Research on Identification Model of Financial Fraud of Listed Company Based on Data Mining Technology

Jiaqi Hu, Xiao Chen

School of Business, Zhejiang University City College, Hangzhou, 310016, China

Keywords: Accounting information distortion, Financial fraud identification, Data mining, Classification, Support vector machine

Abstract: Corporate financial information fraud is a difficult problem and a great hazard since the birth of the listed company system. The classic method of financial fraud detection is of high accuracy, but it requires enormous manpower and material resources, which also needs a relatively long time for investigation. With the development of computer technology, data mining technology can be used as a very important tool to identify the distortion of enterprise financial accounting information. Through the modeling work of different classification and clustering algorithms, a set of identification models can be constructed with low cost and high efficiency. This paper uses the popular support vector machine as a data mining tool to identify financial fraud and achieves good results.

1. Introduction

Although governments are actively dealing with financial fraud in various industries, the situation at home and abroad is still not optimistic [1]. The study of financial fraud discrimination has undergone many changes. The main research subjects in the early years can be considered as "applications of analytical procedure law". In the later period, the extensive application of statistical models and the popularity of data mining algorithms make the research direction has changed. The main direction of this study is to use data mining technology to identify financial frauds [2].

2. Selection of Research Samples and Variables

2.1 Samples Selection

Except some enterprises linked to interim reports, quarterly reports and temporary reports, we chose companies that were publicly punished by the CSRC for their annual reports from 2000 to 2008. The 116 companies have varying degrees of deceit in their financial statements. If a company takes a lot of fraud, it will only calculate its first fake year. We must take the necessary measures in line with the principle that the sample control enterprises cannot be false.

First, to ensure the similarity between the fake sample and the control sample in the industry and the annual distribution, we should choose the same business as the industry and the year of the fake company. Secondly, it is not possible to consider the enterprises of ST, S and PT; finally, within 3 years, the audit report column of the control sample is the standard and unreserved. Opinion. Based on the above principles, after a series of selection, we finally chose the same number of trustworthy enterprises and fake enterprises. Both of them are 116 companies. The Tai'an database provides us with all the sample data.

2.2 Variables Selection

On the choice of financial indicators, this paper mainly uses Relief algorithm to analyze and screen the existing financial indicators.

According to the algorithm requirements, first, a sample is randomly selected, and the sample used for the study is from the sample set. Secondly, two samples, Near Hit and Near Miss, were selected from different groups of samples. Compared with the target samples, the two selected samples belong

DOI: 10.25236/systca.18.045

to the same kind and the same kind. However, it should be noted that these two samples are not randomly selected but are the most relevant samples. Thirdly, to update the weights of each feature, we need the following rules. First, the distinction between class attributes is the core of the algorithm and is positively related to the weight size. Therefore, to know whether the weight should be increased, the key point is to calculate the degree of association with two selected samples on a certain feature of the target sample.

In general, it is to calculate the degree of the correlation between the Near Hit and the feature first, and then calculate the Near Miss in the same way and compare the relative size of the two correlation degrees. If Near Hit is smaller, the weight of the feature can be improved, because it shows that the feature can be used to distinguish between the samples of the same class and the different categories; on the other hand, the weights should be reduced. Secondly, because the average weight is more convincing, it is necessary to calculate the average value many times. For example, to conduct random sampling, we first need to have an overall sample n. The object that is selected by the feature in the algorithm is X={x1,x2,…,xn}. Secondly, m is a random sample number. Then the N characteristic value of the i sample is expressed as xi={xi1,xi2,…,xiN}. Based on the above conditions, we can use the following formula to express the sum of the the weighted value of all the features: $W_{\mathrm{j}}^{\mathrm{i+1}} = W_j^i + diff(j,x,M(x))/m - diff(j,x,H(x))/m$.

Among them, the representative feature is j. [1:N] is the range of its value, and the samples that are randomly selected according to the requirements are represented by i. M (x) represents a heterogeneous nearest neighbor sample, and H (x) represents the nearest neighbor sample of the same kind.

