A Financial Statement Fraud Detection Model Based on Hybrid Data Mining Methods

Jianrong Yao
School of Information Management and Engineering
Zhejiang University of Finance and Economics
Hangzhou, China
e-mail: y6310@163.com

Jie Zhang
School of Information Management and Engineering
Zhejiang University of Finance and Economics
Hangzhou, China
e-mail: zj_4018@163.com

Lu Wang*
School of Information Management and Engineering
Zhejiang University of Finance and Economics
Hangzhou, China
e-mail: wanglu_hit@126.com

Abstract—**Financial statement fraud has been a difficult problem for both the public and government regulators, so various data mining methods have been used for financial statement fraud detection to provide decision support for stakeholders. The purpose of this study is to propose an optimized financial fraud detection model combining feature selection and machine learning classification. The study indicated that random forest outperformed the other four methods. As to two feature selection methods, Xgboost performed better. And according to our research, 2 or 5 variables are more acceptable for models in this paper.**

Keywords-machine learning; financial statements fraud detection; feature selection

I. Introduction

Association of Certified Fraud Examiners (ACFE) defines fraud as "A kind of misrepresentation or deception that an entity or individual makes knowing that the it could result in some unauthorized benefit." According to a study conducted by the ACFE, financial statement fraud accounts for about 10% of white-collar crime. Once fraudulent accounting practices happened, various actions will be taken to maintain a sustainable appearance.

Considering financial statement fraud bring huge property damage to investors, a large number of researches have been conducted on the this area using machine learning methods such as ANN[1-3] ,DT[4] , SVM[5, 6], and text mining[7]. Meanwhile, other fraud problems like credit card fraud[8, 9], internal transaction fraud[10] and insurance fraud[11] have also been investigated. Given the different characteristics of each type of financial fraud, specific methods have been developed[12]. This paper puts forward a hybrid detection model for financial statement fraud, and this model have the advantages of (1) combining the financial and non-financial data, (2) using two feature selection methods, and (3) easy to explain.

During the research, we chose 120 fraudulent financial statements disclosed by CSRC during the period 2007–2016. Then, according to the industry and size of these companies, we found the non-fraudulent company for contract. The rest of this paper is organized as follows. In the next section, we review the previous literature on financial statement fraud detection. Section 3 introduces financial and non-financial variables. In section 4, we describe our experimental setting and procedure and section 5 introduces the methods we used in this research, including classification methods and feature selection. Section 6 gives the results of the prediction. Finally, we discuss the contribution of this study and propose future work in section 7.

II. Literature Review

In the past, people used to use expert analysis to find fraududent financial statements. In this way people may not fully analyzed the report data for its huge amount and wild range, which caused many shortcomings in judgment. In recent years, data mining method has been widely used in fraud detection to reduce the errors caused by experts' judgment, including Internet fraud detection[13-15], telecommunciations fraud detection[16, 17], financial fraud detection[18], and fraud in other areas. These frauds are hidden in huge information, and traditional experts' analysis sometimes fails to take into account the whole, and the application of data mining method solves this problem. These data mining methods include SVM[5, 6, 19], RF[20], DT[4], ANN, LR[21], etc. Most previous stuies use only 1-2 data mining techniques, without comparisons between each other; And most people overlook the importance of feature selection, which is more important than we think to some extent.

As for those variables, previous studies usually picked them from annual reports. These variables need to be comprehensive and representative enough to cover all aspects of the company's operations. For example, quick ratio and liquidity ratio reflect the solvency; sales growth rate and EPS represent the profitability; and we can see operating capacity from inventory turnover and turnover of total capital. In addition, linguistic variables from MD&A are used for emotional analysis[22]. Petr Hajek found that compared with non-fraudulent companies, fraudulent companies reported a slightly higher negative sentiment in their annual report. That is to say fraudulent companies are more likely to use negative words.

The number of fraudulent companies used in previous studies ranged from tens to thousands, and the patterns of research in each country were slightly different. Most studies have adopted a matching method to match the non-fraudulent

978-1-5386-6987-7/18/$31.00 ©2018 IEEE

companies with the fraudulent companies[4]. The year, industry, and scale are mainly used as matching criteria. Fraudulent financial statement detection is a binary classification problem so it has four possible classification results[22]. Of the four results, the cost of two types of misclassification is different, so cost-sensitive learning was used to solve this problem[23].

