Soft Comput (2016) 20:1945–1960
DOI 10.1007/s00500-015-1616-6

METHODOLOGIES AND APPLICATION

An alternative model for the analysis of detecting electronic industries earnings management using stepwise regression, random forest, and decision tree

Fu-Hsiang Chen · Hu Howard

Published online: 18 February 2015

Abstract This study attempts to diagnose the detecting electronic industries' earnings management by integrating suitable soft computing methods. Accounting earnings information is a very crucial element for corporate stakeholders to determine their stock prices and evaluate their supervision and management authority's performance, while it is also essential information for measuring corporate value. Hence, whether an enterprise can faithfully express its true economic meaning over its financial statements and how the management handles its earnings have turned out to be a popular issue widely discussed by researchers. Detecting public companies' earnings management is an important and challenging issue that has served as the impetus in many academic studies over the last few decades. Data mining technique and machine learning methods have also been commonly applied by accounting and financial personnel to other fields of studies. The study used the stepwise regression and random forest techniques to screen the variables in the first place, followed by adopting three kinds of decision trees including Chi-squared automatic interaction detector, classification and regression trees and C5.0 to establish a model and find out if the tested enterprise had extreme earnings manipulation. The results show that the proposed hybrid approach (RF+C5.0) has the optimal classification rate (the accuracy rate is 91.24 %) and the lowest occurrence of Type I error and Type II error. Also, as discovered from the rules set of the final additional testing, an enterprise's operating cash flow, times interest earned ratio and previous period's discretionary accruals play a decisive role in affecting its extreme earnings management.

Keywords Earnings management · Random forest · Stepwise regression · CHAID · CART · C5.0 · Decision tree

Communicated by V. Loia.

F.-H. Chen (✉)
Department of Accounting, Chinese Culture University, No.55, Hwa-Kang Road, Yang Ming Sang, Taipei 11114, Taiwan
e-mail: chenfuhsiang1@gmail.com; cfh@faculty.pccu.edu.tw

H. Howard
Ernst & Young LLP, London, UK

1 Introduction

The accounting earnings information shown on a company's financial statements is usually the concerned focus and decision making indicator of the company's stakeholders covering investors, creditors, analysts and customers, whereas accounting standards allow the management authority to give some degree of professional judgment to enhance usability of financial statements. Hence, the corporate management authority has often tried to influence its corporate earnings information by selection of a variety of accounting methods (Healy 1999). As such, researchers have been concerned about whether an enterprise carries out any concrete earnings management. In this study, we propose an integrated model by infusing soft computing methods to detecting the problem.

The management or managers are likely to use the earnings management method to mislead those who use financial statements to press for the maximum corporate profit and stock value (Armstrong et al. 2013; Greenfield et al. 2008; Jiraporn et al. 2008; Königsgruber and Palan 2014). To be more specific, the management and managers often use discretionary accruals to manipulate earnings and make them meet the specific objective and intent (Ayers et al. 2006; Bergstresser and Philippon 2006). The financial statements glossed over by this kind of conduct and intent may result

in serious outcome for the aforesaid stakeholders and lead to the problem of information asymmetry. Accounting earnings information plays a very important role in determining stock prices and the management authority's performance and supervision, while it is also critical information for measuring the corporate value. However, under the circumstance where the aforesaid information is not asymmetric, the management authority may go undercover to adjust accounting earnings information with discretionary accruals, which may weaken the valuation function of accounting earnings information. In addition, using earnings information to manipulate the cost may result in the risk of damage to the corporate value. Hence, if the management authority uses its accounting discretionary power to manipulate accounting earnings information, accounting earnings information will lose its functions in measuring the management authority's performance and supervision and valuating the corporate value.

Furthermore, the corporate value may decline as a result of using earnings information to manipulate the cost. According to the study conducted by Perols and Lougee (2011), earnings management and financial statement fraud show a positive correlation. For instance, Enron Corp. and WorldCom Corp. were high-profile publicly listed companies, but their financial statement fraud broke out and caused uproar. Those companies were all worldwide acknowledged public companies and their stock prices were very high, but they collapsed overnight. Academics pointed out in the past that investors and creditors often do not quite understand how the management and managers of a company manipulate earnings management (Armstrong et al. 2013; Barua et al. 2010; Chang et al. 2011). In order to achieve their performance goals and implement their reward plans, the management often manipulates accruals trying to increase their rewards, have their companies go public and boost their companies' stock prices (Dechow et al. 1995; Jiraporn et al. 2008).

Under such circumstances, financial statements and existing information used by external users are often asymmetric. However, the real earnings management status is very hard to be measured from actual business activities. Given that the management often uses discretionary accruals to influence the earnings shown on financial statements, Jones (1991) put forth a testing method and suggested using discretionary accruals as the substitute variable to measure earnings management, whereas the conventional regression method was recommended for control of discretionary accruals. Owing to the needs from the detecting earnings management, many methods have been tried to solve the problem, and we roughly divide the used methods into two categories: statistics and computational intelligence. Conventional studies mainly rely on statistical methods; however, statistical models are constrained by certain unrealistic assumptions. Take the regression model for example: the assumption of the independence of variables and linearity relationship are unrealistic (Liou and Tzeng 2012).

Prior studies on earnings management primarily put more focus on identifying earnings management. In general, they are based on the assumption that discretionary accruals by the residual from a linear regression on firm-level observables represent either explicit earnings management or poor quality earnings, but these discretionary accruals have not been be used to directly forecast the level of earnings management and detect the earnings management conducted with conventional statistical techniques, such as univariate statistical methods, factors analysis, discrimination analysis, logit and probit models (DeAngelo 1986; Dechow et al. 1995, 2012; Hribar and Collins 2002; Jones 1991; Marquardt and Wiedman 2004; Kothari et al. 2005). These conventional statistical methods, however, have some restrictions in assumptions such as the linearity, normality, and independence of earnings management variables. Given that the assumptions are often inconsistent with earnings management financial data, the methods have their intrinsic limitations in terms of effectiveness and validity. Those unrealistic assumptions cause limitations in exploring the entwined relationships of complex problems in practice (Shen and Tzeng 2014). In particular, there are few studies examining how to predict the level of manipulating earnings or earnings management.

As for the computational intelligence, many artificial intelligent and data mining techniques developed in the recent years have been applied to the fields of financial banking and accounting, such as diagnosis of financial crisis (Bernardo et al. 2013; Dragotă and Ţilică 2014; Geng et al. 2015; Hsu and Pai 2013; Verikas et al. 2010), bank performance (Fethi and Pasiouras 2010; Shen and Tzeng 2014), stock and investment decision-making (Yan and Clack 2011; Contreras et al. 2012; Zhiqiang et al. 2013), going-concern prediction (Yeh et al. 2014), etc. The rising computational capability of computer makes those machine learning techniques more efficient and effective in handling big financial data set. Many related studies show that most of the discretionary accrual estimation models use a linear approach, which might negatively impact the performance of the models. Also, several studies suggest that the accrual process in fact is non-linear (e.g. Dechow et al. 1995; Jeter and Shivakumar 1999; Kothari et al. 2005).

Data mining approaches, like decision tree (DT), are less vulnerable to the aforesaid violations (Afsari et al. 2013; Bernardo et al. 2013; Ravi and Pramodh 2008). Moreover, data mining aims to identify valid, novel, potentially useful and understandable correlations and patterns in earnings management data, and can be an alternative solution to classification problems. Related studies show that data mining has better predictive capability than conventional statistical methods in detecting earnings management, but it is not with-

out limitations (Hsu and Pai 2013; Hoglund 2012; Malliaris and Malliaris 2014; Nan et al. 2012; Tsai and Chiou 2009).

