ICIC Express Letters
Part B: Applications
Volume **9**, Number **4**, April **2018**

ICIC International ©2018 ISSN 2185-2766
pp. 347–352

DETECTION OF FRAUDULENT FINANCIAL STATEMENTS USING DECISION TREE AND ARTIFICIAL NEURAL NETWORK

Chyan-Long Jan[1,*] and David Hsiao[2]

[1]Department of Accounting
Soochow University
No. 56, Sec. 1, Guiyang Street, Zhongzheng District, Taipei 100, Taiwan
*Corresponding author: janchyanlong@yahoo.com.tw

[2]Ernst & Young Global Limited
7F, No. 419, Sec. 2, Gongdao 5th Road, Hsinchu 300, Taiwan
stage1020@gmail.com

Received October 2017; accepted January 2018

Abstract. *This study adopts decision tree CART, CHAID, C5.0, and artificial neural network to establish rigorous detection models of fraudulent financial statements by considering financial and non-financial variables. The objects of study are 32 companies with fraudulent financial statements and 96 companies without fraudulent financial statements for the period 2008-2014. The research results show that the detection accuracy of the CHAID+CHAID model is as high as 93.47%.*

Keywords: Fraudulent financial statements, Data mining, Decision tree CART, Decision tree CHAID, Decision tree C5.0, Artificial neural network (ANN)

1. **Introduction.** Financial statements are specific representations of a company's operating situation and are also an important basis for stakeholders to evaluate and communicate with the company, particularly so for one conducting an initial public offering, as they are important references for investors to decide whether to buy shares in such a company. The scandals of Enron in the U.S. in 2001 and WorldCom in 2003 are typical examples of fraudulent financial statements (FFS). In Taiwan, FFS-related cases include Procomp Informatics in 2004, Infodisc in 2004, and Summit Computer in 2006.

The United States Congress signed the Sarbanes-Oxley Act in 2002, aiming to strengthen the control and supervision of company operations and accuracy of financial statement information disclosure. The Act specifies the responsibility of corporate senior management for their firm's financial statements and stipulates that those providing fraudulent financial statements will be punished. Therefore, it is very important to effectively detect fraudulent financial statements.

Certified public accountants (CPAs) and auditors in the past just made decisions by following the auditing standards and their own experience, and so it was possible to make a wrong judgment. In recent years, data mining techniques (ANN, DT, SVM, etc.) have been adopted in some studies to detect fraudulent financial statements in order to reduce errors in judgment [1-11]. Except for the multiple data mining techniques used by [10], these previous studies usually employ only 1 or 2 kinds of data mining techniques to detect fraudulent financial statements, and they are not the complete or rigorous detection models. Therefore, this study applies more data mining techniques to establishing a more complete and rigorous detection model for fraudulent financial statements in two stages: through variable selection with decision tree CART and CHAID algorithm; here, the first stage selects the influential variables; in the second stage, after the establishment of FFS detection models such as decision tree C5.0, decision tree CHAID, and artificial neural network, these models are compared in accuracy with Type I error and Type II error, so as to find the optimal model.

This paper proceeds as follows. 1) Introduction states the background, motivation, purpose, and literature of this study. 2) Methodology describes the research methods, samples, and research design and process. 3) Empirical Results and Analysis offer the empirical results and makes a comparison of each model for detection accuracy. 4) Conclusions present the research conclusion, academic contribution, and practical contribution of this study, as well as future research suggestions.

2. Methodology.

2.1. Statistical methods.

This study utilizes decision tree CART, decision tree CHAID, decision tree C5.0, and artificial neural network (ANN).

Decision tree is a decision-support tool that is a tree-like graph or decision model. It is frequently used in operational research and especially in decision analysis, because it can help determine a strategy that has the greatest possibility for achieving a certain goal. If in practice a decision has to be adopted online when no perfect knowledge is available, then a decision tree should be parallel to the probability model and taken as the optimal selection model or online selection model. Another use of the decision tree is to take it as the descriptive means to calculate the conditional probability. Based on the classification of the known facts, the main function of a decision tree is to establish a tree structure and summarize some rules from it. The decision tree formed in this way can be used for a forecast beyond the samples.

