ICIC Express Letters
Part B: Applications
Volume **7**, Number **11**, November **2016**
 ISSN 2185-2766
pp. **2361–2366**

APPLICATION OF CORRELATION-BASED FEATURE SELECTION AND DECISION TREE TO DETECT EARNINGS MANAGEMENT AND ACCOUNTING FRAUD RELATIONSHIP

Kuang-Hua Hu[1], Fu-Hsiang Chen[2] and We-Jie Chang[2]

[1]Accounting School
Nanfang College of Sun Yat-sen University
China Wenquan Town, Conghua, Guangzhou 510970, P. R. China
khhu0622@gmail.com

[2]Department of Accounting
Chinese Culture University
No. 55, Hwa Kang Rd., Yang Ming Shan, Taipei 11114, Taiwan
chenfuhsiang1@gmail.com

Received April 2016; accepted July 2016

Abstract. *In previous studies, the accounting fraud firms are often manipulating earnings management in order to achieve the goals set by the companies. The earnings management has been considered directly related to the accounting fraud. Data mining technique had also been applied to the fields of accounting and financial studies. This research has integrated correlation-based feature selection and decision tree to construct a corporate financial statement earnings management and accounting fraud relationship detection model. The research used the correlation-based feature selection techniques to identify relevant variables in the first place, followed by adopting three kinds of decision trees including REPTree, CART and C4.5 to establish a model to detect earnings management and accounting fraud relationship. The results show that the previous year earnings management of the accounting fraud is importance variable and the CART has the optimal classification rate and the accuracy rate is 75%.*
Keywords: Correlation-based feature selection, Decision tree, Accounting fraud, Earnings management

1. **Introduction.** In recent years, many enterprises have had accounting frauds, and most of the accounting frauds are caused by the illegal activity of the management level. For example, the decline of Arthur Anderson CPA Firm, in 2002, WorldCom burst the accounting fraud scandal that the president and CEO Bernard Ebbers, in order to hide the loss, raised the stock price and listed the income falsely, causing bankruptcy. Enterprises exaggerate the surplus, list assets falsely and hide debts skillfully by various means, causing the corporate reorganization, and even bankruptcy and resulting in great economic impact, which not only makes the investors believe the false information of financial statements by mistake, causing information asymmetry, and investors' loss, but also will damage the image of accountants, making the public doubt the independence and audit report of the accounts and destroying the investors' confidence.

When financial statements involve the intentional modification or false listing of financial statements and accounting fraud has become a serious economic and social problem, its universality will directly affect employees, shareholders and creditors, and indirectly affect the market participants by destroying the reliability of financial statements in the financial market [1]. Earnings management has always been deemed to be closely related with accounting fraud, and some studies also show that the companies participating in the accounting fraud will achieve the objective set by them by means of manipulating earnings management [2]. Academics pointed out in the past that investors often do not

understand how the management of a company manipulates earnings management trying to increase their rewards, have their companies go public and boost their companies' stock prices [3,4]. Perols and Lougee [1] proposed that the companies with accounting fraud will manipulate earnings for accounting fraud by exaggerating the income, but not by manipulating in order to achieve the expectancy of the analysts, so when reaching the extremity, earning management crosses the boundary of objective financial statements and fraud financial statements and should be detected. In order to understand whether a company conducts accounting fraud by manipulating the earnings, in this research, the relation of accounting fraud is estimated by applying earnings management, so as to find out the clues.

In the past, most earnings management and accounting fraud related studies have adopted conventional statistical methods, such as logistic regression and multiple discriminant analysis [1-5]. These conventional statistical methods, however, have some restrictive assumptions such as linearity, normality, and independence of input variables. Considering that the violation of these assumptions occurs frequently within financial data, the methods have intrinsic limitations in terms of effectiveness and validity [6]. Data mining is a process to transform data into knowledge, which is one of the most active ways in research, development and application in the field of data processing. The advantage to exploring earnings management and accounting fraud relationship with the data mining method is to set up a non-linear model which does not require hypotheses as what is required by the conventional method [7]. In order to help corporate stakeholders avoid suffering a great loss in the stock market as a result of manager's earnings management and accounting fraud. This study proposes an integrated model by infusing soft computing methods to resolve the problem. At the first stage, considering the number of earnings management, the correlation-based feature selection techniques (CFS) are used to obtain significant earnings management manipulating period from historical data. At the second stage, decision tree are used to generate meaning rules for earnings management and accounting frauds identify relationship.