Our research proposes a hybrid fraudulent financial statement detection model combining the PCA and Xgboost to do feature selection, and then, SVM, RF, DT, ANN and LR were applied to construct the fraud detection model, and the classification accuracy of each model was compared to determine the optimal model.

III. Data

Our cases of fraudulent financial reports were disclosed by the China Securities Regulatory Commission (CSRC). 120 listed companies were involved during the period of 2007–2016, so a set of 120 annual reports were used as our sample. We extracted the financial and non-financial information from the annual report for these may cover all aspects of a company. Then we identified the matched sample of non-fraudulent companies based on their industry and size. Thus, we finally have 240 firms (120 fraudulent and 120 non-fraudulent).

A. *Financial Variables*

In order to identify the various types of financial reporting fraud, the selected collection of financial variables should cover as many aspects as possible. [24]provides a theoretical support to use financial variables. Table I shows our choice of financial variables.

TABLE I. Financial Variables

Variables	Variable Description
x1	Quick ratio
x2	Sales growth rate
x3	Liquidity ratio
x4	Operation revenue / average account receivable
x5	Rate of return on total assets
x6	Inventory turnover
x7	Operating profit / income before tax
x8	Net cash content of operating profit
x9	NCFPS
x10	Turnover of total capital
x11	Return on assets
x12	Operating profit ratio
x13	EPS
x14	Asset quality index
x15	Accounts receivable / current assets
x16	Growth rate of net profit
x17	Growth rate of net cash flow of operating activities

B. *Non-financial Variables*

Non-financial variables, different from financial data, are related to the corporate governance structure. Table II shows our selection of the non-financial variables.

TABLE II. NON-Financial Variables

Variables	Variable Description
x18	The proportion of the largest shareholder
x19	Board of directors
x20	Board of supervisors
x21	the proportion of independent directors
x22	LHSR

C. *Descriptive Statistics*

TABLE III. Descriptive Statistics on All Variables

	Min	Max	Mean ± S.D.
x1	0.02704	8.78993	1.06034 ± 1.05680
x2	-81.28303	1160.93152	24.42644 ± 105.28405
x3	0.16113	8.79113	1.54670 ± 1.16503
x4	0.36027	3324.42767	54.67923 ± 244.50699
x5	7.63125	220.50679	54.05857 ± 24.27745
x6	0.00365	974.98648	11.94133 ± 70.19266
x7	-20.83488	4.02214	0.47371 ± 2.30085
x8	-1405.53898	95.37726	-7.93551 ± 118.54589
x9	-3.33460	4.32000	0.27264 ± 0.81329
x10	0.00431	9.10804	0.77142 ± 0.86311
x11	-46.80306	46.18315	5.42923 ± 7.95734
x12	-475.60620	58.76700	-0.74138 ± 49.67481
x13	-1.20000	2.32000	0.20917 ± 0.43805
x14	-2.50919	0.86654	0.22006 ± 0.29058
x15	0.00008	0.65354	0.16174 ± 0.13355
x16	-5534.31950	3809.30900	32.96119 ± 554.23800
x17	-4077.52524	2177.45421	-65.03040 ± 465.20060
x18	1.27555	78.97457	33.39915 ± 15.25724
x19	1.00000	11.00000	5.15000 ± 1.56932
x20	1.00000	5.00000	2.04583 ± 0.73892
x21	0.12500	0.80000	0.38819 ± 0.08750
x22	0.00000	78.84930	22.01327 ± 20.89372

Table III shows basic descriptive statistics of these chosen companies, from which we can see that the range of several indicators is large, such as x8, x16 and x17.Due to the reasons of high dimension, high repetition, and having noise, original data need to be preprocessed, which includes data cleaning, data transformation and data reduction. In this paper we removed all the records which contain missing value and used sigmoid function to solve the problem of the inconsistent orders of magnitude.

IV. Experimental setting

First, we used feature selection to reduce its dimensionality. As we known, feature selection play an important role in pre-processing[24]. In this paper, we used PCA and Xgboost to do it. Principal component analysis (PCA) is a statistical method. By orthogonal transformation, a set of observations of possibly correlated variables converted into a set of linearly independent variables, which called principal component. As you can see from Fig. 1, six variables among all are of great importance.

