An Effective Financial Statements Fraud Detection Model

Suduan CHEN\textsuperscript{1,} and Alex YANG\textsuperscript{2}

\textsuperscript{1}Department of Accounting Information, National Taipei University of Business, No. 321, Sec. 1, Jinan Rd., Zhongzheng District, Taipei 10051, Taiwan
\textsuperscript{2}Ernst & Young Global Limited, 9F, No. 333, Sec.1 Keelung Rd, Taipei 11012, Taiwan

*Corresponding author

Keywords: Financial statements fraud, Decision tree CART, CHAID, Artificial neural network (ANN), Support vector machine (SVM).

Abstract. The purpose of this study is to establish an effective and rigorous financial statements fraud (FSF) detection model. The samples of this study are the listed companies in Taiwan, totaling 220 companies and including 55 companies with financial statements fraud and 165 normal companies. The data are from the Taiwan Economic Journal (TEJ) during the period from 2006 to 2015. In Stage I, decision tree CART and artificial neural network (ANN) are applied to select the important variables. In Stage II, decision tree CHAID, ANN, and support vector machine (SVM) are used for modeling. The results show that the ANN-CHAID model has the highest FSF detection accuracy of 87.41%.

Introduction

Since the Asian Financial Crisis in 1997, many significant financial statement fraud cases have occurred in the U.S. and Taiwan, causing heavy losses to investors and seriously damaging the capital market, such as Enron in 2001, WorldCom in 2003, and AIG in 2005 in the U.S.; Procomp Informatics, Infodisc Technology, Summit Technology, ABIT Computer in 2004, and XPEC Entertainment in 2016 in Taiwan, as well as other significant financial statement fraud cases. Therefore, it is a very important issue for enterprises to detect financial statement fraud. The United States Congress also passed the Sarbanes-Oxley Act of 2002 to strengthen the supervision of competent authorities, the implementation of corporate governance, and the independence of accountants, as well as to enhance the responsibilities of the administering authority (CEO and CFO) and certifying CPAs. SAS No.99 (Statement on Auditing Standards No.99) incorporates fraud theory and practices to develop reasonable assurances that an entity's financial statements are free of material misstatements whether by fraud or error. SAS 99 defines fraud as “an intentional act that results in material misstatements in financial statements that are the subject of an auditor”. Many researches have discussed financial statement fraud in the past, which mainly used the traditional regression analysis method. In recent years, many scholars have used data mining techniques to detect financial statement fraud, which has improved the accuracy rate of judgment [1-7]. The commonly used data mining techniques include artificial neural network (ANN), decision tree, Bayesian belief networks (BBN), and support vector machine (SVM). The purpose of this study is to establish an effective and rigorous financial statements fraud (FSF) detection model. In Stage I of this study, decision tree CART and ANN are applied to select the important variables. In Stage II, decision tree CHAID, ANN, and SVM are used for modeling.

Methodology

Statistical Methods

The statistical methods used in this study include decision tree CART, ANN, CHAID, and SVM. CART (Classification and regression tree), as proposed by Breiman et al. [8], is a kind of decision-tree algorithm where binary partitioning is used; partitioning conditions are determined according to the gini rules, and the data are divided into 2 subsets for each portion. CHAID
(Chi-square automatic interaction detector) is also a kind of decision-tree algorithm, as proposed by Kass [9], and this algorithm applies Chi-square testing to calculate the p-value of splitting the nodes of branches and leaves in the decision tree to determine whether the portioning is continued, and seeks the best branch node during the process of portioning and trimming. ANN is a system that imitates the calculation ability of a biological neural network to counterfeit the simplest nerve cell mode, while the information processing system simulating the biological neuron is used to receive multiple input values to establish the system model (relationship between input and output), which can be used for estimation, prediction, decision, and diagnosis. As early as 1943, Mcculloch and Pitts [10] mentioned the application of ANN in their published articles, which is mainly a kind of simplified mathematical model for nerve cells to simulate the human brain’s ability to handle calculations. SVM is a set of AI learning method, as proposed by Vapnik [11], and in terms of its model, the input vector is mapped to a high-dimensionality feature space from the training data in a manner of linear or non-linear kernel function through the learning mechanism, in order to determine an optimal separating hyperplane that can distinguish 2 or more different classes of data.

