ATLANTIS PRESS
Advances in Engineering Research, volume 136
7th International Conference on Advanced Design and Manufacturing Engineering (ICADME 2017)

Research on the Detection of Financial Fraud Using Data Mining Techniques

Li Yanling [1, a], Li Nan[2,b] and Yang Mingpei[3,c]
[1]Dalian Neusoft University of Information,Dalian, China
[2]Dalian Neusoft University of Information,Dalian, China
[3]University of Connecticut, CT, USA
[a]liyanling@neusoft.edu.cn
[b]linan_xg@neusoft.edu.cn
[c] 374110380@qq.com

Keywords: Financial information fraud detection; Data mining; Ada Boost method; Rattle package
Abstract. Financial information plays a crucial role for future investors to make important decisions, and how to provide true, reliable and accurate financial information becomes a top mission for enterprises. To effectively identify financial fraud information, we first select the relative indicators by reviewing the financial information of previous studies, and the indicators related to false information are prepared for data modeling using data mining tool. Furthermore, we analyze these relative indicators through the rattle package in the R and Ada Boost method. The results we obtained demonstrate that a company's solvency is the primary factor in determining whether a company has financial information fraud. Meanwhile, key factors like profitability, operating capacity, accounts receivable turnover days, business debt ratio, and financial debt ratio are useful when detecting financial information fraud.

Introduction

Global economic integration has been further strengthened and information technology has developed rapidly since the 21st Century. The quality of financial information is becoming increasingly significant in financial market. However, financial scandals have occurred in many eras, countries and industries: the United States Enron Corp collapse and China Yinguangxia trap. The frequent occurrence of financial scandals not only raises the risk of bankruptcy of companies, but also affects the stability of financial markets negatively. Corporate or management manipulate financial data deliberately and covertly through violation of accounting standards, disclosure of information and breaking laws or regulations to obtain improper benefits, resulting in the distortion of financial information. The broad technical accounting effects involved in scandals have been fairly limited - asset valuation, off-balance sheet transactions or group accounting issues, revenue recognition, capitalisation of costs. Therefore, how to identify the false financial information has become the priority of public attention, and this paper studies how to use data mining techniques to detect financial fraud, and to establish a better financial system for the screen of false information.

Literature Review

Foreign countries start early about data mining research in financial false information screen. Lin (2003) built a fuzzy neural network model [1]. A comparative analysis with decision tree model, Bayesian network and artificial neural network model, which has been agreed by Kirkos (2007), found that the Bayesian network model shows the best outcome [2]. The domestic research starts relatively later, mainly referring to foreign ideas. Geng (2002) detected effective financial frauds through the difference between net profit and cash flow by T test for the 36 financial fraud companies [3]; with the decision tree method, Yao (2010) established fraud identification model from the 227

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

listed companies punished for financial fraud [4]; Han (2011) sent an early warning to false information with quantitative and qualitative indexes[5];Yu (2011) pointed out that there is a negative correlation between the seriousness of financial fraud assets net profit margin and sales expense ratio [6]; Liu (2011) and Zhang (2014) used the data mining to identify financial fraud [7-8]; Huang (2016) introduced data mining algorithm into online audit model [9].

Through the study of 50 illegal enterprises in the 2000 -2014 two stock markets together with 50 healthy listed companies, this paper first analyzes data based on false information financial motives, methods and related literature, through the application of data mining technology [10], and the Ada Boost method modeling. Additionally, we select the appropriate financial indicators from the two times of iteration, and finally, this paper will give some suggestions on how to prevent false financial information.

Empirical analysis of false financial information screening model

Firstly, we compare the data from financial fraud company punished by CSRC with the data from the same number of companies not been punished by CSRC. Secondly, due to the excessive number of variables seriously affect the accuracy of the model, we will select the indicators to be analyzed through two times of iterations. Finally, the Ada Boost is used to construct the model of enterprise financial false information and to analyze results.