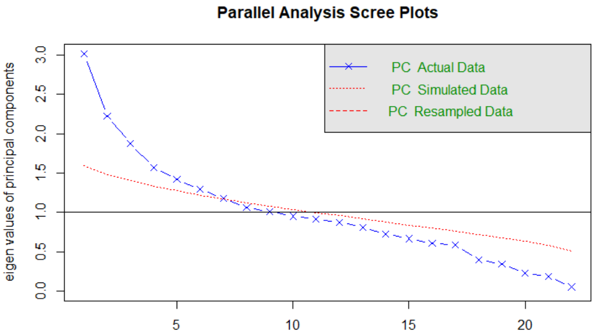

Figure 1. The result of feature selection using PCA.

Xgboost usually known as a trump card in a variety of competitions in the area of data mining. It also has the function of feature selection. The order of importance of all the variables is showing followed in Fig. 2.

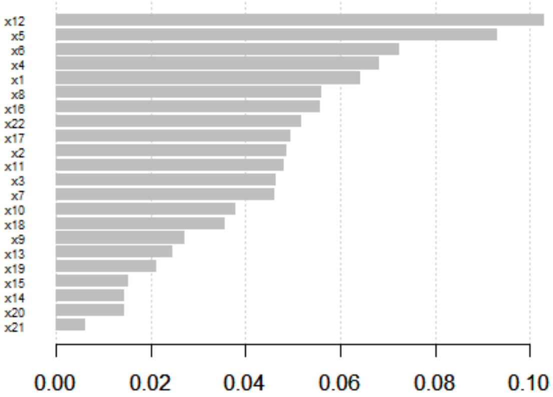

Figure 2. The result of feature selection using Xgboost.

Second, we applied the five classification methods to two feature subsets selected by PCA and Xgboost, therefore ten models were created. At last, the results were compared and analyzed.

V. Classification Methods

There are many classifiers available in data mining. Different methods can be used for more comprehensive and multi-angle analysis of data.

Logistic regression is the regression analysis for binary variables. From the angle of implementation, logic regression is simple, easy to understand and realize. Its run fast and the computational cost is relatively low. SVM puts the original data into a higher dimension one using nonlinear mapping, and then an optimal decision hyperplane is established to maximize the distance between the two closest samples of the plane on both sides of the plane. Compared with other non-linear methods, SVM requires relatively few samples, and its goal is to minimize structural risk. SVM performs well in noisy financial data[5]. Decision tree model is a tree structure that describes the classification of an instance, consisting of nodes and directed edges. The biggest advantage of decision trees is their interpretability to the model. ANN is inspired by the central nervous system of animals, which is a model of information processing of applied neurons. The distributed structure makes it have the same robustness as human brain, while ANN also good at self-learning, self-organizing. Random forest improves the decision tree by using the voting mechanism of multiple decision tree, it is an ensemble algorithm in essence. For example, if there are N samples and M variables (dimensions), the specific process of random forest is as follows: (1) Determine a value m, which is used to indicate how many variables each tree classifier chooses. (2) Collect k samples from the data set and use them to create k tree classifiers. In addition, k bags of external data are generated to be used for testing later.(3) After entering the classified sample, each tree classifier will classify it, and then all classifiers will determine the classification result according to the majority rule.

VI. Experimental Results

At the beginning, we use machine learning methods to explore with all the variables, the accuracy of SVM, RF, DT, ANN, and LR are show in Fig.3. It is obviously that SVM was better than others in this condition. Random forest has the lowest accuracy among these methods.

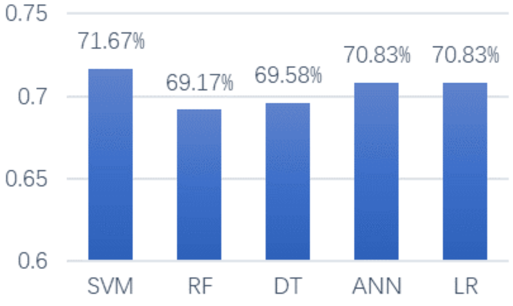

Figure 3. Accuracy with all the variables.