The most commonly used decision tree algorithms currently include CART, CHAID, C5.0, etc. CART (classification and regression trees) is a data mining and prediction algorithm developed by [12]. It is a kind of decision tree technology of binary division. It is applied to non-parameter data continuously or classified in attributes, and the selection of division conditions is determined based on the number of data classification and data attributes. According to gini rules, the division conditions are determined, and each time the data are divided into two subsets. The requirement is to repeatedly find the conditions of the next division from each subset. The classification rules target to obtain the maximum value from the following formula: $\Delta i(s,t) - i(t) - p_{L_i}(t_L) - p_{R_i}(t_R)$. When building a decision tree, the criterion variable is the categorical variable, and the impurity level measurement may be subject to the gini index and Twoing rules [14].

CHAID (chi-squared automatic interaction detection) is a method to calculate the p-value of a branch or leaf split node in a decision tree, so as to decide whether to continue the division. CHAID may prevent mechanically using excessive data and a stop division of the decision tree – that is to say, CHAID will complete trimming before model establishment. C5.0 is developed on the basis of ID3, as put forward by [13]. Taken on the back of ID3, C5.0 improves the handling of the continuous attribute problem, which cannot be treated by ID3, and is applicable to processing a large dataset. The establishment of a decision tree includes the following 3 steps: establishment of a decision tree based on training data, trimming the decision tree, and generating learning rules from the decision tree.

ANN (artificial neural network) is a kind of mathematic model or calculation model simulating the structure and function of a biological neural network. The modern neural network is a kind of non-linear statistical data modeling tool and often simulates the complicated relationship between input and output or to explore the data model. Artificial neuron is the most basic artificial neural network unit and also can be called the processing element. The output of each processing element is the input of other processing elements. The signal between processing elements is transmitted by a link, and there is a weight W_{ij} on each connection, indicating the intensity of influence of the number i processing element on the number j processing element.

The relationship between input and output signals in the artificial neuron model may be represented by the function of a weighted product of input signals, shown as [14]:

$$Y_j = f\left(\Sigma_i W_{ij} X_i - \theta_j\right)$$

Here, Y_j is the output signal of the imitation of a biological neuron model; f is the transfer function of the imitation of a biological neuron model; W_{ij} is the ganglion intensity of the imitation of a biological neuron model and is also called the weighted value of connection; X_i is the input signal of the imitation of a biological neuron model; θ_j is the input signal of the imitation of a biological neuron model and is also called the threshold value of the imitation of a biological neuron model.

2.2. **Samples.** The samples in this study are selected from the Taiwan Economic Journal (TEJ). The research period is from 2008 to 2014. During this period, 32 companies with fraudulent financial statements and 96 companies without fraudulent financial statements are paired based on the ratio of 1:3, for a total of 128 companies.

2.3. **Variables.** The dependent variable is expressed by a dummy variable. The classification is done on the basis of whether a company has fraudulent financial statements. A company with fraudulent financial statements is set as 1, while a company without fraudulent financial statements is set as 0. In this study, there are 29 variables to measure the fraudulent financial statements, including 22 financial variables and 7 non-financial variables. Financial variables include X1: Current ratio, X2: Quick ratio, X3: Accounts receivable turnover, X4: Inventory turnover, X5: Debt ÷ equity ratio, X6: Net profit rate, X7: Return on assets (ROA), X8: Return on equity (ROE), X9: Interest paid ÷ total liabilities, X10: Inventory ÷ net sales, X11: Average inventory ÷ total assets, X12: Net sales ÷ total assets, X13: Gross profit ÷ total assets, X14: Current liabilities ÷ total assets, X15: Operating cash flow ÷ net sales, X16: Operating cash flow ÷ current liabilities, X17: Debt ratio, X18: Operating revenue growth rate, X19: Total assets growth rate, X20: Gross profit rate, X21: Operating expenses ratio, X22: Fixed assets ÷ total assets; non-financial variables include X23: Audited by BIG 4, X24: Ratio of stocks held by directors and supervisors, X25: Number of directors and supervisors, X26: Audit committee, X27: Number of audit committee members, X28: Independence of directors and supervisors, X29: Restatement of financial statements.