Sample data selection

In the sample selection of false information for financial enterprises, we select the top 50 of illegal listed companies through Tai'an's data center in the CSMAR database in ascending order according to their penalty amount. In the sample selection of healthy firms, we choose 50 firms excluding the ones punished by CSRC and ST, PT companies. Whether the company breaks the rules is set as a categorical variable, all other kinds of indexes as independent variables. Finally, we export the EXCEL format file into CSV file to facilitate the analysis of R software.

Index Selection

1) Index first selection

According to the motivations and methods of enterprise, in combination with related literature review, this article first selects 19 financial indicators as primary index as TABLE I shown from the debt-paying ability, profitability, cash flow and operating capacity 4 aspects.

TABLE I. INDEX PRIMARY SELECTION

Index primary selection		
Number	***Name***	***Formula***
F031201A	Retained earning assets ratio	(surplus reserve +returned earnings)/ total assets
F031301A	Long term assets rate	(equity plus long-term debt)/(fixed assets + long-term investment)
F031501A	Current liability ratio	total current liabilities/total liabilities
F031601A	Business debt ratio	(total current liabilities - short-term borrowing - non-current liabilities due within one year-tradable financial liabilities-derivativefinancial liabilities)/(total liabilities)
F031701A	Financial debt ratio	(non-current liabilities due within 1 year + + short-term borrowing tradable financial liabilities + total current liabilities + derivative financial liabilities)/ (total liabilities)
F040301B	Accounts receivable turnover days	calculation period days/accounts receivable turnover
F010801B	Net business activities generated cash flow/current liabilities	net business activities generated cash flow/current liabilities combined the denominator for the total current liabilities)
F010301A	Conservative quick ratio	(short-term investment + tradable financial assets + notes receivable + money + net receivables)/current liabilities
F011401A	Tangible assets ratio	(total liabilities)/ (total asset- intangible assets-goodwill)
F011501A	Tangible assets interest-bearing debt ratio	(total current liabilities + short-term borrowings + non-current liabilities due within 1 year)/(total assets- net intangible assets-goodwill)
F040101B	Accounts receivable to revenue ratio	accounts receivable/revenue
F010201A	Quick ratio	(current assets - inventory)/current liabilities
F053201B	Long-term return on capital	(current assets - inventory)/current liabilities
F060301B	Cash net operating income levels	(net business activities generated cash flow)/(operating income)
F061701B	Toal cash recovery rate	(net business activities generated cash flow)/(total assets) ending balance
F050601B	Earnings Before Interest and Tax	net income + income tax expense+ financial expenses
F050701B	Profit before interest and after tax	EBIT * (1 - income tax)
F050801B	Income before interest, tax, depreciation and amortization	Net income+income tax expenses + financial expenses + fixed assets depreciation+depletion of oil and gas assets+ productive biological assets depreciation+amortization of intangible assets + long-term prepaid expenses amortization
F040801B	Accounts payable turnover ratio	operating cost/accounts payable ending balance

2) Index secondary selection

Primary index may not provide an accurate result due to some human-caused factors. Thus, we will first implement data mining to all listed companies from Guo Tai'an's data with Ada Boost method, and then get the significance of each indicators. Eight indexes are selected: retained earnings assets ratio, current debt ratio, business debt ratio, financial debt ratio, business activities generated cash flow/current liabilities, accounts receivable to income ratio, accounts receivable turnover days and accounts receivable turnover ratio.

TABLE II. INDEX SECONDARY SELECTION

Index secondary selection		
Number	***Name***	***Formula***
F031201A	Retained earning assets ratio	(surplus reserve +returned earnings)/ total assets
F031501A	Current liability ratio	total current liabilities/total liabilities
F031601A	Business debt ratio	(total current liabilities - short-term borrowing - non-current liabilities due within one year-tradable financial liabilities-derivative financial liabilities)/(total liabilities)
F031701A	Financial debt ratio	(non-current liabilities due within 1 year + + short-term borrowing tradable financial liabilities + total current liabilities + derivative financial liabilities)/(total liabilities)
F010801B	Net business activities generated cash flow/current liabilities	net business activities generated cash flow/current liabilities combined the denominator for the total current liabilities)
F040101B	Accounts receivable to revenue ratio	Accounts receivable/revenue
F040301B	Accounts receivable turnover days	calculation period days/accounts receivable turnover
F040801B	Accounts payable turnover ratio	operating cost/accounts payable ending balance