**Samples**

Based on companies with claim types belonging to the misrepresentation of financial statements, misrepresentation of prospectus, stock price manipulation, and insider trading, as listed by the Securities and Futures Investors Protection Center (SFIPC), this research quotes Article 20, Article 20.1, Article 32, Article 155, and Article 157.1 of the Exchange Act, which is supplemented by lawsuits and judgment situations against significant securities crime, and published by the securities futures bureau. Samples and data are from the Taiwan Economic Journal (TEJ) during the period from 2006 to 2015, including 55 listed companies with financial statements fraud and 165 normal listed companies (1:3 pairing), thus, the total samples include 220 companies.

**Variables**

Regarding the companies with financial statement fraud, the dependent variable is set as 1; regarding the companies with no financial statement fraud, the dependent variable is set as 0. A total of 30 independent variables are used to measure financial statement fraud in this research, including 22 financial variables and 8 non-financial variables. The financial variables include X1: Current ratio, X2: Quick ratio, X3: Accounts receivable turnover, X4: Debt ÷ equity ratio, X5: Net profit rate, X6: Return on assets (ROA), X7: Return on equity (ROE), X8: Interest paid ÷ total liabilities, X9: Inventory ÷ net sales, X10: Net sales ÷ total assets, X11: Gross profit ÷ total assets, X12: Current liabilities ÷ total assets, X13: Operating cash flow ÷ net sales, X14: Operating cash flow ÷ current liabilities, X15: Total assets growth rate, X16: Gross profit rate, X17: Operating expenses ratio, X18: Fixed assets ÷ total assets Inventory turnover, X19: Inventory ÷ total assets, X20: Inventory ÷ current assets, X21: Net income ÷ fixed assets, X22: Cash ÷ total assets; while the non-financial variables include X23: The ratio of stocks held by directors and supervisors, X24: The number of independent directors, X25: The ratio of pledged stocks held by directors and supervisors, X26: The major stockholders' stockholding ratio, X27: Audited by BIG 4, X28: whether the president or CEO is changed within 3 years, X29: whether the CFO is changed within 3 years, X30: whether the internal auditing supervisor is changed within 3 years.

**Research Design and Procedures**

Stage I of this study uses CART and ANN to screen out the important variables, including the financial and non-financial variables. In Stage II, this study uses CHAID, ANN, and SVM to establish the models. Then, the detection accuracy of the models are compared to determine the best FSF detection model. The research design and procedures are illustrated in Figure 1.
Results and Analysis

Variables Screened by CART

A total of 6 variables are screened out, with the order of importance of the variables being (importance value $\geq 0.05$): $X_{18}$: Fixed assets $\div$ total assets Inventory turnover (0.19), $X_{13}$: Operating cash flow $\div$ net sales (0.16), $X_{23}$: The ratio of stocks held by directors and supervisors (0.15), $X_{3}$: Accounts receivable turnover (0.07), $X_{2}$: Quick ratio (0.06), and $X_{19}$: Inventory $\div$ total assets (0.06).

Variables Screened by ANN

A total of 6 variables are screened out, with the order of importance of the variables being (importance value $\geq 0.05$): $X_{21}$: Net income $\div$ fixed assets (0.08), $X_{3}$: Accounts receivable turnover (0.06), $X_{5}$: Net profit rate (0.06), $X_{12}$: Current liabilities $\div$ total assets (0.05), $X_{25}$: The ratio of pledged stocks held by directors and supervisors (0.05), and $X_{7}$: Return on equity (ROE) (0.05).