2.4. **Research design and process.** This study establishes FFS detection models in two stages. CART and CHAID are very suitable for selecting important variables; C5.0, ANN, and CHAID are very suitable for classifying, predicting, and detecting [9,10,14]. For Stage I, CART and CHAID are respectively used to select a few influential and important variables from 29 financial and non-financial variables. In Stage II, C5.0, ANN, and CHAID are used to establish the fraudulent financial statements detection models and gain detection accuracy, Type I error rate, and Type II error rate. Their detection accuracy is then compared, so as to get the best model with the highest accuracy. Figure 1 shows the research design and process.

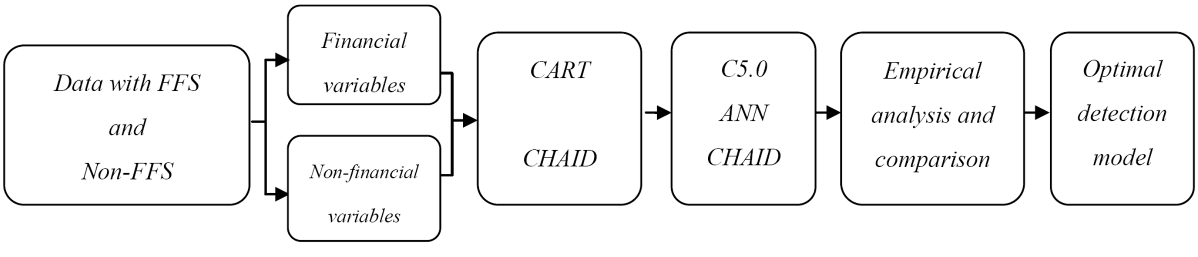

Figure 1. Research design and process

3. **Empirical Results and Analysis.**

CART and CHAID are used as variable selection methods in Stage I. The selection results are described as follows.

1) Variable selection results of CART. A total of eight influential variables are selected. The order of importance of the variables is: X10 (0.210): Inventory ÷ net sales, X9 (0.164): Interest paid ÷ total liabilities, X13 (0.079): Gross profit ÷ total assets, X7 (0.039): ROA, X2 (0.039): Quick ratio, X29 (0.039): Restatement of financial statements, X1 (0.039): Current ratio, and X25 (0.039): Number of directors and supervisors.

2) Variable selection results of CHAID. A total of five influential variables are selected. The order of variable importance is: X23 (0.310): Audited by BIG 4, X10 (0.263): Inventory ÷ net sales, X16 (0.232): Operating cash flow ÷ current liabilities, X14 (0.120): Current liabilities ÷ total assets, and X13 (0.075): Gross profit ÷ total assets.

After the significant variables are screened out in Stage I, C5.0, ANN, and CHAID are used for modeling in Stage II. After normalization of the selected variables, a random non-repeated sampling is conducted. This study adopts SPSS Clementine to conduct ten-fold cross validation which is recognized by academic circles as a more rigorous cross validation to gain the detection accuracy rate [10,11,14], Type I error rate and Type II error rate. The dataset is divided into 10 parts: 9 parts are used as a training group in turn; 1 part is used as the test group to be experimented with one by one, and then the average of the detection accuracy rate is gained.

1) CART models. As shown in Table 1, after the CART models go through a ten-fold cross validation, CART+C5.0 has the highest detection accuracy of FFS (87.75%). Regarding overall accuracy, CART+C5.0 also has the highest detection accuracy (86.74%). As shown in Table 2, CART+C5.0 presents the lowest Type I error rate (12.25%) and Type II error rate (14.27%).

Table 1. CART models' accuracy using ten-fold cross validation

Model	FFS detection accuracy	Non-FFS detection accuracy	Overall accuracy
CART+C5.0	87.75%	85.73%	86.74%
CART+ANN	77.27%	75.87%	76.57%
CART+CHAID	76.04%	72.25%	74.15%

Table 2. Type I error and Type II error

Model	Type I error rate	Type II error rate	Overall error rate
CART+C5.0	12.25%	14.27%	13.26%
CART+ANN	22.73%	24.13%	23.43%
CART+CHAID	23.96%	27.75%	25.85%

2) CHAID models. As shown in Table 3, after the CHAID models go through ten-fold cross validation, CHAID+CHAID has the highest detection accuracy of FFS (93.54%). Regarding overall accuracy, CHAID+CHAID also has the highest detection accuracy (93.39%). As shown in Table 4, CHAID+CHAID presents the lowest Type I error rate (6.46%) and Type II error rate (6.61%).