The structure of the study is divided into five parts, in which earnings management and accounting fraud relationship's study motivation are first explained, followed by exploration of the literature of earnings management, accounting fraud, CFS and the decision tree (DT) applied in relevant fields. Then, our study discourses on the adopted methodology before analysis of the empirical results of the earnings management and accounting fraud. Finally, conclusion, recommendations and possible future direction are proposed.

2. **Preliminary.**

This study infuses several computational methods to resolve the earnings management and accounting fraud relationship, and this section briefly reviews earnings management, accounting fraud, and the origins and concepts of the used methods.

2.1. **Earnings management.**

Earnings management is to manipulate earnings through certain methods or processes in an attempt to meet managers' manipulation purposes. The core issue of earnings management is the leverage between relevance and reliability of accounting information. Even though relevance and reliability do not completely exclude each other, the reliability of accounting information often rises as increase of relevance, whereas the financial information with relevance often requires the management to make various estimations for accounting items [6].

2.2. **Financial statement fraud.**

According to the U.S.'s Statement on Auditing Standards (SAS) No. 99, financial statement fraud can be defined and described as accounting fraud or error resulting from unfaithful representation of financial statements. It refers to deliberate misstatement, amount omission or disclosure, or preparation of the financial statements causing others to misunderstand. According to Taiwan's Statement on Auditing Standards No. 43, unfaithful representation of financial statements may result

from accounting fraud or error. The difference between accounting fraud and error lies in whether the motivation of the unfaithful representation of financial statements is deliberate. The accounting fraud refers to the behavior that the management or one or more governance unit members deliberately conduct with the deceitful method in order to acquire ill-gotten or illegal benefits. The standard emphasizes that the CPA shall follow the generally accepted auditing standards to execute their auditing tasks, and reasonably confirm no material misstatement of the financial statements.

2.3. **Feature selection approach.** CFS is an algorithm that couples this evaluation formula with an appropriate correlation measure and a heuristic search strategy. CFS can quickly identify and screens irrelevant, redundant, and noisy features, and identifies relevant features as long as their relevance does not strongly depend on other features [8]. Due to CFS's excellent performance in classification tasks, it was adopted by the study as obtaining significant earnings management manipulating period.

2.4. **Classification approach.** Decision tree is one of common DM (data mining) methods which simultaneously have both classification and predictive functions. By focusing on the data provided, it could produce a model of tree-shaped structure using inductive reasoning [5]. DT does not require any statistical assumptions concerning the data in a training sample and can handle incomplete and qualitative data. The study used DT to generate meaning rules for earnings management and accounting frauds identify relationship [8].

3. Research Methodology.

3.1. **Method to measure earnings management.** Prior studies usually used the following two methods to estimate the accrual earnings management: the cash flow statement method and balance sheet method [10]. As indicated by Collins and Hribar [10], the deviation of total accruals estimated by the cash flow statement method is smaller. Hence, the research adopted the cash flow statement method proposed by Collins and Hribar [10] to assess earnings management.

The cash flow statement method is shown as per Formula (1) below:

$$TAC_{i,t} = EBXI_{i,t} - CFO_{i,t} \tag{1}$$

The variables are defined as follows.

$TAC_{i,t}$: the total accruals of company i in the t^{th} year.

$EBXI_{i,t}$: the continuing operations' income of company i in the t^{th} year.

$CFO_{i,t}$: the cash flow of company i's operating activities in the t^{th} year.

In this research, subject to the year of accounting fraud determined by the companies with accounting fraud, the TAC in the year of accounting fraud, the previous one year and two years is calculated according to Equation (1), and CFS is used to judge the importance of TAC and accounting fraud identity in the 3 years above.

3.2. **Decision tree.** Decision tree is a tool to establish classification models and give predictions [6]. This research adopted three methods of REPTree, CART and C4.5, which are described respectively as below.

