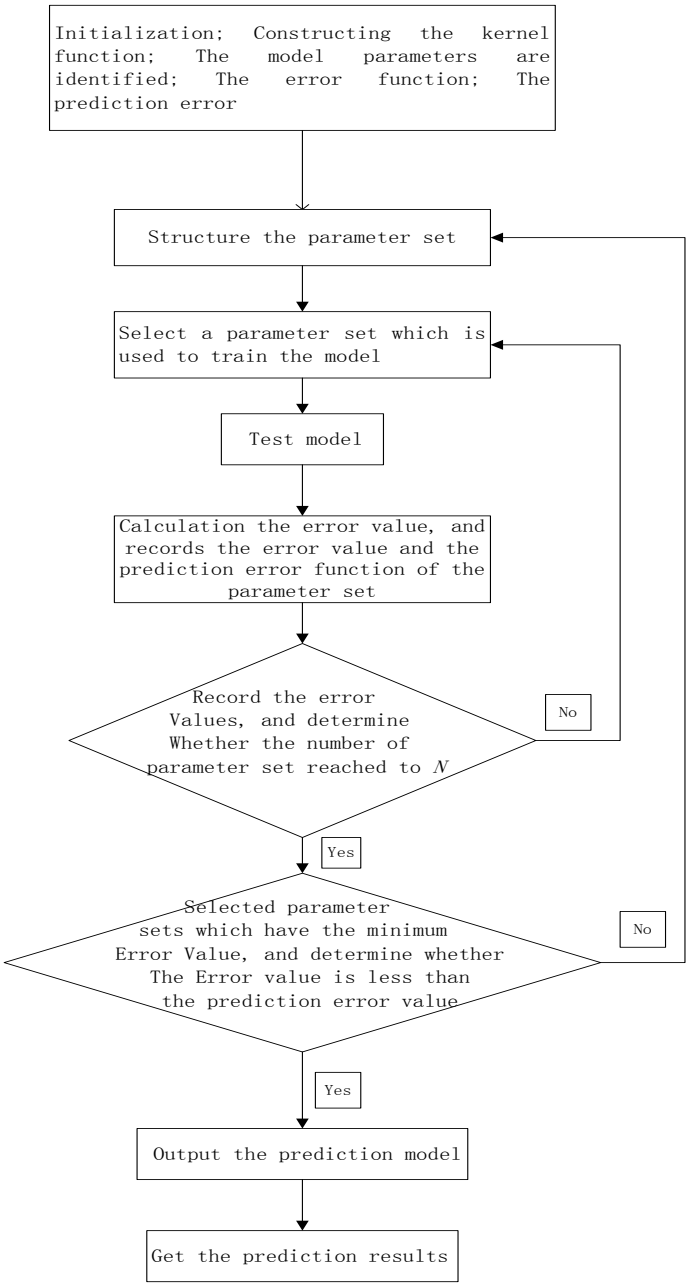

Figure 3. The Flow Figure of Improved Support Vector Machine Model

5. The Simulation and Conclusion

5.1 Sample selection

Taking into account the timeliness of research and the availability of data, we use the listing Corporation which is continuous operation and not cancelled the qualifications as the basis for modeling samples between the 2005 and 2008. Then, we choice the listing Corporation as a basis for testing sample between 2009 and 2012. Because China securities market started late, company was found to have no financial fraud was few, and many

 Copyright ⓒ 2014 SERSC

samples were occurred in many years ago. That is not conducive to the correct understanding of the characteristics of financial fraud. Therefore, we choose the sample of false financial reports according to the following standards:

(1) the audit opinion as "declined to comment" or "negative opinion" or "disclaimer of opinion"

(2) the time is not earlier than 2005.

5.2 Select the index variable

We select the 15 index sets can best embody the listing Corporation's financial ability. They can be expressed by the following table:

Table 1. Variable ordering table

Serial number	Index	Serial number	Index
1	The actual income tax rate	9	Assets and financial expense rate
2	Asset growth rate	10	Liquidity financial expense rate
3	Net income expense rate	11	Cash received growth rate
4	Current asset growth rate	12	The cumulative rate of retained earnings
5	The rate of assets and liabilities	13	Return on assets
6	Main business net income growth rate	14	The cost of assets rate
7	The assets turnover	15	Increase the amount of working capital ratio
8	The ratio of working capital		

According to the related knowledge of support vector machine which is described in paragraph second, we express the false financial statements samples by 1, and the normal samples are -1. So you can according to the test result is 1 or -1 to judge whether the test sample is false samples or normal samples.

5.3 Results and conclusion

We use 100 companies as the training samples for training which from 2005 to 2008. Then, we choose the 200 companies training samples from 2009 to 2012 for testing. We are training and testing the all 15 identification variables, the first 10 identification variables, the first 6 identification variables, and the first 3 Identification variables. We first give the program code of the support vector machine which thought the cross examination.