CART Models’ Detection Accuracy

As shown in Table 1, after the CART models undergo ten-fold cross-validation, the CART-SVM model has the highest detection accuracy of 84.50%. The type I error rate and type II error rate are also shown in Table 1.

| Model       | FSF detection accuracy | Type I error rate | Type II error rate |
|-------------|------------------------|-------------------|-------------------|
| CART-CHAID  | 80.75%                 | 4.55%             | 17.27%            |
| CART-ANN    | 77.96%                 | 8.64%             | 14.55%            |
| CART-SVM    | 84.50%                 | 5.45%             | 13.18%            |

ANN Models’ Detection Accuracy

As shown in Table 2, after the ANN models undergo ten-fold cross-validation, the ANN-CHAID model has the highest detection accuracy of 87.41%. This is also the model with the highest accuracy rate among the 6 models established by this research. The type I error rate and type II error rate are also shown in Table 2. It is worth mentioning that the type I error rate and type II error rate in ANN-CHAID model are also the lowest among the 6 models.

| Model       | FSF detection accuracy | Type I error rate | Type II error rate |
|-------------|------------------------|-------------------|-------------------|
| ANN-CHAID   | 87.41%                 | 4.55%             | 10.45%            |
| ANN-ANN     | 84.59%                 | 7.27%             | 11.36%            |
| ANN-SVM     | 81.09%                 | 6.36%             | 14.55%            |

Conclusions

For more than 10 years, significant financial statement fraud cases have emerged one after another, thus, it is very important for enterprises to detect financial statement fraud. The purpose of this
study is to establish an effective and rigorous financial statements fraud (FSF) detection model. This research successfully used several data mining techniques, including CART, ANN, CHAID, and SVM, and established an effective and rigorous two-stage financial statements fraud (FSF) detection model. The results show that the ANN-CHAID model has the highest FSF detection accuracy rate of 87.41%, which can be used as reference by auditors and CPAs, in order to reduce auditing risks, avoid losses to investors, and maintain the development of the capital market. This study has also contributed to the academic research and practice of FSF. This research has the following limitations: First, as the size of Taiwan’s capital market is not big, FSF companies (sample size) are limited. Second, if the financial reports of enterprises in previous years are true, top management and financial top management colluded with each other, as based on certain or some reasons this year, to deliberately conceal or forge their financial statements; however, as their techniques are so meticulous that the flaws cannot be found in credit investigation, it is very difficult to make the correct auditing judgment and conduct the FSF detection. In fact, this is also an unavoidable limitation.

References

[1] K. Fanning, K. Cogger. Neural network detection of management fraud using published financial data, Int. J. Intell. Sys. Accoun. Financ. Mgt. 7 (1998) 21-24.

[2] S. Kotsiantis, E. Koumanakos, D. Tzelepis, V. Tampakas, Forecasting fraudulent financial statements using data mining, Int. J. Comput. Intell. 3 (2006) 104-110.

[3] C.C. Yeh, D.J. Chi and M.F. Hsu, A hybrid approach of DEA, rough set and support vector machines for business failure prediction, Exp. Syst. Appl., 37 (2010) 1535-1541.

[4] S. Chen, Y.J. Goo, Z.D. Shen, A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements, Sci. World J. (2014), doi:10.1155/2014/968712.

[5] C.C. Yeh, D.J. Chi, T.Y. Lin, S.H. Chiu, A hybrid detecting fraudulent financial statements model using rough set theory and support vector machines, Cybern. Syst., 47 (2016) 261-276.

[6] S. Chen, Detection of fraudulent financial statements using the hybrid data mining approach. SpringerPlus 5 (2016), doi: 10.1186/s40064-016-1707-6.

[7] Y.J. Goo, D.J. Chi and Z.D. Shen, Improving the prediction of going concern of Taiwanese listed companies using a hybrid of LASSO with data mining techniques. SpringerPlus 5 (2016), doi: 10.1186/s40064-016-2186-5.

[8] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. Classification and Regression Trees. Chapman & Hall/CRC. (1984)

[9] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. Appl. Stat. 29 (1980) 119-127.

[10] W. S. McCulloch, W.H. Pitts: A Logical Calculus of the Ideas Immanent in Nervous Activity, Bull. Math. Biophys. 7 (1943) 115-133.

[11] V. Vapnik, The nature of statistical learning theory. Springer-Verlag: New York, NY, USA, 1995.