ICIC Express Letters
Part B: Applications
Volume **7**, Number **12**, December **2016**
 ISSN 2185-2766
pp. **2677–2682**

AN APPLICATION OF DATA MINING TECHNIQUES ON EARNINGS MANAGEMENT DETECTION

Yung-Ming Hsieh[1,*] and Yi Kao[2]

[1]Department of Accounting
Soochow University
No. 56, Kuei-Yang Street, Section 1, Taipei 10048, Taiwan
*Corresponding author: armin@scu.edu.tw

[2]Department of Accounting
Chinese Culture University
Rm. 818, 8F., No. 55, Huagang Rd., Shilin Dist., Taipei City 111, Taiwan
eddieku2@hotmail.com.tw

Received May 2016; accepted August 2016

Abstract. *This study applies data mining techniques to creating more effective models for detecting corporate earnings management. The data mining techniques, including the artificial neural network (ANN) combined with the decision tree CHAID and decision tree C5.0, are used to establish a two-stage approach for developing the earnings management detection models with relatively higher accuracy. The research variables include both financial variables and non-financial variables. Our empirical result shows that the ANN/C5.0 model generates the best accuracy rate for earnings management detection.*
Keywords: Earnings management detection model, Artificial neural network, Decision tree CHAID, Decision tree C5.0, Stepwise regression analysis

1. **Introduction.** In the worldwide capital markets, since earnings on financial statements express operating performance, investors deem earnings as the decision-making indicator which induces incentives for earnings management. The bonus plan hypothesis states that, a bonus plan can bring about moral crisis in enterprises that use accounting earnings as the bonus mechanism [1]. Manager remunerations are closely related to earnings. For maximizing their self-interests, top management or managers have higher opportunity of using earnings management to achieve business objectives [2], namely, behavior under self-interest motivation. Management affects the preparation of financial statements, and thus, asymmetry between the managers and the information users may occur. In addition, management motivated by the performance threshold manipulates discretionary accruals to increase their own remunerations and boost shares price [3,4]. To constrain those opportunity behaviors, regulators need a more effective model to detect earnings management especially on corporate top management or managers.

In fact, it is difficult to measure earnings management from commercial activities, and thus, [5] pointed out that, most earnings management affects earnings on financial statements through discretionary accruals (DA). Management can disclose more accurate or misleading information through earnings management. Therefore, it is very important to establish effective earnings management detection models. Previous studies have proposed data mining techniques to detect earnings management and financial frauds [6-8]. However, the past research only addressed on one-stage approach to build the detection models of earnings management. This encourages current study to explore a two-stage (variable selection and model building) approach for developing more accurate detection models.

According to [9], the Jones model and the modified Jones model, which are used to estimate discretionary accruals, have errors [3,10], and thus, the performance adjustment method is proposed to correct the estimates of discretionary accruals. Return on assets (ROA) is added as a variable to correct errors caused by abnormal performance. As compared to the Jones model and the modified Jones model, the Kothari model has lower probability regarding type I and type II errors. The non-discretionary accrual, as presented by [9], is as shown in Equation (1).

$$\frac{\mathrm{TA}_{i,t}}{\mathrm{A}_{i,t-1}} = \alpha + \beta_1 \left(\frac{1}{\mathrm{A}_{i,t-1}}\right) + \beta_2 \frac{(\Delta\mathrm{REV}_{i,t} - \Delta\mathrm{REC}_{i,t})}{\mathrm{A}_{i,t-1}} + \beta_3 \left(\frac{\mathrm{PPE}_{i,t}}{\mathrm{A}_{i,t-1}}\right) + \beta_4 \mathrm{ROA}_{i,t} + \varepsilon_{i,t} \quad (1)$$

where, $\mathrm{TA}_{i,t}$: total accruals of firm i during period t; $\mathrm{A}_{i,t-1}$: total assets of firm i during period $t-1$; $\Delta\mathrm{REV}_{i,t}$: changes in net sales revenue of firm i during period t; $\Delta\mathrm{REC}_{i,t}$: changes in net accounts receivables of firm i during period t; $\mathrm{PPE}_{i,t}$: total depreciable fixed assets of firm i during period t; $\mathrm{ROA}_{i,t}$: return on assets of firm i during period t; α: constant term; β_i: regression coefficient terms; $\varepsilon_{i,t}$: residual term.