Copyright ⓒ 2014 SERSC

Table 2. Parameter optimization algorithm

```
PARAMETER-OPTIMIZATION-ENUMERATE( train set )
1     bestMse = 0
2     bestc = 0
3     bestg = 0
4     for c = 2 ^ ( cmin ) : 2 ^ ( cmax )
5         for g = 2 ^ ( gmin ) : 2 ^ ( gmax )
6             for run = 1 : k
7                 Train( run ) as validate set
8                 others as train set
9                 record Mse( run )
10            end
11            cv = ( Mse( 1 ) + Mse( 2 ) + ··· + Mse( k ) ) / k
12            if ( cv < bestMse )
13              bestMse = cv
14                  bestc = c
15                  bestg = g
16            end
17        end
18    end
```

Secondly, we give experimental results of four experiments.
Select 15 variables, sample recognition results as shown in Table 3.

Table 3. Results table of 15 variables

The sample	the number of T samples	the number of Correct recognition	accuracy of recognition	the number of F samples	the number of Correct recognition	accuracy of recognition	Overall accuracy
The training sample	100	99	99%	100	99	99%	99%
2009	50	48	96%	50	47	94%	95%
2010	50	47	94%	50	48	96%	95%
2011	50	49	98%	50	49	98%	98%
2012	50	48	96%	50	46	92%	94%
Overall	200	192	96%	200	190	95%	95.5%

Select 9 variables, sample recognition results as shown in Table 4.

Copyright ⓒ 2014 SERSC

Table 4. Results table of 9 variables

The sample	the number of T samples	the number of Correct recognition	accuracy of recognition	the number of F samples	the number of Correct recognition	accuracy of recognition	Overall accuracy
The training sample	100	99	99%	100	99	99%	99%
2009	50	47	94%	50	45	90%	92%
2010	50	46	92%	50	47	94%	93%
2011	50	47	94%	50	42	84%	89%
2012	50	44	88%	50	45	90%	89%
Overall	200	184	92%	200	179	89.5%	90.75%

Select 6 variables, sample recognition results as shown in Table 5.

Table 5. Results table of 6 variables

The sample	the number of T samples	the number of Correct recognition	accuracy of recognition	the number of F samples	the number of Correct recognition	accuracy of recognition	Overall accuracy
The training sample	100	99	99%	100	99	99%	99%
2009	50	42	84%	50	43	86%	85%
2010	50	43	86%	50	45	90%	88%
2011	50	41	82%	50	42	84%	83%
2012	50	40	80%	50	44	88%	84%
Overall	200	166	84%	200	174	87%	85%

Select 3 variables, sample recognition results as shown in Table 6.

Copyright ⓒ 2014 SERSC

Table 6. Results table of 3 variables

The sample	the number of T samples	the number of Correct recognition	accuracy of recognition	the number of F samples	the number of Correct recognition	accuracy of recognition	Overall accuracy
The training sample	100	99	99%	100	99	99%	99%
2009	50	39	78%	50	37	74%	76%
2010	50	40	80%	50	38	76%	78%
2011	50	37	74%	50	39	78%	76%
2012	50	40	80%	50	36	72%	76%
Overall	200	156	78%	200	150	75%	76.5%

The above the four cases, test sample rate of the 15 variables is the highest recognition model which can reach the 95.5%. It is the highest recognition accuracy rate of classification methods used in this paper. We can learn from the results that with the decrease in the identification of variables, the recognition accuracy will decline. It confirms the importance of 15 financial indicators to identify the financial fraud.

References

[1] Friedman, "Case study: Pay attention to warning signals", Journal of Accountancy, vol. 10, (**1995)**, pp. 65-80.
[2] Kyoung, "Financial time series forecasting using support vector machines", Neurocomputing, vol. 55, **(2003)**, pp. 307-319.
[3] L. j. Cao and F. E. H. Tay, "Financial time series forecasting using support vector machines", Neural Computing and Applications, vol. 10, **(2001)**, pp. 184-192.
[4] M. Firth and P. L. L. Mo, "Financial Statement Frauds and Auditor Sanetions: An Analysis of Enforcement Acions in China", Journal of Business Ethics, vol. 62, **(2005)**, pp. 367-381.
[5] L. j. Cao and F. E. H. Tay, "Financial time series forecasting using support vector machines", Neurocomputing, vol. 48, **(2002)**, pp. 847-861.
[6] Kononenko, "Estimating Attributes: Analysis and Extensions", Fealty of Elect. Eng. & Comp. Sci., **(2005)**.
[7] J. T. Wells, "Nothing But the Truth: Uncovering Fraudulent Disclosures", Journal of Accnt, vol. 7, **(2001)**.
[8] P. Ravi, Kumar and V. Ravi, "Bankrupted prediction in banks and firms via statistical and intelligent techniques-A review", European Journal of Operational Research, vol. 180, **(2007)**, pp. l-28.
[9] S. Kotsiantis and E. Kouma, D. TzelePis and V. T. Pakas, "Forecasting Fraudulent Financial Statements using Data Mining", International Journal of Computational Intelligence, **(2007)**.
[10] T. Van Gestel and B. Baesens, "Bayesian kernel based classification for financial distress detection", European Journal of Operational Research, vol. 172, **(2006)**, pp. 979-1003.

Authors

Zhang Chen. She received her Bachelor Degree in Accounting (1998) from Finance Department of Changsha University of Science and Technology. Now she is a teacher of Accounting at Logistics Management Department of Hunan Communication Polytechnic. Her current researches focus on the corporation finance and capital market.

 Copyright ⓒ 2014 SERSC