3.2.1. *REPTree.* Basically reduced error pruning tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information result or decreasing the variance. REPTree uses the regression tree logic and creates multiple trees in different iterations. After that, it selects the best one from all generated trees. That will be considered as their presentative. In pruning the tree, the measure used is the mean square error on the predictions made by the tree [11].

3.2.2. *CART.* Classification and regression tree (CART) is a binary splitting decision tree technique, and applied to the attribute where the data are continuous or classified non-parameters, and its selection of splitting terms is determined by the data's classification and attribute. CART will test the attributes of all the data, and split them into two subsets according to their respective attribute values, followed by calculating the Gini value divided from each attribute [7]. CART's advantage is that CART can automatically inspect models, and identify the optimized general models, and is used to establish a very complex tree that is trimmed into an optimized one according to the results of interactive tests and inspection.

3.2.3. *C4.5.* C4.5 is an algorithm used to generate a decision tree developed by Quinlan [12]. For consecutive value segmentation criteria, the target sets are sequenced and the medium-value of the attributes of two neighboring objects is identified as the segmentation point. In the case of missing or uncertain attribute values, replacement of the most common attribute value or optimistic estimation of probability is generally used. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

4. Empirical Results and Analysis.

4.1. **Data and sample.** The research takes the companies which had identified the accounting fraud in Taiwan as samples from 2003 to 2012, and the pairing of sample companies is selected by the same industry and the similar total assets. In this research, we take the ratio of 1:2 to reduce the oversampling shortcomings of the accounting fraud companies.

4.2. **Accounting fraud companies.** It takes many years to judge and determine an accounting fraud case, so the samples in this research are selected based on the great event of listed companies in Taiwan economic journal (TEJ) from 2003 to 2012, and based on the companies that conform to the definition of Taiwan Generally Accepted Auditing Standards No. 43 with regard to accounting fraud. In the case of incomplete data due to a lack of subvalue, the samples are deleted, and finally 20 companies with accounting fraud are selected as the samples finally. The selection conditions are as follows: 1) the accounting fraud companies conforming to the matching with healthy companies; 2) the companies that are or were listed companies; 3) determined through judgment; 4) during 2003-2012, with the year of accounting fraud as the benchmark, it is possible to calculate the figures of earnings management in the year of accounting fraud, the previous one year and two years. At last, this research adopts 20 accounting fraud companies and 40 financial healthy companies as our samples. The sample selection is shown in Table 1 below.

Table 1. This study's sample selection

Sample selection	**The number of samples**
2003-2012 accounting fraud companies	20
2003-2012 healthy companies	40
Final number of samples	60

The research adopted WEKA to execute CFS and decision tree. At the same time, the research also further disclosed type I and type II errors of each model. Type I error shows the situation where accounting fraud is actually in serious error but is classified to no accounting fraud error rate, whereas type II error is the no accounting fraud, but is classified to having accounting fraud error rate. With 20 accounting fraud companies and 40 healthy companies, the TAC in the year of accounting fraud, the previous one year

Table 2. Total accrual (TAC) descriptive statistics

Max	Min	Standard deviation	Number of samples	Median	Mean
0.98036	−0.94877	0.18502	180	0.010636	0.024749

Table 3. Accuracy and error rate of REPTree, CART and C4.5

	Overall accuracy	Overall error	
REPTree	65%	Type I error	15.00%
		Type II error	32.94%
CART	75%	Type I error	12%
		Type II error	28.57%
C4.5	63%	Type I error	5.00%
		Type II error	2.5%

and two years, is calculated according to Equation (1) in 3.1 above, totally 180 data, and the descriptive statistics of TAC are as shown in Table 2.