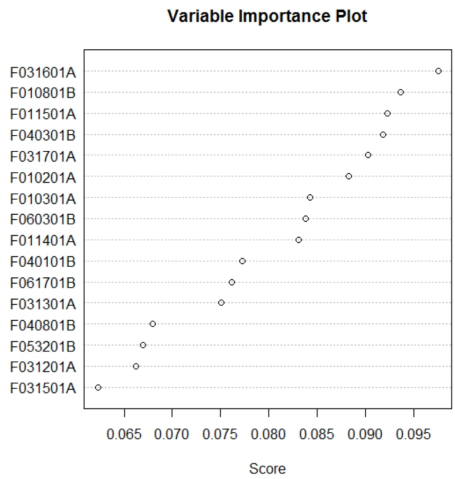

Figure 1 Preliminary results of index importance through Ada Boost method

From Fig 3.1, it could be shown that there are eight indexes including F031201A got more than 0.08 points, while index F040101B and later got relatively lower points with respect to importance to test result. Therefore, we only choose eight indicators with highest scores as our secondary indexes as shown in TABLE II.

Modal selection

Financial false information screening itself belongs to a classification problem, and Ada Boost is an integration algorithm based on base classification. Therefore, we choose Ada boost method to establish the false financial information model, analyze the result and summarize the general rules

1) Ada boost model construction

Sample rate allocation

Various samples will affect the accuracy of model recognition. In the case of the random seed for 42, according to the different proportion, inspection and testing samples, the identifying error rates are shown in TABLE III.

TABLE III. ADA BOOST MODEL IDENTIFYING DIFFERENT SAMPLES ERROR RATE

Sample ratio	Training sample	Examine sample	Testing sample	All samples
50/25/25	14%	20%	8%	14%
60/20/20	10%	15%	25%	14%
70/15/15	7.14%	6.67%	20%	9%
80/10/10	5%	0%	0%	4%

It shows from table that the sample size has a positive relationship with the accuracy rate. However, too many big size of sample may lead to excessive training even if the accuracy is high in training sample, the resulting performance is not stable when applied to other samples, no much significance for practical application as well. Therefore, this experiment adopts the proportion of “70/15/15” on training to achieve the best outcome.

Random seed setting

Shadow package gives different distributions of the sample by setting the random seed value, which may affect the test results. Because sometimes the same kind of samples may be chosen for training, while test is the other different category, leading to the high error rate. Thus, to guarantee the reliability of the test sample, under the condition of the sample proportion to “70/15/15”, we change the random seed, to find the most suitable sample set, and the random seed error rates are shown in TABLE IV:

TABLE IV. ADA BOOST MODEL ERROR RATE UNDER DIFFERENT RANDOM SEED

Random seed	Training sample	Examine sample	Testing sample	All sample
10	7.14%	20%	46%	15%
20	7.14%	20%	20%	11%
42	7.14%	6.67%	20%	9%
80	5.70%	33%	13.30%	11%

From the above table, when the size of random seed is 42, the optimal screening outcome either all samples error rate or other sample error rate, is the lowest. Therefore, this experiment sets random seed to be 42.

Complexity set

Complexity degree can affect the precision of model precision, the greater the complexity degree, the more complicated the model calculation, the greater the precision, and the lower the error rate. But too high complexity can increase the calculation cost, need to consider the balance of the accuracy and computational cost. Hence, in the sample proportion is the proportion of "15" 70/15 /, random seed set to 42, model discernment error rate as shown in TABLE V:

TABLE V. ADA BOOST MODEL ERROR RATE UNDER DIFFERENT COMPLEXITY

Complexity	Training sample	Examine sample	Testing sample	All sample
0.08	8.57%	6.67%	13%	9%
0.05	8.57%	6.67%	20%	10%
0.01	7.14%	6.67%	20%	9%
0.001	7.14%	6.67%	20%	9%

We can see from above table that when the complexity is 0.08, the comprehensive error rate of model is lowest. Because the quantity is not much, and operation is faster. In this case, the model can be more effectively to identify false information, so this study sets the complexity of 0.08.