The equation used to estimate discretionary accruals is expressed by Equation (2)

$$\begin{aligned} \mathrm{DA}_{i,t} = \frac{\mathrm{TA}_{i,t}}{\mathrm{A}_{i,t-1}} - \Bigg(\hat{\alpha} + \hat{\beta}_1 \left(\frac{1}{\mathrm{A}_{i,t-1}}\right) + \hat{\beta}_2 \frac{(\Delta\mathrm{REV}_{i,t} - \Delta\mathrm{REC}_{i,t})}{\mathrm{A}_{i,t-1}} \\ + \hat{\beta}_3 \left(\frac{\mathrm{PPE}_{i,t}}{\mathrm{A}_{i,t-1}}\right) + \hat{\beta}_4 \mathrm{ROA}_{i,t}\Bigg) \end{aligned} \quad (2)$$

where, $\mathrm{DA}_{i,t}$: estimated discretionary accruals (an indicator of earnings management) of firm i during period t; $\hat{\alpha}$: estimated constant term; $\hat{\beta}_i$: estimated coefficient terms.

In this study, the mean value and standard deviation of the absolute value of discretionary accruals ($\mathrm{DA}_{i,t}$) calculated are 0.231 and 0.624, respectively. When the absolute $\mathrm{DA}_{i,t}$ is greater than 0.543 (about 30% of sample) it is defined as 1, and there is serious earnings management through the manipulation of discretionary accrual. Then, when it is smaller than 0.543 it is defined as 0, and there is minor earnings management.

This paper is organized as follows: Section 1 states the background, motivation, and purpose of this study, and explains the uses of stepwise regression analysis and data mining techniques (ANN, CHAID, and C5.0) to establish the earnings management detection models; Section 2 is the methodology section, which describes the research methods and samples used, as well as the research procedure followed in this study; Section 3 is the results and discussion section, which explains and discusses the comparison of the models' accuracy; finally, Section 4 is the conclusion section, which is drawn from the results of the research, and concludes the research findings for academic literature and practice.

2. **Methodology.**

2.1. **Research method.** In addition to stepwise regression analysis (SRA), this study utilizes several data mining techniques, such as artificial neural network (ANN), decision tree CHAID, and decision tree C5.0. ANN is a parallel computational model of an artificial neural network, which is a processing technology inspired by the study of the brain and nervous system, often called the Parallel Distributed Processing Model. The artificial neural network theory was proposed in the 1950s, when scientists put forward the perceptron neuron model by simulating human brain organization and operation mode, which is the simplest and earliest artificial neural model. The perceptron is often used as a classifier. Before 1980, the artificial neural network was not taken seriously as an expert system, was the most popular artificial intelligence and the artificial neural network was not mature. Decision tree CHAID is a chi-square automatic cross-validation (CHAID) method used to calculate the P values of the branch nodes of the tree, and determines node splitting, where the CHAID advantage is to prevent data abuse and force the decision tree to stop splitting, meaning CHAID can finish trimming before modeling. Decision

tree C5.0 is developed by modifying and adjusting the ID3 (Iterative Dichotomiser 3), as proposed by Quinlan in 1986, because ID3 does not have good capability for processing continuous data. C5.0 first classifies data into the same area, selects the criteria of branch attributes, and then calculates the information gain of each attribute in order to facilitate choosing the optimum attributes. As compared to other decision trees, C5.0 is more stable in processing data with complicated attributes and numerous input fields.