Second, in order to test five methods, we added the variables to the model one by one follow the order of importance from high to low. Thus, we do experiments with the most two important variables first, and then added the third, and go on. The results based on the importance of variables provided by PCA (Fig.4) bellow indicate that with variables' growing, RF performs better and more stable.

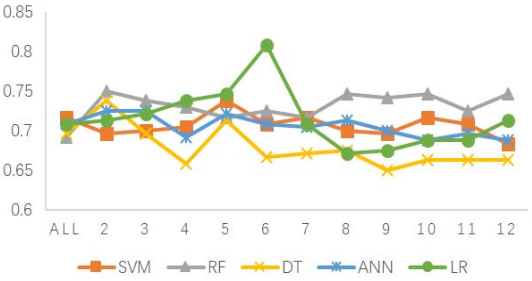

Figure 4. Results based on PCA.

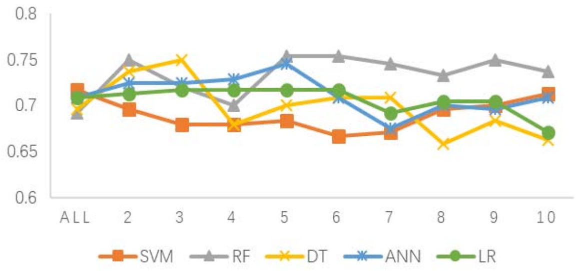

Figure 5. Results based on Xgboost.

In the same way, we test data based on the variables provided by RF, and results are in Fig. 5.

Random forest still performs good when the number of selected variables get bigger. LR reach the highest accuracy when the number of selected variables is 6, but it's not stable.

Third, in order to test two feature selection functions and to find which variables are of great importance in machine learning, we plot the average value of five accuracy (Fig. 6). We can find that when the number is 2 or 5, we may get a satisfied result.

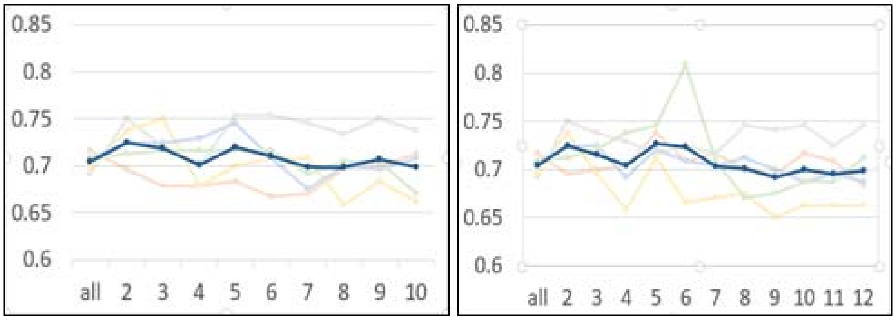

Figure 6. Average value of Fig.4 and Fig.5.

At last, to compare these five methods more intuitively, we put all variables' combination into Fig. 7, from which we find that RF do the best performance.

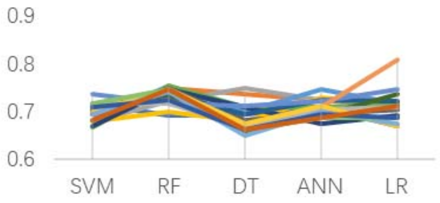

Figure 7. Results of all variables' combination.

VII. Discussion

This article has three contributions: First, to calculate and analyze the factors that affect the fraud behavior. According to our research, the effect of x12 and x13 on the model is confirmed by both two feature selection methods. Operating profit margin is the ratio of an enterprise's operating profit and operating income. There's a big risk of financial fraud when there's a big loss in the firm and a lower profit margin, it would be reflected from indicators related to profit. EPS is generally used to measure common profit level and investment risk, and to reflect the operating results of an enterprise. In order to cover losses or exaggerate earnings, the companies have a tendency to increase EPS. Second, consider the influence of the number of variables on the model. The results indicate that 2 or 5 variables are better than the others. Third, we compared the performance of five machine learning methods and found that among them, random forest has the following advantages: (1) it is good at processing high-dimensional data; (2) it may avoid overfitting to some extent; (3) it has good robustness and stable results.