The rest of the parameters set

After adjustment, the core of the final model parameter is set to the decision tree number of 100, the maximum depth of 30, the minimum divided into 1, X value of 10. After trial and error, the computation accuracy of the model is the highest in this situation.

2) Ada Boost screening rules and results

Ada Boost model in this paper is ultimately chosen eight financial indicators as follows: sample distribution as "70/15/15", a random seed for 42, decision tree number of 100, the maximum depth of 30, the minimum divided into 1, complexity of 0.08, X value of 10.

TABLE VI. ADA BOOST MODEL ERROR RATE

Training sample	Examine sample	Testing sample	All sample
0%	13.30%	13.30%	4%

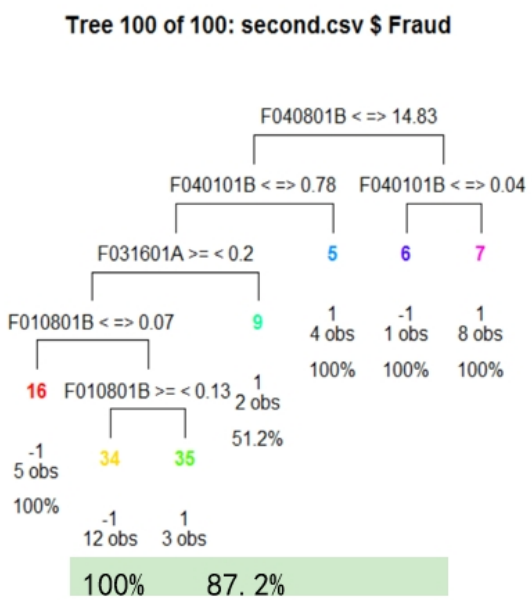

Figure 2 Ada Boost judgment structure model (one hundred tree)

This picture shows in the case of F040801B (accounts payable turnover ratio) greater than 14.83, if F040101B (accounts receivable and revenue ratio) is greater than 0.04, the sample firms have a 100% chance of false financial information behavior; if F040101B (accounts receivable and revenue ratio) value is less than or equal to 0.04, healthy enterprise percentage is 100%. In the case of F040801B (accounts payable turnover ratio) less than or equal to 14.83, if F040101B (accounts receivable and revenue ratio) is greater than 0.78, the enterprise should have fraud information; if F040101B (accounts receivable and revenue ratio) is less than or equal to 0.78 and F031601A business debt ratio is less than 0.2, then the enterprises have the possibility of 51.2% for the false information. If F031601A (business debt ratio) is greater than or equal to 0.2 and F010801B (net business activities generated cash flow/current liabilities) is less than or equal to 0.07, the enterprise is healthy enterprise; while if F010801B (net business activities generated cash flow/current liabilities) is less than 0.13, the company has a 87.2% chance of false financial information behavior, otherwise health enterprise.

Ada Boost model on this experiment built 100 decision tree, we will not show them one by one. Through secondary index distribution of important variables in the Ada Boost (Fig. 3.3), it can be seen that F040301B (accounts receivable turnover days), F031701A (financial liabilities ratio) and F031601A (business debt ratio) are best three to identify whether there is false financial information, these indicators and worth more attention for auditors.