2.2. **Sample selection.** In this study, the research sample includes all companies in IC design and wafer manufacturing industry and listed in Taiwan Stock Exchange or Over-the-Counter (OTC) markets during years of 2009-2014. All data are collected from the Taiwan Economic Journal (TEJ) database. If the companies have missing data on dependent or independent variables affecting earnings management through accruals, they are deleted. Finally, a total of 748 effective observations (firm/year) are obtained in the research sample.

2.3. **Variables.** The variables used in this study include financial variables and non-financial variables which were selected to measure the magnitude of earnings management. The financial variables include 17 variables (X1 to X17) as shown in Table 1. The non-financial variables include X18: the ratio of stocks held by board directors and supervisors, X19: the ratio of stocks held by institutional investors, X20: the pledged ratio of stocks held by board directors and supervisors, and X21: audited by BIG 4 auditors' firms.

Table 1. Summary of selected financial variables

No.	*Variables*	*No.*	*Variables*
X1	ROA	X10	Cash flow per share
X2	Gross profit rate	X11	Current ratio
X3	Operating income rate	X12	Quick ratio
X4	Net income rate	X13	Debt ratio
X5	Continuous net income rate	X14	Debt/equity ratio
X6	Continuous earnings per share (EPS)	X15	Inventory turnover
X7	Income before tax per share	X16	Accounts receivable turnover
X8	Operating income per share	X17	Net assets turnover
X9	Operating revenue per share		

2.4. **Research procedure.** The data of this study are from the TEJ, and input variables include financial and non-financial variables. The earnings management detection model is established in two stages. In the first stage, SRA and ANN, respectively, were used to select the variables by a tool of SPSS Molder14.1. All variables were normalized first before running SRA and ANN. In the second stage, CHAID and C5.0 were used for modeling, the detection accuracies of the models were compared, and the optimal model was obtained. The research procedure is as shown in Figure 1.

3. **Results and Discussion.** As there are a number of variables to be input, SRA and ANN methods were used to select the variables that could possibly improve the prediction accuracy. After the first stage of variable selection, the second stage of model building with two decision tree methods (CHAID and C5.0) established the earnings management detection models before comparing the detection accuracy rates of four models ($2 \times 2 = 4$).