Some scholars indicated that feature selection would help remove interference features, reduced computation time, reduce computation time, classification errors, the dimensionality of data sets by deleting unsuitable attributes and reduce risk of over-fitting which would, therefore, further improve the performance of data mining algorithms (Jensen et al. 2014; Jing 2014; Ravisankar et al. 2011; Shu and Shen 2014; Vatolkin 2012; Vatolkin et al. 2011). Moreover, this study used the random forest (RF) method and stepwise regression (STW) method to determine and select important independent variables in development of an earnings management detecting model. RF is a relatively newer ensemble method that combines trees grown on bootstrap samples of data and a random subset bagging of predictor variables (Breiman 2001; Yeh et al. 2014). During the randomization of features, RF can provide an importance index of independent variables by accurate calculation and the Gini index. Furthermore, the importance index captures the interactions among predictors through the randomizations of predictors (Cugnata and Salini 2014; Cadenas et al. 2012). Stepwise regression has an advantage to avoid collinearity. It is a type of multiple linear regression that can select the fittest combination of independent variables for dependent variable prediction with forward-adding and backward-deleting variables (Chang and Wu 2014; Huang and Cheng 2013).

The main purpose of this study was to explore if an enterprise has any earnings management and find out the degree of the earnings management, so as to make most use of the advantages of RF and STW in preprocessing the earnings management financial data, and further improve classification accuracy of the decision tree predictor model. An integrated model that can leverage different technique's advantages is still under explored. This study is not to be constrained by a single approach; thus, the researchers decompose the detecting earnings management problem into five stages, and devise a reasonable infusion model to solve it. First, RF and STW have been used for variable selection because of its reliability in obtaining the significant independent variables. Second, the significant independent variables obtained from RF and STW have been used as the input for the DT model. Third, this study has generated meaningful rules using DT for earnings management detection. Fourth, to activate the effectiveness of our model, comparative experiments have been conducted. Finally, the model having the best performance and the highest accuracy in the test group as evaluated according to the model evaluation list has come up with the rules set.

The structure of the study is divided into five parts, in which earnings management's study objectives and motivation are first explained, followed by exploration of the literature of the used method and the decision tree applied in relevant fields. Then, the study discourses on the adopted methodology before analysis of the empirical results of the earnings management model. Finally, conclusion and recommendations are proposed.

2 Literary review

This section mainly explores this study infusing several computational methods to resolve the detecting of earnings management and reviews the origins and concepts of the used methods.

2.1 Stepwise regression

Stepwise regression procedure is proposed for evaluation of the relative importance of variants at different sites and is a modification of the forward selection procedure to select useful subsets of variables and evaluate the order of importance of variables (Huang and Cheng 2013). A step is added in which, after each independent variable enters the included group, the critical F value is used to check the eligibility of the added variable. With an added new variable, the previous variables in the model may lose their predictive ability. Thus, stepping criteria are used to check the significance of all the included variables. If the variable is insignificant, the backward method will be used to delete it, and each of the included groups will be re-investigated to see whether it is still worth inclusion. That is, an included independent variable may be discarded later if, at any future step, a subset of the included independent variables contains most of its predictive value. This analysis includes the procedures in which the choice of earning management variables is carried out by automatic procedure in the form of either forward or backward stepwise, and then stepwise regression in selection of earning management variables with the best identification and predictability is performed.

2.2 Random forest

Random forest is a combination of tree predictors, in which each tree depends on the value of a random vector sampled independently and has the same distribution as all of other trees in the forest. They are a relatively newer ensemble method that combines trees grown on bootstrap samples of data and a random subset bagging of predictor variables (Breiman 2001; Lunetta et al. 2004). Each classification of the trees is built based on a bootstrap sample of the data, while the candidate set of variables is a random subset of the variables in each split. Each tree is unpruned (grown fully), so as to obtain low-bias trees; at the same time, bagging and random variable selection lead to low correlation of the individual trees. During the randomization of features, random

forests may provide an importance index of independent variables by accurate calculation and the Gini index, whereas the importance index may capture the interactions among predictors through the randomizations of predictors (Vatolkin et al. 2012). Due to random forest's excellent performance in fulfilling classification tasks, it was adopted by the study as an important indicator to judge the variables of earning management.

2.3 Decision tree related research literatures

Decision tree is one of common data mining (DM) 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 (Chang and Chen 2009; Eskandarzadeh and Eshghi 2013; Hsu and Pai 2013; Ravi and Pramodh 2008). By using the multilayer perceptron (MLP), Tsai and Chiou 2009 implemented a classification test to probe the earnings management conducted by Taiwan's TSEC/OTC electronic companies, in which 11 earnings management-related variables were selected from the database of Taiwan Economic Journal (TEJ) and the research period was from 2002 to 2005. In addition, the cross-sectional Jones model (Jones 1991) was used to calculate discretionary accruals as the proxy variable of earnings management. In the aspect of the design of the MLP model, the study established 20 models using various hidden nodes and learning epochs. These models provide the highest prediction rate of 81 % in the cases of manipulating earnings upwards. Lu and Chen (2009) used CART and C5.0 decision tree technique to investigate the information disclosure degree of corporate financial statements and explore the binary issue (disclosure of good information vs. deficient information disclosure), in which 17 variables such as the EPS, company size, institutional investor's shareholding ratio, etc. were input and classification performance was evaluated, followed by identifying the 14 rules of "good information disclosure" of each variable through C5.0 decision tree where the probability of the company's "good information disclosure" should reach 91 %.

For the first stage of the decision tree technique applied by Delen et al. (2013) to measure a company's operating performance, the sample observation value was used to process feature analysis through EFA (exploratory factor analysis) and screen out the variables having greater influence on the target variable before entering the second stage for the model building procedure. When it came to the second stage, four decision trees of CHAID, C5.0, CART and QUEST were used for model building. The study results show that, when using ROE (return on equity) or ROA (return on assets) as the dependent variable, pre-tax earnings have very critical influence on two variables of the earnings before tax-to-equity ratio and net profit margin, and when the dependent variable is ROE, the CHAID model has the highest accuracy rate at 92.1 %, followed by C5.0, and QUEST has the lowest accuracy rate at 73.2 %. On the other hand, when using ROA as the dependent variable, the C5.0 model has the best performance.

2.4 Discussion

By reviewing the prior earnings management studies, the focus was on identifying some related factors which could significantly affect earnings management, i.e. we could only figure out the correlation between those factors and earnings management or poor quality earnings (Barua et al. 2010; Chang et al. 2011; Dechow et al. 1995, 2012; Jiraporn et al. 2008), but those factors were not directly used to forecast the level of earnings management. In order to help corporate stakeholders better understand the degree of earnings management and offer auditors a new method to probe earnings management and understand how an enterprise manipulates its earnings management, it is necessary to develop a model which is able to predict the level of earnings management. Nevertheless, a majority of studies only examine their models' average prediction performance without considering the Type I and Type II errors.

Therefore, this paper proposes a novel hybrid model for earnings management prediction which integrates the RF, STW and DT (including CHAID, CART and C5.0) techniques. The RF and STW methods were used for variable selection so as to obtain the significant independent variables, whereas DT could generate meaningful rules of earnings management. In order to evaluate the performance of the proposed framework, comparative experiments were conducted and the Type I and Type II errors were taken into consideration.

3 The integrated soft computing model

The integrated model comprises of three stages, and the three stages should be conducted in sequence. The first stage focuses on exploring the level of earnings management from the historical data. The second phase starts with RF and STW approach to screening variables and the third stage, adopt decision tree including CHAID, CART and C5.0 to establish the detection model. This section introduces earnings management's proxy variables. In other words, it covers discretionary accrual's algorithmic process and the process to divide the levels of earnings management into "extreme earnings management" and "slight earnings management". In addition, this section also elaborates on the theorem of three decision trees of CHAID, C5.0 and CART, while others, such as selection of samples and variables and the process of model establishment, are also included in the section.