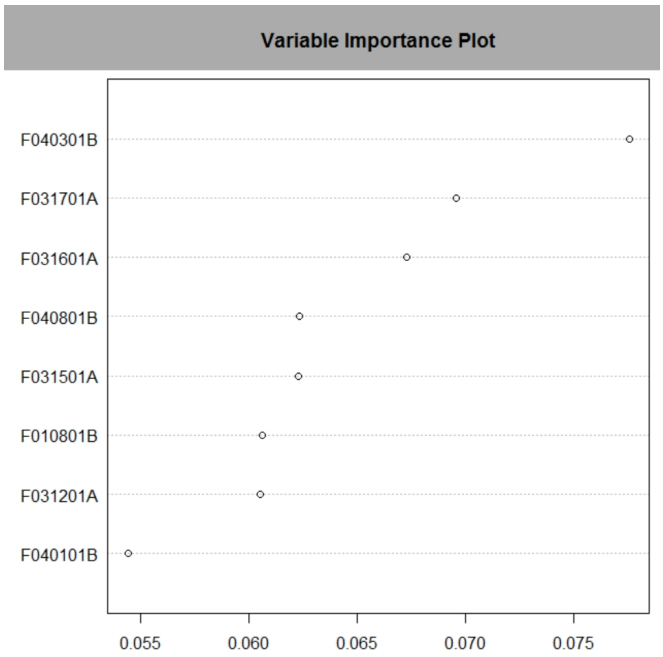

Figure 3 importance index check result of Ada Boost method

Conclusion

How to effectively identify financial false information is necessary for investors and creditors. This paper studies 2000-2014 Shanghai and Shenzhen stock market listed illegal companies with 50 healthy companies as reference, together with the motivation, means and related documents of false financial information, using Ada Boost model, to analyze the data. Through two times of iteration to select appropriate financial indicators, we found eight key indicators that can effectively screen the false financial information; accounts receivable turnover days, financial debt ratio and operating gearing ratio are the best three ones. What's more, in comparison with the decision tree, random forests and support vector machine (SVM), we found that Ada Boost model has the highest accuracy. Besides, the solvency is the primary factor to judge whether the enterprises have false financial information. This is because how to raise money is one of the most key issues for Chinese listed companies. To raise more money, companies need to prove their good solvency ratio, and therefore they are more incentive to disclose false information. Meanwhile, the profitability and operating capacity of enterprises are two indicators which could also reflect management desire for profit, making the enterprise continuously to whitewash profits to attract the eye of investors.

Because of the limit research time and knowledge, in this paper, there are still many shortcomings: small sample size may not suitable for all company's financial statements; no consideration for extreme value or null value; no test on other companies only applicable to companies after filter. In the future, we hope to broaden the research scope into non-financial area, through more comprehensive study of data mining technology to get a more widely used, more accurate financial false information screen model.

Acknowledgement

Thanks to the General Project of Education Humanities and Social Sciences in Liaoning Provincial Department of Education " Enterprise Financial Crisis Early Warning Research Based on the Intelligent Multi-Classification Method"（ Project Number W2014383）.

References

[1] Lin J.W,Hwang M.I,Becker J.D. A fuzzy neural network for assessing the risk of fraudulent financial reporting[J]. Managerial Auditing Journal,2003(18),657-665

[2] E Kirkos, C Spathis, Y Manolopulo. Data mining techniques for the detection of fraudulent financial statement[J]. Expert Systems with Application,2007(32),995-1003

[3] Geng Jianxin, Xiao Zezhong, Xu Qin. An empirical analysis of the relationship between earnings and cash flow data – false information warning signals of the company[J].Accounting research, 2002（12）：28-34

[4] Yao Zhixuan, The decision tree technology in the application of financial reporting fraud of listed company recognition research [D]. Tianjin university of finance and economics, 2010

[5] Han Ying. Financial fraud warning signs indicating study [J]. The northern economy and trade, 2011 (2), 55-56

[6] Yu Chunbo. Classification based on Logistic regression model of the factors affecting accounting fraud research [D]. Jilin university, 2011

[7] Liu Shulei, Li Qiang. Financial reporting fraud identification of research: study based on the empirical data of listed companies manufacturing [J].Accounting communications: comprehensive (below), 2011 (1)

[8] Zhang Qiusan, Zhang lei, Zhang Ning, et al. Based on data mining of the listed company financial fraud identification research, technology and industry. 2014,11（14）

[9] Huang Zhiyan. The online auditing model based on data mining design [J], electronic technology and software engineering, 2016,11,168-169

[10] Han Jiawei. Data mining: concepts and techniques (the original book the third edition) [M], mechanical industry publishing house publishing house, 2014,2,2-30