D'Amico and Mafrolla [2] pointed out that the data of earnings management show that the total accrual of the accounting fraud companies is much higher than that of non-accounting fraud companies, so in this research, CFS is used as the TAC in the year of accounting fraud, the previous one year and two years, to determine the importance of accounting fraud; in other words, the research variables in this research are X_1 (TAC of the accounting fraud companies and healthy companies in the year of accounting fraud calculated based on Equation (1)), X_2 (TAC of the accounting fraud companies and healthy companies in the previous one year of accounting fraud calculated based on Equation (1)) and X_3 (TAC of the accounting fraud companies and healthy companies in the previous two years of accounting fraud calculated based on Equation (1)), and the result is TAC (X_2) of one year prior to accounting fraud, with importance; in other words, it is important to detect the accounting fraud by observing the earnings management occurring one year prior to accounting fraud. Table 3 shows the accuracy and error rate, the accuracy of REPTree, CART and C4.5 is above 63%, and CART has the highest overall classification accuracy at 75%. As shown in Table 3, C4.5 has the lowest type I error at 5.00%.

5. **Conclusion and Recommendations.** The research adopts Taiwan's accounting fraud companies listed in the TSEC/OTC market from 2003 to 2012 as the samples to detect earnings management and accounting frauds identify relationship. It first tried the CFS to screen the impartment variables, followed by decision tree to detect earnings management and accounting frauds identify relationship. The empirical results show the previous year earnings management of the accounting fraud caused (X_2) is importance variable and CART can have the best performance, and C4.5 have the lowest type I error. The research overcame the restriction of the conventional statistic method and could be more applicable to the real world.

Despite the contributions of this study, there are some limitations: First, only the CFS was adopted for the feature selection, and the obtained decision rules or accuracy of approximation might be different by using the other methods. Future studies may incorporate some other machine learning techniques to find the optimal feature selection. Second, the model only used 2 period-lagged earnings management data to identify accounting fraud relationship (i.e., associate the data of earnings management in t, $t-1$, $t-2$ periods with its decision class in period t). Some latent tendency in relatively long-lagged periods (e.g., more than 3 years) might not be captured in the model. Future studies are recommended to pursue solutions in this direction.

REFERENCES

[1] J. L. Perols and B. A. Lougee, The relation between earnings management and financial statement accounting fraud, *Advances in Accounting, Incorporating Advances in International Accounting*, vol.27, pp.39-53, 2011.

[2] E. D'Amico and E. Mafrolla, The importance of earnings management detection models to identify fraud: A case from Italian listed firms, *Journal of Modern Accounting and Auditing*, vol.9, no.1, pp.68-75, 2013.

[3] C. S. Armstrong, D. F. Larcker, G. Ormazabal and D. J. Taylor, The relation between equity incentives and misreporting: The role of risk-taking incentives, *Journal of Financial Economics*, vol.109, no.2, pp.327-350, 2013.

[4] S. Kedia, K. Koh and S. Rajgopal, Evidence on contagion in earnings management, *The Accounting Review*, vol.90, no.6, pp.2337-2373, 2015.

[5] V. W. Fang, A. H. Huang and J. M. Karpoff, Short selling and earnings management: A controlled experiment, *The Journal of Finance*, vol.71, no.3, pp.1251-1294, 2016.

[6] F. H. Chen, D. J. Chi and Y. C. Wang, Detecting biotechnology industry's earnings management using Bayesian network, principal components analysis, back propagation neural network, and decision tree, *Economic Modeling*, vol.46, pp.1-10, 2014.

[7] F. H. Chen and H. Hu, An alternative model for the analysis of detecting electronic industries earnings management using stepwise regression, random forest, and decision tree, *Soft Computing*, vol.20, no.5, pp.1945-1960, 2016.

[8] K. O. Akande, T. O. Owolabi and S. O. Olatunji, Investigating the effect of correlation-based feature selection on the performance of support vector machines in reservoir characterization, *Journal of Natural Gas Science and Engineering*, vol.22, pp.515-522, 2015.

[9] H. Ozturk, E. Namli and H. I. Erdal, Modelling sovereign credit ratings: The accuracy of models in a heterogeneous sample, *Economic Modelling*, vol.54, pp.469-478, 2016.

[10] D. W. Collins and P. Hribar, Errors in estimating accruals: Implications for empirical research, *Journal of Accounting Research*, vol.40, no.1, pp.105-134, 2002.

[11] S. Kalmegh, Analysis of WeKA data mining algorithm REPTree, simple cart and random tree for classification of Indian news, *International Journal of Innovative Science, Engineering & Technology*, vol.2, no.2, pp.438-446, 2015.

[12] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.