3.1 Earning management's proxy variables

When calculating discretionary accruals, the total accruals shall be first calculated, following by eliminating non-discretionary accruals to come up with discretionary accruals. In view of the contents of prior studies, the following two methods are generally recommended for estimations of total accruals: (1) the balance sheet method and (2) the cash flow statement method. To avoid the deviation and extreme value brought by acquisition, asset disposal and foreign currency conversion, the study selected the cash flow statement method to calculate total accruals (Hribar and Collins 2002). Formula (1) shows the balance sheet method, whereas Formula (2) is the cash flow statement method, in which the factor TACC_{it} is the total accruals of company i in the t period, $\Delta\mathrm{CA}_{it}$ is the change of the current assets of company i in the t period, $\Delta\mathrm{CL}_{it}$ is the change of current liabilities of company i in the t period, $\Delta\mathrm{CASH}_{it}$ is the change of cash of company i in the t period, $\Delta\mathrm{STDEBT}_{it}$ is the change of company$i$'s long-term liabilities expiring within one year in the t period, DEP_{it} is the depreciation and depletion expenses of company i in the t period, EXBI_{it} is income from continuing operating department of company i in the t period and CFO_{it} is the operating cash flow of company i in the t period.

$$\mathrm{TACC}_{it} = \Delta\mathrm{CA}_{it} - \Delta\mathrm{CL}_{it} - \Delta\mathrm{CASH}_{it} + \Delta\mathrm{STDEBT}_{it} - \mathrm{DEP}_{it} \quad (1)$$

$$\mathrm{TACC}_{it} = \mathrm{EXBI}_{it} - \mathrm{CFO}_{it} \quad (2)$$

After the total accruals were calculated with the cash flow statement method, the study adopted the cross-sectional modified Jones Model to estimate non-discretionary accruals. The calculation method is shown as Formula (3), in which factor NDA_{it} is the non-discretionary accruals of company i in the t period deducting the total asset amount of the previous period, TA_{it-1} is total asset amount of company i in the t–1 period, $\Delta\mathrm{REV}_{it}$ is the income of company i in the t period deducting that in the $t-1$ period, $\Delta\mathrm{REC}_{it}$ is the account receivable (net amount) in the t period deducting that in the t-1 period and PPE_{it} is the total amount of company i's property, buildings and equipment in the t period.

$$\mathrm{NDA}_{it} = \alpha_{0it}\left(\frac{1}{\mathrm{TA}_{it-1}}\right) + \alpha_{1it}\left(\frac{\Delta\mathrm{REV}_{it} - \Delta\mathrm{REC}_{it}}{\mathrm{TA}_{it-1}}\right) + \alpha_{2it}\left(\frac{\mathrm{PPE}_{it}}{\mathrm{TA}_{t-1}}\right) \quad (3)$$

Estimates of the firm-specific parameters, α_0, α_1, andα_2 are generated using the follow model in the estimation period:

$$\mathrm{TAC}_{it} = a_{0it}\left(\frac{1}{\mathrm{TA}_{it-1}}\right) + a_{1it}\left(\frac{\Delta\mathrm{REV}_{it} - \Delta\mathrm{REC}_{it}}{\mathrm{TA}_{it-1}}\right) + a_{2it}\left(\frac{\mathrm{PPE}_{it}}{\mathrm{TA}_{t-1}}\right) + \varepsilon_{it} \quad (4)$$

The regression coefficients a_0, a_1, anda_2 are the estimators of α_0, α_1, andα_2 (denote the Ordinary Least Square estimate), and TAC_{it} refers to total accruals scaled by lagged total assets in period t and ε_{it} is the residual terms.

After the total accruals and non-discretionary accruals were calculated using the preceding model, and the non-discretionary accruals were eliminated from the total accruals, earnings management's proxy variable discretionary accruals were obtained as shown in Formula (5), in which DA_{it} is the discretionary accruals of company i in the t period (Marquardt and Wiedman 2004).

$$\mathrm{DA}_{it} = \frac{\mathrm{TACC}_{it}}{\mathrm{TA}_{it-1}} - \mathrm{NDA}_{it} \quad (5)$$

3.2 Decision tree

Decision tree is a tool to establish classification models and give predictions. It can process continuous and non-continuous variables (Ulutagay et al. 2014). This kind of algorithm can create a dendritic structure model according to utilization and induction of the specifically given data, which may give prediction analysis for the scattered or continuous attribute, e.g. "whether an enterprise has extremely manipulating earnings", and each branch represents a possibility of the attribute, for example "yes" or "no", whereas the leaf node at the tree tip represents a category or the attribute of a category, e.g. the "extreme earnings management company" or "slight earnings management company". In order to classify the input data, each node of the decision tree is a judgment formula, and the node is the data divided according to varying classification rules. The judgment formula will judge if the input data fall within the value of an attribute according to a specific variable. As such, each node can divide the input data into several categories, and, thus, a dendritic structure starts taking shape. The topmost node of a tree is called the root, and each path forming from the root to a leaf node represents a rule. This study adopted three methods of CHAID, CART and C5.0, which are described, respectively, as follows:

3.2.1 CHAID

Chi-squared automatic interaction detector (CHAID) is an extremely effective statistical technique developed by Kass (1980). It uses the Chi-square test to calculate the size of P-value at the leaf splitting node of the decision tree so as to determine if continuous division is required. Differing from other decision tree techniques, CHAID can produce more than two categories at any level in the tree; therefore, it is not a binary tree method. Its output is highly visual and easy to interpret since it uses multi-way splits by default. The main advantage of CHAID is to prevent data from being over-copied and stop the decision tree from continuous splitting.

In other words, CHAID can complete pruning before a model is established, whereas CART and C5.0 have to assess if pruning is required after the model is established.

3.2.2 CART

Classification and regression trees (CART) were established by Breiman et al. (1984). 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. The splitting terms are decided by the Gini rule. The data are divided into two subsets in each split. By repeating the process, the next splitting terms are searched from each subset. The tree is continuously constructed by the way of incessant data splitting into two subsets until there is no room for further splitting. 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. In the end, the minimum Gini value is used to determine the spitting attribute and attribute value. The category having the largest number of pieces in a node is then separated from other categories according to the Gini rule. Assume that data S covers N categories $C_1, C_2, \ldots, C_N$; if attribute A's value V is the splitting term, S will be split into $\{S_L, S_R\}$. l_i and r_i, respectively, representing the numbers which either belong or do not belong to category C_i in the subsets of S_L and S_R, $i = 1, 2, \ldots, N$. If C_n is the largest category in S, the calculation of the Gini value is shown as Formula (5) below:

$$\text{Gini}(A, v) = \frac{|S_L|}{|S|}\left[1 - \sum_{i=1}^{n}\left(\frac{l_i}{|S_L|}\right)^2\right] + \frac{|S_R|}{|S|}\left[\sum_{i=1}^{n}\left(\frac{r_i}{|S_R|}\right)^2\right] \quad (5)$$

CART algorithm shows the following characteristics: it is a non-parametric process, so there is no need to consider the data distribution type; the splitting rules are determined by the stepwise method; the possible splitting of all the parameters shall all be considered; dependent variables can be converted in a simple way; complicated and multivariable data structure can be processed; the outlier in the data will not affect the calculation of the algorithm; and there is no need to convert the data into category-type data in advance.

3.2.3 C5.0

C5.0 is a flow-chart-like tree structure constructed by a recursive divide-and-conquer algorithm which will generate a partition of the data. For the continuous numeric attribute node division method, C5.0 first gathers the objects and sorts them according to the attribute, followed by finding out the attribute value midpoint of two neighboring objects, which is called the cut point. Those that can obtain the optimal value after calculation of the evaluation function can follow the attribute's midpoint to make binary division. As for the defective and uncertain attribute values, they are commonly replaced by the most frequent attribute values or solved by the optimistic estimate probability method.