4. **Conclusions.**

There has been an unceasingly increase in the number of cases of fraudulent financial statements by listed firms worldwide. When a company issues serious fraudulent financial statements, it not only damages its reputation, but also causes heavy losses to investors, and society then must pay a great cost to make up for these damages. Therefore, how to establish an effective detection model of fraudulent financial statements

Table 3. CHAID models' accuracy using ten-fold cross validation

Model	FFS detection accuracy	Non-FFS detection accuracy	Overall accuracy
CHAID+C5.0	87.02%	86.00%	86.51%
CHAID+ANN	81.99%	81.30%	81.65%
CHAID+CHAID	93.54%	93.39%	93.47%

Table 4. Type I error and Type II error

Model	Type I error rate	Type II error rate	Overall error rate
CHAID+C5.0	12.98%	14.00%	13.49%
CHAID+ANN	18.01%	18.70%	18.35%
CHAID+CHAID	6.46%	6.61%	6.53%

has become a very important topic. CART and CHAID are very suitable for selecting important variables; C5.0, ANN, and CHAID are very suitable for classifying, predicting, and detecting. This study utilizes CART and CHAID in the first stage for the selection of important variables and in the second stage, in combination with C5.0, CHAID, and ANN, establishes fraudulent financial statements detection models. The empirical results indicate that the CHAID+CHAID model has the highest detection accuracy for fraudulent financial statements (FFS: 93.54%, Non-FFS: 93.39%), while the same model also has a lower rate of Type I error and Type II error. The research results provide a reference for related academic research personnel, auditors, CPAs, investment bank analysts, and credit rating agencies in the accounting, finance, and operating management fields. It is suggested herein that other data mining techniques can be adopted to establish fraudulent financial statements detection models in future studies.

REFERENCES

[1] H. C. Koh, Going concern prediction using data mining techniques, *Managerial Auditing Journal*, vol.19, pp.462-476, 2004.

[2] S. Kotsiantis, E. Koumanakos, D. Tzelepis and V. Tampakas, Forecasting fraudulent financial statements using data miming, *International Journal of Computational Intelligence*, vol.3, no.2, pp.104-110, 2006.

[3] E. Kirkos, C. Spathis, A. Nanopoulos and Y. Manolopoulos, Identifying qualified auditors' opinions: A data mining approach, *Journal of Emerging Technologies in Accounting*, vol.4, no.1, pp.183-197, 2007.

[4] P. F. Pai, M. F. Hsu and M. C. Wang, A support vector machine-based model for detecting top management fraud, *Knowledge-Based Systems*, vol.24, pp.314-321, 2011.

[5] P. Ravisankar, V. Ravi, G. R. Rao and I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, *Decision Support Systems*, vol.50, pp.491-500, 2011.

[6] W. Zhou and G. Kapoor, Detecting evolutionary financial statement fraud, *Decision Support Systems*, vol.50, pp.570-575, 2011.

[7] R. Gupta and N. S. Gill, Analysis of data mining techniques for detection of financial statement fraud, *International Journal of Computer Applications*, vol.50, no.8, pp.7-14, 2012.

[8] M. Salehi and F. Z. Fard, Data mining approach to prediction of going concern using classification and regression tree (CART), *Global Journal of Management and Business Research*, vol.13, no.3, pp.24-30, 2013.

[9] S. Chen, Z. D. Shen and Y. J. Goo, A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements, *The Scientific World Journal*, vol.2014, 2014.

[10] S. Chen, Detection of fraudulent financial statements using the hybrid data mining approach, *SpringerPlus*, vol.5, no.89, 2016.

[11] C. C. Yeh, D. J. Chi, T. Y. Lin and S. H. Chiu, A hybrid detecting fraudulent financial statements model using rough set theory and support vector machines, *Cybernetics and Systems*, vol.47, no.4, pp.261-276, 2016.

[12] L. Breiman, J. H. Friedman, R. A. Olshen and C. I. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.

[13] J. R. Quinlan, *C5.0: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., 1986.

[14] S. Chen and J. Lee, Going concern prediction using data mining, *ICIC Express Letters, Part B: Applications*, vol.6, no.12, pp.3311-3317, 2015.