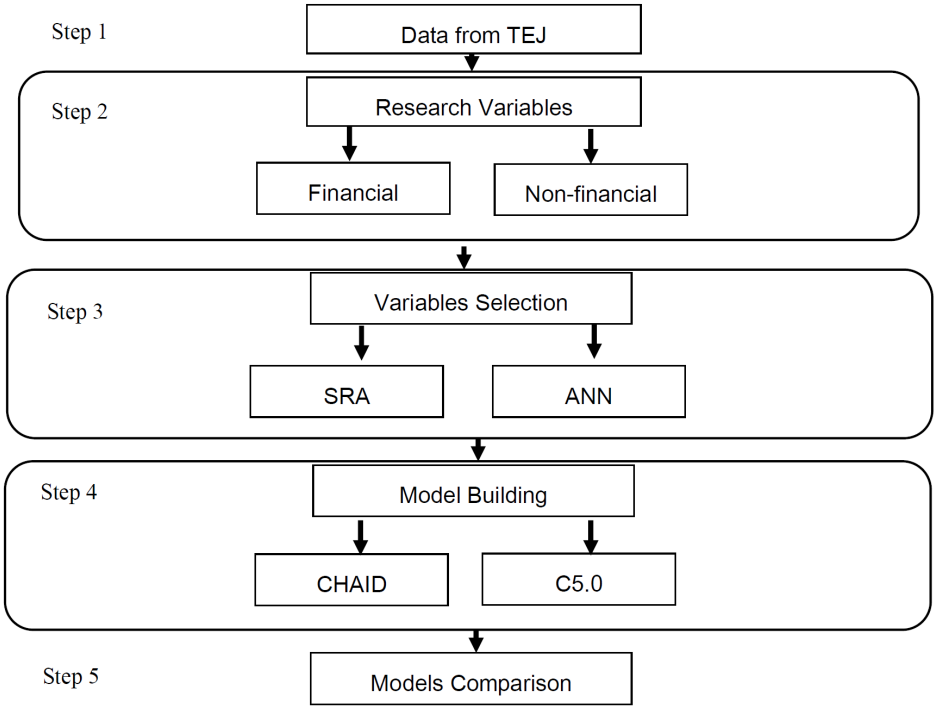

FIGURE 1. Research procedure

3.1. **Variables selection by stepwise regression analysis.** After the selection process of the SRA, there are 8 of 21 most important variables remaining. The eight variables (sorted by their importance) include: X19 (0.37): the ratio of stocks held by institutional investors, X11 (0.21): current ratio, X9 (0.13): operating revenue per share, X7 (0.08): income before tax per share, X13 (0.07): debt ratio, X14 (0.05): debt/equity ratio, X6 (0.04): continuous EPS, and X20 (0.03): the pledged ratio of stocks held by board directors and supervisors. The results indicated that SRA selected higher important factors from both financial and non-financial variables which implied management motivated by the performance target tends to manipulate earnings to increase their own interests and boost shares price.

3.2. **Variables selection by artificial neural network.** Through the ANN method, there were 10 of 21 important variables screened. The ten variables (sorted by their importance) include: X7 (0.11): income before tax per share, X12 (0.08): quick ratio, X16 (0.08): accounts receivable turnover, X19 (0.08): the ratio of stocks held by institutional investors, X6 (0.07): continuous EPS, X17 (0.07): net assets turnover, X5 (0.05): continuous net income rate, X14 (0.05): debt/equity ratio, X20 (0.05): the pledged ratio of stocks held by board directors and supervisors, and X9 (0.04): operating revenue per share. The ANN method selected more variables but with lower importance weights than SRA method. However, ANN also selected both financial and non-financial variables to predict the possibility of earnings management.

3.3. **SRA models.** This part illustrates the results when SRA selected variables in the first stage, and then two models were constructed with CHAID and C5.0 in the second stage. It was found that, the SRA/C5.0 model has the higher detection accuracy rate of 96.61%, while SRA/CHAID is 95.37%, as shown in Table 2. The results show that the SRA/C5.0 model can predict better than the SRA/CHAID model from the overall accuracy or the point of view of both type I and II errors.

Table 2. Detection accuracy of models – selected by SRA

Model	Overall detection accuracy	Subgroups accuracy		Type I Error & Type II Error	
SRA/CHAID	95.37%	Training groups	94.08%	Type I Error	1%
				Type II Error	49%
		Testing groups	96.67%	Type I Error	0%
				Type II Error	19%
SRA/C5.0	96.61%	Training groups	95.46%	Type I Error	2%
				Type II Error	23%
		Testing groups	97.75%	Type I Error	1%
				Type II Error	12.5%

Table 3. Detection accuracy of models – selected by ANN

Model	Overall detection accuracy	Subgroups accuracy		Type I Error & Type II Error	
ANN/CHAID	94.64%	Training groups	93.37%	Type I Error	4%
				Type II Error	26%
		Testing groups	95.91%	Type I Error	2%
				Type II Error	5%
ANN/C5.0	**96.97%**	Training groups	96.76%	Type I Error	1%
				Type II Error	21%
		Testing groups	97.18%	Type I Error	1%
				Type II Error	18%

3.4. **ANN models.** After the first stage with ANN and the second stage of the model constructed with CHAID and C5.0, it was found that, the ANN/C5.0 model has higher detection accuracy of 96.97%, while ANN/CHAID is 94.64%, as shown in Table 3. For the Type I error, the ANN/C5.0 model is the lowest at 1%; and the overall error rate (Type I Error & Type II Error) is also the lowest at 3.03%. The results indicate that the ANN/C5.0 model outperformed the ANN/CHAID model with highest overall detection accuracy of 96.97%, which suggests the ANN/C5.0 model is the best model to detect company's manipulation of earnings management.