4 Experimental design

4.1 Data and samples

The study's samples were all selected from the publicly listed electronic companies for 2008 through 2012 covered in Taiwan Economic Journal (TEJ) and on a quarterly basis. The electronic industry selected as empirical dataset by this study is an outstanding industry as it has constituted about 45–80 % of total stock trading amount and volume in Taiwan every day (Taiwan Stock Exchange Corporation 2012); the study used the 3rd quarter of 2012 as the base period. Out of the dependent and independent samples influencing earnings management, those that lack any numerical values were deleted. As a result, 307 valid samples were obtained. The sample selection process is shown in Table 1 below:

4.2 Potential predictive variables

To apply prediction methods to earnings management prediction, first, potential predictive variables should be selected. In terms of variable selection, the study selected 17 variables, which could affect earnings management, from past earning management research papers for model building. These variables included financial indicators, corporate governance indicators and various kinds of performance and threshold indicators. The variables used by the study are shown in Table 2 below. The names of the variables, calculation methods and references are all indicated in the table (Abowd 1990;

Table 1 The study's sample selection process

Sample selection process	The number of samples
Dependent variables, discretionary accruals, the third quarter of 2012	390
Independent variables, from 2008 to the third quarter of 2012	390
Samples which were left out or with incomplete data	(83)
The final number of samples	307

Table 2 Variables used by the study

Variable code	Variable name	Calculation method
X1	(*%INST*)	$\%INST_{iq} = TRUST_{iq}/SHARES_{iq}$
	External supervision percentage	$\%INST_{iq}$: external supervision percentage of company i in the q quarter
		$TRUST_{iq}$: trust institute's shareholding of company i in the q quarter
		$SHARES_{iq}$: outstanding shares of company i in the q quarter
X2	(*THOD*)	$THOD_{iq} = NDA_{iq} - NDA_{iq-4}$
	Performance threshold	$THOD_{iq}$: performance threshold of company i in the q quarter
		NDA_{iq}: non-discretionary accruals of company i in the q quarter
		NDA_{iq-4}: non-discretionary accruals of company i in the q-4 quarter
X3	(*PPS*)	$PPS_{in} = (ROE_{in} - ROE_{mn}) \cdot (COMP_{in} - COMP_{mn})$
	Pay-performance sensitivity	PPS_{in}: pay-performance sensitivity of company i in the n year.
		ROE_{in}: shareholder's ROE of company i in the n year
		ROE_{mn}: average value of shareholders' ROE of company i in the n year.
		$COMP_{in}$: pay of high-rank managers of company i in the n year
		$COMP_{mn}$: average value of pay of high-rank managers of company i in the n year.
X4	(*LEV*)	$LEV_{iq} = TL_{iq}/TA_{iq}$
	Leverage coefficients	LEV_{iq}: leverage coefficient of company i in the q quarter.
		TL_{iq}: total liability of company i in the q quarter.
		TA_{iq}: total asset of company i in the q quarter.
X5	(*RISK*)	$RISK_{iq} = \beta_{iq}$
	Management risk	$RISK_{iq}$: management risk of company i in the q quarter.
		β_{iq}: value at risk of company i in the q quarter.
X6	(DA_{n-1})	$DA_{in-1} = DA_{iq-3}$
	Discretionary accruals of the previous period	DA_{in-1}: discretionary accruals of company i in the n–1 year
		DA_{iq-3}: discretionary accruals of company i at the latest same quarter
		q of the predicting quarter.
X7	(*PERS*)	$PERS = \frac{\sum_{q=2}^{16}(UE_q - \overline{UE})(UE_{q-1} - \overline{UE})}{\sum_{q=1}^{16}(UE_q - \overline{UE})^2}$
	Earnings persistence	$UE_q = e_q - e_{q-4}$; e_q refers to earnings in the quarter q.
		$\overline{UE} = \frac{\sum_{q=1}^{16} UE_q}{16}$
X8	(*SIZE*)	$SIZE_{in} = \ln(\sum_{p=q-3}^{q} SALES_{ipn})$
	Corporate size	$SIZE_{in}$: corporate size of company i in the n year
		$\sum_{p=q-3}^{q} SALES_{ipn}$: the summation of sales from operations quarterly amounts in firm i in the latest four quarters of year n.
X9	(*CP*)	$CP_{in} = \frac{\sum_{p=q-3}^{q} OCF_{ipn}}{ASSETS_{ing-4}}$
	Corporate performance	$\sum_{p=q-3}^{q} OCF_{ipn}$: the summation of cash from operations quarterly amounts in firm i in the latest four quarters of year n.
		$\boldsymbol{ASSETS}_{inq-4}$: total assets in firm i at the same date of four quarters ago.
X10	(*SHARVAR*)	$SHARVAR_{in} = 1$: when the change (increase or decrease) of the number of outstanding shares is more than 10 %.
	Financing activities	$SHARVAR_{in} = 0$: when the change (increase or decrease) of the number of outstanding shares is less than 10 %.
X11	(*TIE*)	$TIE_{iq} = (OI_{iq} + IN_{iq})/IN_{iq}$
	Times interest earned ratio	TIE_{iq}: times of interest earned ratio by company i in the q quarter.
		OI_{iq}: pre-tax net profit of company i in the q quarter.
		IN_{iq}: interest expense of company i in the q quarter.
X12	(*ROE*)	$ROE_{iq} = NI_{iq}/EQ_{iq}$

Table 2 continued

Variable code	Variable name	Calculation method
	Return on equity	ROE_{iq}: return on equity of company i in the q quarter. NI_{iq}: net income of company i in the q quarter. EQ_{iq}: average equity of company i in the q quarter.
X13	(*ARC*) Account receivable collection days	$ARC_{iq} = 365/RATE_{iq}$ ARC_{iq}: account receivable collection days of company i in the q quarter. $RATE_{iq}$: account receivable turnover rate of company $i$in the q quarter.
X14	(P/E) Price earnings ratio	$P/E_{iq} = PRICE_{iq}/EPS_{iq}$ P/E_{iq}: the P/E of company i in the q period $PRICE_{iq}$: the stock price of company i in the q period EPS_{iq}: the EPS of company i in the q period
X15	(P/B) Price/book value ratio	$P/B_{iq} = PRICE_{iq}/BVPS_{iq}$ P/B_{iq}: the P/B ratio of company i in the q period. $PRICE_{iq}$: the market price per share of company i in the q period $BVPS_{iq}$: the book value per share of company i in the q period
X16	(*ROA*) Return on assets	$ROA_{iq} = \{NI_{iq} + IN_{iq}(1-t)\}/ASSET_{iq}$ ROA_{iq}: the ROA of company i in the q quarter. NI_{iq}: the net income of company i in the q quarter. IN_{iq}: interest expense of company i in the q quarter. T: tax rate of company i in the q quarter. $ASSET_{iq}$: average total asset of company i in the q quarter
X17	(*CFO*) Operating cash flow	CFO_{iq}: company i's cash flow from operating activities in the q quarter

Source: 1.Abowd (1990); 2. Becker et al. (1998); 3. Chan et al. (2004); 4. Hoglund (2012); 5. Nan et al. (2012); 6. Tsai and Chiou (2009)

Becker et al. 1998; Chan et al. 2004; Hoglund 2012; Nan et al. 2012; Tsai and Chiou 2009).

4.3 Degree classification of earnings management

In order to specifically identify serious cases of earnings management, the study properly classified the proxy variables of earnings management, i.e. discretionary accruals (DA) by means of the statistic method, in which the average value and standard deviation of all the samples were calculated in the first place, followed by setting the value calculated by adding a notch of standard deviation value to the average value as the ceiling and the one calculated by deducting a notch of standard deviation from the average value as the floor. If the value of the discretionary accrual was over the ceiling value or below the floor value, it would be defined as extremely upward or downward earnings management, whereas other sample observation values falling in the area between the ceiling and floor would be deemed to be slight earnings management. Using the aforesaid method to classify the intervals, the numbers of samples and descriptive statistics, Table 3 shows that the average value of discretionary accruals calculated from the total sample observation value is 0.001961, whereas the average value deducting a notch of standard deviation is the floor value, which is at −0.035402. When the discretionary accrual value is >0.039324 (the average value plus a notch of standard deviation) or smaller than −0.035402 (the average value minus a notch of standard deviation), it would be defined to be serious accrual earnings management behavior. However, the value falling in the area between the ceiling and floor would be deemed to be slight accrual earnings management behavior. A total of 29 samples are below the floor value and their average value is −0.0725. Thus, those samples are defined as extremely downward earnings management. On the other hand, the value calculated by adding a notch of standard deviation to the average value is 0.039324, which exceeds the ceiling value and is defined as extremely upward earnings management. The extremely upward earnings management has 28 observation values in total, and their average value is 0.062.