There are still shortcomings in this article: the data is not large enough that only Chinese companies contained; variables can be more various and more innovative. The practical problems are far more complicated than the ones in the article, and more factors should be taken into consideration when designing variables.

References

[1] Kirkos, E., C. Spathis, and Y. Manolopoulos, Data Mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 2007. 32(4): p. 995-1003.

[2] Saha, S.C., C. Lei, and J.C. Patterson, Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. Knowledge-Based Systems, 2015. 89(9): p. 459-470.

[3] Wang, L. and C. Wu, A Combination of Models for Financial Crisis Prediction: Integrating Probabilistic Neural Network with Back-Propagation based on Adaptive Boosting. International Journal of Computational Intelligence Systems, 2017. 10(1): p. 507.

[4] Kotsiantis, S., et al., Forecasting fraudulent financial statements using data mining. Enformatika, 2006. 3(2): p. 104-110.

[5] Pai, P.F., M.F. Hsu, and M.C. Wang, A support vector machine-based model for detecting top management fraud. Knowledge-Based Systems, 2011. 24(2): p. 314-321.

[6] Huang, S.Y., Fraud Detection Model by Using Support Vector Machine Techniques. International Journal of Digital Content Technology & Its Applic, 2013.

[7] Cecchini, M., et al., Making words work: Using financial text as a predictor of financial events. Decision Support Systems, 2011. 50(1): p. 164-175.

[8] Bhattacharyya, S., et al., Data mining for credit card fraud: A comparative study. Decision Support Systems, 2011. 50(3): p. 602-613.

[9] Olszewski, D., Fraud detection using self-organizing map visualizing the user profiles. Knowledge-Based Systems, 2014. 70(C): p. 324-334.

[10] Jans, M., et al., A business process mining application for internal transaction fraud mitigation. Expert Systems with Applications, 2011. 38(10): p. 13351-13359.

[11] Bermúdez, L., et al., A Bayesian dichotomous model with asymmetric link for fraud in insurance. Insurance Mathematics & Economics, 2008. 42(2): p. 779-786.

[12] Ngai, E.W.T., et al. The application of data mining techniques in financial fraud detection:A classification framework and an academic review of literature. 2013.

[13] Hoque, N., et al., Review: Network attacks: Taxonomy, tools and systems. Journal of Network & Computer Applications, 2014. 40(1): p. 307-324.

[14] Kumar, P.A.R. and S. Selvakumar, Detection of distributed denial of service attacks using an ensemble of adaptive and hybrid neuro-fuzzy systems. Computer Communications, 2013. 36(3): p. 303-319.

[15] Zhang, B., Y. Zhou, and C. Faloutsos. Toward a Comprehensive Model in Internet Auction Fraud Detection. in Hawaii International Conference on System Sciences, Proceedings of the. 2008.

[16] Olszewski, D., A probabilistic approach to fraud detection in telecommunications. Knowledge-Based Systems, 2012. 26: p. 246-258.

[17] Olszewski, D., Fraud Detection in Telecommunications Using Kullback-Leibler Divergence and Latent Dirichlet Allocation. Intelligent Data Analysis, 2011. 16(3): p. 467-485.

[18] Wang, L. and C. Wu, Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map. Knowledge-Based Systems, 2017. 121: p. 99-110.

[19] Yeh, C.C., et al., A Hybrid Detecting Fraudulent Financial Statements Model Using Rough Set Theory and Support Vector Machines. Journal of Cybernetics, 2016. 47(4): p. 261-276.

[20] Liu, C., et al., Financial Fraud Detection Model: Based on Random Forest. International Journal of Economics & Finance, 2015. 7(7): p. 178-188.

[21] Dechow, P.M., et al., Predicting Material Accounting Misstatements *. Social Science Electronic Publishing, 2011. 28(1): p. 17–82.

[22] Petr Hajek, R.H., Mining corporate annual reports for intelligent detection of financial statement fraud –A comparative study of machine learning methods. Knowledge-Based Systems, 2017. 128: p. 139-152.

[23] Kim, Y.J., B. Baik, and S. Cho, Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. Expert Systems with Applications, 2016. 62: p. 32-43.

[24] Guyon, I. and A. Elisseeff, An Introduction to Feature Extraction. 2006: Springer Berlin Heidelberg. 1-25.