4. **Conclusions.** The relevancy and reliability of financial statements, as those reports issued by enterprises and earnings management by managers through manipulation of discretionary accruals, have been addressed in accounting field for many years. Earnings management is motivated by two types of motivation. The first type is the investment in subsidiary companies, where the purpose is to conceal worse operating performance or obtain self-interest or corporate benefits. The other type is the information misleading, where earnings without management have difference in terms of accrual basis and generally accepted accounting principles (GAAP). Management can disclose more accurate or misleading information through earnings management. When financial statements involve the intentional modification or false listing of financial statements and accounting fraud has become a serious economic and social problem, it in turn needs a better model to detect earnings management and accounting frauds [11]. Therefore, it is very important to establish effective earnings management detection models.

This study establishes a two-stage approach for building earnings management detection models. In the first stage, SRA and ANN are used to screen the variables. In the second

stage, CHAID and C5.0 are used for model building. The empirical results show that the accuracy of earnings management detection is prioritized as follows: ANN/C5.0 (96.97%), SRA/C5.0 (96.61%), SRA/CHAID (95.37%), and ANN/CHAID (94.64%). This result indicates the ANN/C5.0 model outperforms other three models to detect company's earnings management.

The results of this study provide effective tools for detecting earnings management and serve as important references for stakeholders, such as auditors, CPAs (certified public accountants), management, investors (shareholders), creditors, credit rating agencies, and future academic studies. There are two suggestions for future research. First, researchers can apply the C5.0 as the start point to building models with different classifiers or statistical methods. Second, the momentum or changes of financial and non-financial indicators between different years may be good explanation or prediction variables so that future research should consider the effects of those momentums or changes for increasing the explaining power of future prediction models. We leave both issues for future research in new direction.

Acknowledgment. The authors gratefully acknowledge the comments and suggestions of the anonymous reviewers and conference participants, especially the suggestions from the seating chair, Dr. Huey-Ming Lee.

REFERENCES

[1] R. L. Watts and J. L. Zimmerman, *Positive Accounting Theory*, Prentice-Hall Company, London, 1986.

[2] P. Jiraporn, G. A. Miller, S. S. Yoon and Y. S. Kim, Is earnings management opportunistic or beneficial? An agency theory perspective, *International Review of Financial Analysis*, vol.17, no.3, pp.622-634, 2008.

[3] P. M. Dechow, R. G. Sloan and A. P. Sweeney, Detecting earnings management, *The Accounting Review*, vol.70, no.2, pp.193-225, 1995.

[4] P. M. Healy and J. M. Wahlen, A review of earnings management literature and its implications for standard setting, *Accounting Horizon*, vol.13, no.4, pp.365-384, 1999.

[5] J. J. Jones, Earnings management during import relief investigations, *Journal of Accounting Research*, vol.29, no.2, pp.193-228, 1991.

[6] X. L. Nan, X. Sun, Y. X. Li and T. S. Hou, Weighted-support vector machine based earnings management detection during IPOs, *Journal of Information & Computational Science*, vol.9, no.9, pp.2607-2617, 2012.

[7] 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 Modelling*, vol.46, pp.1-10, 2014.

[8] 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.

[9] S. P. Kothari, A. J. Leone and C. E. Wasley, Performance matched discretionary accrual measures, *Journal of Accounting and Economics*, vol.39, no.1, pp.163-197, 2005.

[10] W. R. Guay, S. P. Kothari and R. L. Watts, A market-based evaluation of discretionary accrual models, *Journal of Accounting Research*, vol.34 (supplement), pp.83-105, 1996.

[11] K. H. Hu, F. H. Chen and W. J. Chang, Application of correlation-based feature selection and decision tree to detect earnings management and accounting fraud relationship, *ICIC Express Letters, Part B: Applications*, vol.7, no.11, pp.2361-2366, 2016.