To further explore if an enterprise showed extreme earnings management, the study set "1" for the levels of earning management which were extremely upward and extremely downward, whereas other levels of earnings managements

Table 3 Earnings management classification intervals and descriptive statistics

Classification name	Classification interval	Number of samples	DA classification descriptive statistics	
			Median	Mean
Extremely downward earnings management	DA < −0.035402	29	−0.0557	−0.0725
Slightly downward earnings management	0.001961 >DA ≥ −0.035402	123	−0.01	−0.0122
Slightly upward earnings management	0.001961 <DA≤0.039324	127	0.0133	0.0147
Extremely upward earnings management	DA > 0.039324	28	0.0482	0.062
Total		307	−0.0042	−0.008

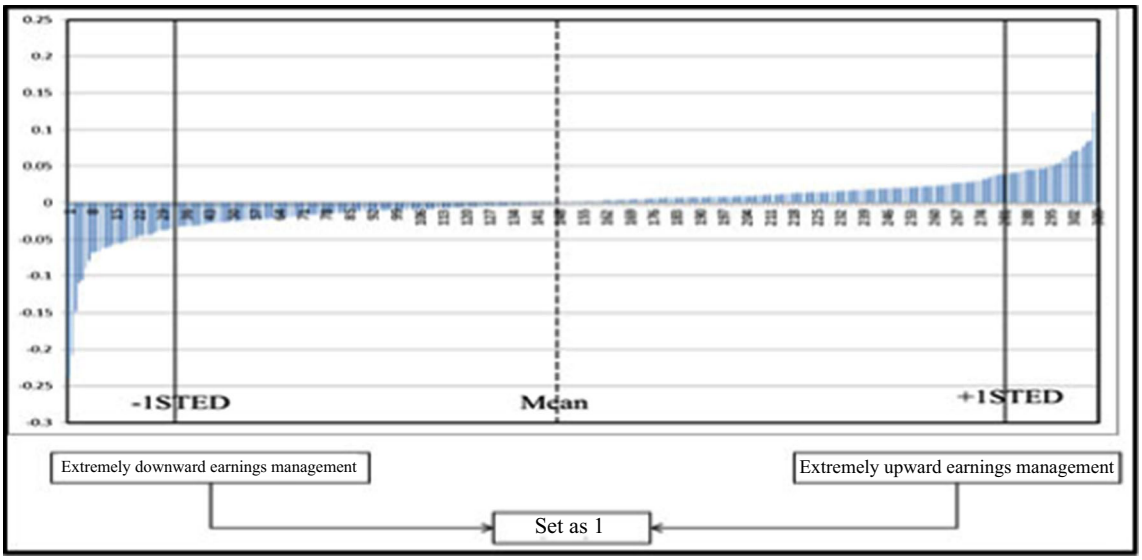

Fig. 1 Illustration of earnings management classification

were set as "0". In this way, it became the issue of binary judgment. In other words, the study tried to use the decision tree to probe if an enterprise had "extreme earnings management". Figure 1 is the illustration of the setting of binary classification, in which the histogram shows the ranking of discretionary accruals in order of ascendance.

4.4 Experimental process

After collecting respective variables and all the observation values, the study screened the variables to select influential variables. However, given the fact that the calculation and measurement bases of respective variables are different, the range of source data could be too big or too small. Hence, the study normalized all the independent variables to be in the range between 0.1 and 0.9. The purpose for doing so was to re-scale the data into a proper range, so the decision tree could give more accurate classification. Formula (6) shows the normalization process:

$$N = 0.1 + \left[\frac{(f(x) - \min f(x))(0.9 - 0.1)}{\max f(x) - \min f(x)}\right], \tag{6}$$

in which N is the eigenvalue after normalization, $f(x)$ is the respective samples of the variable in question, max $f(x)$ is the maximum value of the variable in question and min $f(x)$ is the minimum value of the variable in question. Then, classifier model building and comparison of experimental results were processed. The study used three methods of CART, C5.0 and CHAID to establish the model, and the process is shown as Fig. 2 below:

4.5 Performance evaluation

Prediction accuracy and Type I and II errors should be taken into consideration when evaluating the performance of developed earnings management prediction models. In addition, the study also further disclosed Type I and Type II errors of each model. Type I error shows the situation where earnings management is actually in serious error, but is classified to have the slight error rate of earnings management, whereas Type II error represents that earnings management actually has no serious error, but is classified to have the serious error rate of earnings management. As defined by this study, Type

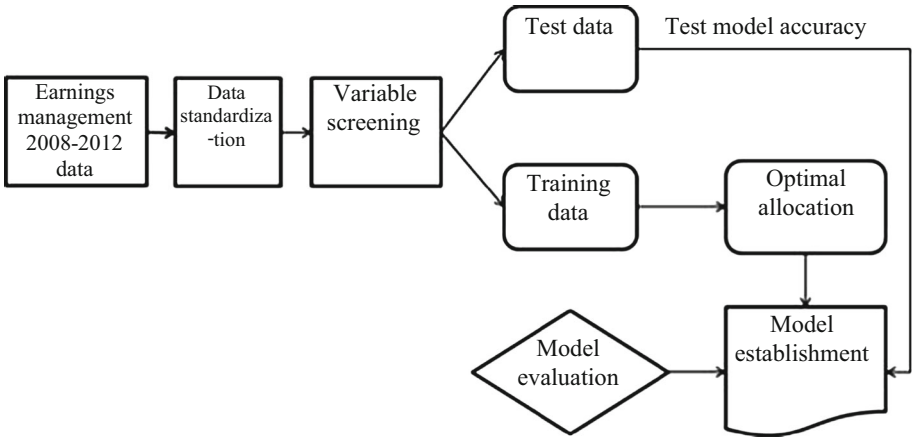

Fig. 2 The study's mold building process

I error represents more serious error. As such, in addition to comparing the test group's accuracy, the study also took model Type I error rate and Type II error into account.

5 Empirical results and analysis

This section first screens the 17 normalized variables selected by the study, in the hope of obtaining the variables having greater influence on earnings management, followed by stepping into the second stage with the variables selected from screening to proceed with model building and testing of classification performance. The classification results and training as well as testing accuracy are shown in a matrix. Finally, two screening methods are compared and the classification accuracy output by the three models is paired.

5.1 Variable screening

Given that the study obtained many variables, two methods were used to find out important and representative variables before establishing the three decision tree models. The study used two variable screening methods of stepwise regression (STW) and random forest (RF), one of the data mining methods. The screening results are respectively listed as below.

5.1.1 Stepwise regression (STW) screening

STW has been extensively applied in the studies of social science. It screens variables by using the t value (and its significant level α value) as the reference indicator for selecting an independent variable. Table 4 below shows the results of the STW screening used by the study, in which after screening through STW analysis, three variables are left from the 17 variables. The three variables in order of selection are X9 for corporate performance, X16 for ROA and X15 for P/B, respectively.

Table 4 STW screening results

Model	Selected variable	Method
Selected/deleted variables		
1	X9	STW analysis method (rule: F-selected probability rate $\leq$ .050, F-deleted probability rate $\geq$.100)
2	X16	STW analysis method (rule: F-selected probability rate $\leq$ .050, F-deleted probability rate $\geq$.100)
3	X15	STW analysis method (rule: F-selected probability rate $\leq$ .050, F-deleted probability rate $\geq$.100)

[a] Dependent variable: DA (discretionary accruals)

5.1.2 Random forest (RF) screening

The RF technique has the advantages of classification, detection of variable correlation and evaluation of variable importance (Breiman 2001). The RF adopted by the study uses the mean decrease in the Gini coefficient as the important indicator to determine variables. When the value of the mean decrease in the Gini coefficient is greater, the influence of its variable on the level of earnings management will be higher. Table 5 below shows the mean decreases in Gini coefficients of the variables screened out by the study, the selected variables and the parameters estimated from the dependent variables of respective categories. The variables screened out by RF are listed in order of their importance, which is X17, X6, X5, X9, X2, X12 and X11, i.e. in the order of operating cash flow, previous period's discretionary accruals, management risk, corporate performance, performance threshold, return on equity and times-interest-earned ratio.

5.2 Decision tree model

When constructing the two-stage model for the three decision tree models, the study normalized its selected vari-

Table 5 RF screening results

Variable	0	1	Mean decrease accuracy	Mean decrease gini
X17	22.1041	18.7747	25.8260	13.2092
X6	9.5264	4.1731	9.5819	7.9475
X5	4.3780	4.6302	5.7660	6.7198
X9	10.3037	−1.9901	8.6297	6.3858
X2	3.9693	2.7224	4.9346	5.6379
X12	10.2703	−0.7145	10.0672	5.4016
X11	3.8642	5.0500	5.8770	5.2649

ables before processing random non-repetitive sampling. The training group and test group were trained and tested at a ratio of 9 to 1, i.e. 90 % of the total samples were used for training and model establishment, whereas 10 % of the data were tested and calculated for accuracy. This kind of division ratio was also recommended in research papers (Huang et al. 2007)

5.2.1 STW+ decision tree model

The second stage of detection model of earnings management could be established with the three decision tree; in addition, this study used tenfold cross validation. By combining stepwise regression with three decision trees, Table 6 below shows accuracy of classification models of earnings management. As shown in Table 6, in terms of classification accuracy, C5.0 is 88.31 %, which was higher than CART and CHAID. As for Type I error shown in Table 7, C5.0 has lower Type I error at 20.70 %, which was lower than the result of CART and CHAID. The Type II error is shown in Table 8; C5.0 has lower Type II error at 9.64 %.