To eliminate the influence of random sampling, we conducted five experiments in this experiment. And when the weights of each experiment are more than zero at the same time, we can reach the following eight indicators: (1) asset liability ratio, (2) cash ratio of operating income, (3) asset reward rate (4) total net profit rate (ROA), (5) receivable turnover, (6) inventory turnover, (7) the rate of cost during sales. (8) the turnover of mobile assets [3].

3. Basic Theory of Support Vector Machine

To improve generalization ability of classification machines, we generally believe that small sample size has absolute advantages. In the past, however, to make the expected risk meet the requirements, we often speak with experience that the experience risk determines the expected risk, but this is only applicable to the case with large sample content. In the support vector machine (SVM) idea proposed by Vapnik et al., the fixed experience risk has abandoned the previous view. The result is not only to reduce the confidence range to a certain extent, but also to minimize the impact of the sample content [4].

For the learning of machine generalization ability, the traditional concept is different from SVM [5]. The former focuses on all training samples, and the latter focuses on collecting small samples. This part is often the support vector at the boundary of different classes of samples, called SV.

For the source of support vector machines, the first is linear separable, followed by the optimal classification surface; the following graph can basically reflect the two ideas.

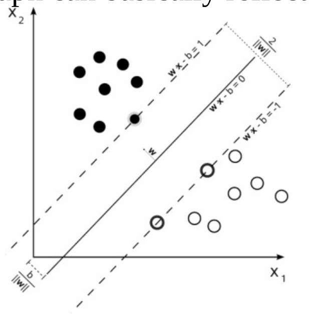

Figure 1 Schematic diagram of SVM model

Solid point: sample 1; hollow point: Sample 2; solid line (except coordinate axis): classification line; dotted line: meet three conditions: (1) through samples, (2) parallel to the solid line, (3) and the minimum spacing of classified samples. The classification interval refers to the distance between two dotted lines, and the real line in the graph is also the optimal line. One reason: it successfully separates two samples, and the reason is two: the distance of the two dotted lines is maximized. For the equation where the classification line is located, the expression can be expressed, and then the equation is normalized. The purpose is to make the equation set for a linearly separable set of samples $(x_i, y_i), i = 1, ..., n, x \in R^d, y_i \in \{+1, -1\}$. When we have the condition of $y_i[(w \bullet x_i) + b] - 1 \geq 0, i = 1, ..., \nu$, the formula is also correct.

Under this condition, the classification interval we calculate is $2 / \| w \|$. It is not difficult to find the maximum classification interval we require is set up at the minimum value of $\| w^2 \| / 2$.That is, the two items are inverse ratio. In the above picture, the dotted line is the set of samples on the class boundary (SV), and the distance reflects the classification interval, and the optimal classification surface is the classification surface when the classification interval is the largest.

According to the above analysis, the optimal classification surface is related to the maximum classification interval, and is also taken at the minimum value, so we have the inequality: $\min \Phi(W) = \frac{1}{2} \| W \|^2 = \frac{1}{2}(W \bullet W)$.

It is difficult to solve the problem directly, so in practical application, the problem of the complex spherical optimal classification surface is converted to a simple dual problem (Formula 1). This method is also called the Lagrange optimization method. But at the same time, the solution of this formula has its restrictive conditions (Formula 2).

Formula 1: $W(\alpha) \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \bullet x_j)$

Formula 2: $\sum_{i=1}^{n} y_i \alpha_i = 0, \alpha_i \geq 0, i = 1, 2, ..., n$

Combined with the two-type solution, we get the formula 3, which represents the optimal classification function:

Formula 3: $f(x) = \operatorname{sgn}\{(w \cdot x) + b\} = \operatorname{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\}$

Linear separable is the basic idea of support vector machine and is also a prerequisite for solving the optimal classification plane. However, to fully reflect the advantages of SVM, we can not only solve the linear separable problem, but also consider the linear inseparable condition, which means that the linear constraint conditions will not be applied. Therefore, we have introduced the linguistic symbols in hyperplanes, which are called relaxation variables. It is hoped that all samples can be correctly classified by hyperplanes, so that we can get $y_i[(w \cdot x_i) + b] - 1 + \xi_i \geq, i = 1, ..., n$. In this way, we go back to the steps mentioned earlier, but the equation becomes into Formula 5: $\min \phi(W) = \frac{1}{2}(W \cdot W) + C \sum_{i=1}^{n} \xi_i$. The constraint condition is $[(W \cdot X_i) + b] - 1 + \xi_i \geq 0, i = 1, 2, ..., n$.