5.2.2 RF+ decision tree models

The classification accuracy of RF and three decision tree models is shown in Table 9. The accuracy rate C5.0 has the best classification effect 91.24 %. For the serious Type I error, C5.0 is at 15.26 %, which is lower than CHAID and CART and shown in Table 10. The Type II error is shown in Table 11; C5.0 has lower Type II error at 7.28 %. Comprehensive comparison shows that RF + C5.0 overall accuracy is 91.24 %, followed by STW+ C5.0 model's 88.31 %, as shown in Table 12.

5.3 Model evaluation and additional testing

5.3.1 Model statistical test

For the sake of prudence, to verify whether the above models are statistically significant, we conduct the statistical test on the above-mentioned results to confirm whether the differences in between models are significantly. The analysis

Table 6 STW+ three decision tree models cross-validation results

Cross-validation (CV)	Model comparison (overall accuracy) (%)		
	STW+C5.0	STW+CART	STW+CHAID
CV1	90.23	87.62	86.64
CV2	86.64	84.04	85.99
CV3	89.90	88.93	92.51
CV4	90.23	85.34	85.67
CV5	83.39	82.08	85.02
CV6	87.30	83.71	82.41
CV7	87.30	84.69	81.43
CV8	88.93	82.08	82.74
CV9	87.95	82.74	86.32
CV10	91.21	85.02	84.69
Average	88.31	84.63	85.34

Table 7 STW+ three decision tree models cross-validation Type I error results

Cross-validation (CV)	Model comparison (Type I error) (%)		
	STW+C5.0	STW+CART	STW+CHAID
CV1	21.05	26.32	22.81
CV2	22.81	29.82	28.07
CV3	22.81	28.07	26.32
CV4	21.05	22.81	26.32
CV5	22.81	26.32	24.56
CV6	17.54	19.30	22.81
CV7	21.05	21.05	33.33
CV8	24.56	28.07	28.07
CV9	17.54	35.09	22.81
CV10	15.79	33.33	28.07
Average	20.70	27.02	26.32

Table 8 STW+ three decision tree models cross-validation Type II error results

Cross-validation (CV)	Model comparison (Type II error) (%)		
	STW+C5.0	STW+CART	STW+CHAID
CV1	7.20	9.20	11.20
CV2	11.20	12.80	10.80
CV3	7.20	7.20	3.20
CV4	7.20	12.80	11.60
CV5	15.20	16.00	12.80
CV6	11.60	15.60	16.40
CV7	10.80	14.00	15.20
CV8	8.00	15.60	14.80
CV9	10.80	13.20	11.60
CV10	7.20	10.80	12.40
Average	9.64	12.72	12.00

Table 9 RF+ three decision tree models cross-validation results

Cross-validation (CV)	Model comparison (overall accuracy) (%)		
	RF + C5.0	RF + CART	RF + CHAID
CV1	92.51	82.74	87.62
CV2	85.99	82.74	87.95
CV3	90.88	85.67	83.39
CV4	92.18	85.34	85.67
CV5	93.16	83.71	79.15
CV6	91.21	82.41	82.08
CV7	91.53	84.69	87.62
CV8	91.86	87.30	81.76
CV9	91.86	88.93	84.69
CV10	91.21	84.36	86.32
Average	91.24	84.79	84.63

Table 10 RF + three decision tree models cross-validation Type I error results

Cross-validation (CV)	Model comparison (Type I error) (%)		
	RF + C5.0	RF + CART	RF + CHAID
CV1	15.79	26.32	35.09
CV2	22.81	21.05	33.33
CV3	15.79	19.30	28.07
CV4	10.53	17.54	22.81
CV5	15.79	26.32	29.82
CV6	17.54	31.58	28.07
CV7	14.04	33.33	33.33
CV8	12.28	35.09	28.07
CV9	12.28	28.07	26.32
CV10	15.79	22.81	22.81
Average	15.26	26.14	28.77

Table 11 RF + three decision tree models cross-validation Type II error results

Cross-validation (CV)	Model comparison (Type II error) (%)		
	RF + C5.0	RF + CART	RF + CHAID
CV1	5.60	15.20	7.20
CV2	12.00	16.40	7.20
CV3	7.60	13.20	14.00
CV4	7.20	14.00	12.40
CV5	4.80	14.00	18.80
CV6	6.80	14.40	15.60
CV7	7.20	11.20	7.60
CV8	7.20	7.60	16.00
CV9	7.20	7.20	12.80
CV10	7.20	14.00	11.60
Average	7.28	12.72	12.32

Table 12 Summary of classification results

Model	Type I error (%)	Type II error (%)	Overall accuracy (%)
STW+C5.0	20.70	9.64	88.31
STW+CART	27.02	12.72	84.63
STW+CHAID	26.32	12.00	85.34
RF+C5.0	15.26	7.28	91.24
RF+CART	26.14	12.72	84.79
RF+CHAID	28.77	12.32	84.63

results are shown in Table 13. The proposed hybrid model (RF + C5.0) performs the best in terms of prediction accuracy.

5.3.2 Additional testing

For the final part of the empirical analysis, the study used the rules set coming out from the RF+C5.0 model, which has the highest accuracy. Table 14 shows the rules set of the serious earnings management level "1" generated from C5.0, whereas Fig. 3 shows the decision tree chart brought from the C5.0 decision tree. As shown in Table 14, there are two rules for serious earnings management, in which rule 1 is that it is likely to result in the status of serious earnings management when the standardized X11 (times interest earned ratio) is >0.449, standardized X17 (operating cash flow) is smaller than or equal to 0.174 and the prediction accuracy rate is 88.889 % (please see Fig. 3), whereas rule 2 is simpler, in which extreme earning management is likely to occur when standardized X6 (previous period's discretionary accruals) is >0.703.

5.4 Discussion and findings

According to the experiments discussed above, the analysis results and implications of earnings management prediction are presented below:

Numerous predictive variables should be covered for consideration. As such, finding important predictive variables would be crucial, as it would affect accuracy and classification of the model developed. Instead of selecting variables with domain knowledge, the study selected the variables according to their importance as calculated by STW and RF. Tables 4, 5,6, 7, 8, 9, 10, 11, 12 and 13, and proved that the RF+C5.0 method could effectively improve the accuracy rate of earnings management detection, regardless of the methods and variables used. The analysis presented above suggests that variable selection can enable researchers to give earnings management prediction without having any special domain knowledge. Compared to the models adopted by other scholars (e.g. Tsai and Chiou 2009), the model selected by the

Table 13 Paired-samples t test

Model	Overall accuracy P value	Type I error P value	Type II error P value
$\mu_{RF+C5.0}$ vs $\mu_{RF+CART}$	$P = 0.000$	$P = 0.001$	$P = 0.001$
$\mu_{RF+C5.0}$ vs $\mu_{RF+CHAID}$	$P = 0.001$	$P = 0.000$	$P = 0.013$
$\mu_{RF+C5.0}$ vs $\mu_{STW+C5.0}$	$P = 0.011$	$P = 0.003$	$P = 0.047$
$\mu_{RF+C5.0}$ vs $\mu_{STW+CART}$	$P = 0.000$	$P = 0.000$	$P = 0.001$
$\mu_{RF+C5.0}$ vs $\mu_{STW+CHAID}$	$P = 0.001$	$P = 0.000$	$P = 0.008$

study has better accuracy. The results of the experiments give insight into the reason why the proposed model is optimal in this study. They also prove that the proposed hybrid model is stable in terms of accuracy because it is the optimal model in all aspects of accuracy, Type I error, Type II error and predictive variables.

Table 14 Rules set of the RF + C5.0 model

Rules Set for RF + C5.0 Model Rules set in the circumstance where the dependent variable is "1" (serious earnings management)	
Rule 1	if X11 > 0.449 and X17 ≤ 0.174 then 1
Rule 2	if X6 > 0.703 then 1

6 Conclusion and recommendations

Corporate earnings management prediction plays a significant role among corporate stakeholders covering investors, creditors, analysts and customers. In addition, auditing time, human resources and costs are limited to traditional reviews and auditing processes, so it is hard to identify any abnormal behavior out of huge and complex financial information (Calderon and Cheh 2002). Under such circumstances, development of an earnings management predictive model can be very helpful for auditors to find out the degree of manipulation in financial statements.

This study proposed an integrated soft computing model to resolve the earnings management detecting problem. The complexity of the financial reports impedes decision makers to conclude useful patterns from large and imprecise data set; therefore, this study chose RF and STW to induct the patterns and critical variables for detecting earnings management. This study successfully selected seven critical ratios from the original 17 financial variable with the capability to detecting earnings management by RF and focuses on the development of RF and DT models to predict the level of earnings management. A new procedure, based on a hybrid model combining RF, STW and DT, has been developed not only to enhance classification accuracy but also elicit meaningful rules for earnings management prediction. To demonstrate the proposed approach, the study used RF+ C5.0, RF+ CART, RF+ CHAID, STW+ CART, STW+ C5.0 and STW+ CHAID models as benchmarks. Based on the experiments, the results of this study are summarized as follows:

First, using three decision tree models of CHAID, CART and C5.0, the study combined two variable screening methods of STW and RF to make variable importance selection and further explored if an enterprise had extremely serious earnings management. Second, the empirical results show that combining the RF with C5.0 could best investigate the status of extreme earnings management; its accuracy rate in the test group is 91.24 %, which is higher than that of CHAID and CART. In addition, it has the lowest Type I error at 15.26 % and Type II error at 7.28 %. It is believed that the predictive models could help the users of financial statements make decisions in accordance with the earnings information. Besides, building a prediction model to explore the level of earnings management in advance is a new hybrid model application for RF, STW and DT. Finally, for the additional testing, the rules generated by C5.0 against extreme earnings management, an enterprise's operating cash flow, times-interest-earned ratio and previous period's discretionary accruals play a decisive role in affecting its extreme earnings management.

Despite the contributions of this study, we can positively conclude that the proposed hybrid approach using RF and C5.0 is more efficient than the listed approaches for detecting earnings management, there are still several limitations. First, STW and RF were adopted for the variables screening, and the obtained critical ratios might be different using the other feature selection methods. Future studies may incorporate some other machine learning techniques to find the optimal feature selection. Second, the hybrid model combining RF, STW and DT only used one period-lagged data to detecting earnings management. Some latent tendency in relatively long-lagged periods (e.g., more than 2 years) might not be captured in the model.

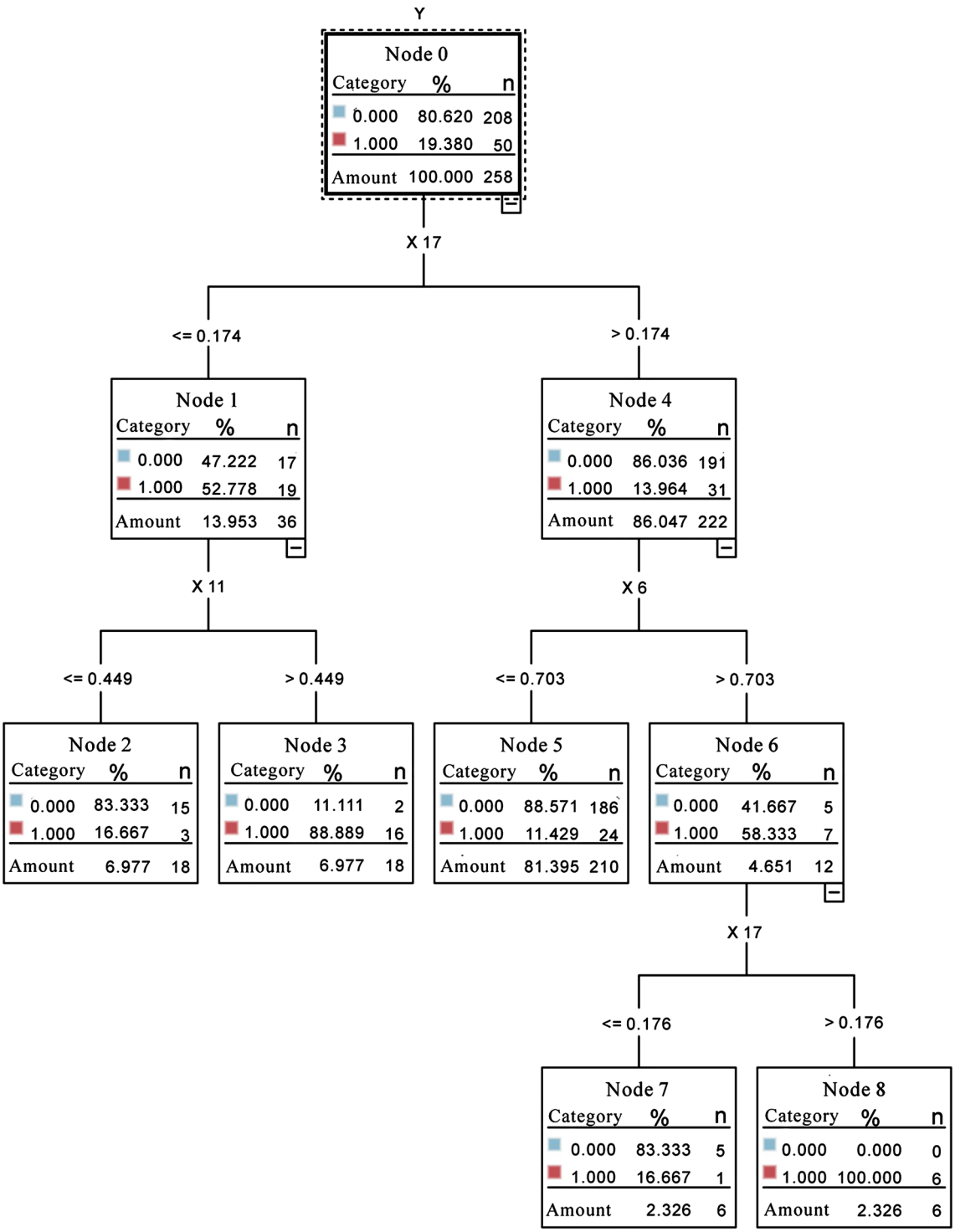

Fig. 3 RF+C5.0 Decision Tree Chart

Acknowledgments The authors would like to thank the Editor-in-Chief and reviewers for their useful comments and suggestions, which were very helpful in improving this manuscript.

References

Abowd JM (1990) Does performance-based managerial compensation affect corporation performance? Ind Labor Relat Rev 43:52–73