In the process of machine learning and popularization, the error of the classification error often appears, which belongs to the system error; and because the system error has some regularity, the constant C is introduced, the purpose is to make up the error. On the principle of simplification of complex problems, the above-mentioned problems are transformed into dual problems. The maximum of the following formula $W(\alpha) \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \bullet x_j)$ when solving the optimal classification surface. The constraint condition is $\sum_{i=1}^{n} y_i \alpha_i = 0, 0 \le \alpha_i \le C, i = 1, 2, ..., n$.

From the three cases above, we know that the constraint conditions must exist regardless of the linear nonlinear solution. For this problem, a condition that satisfies all cases - Kuhn Tucker condition (KKT) appears. In terms of concrete practice, linear classification is not very common. Therefore, to better solve the problem, how to use nonlinear hyperplane becomes the key.

4. Model Construction and Analysis Based on Support Vector Machine

This experiment is implemented with R language as the tool and relying on the e1071 package. Identifying variables still selects the variables selected before, representing 1 false financial statements, and -1 indicating normal samples. After the completion of data standardization, considering that any experiment has errors, the penalty parameter C is introduced and its value is 10; and in many types of kernel functions of SVM, we choose the Gauss kernel function and make =2, according to the above conditions, the following results are obtained:

Table 1. Identification effect of support vector machine

Sample	T sample size	Number of correct recognition	Accuracy of recognition	F sample size	Number of correct recognition	Accuracy of recognition
Training sample	116	116	100.00%	116	116	100.00%
Test sample of 2013	27	26	96.30%	15	12	80.00%
Test sample of 2014	26	26	100.00%	12	9	75.00%
Test sample of 2015	23	18	78.26%	23	21	91.30%
Test sample of 2016	23	21	91.30%	23	14	60.87%
Test sample collection	99	91	91.92%	73	56	76.71%

From the table, we can see that the total recognition accuracy of T sample is higher than that of F sample, while the total recognition accuracy fluctuates at 85%. The effect is still available. However, the recognition of class F samples and the recognition of T class samples always show great differences. The recognition rate of data samples in 2014 is the highest, and it is possible to generate the possibility of model contingency. The overall recognition rate can reach 85%, which reflects the recognition effect of the model.

5. Conclusion

Using data mining technology can speed up the judgement of financial reports and its main tool is the data mining model. The support vector machine model used in this paper has obtained the better experimental results with an average recognition rate of approximately 80%. It can show that data mining technology has good work effect in the field of financial fraud identification. Compared with the traditional high labor cost fraud inspection, data mining technology has a significant advantage in marginal cost. Regulators can use cloud technology to conduct real-time monitoring and checking of financial data and strengthen the checking of financial fraud.

References

[1] Marcel, Jeremy J., and Amanda P. Cowen. "Cleaning house or jumping ship? Understanding board upheaval following financial fraud." Strategic Management Journal 35.6 (2014): 926-937.

[2] Throckmorton, Chandra S., et al. "Financial fraud detection using vocal, linguistic and financial cues." Decision Support Systems 74 (2015): 78-87.

[3] Yang D, Jiao H, Buckland R. The determinants of financial fraud in Chinese firms: Does

corporate governance as an institutional innovation matter?[J]. Technological Forecasting and Social Change, 2017, 125: 309-320.

[4] Albashrawi M, Lowell M. Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015[J]. Journal of Data Science, 2016, 14(3): 553-569.

[5] Petraşcu D, Bucur M A, Dobre E. Analysing the Management of Human Resource in Economic-Financial Fraud Investigation[J]. Procedia Economics and Finance, 2015, 27: 209-215.