Afsari F, Eftekhari M, Eslami E, Woo PY (2013) Interpretability-based fuzzy decision tree classifier a hybrid of the subtractive clustering and the multi-objective evolutionary algorithm. Soft Comput 17(9):1673–1686
Armstrong CS, Larcker DF, Ormazabal G, Taylor DJ (2013) The relation between equity incentives and misreporting: the role of risk-taking incentives. J Financ Econ 109(2):327–350
Ayers BC, Jiang J, Yeung PE (2006) Discretionary accruals and earnings management: an analysis of pseudo earnings targets. Account Rev 81(3):617–652
Barua A, Lin S, Sbaraglia AM (2010) Earnings management using discontinued operations. Account Rev 85(5):1485–1509
Becker CL, Defond ML, Jiambalvo J, Subramanyam KR (1998) The effect of audit quality on earnings management. Contemp Account Res 15(1):1–24
Bergstresser D, Philippon T (2006) CEO incentives and earnings management. J Financ Econ 80(3):511–529
Bernardo D, Hagras H, Tsang E (2013) A genetic type-2 fuzzy logic based system for the generation of summarised linguistic predictive models for financial applications. Soft Comput 17(12):2185–2201
Breiman L (2001) Random forests. Mach Learn 45:5–32
Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall/CRC, New York
Cadenas JM, Garrido MC, Martínez R, Bonissone PP (2012) Extending information processing in a fuzzy random forest ensemble. Soft Comput 16(5):845–861
Calderon TG, Cheh JJ (2002) A roadmap for future neural networks research in auditing and risk assessment. Int J Account Inf Syst 3:203–236
Chan K, Jegadeesh N, Sougiannis T (2004) The accrual effect on future earnings. Rev Quant Financ Account 22(2):97–121
Chang CL, Chen CH (2009) Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Syste Appl 36(2):4035–4041
Chang RD, Tseng YC, Chang CP (2011) The issuance of convertible bonds and earnings management: evidence from Taiwan. Rev Account Financ 9(1):65–87
Chang PC, Wu JL (2014) A critical feature extraction by kernel PCA in stock trading model. Soft Comput, in Press. doi:10.1007/s00500-014-1350-5
Contreras I, Jiang Y, Hidalgo JI, Núñez-Letamendia L (2012) Using a gpu-cpu architecture to speed up a ga-based real-time system for trading the stock market. Soft Comput 16(2):203–215
Cugnata F, Salini S (2014) Model-based approach for importance-performance analysis. Qual Quant 48(7):3053–3064
DeAngelo LE (1986) Accounting numbers as market valuation substitutes: a study of management buyout of public stockholders. Account Rev 61(3):400–420
Dechow PM, Sloan RG, Sweeney AP (1995) Detecting earning management. Account Rev 70(2):193–225
Dechow PM, Hutton AP, Kim JH, Sloan RG (2012) Detecting earnings management: a new approach. J Account Res 50(2):275–334
Delen D, Kuzey C, Uyar A (2013) Measuring firm performance using financial ratios: a decision tree approach. Expert Syst Appl 40(10):3970–3983
Dragotă V, Tilică EV (2014) Market efficiency of the Post Communist East European stock markets. Cent Euro J Operat Res 22(2):307–337
Eskandarzadeh S, Eshghi K (2013) Decision tree analysis for a risk averse decision maker: CVaR criterion. Europ J Operat Res 231(1):131–140
Fethi MD, Pasiouras F (2010) Assessing bank efficiency and performance with operational research and artificial intelligence techniques: a survey. Euro J Operat Res 204(2):189–198
Geng R, Bose I, Chen X (2015) Prediction of financial distress: an empirical study of listed Chinese companies using data mining. Euro J Operat Res 241(1):236–247
Greenfield AC, Norman CS, Wier B (2008) The effect of ethical orientation and prefessional commitment on earnings management behavior. J Bus Ethics 83:419–435
Healy PM (1999) Discussion of a market-based evaluation of discretionary accrual models. J Account Econ 26:143–147
Hoglund H (2012) Detecting earnings management with neural networks. Expert Syst Appl 39(10):9564–9570
Hribar P, Collins DW (2002) Errors in estimating accruals: implications for empirical research. J Account Res 40(1):105–134
Hsu MF, Pai PF (2013) Incorporating support vector machines with multiple criteria decision making for financial crisis analysis. Qual Quant 47(7):3481–3492
Huang CL, Chen MC, Wang CJ (2007) Credit scoring with a data mining approach based on support vector machines. Expert Syst Appl 33(4):847–856
Huang SF, Cheng CH (2013) GMADM-based attributes selection method in developing prediction model. Qual Quant 47(6):3335–3347
Jensen R, Tuson A, Shen Q (2014) Finding rough and fuzzy-rough set reducts with SAT. Inf Sci 255:100–120
Jeter DC, Shivakumar L (1999) Cross-sectional estimation of abnormal accruals using quarterly and annual data: effectiveness in detecting event-specific earnings management. Account Bus Res 29(4):299–319
Jing SY (2014) A hybrid genetic algorithm for feature subset selection in rough set theory. Soft Comput 18(7):1373–1382
Jiraporn P, Miller GA, Yoon SS, Kim YS (2008) Is earnings management opportunistic or beneficial? An agency theory perspective. Int Rev Financ Anal 17(3):622–634
Jones J (1991) Earnings management during import relief investigations. J Account Res 29(2):193–228
Kass G (1980) An exploratory technique for investigating large quantities of categorical data. Appl Stat 29(2):119–127
Königsgruber R, Palan S (2014) Earnings management and participation in accounting standard-setting. Cent Euro J Operat Res. In press. doi:10.1007/s10100-013-0326-3
Kothari SP, Leone AJ, Wasley CE (2005) Performance matched discretionary accrual measures. J Account Econ 39:163–197
Liou JH, Tzeng GH (2012) Comments on multiple criteria decision making (MCDM) methods in economics: an overview. Technol Econ Dev Econ 18(4):672–695
Lu CL, Chen TC (2009) A study of applying data mining approach to the information disclosure for Taiwan's stock market investors. Expert Syst Appl 36(2):3536–3542
Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 5(1):32
Malliaris AG, Malliaris M (2014) N-tuple S&P patterns across decades, 1950–2011. Cent Europ J Operat Res 22(2):339–353
Marquardt CA, Wiedman CI (2004) The effect of earnings management on the value relevance of accounting information. J Bus Financ Account 31(3/4):297–332
Nan X, Sun X, Li Y, Hou T (2012) Weighted-support vector machine based earnings management detection during IPOs. J Inf Computat Sci 9(9):2607–2617
Perols JL, Lougee BA (2011) The relation between earnings management and financial statement fraud. Adv Account 27:39–53
Ravi V, Pramodh C (2008) Threshold accepting trained principal component neural network and feature subset selection: application to bankruptcy prediction in banks. Appl Soft Comput 8(4):1539–1548

Ravisankar P, Ravi V, Bose I (2011) Detection of financial statement fraud and feature selection using data mining techniques. Decis Support Syst 50(2):491–500

Shen KY, Tzeng GH (2014) A decision rule-based soft computing model for supporting financial performance-improvement of the banking industry. Soft Comput. in Press. doi:10.1007/s00500-014-1413-7

Shu W, Shen H (2014) Incremental feature selection based on rough set in dynamic incomplete data. Pattern Recogn 47(12):3890–3906

Taiwan Stock Exchange Corporation (2012) The Taiwan Stock Exchange Corporation web site in Taiwan. http://www.tse.com.tw. Accessed 4.10. 2013

Tsai CF, Chiou YJ (2009) Earnings management prediction: a pilot study of combining neural networks and decision trees. Expert Syst Appl 36(3):7183–7191

Ulutagay G, Ecer F, Nasibov E (2014) Performance evaluation of industrial enterprises via fuzzy inference system approach: a case study. Soft Comput. in Press. doi:10.1007/s00500-014-1263-3

Vatolkin I (2012) Multi-objective evaluation of music classification. In: Gaul W, Geyer-Schulz A, Schmidt-Thieme L, Kunze L (eds) Challenges at the interface of data analysis, computer science, and optimization. In: Proceedings of the 34th annual conference of the Gesellschaft für Klassifikation e. V. Springer, Berlin, pp 401–410

Vatolkin I, Preuß M, Rudolph G (2011) Multi-objective feature selection in music genre and style recognition tasks. In: Krasnogor N, Lanzi PN (eds) Proceedings of the 2011 genetic and evolutionary computation conference (GECCO). ACM Press, New York, pp 411–418

Vatolkin I, Preuß M, Rudolph G, Eichhoff M, Weihs C (2012) Multi-objective evolutionary feature selection for instrument recognition in polyphonic audio mixtures. Soft Comput 16(12):2027–2047

Verikas A, Kalsyte Z, Bacauskiene M, Gelzinis A (2010) Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. Soft Comput 14(9):995–1010

Yan W, Clack CD (2011) Evolving robust GP solutions for hedge fund stock selection in emerging markets. Soft Comput 15(1):37–50

Yeh CC, Chi DJ, Lin YR (2014) Going-concern prediction using hybrid random forests and rough set approach. Inf Sci 254:98–110

Zhiqiang G, Huaiqing W, Quan L (2013) Financial time series forecasting using LPP and SVM optimized by PSO. Soft Comput 17(5):805–818