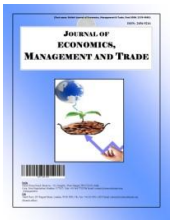

Journal of Economics, Management and Trade

26(4): 77-91, 2020; Article no.JEMT.58568
ISSN: 2456-9216
(Past name: British Journal of Economics, Management & Trade, Past ISSN: 2278-098X)

An Effective Enterprise Earnings Management Detection Model for Capital Market Development

Suduan Chen[1] and Zong-De Shen[2*]

[1]Department of Accounting Information, National Taipei University of Business, No.321, Sec. 1, Jinan Rd. Zhongzheng District, Taipei, 100, Taiwan.
[2]Department of Accounting, Chinese Culture University, No.55, Hwa-Kang Rd., Yang-Ming-Shan, Taipei, 11114, Taiwan.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JEMT/2020/v26i430250
Editor(s):
(1) Dr. Ebere Ume Kalu, University of Nigeria, Nigeria.
Reviewers:
(1) Lemma Shallo Hunde, Wolkite University, Ethiopia.
(2) Majdi A. Quttainah, Kuwait University College of Business, Saudi Arabia.
Complete Peer review History: http://www.sdiarticle4.com/review-history/58568

Original Research Article

Received 26 April 2020
Accepted 02 July 2020
Published 10 July 2020

ABSTRACT

This study focuses on accrual-based earnings management. The purpose of this study is to establish an innovative and high-accuracy model for detecting earnings management using hybrid machine learning methods integrating stepwise regression, elastic net, logistic regression (Logit regression), and decision tree C5.0. Samples of this study are the electronic companies listed on the Taiwan Stock Exchange, and data are derived from the Taiwan Economic Journal (TEJ) for a period of ten years from 2008 to 2017. Results show that the earnings management detection model, as established by elastic net and C5.0, provides the best classification performance, and its average accuracy reaches 97.32%.

Keywords: Earnings management; machine learning elastic net; C5.0.

1. INTRODUCTION

Earnings management is also known as earnings manipulation [1–3]. Under the Generally Accepted Accounting Principles (GAAP), a competent authority is still entitled to certain manipulation over the accounting information of companies under the following four earnings

**Corresponding author: E-mail: szd2@ulive.pccu.edu.tw;*

management methods [4,5]: (1) selection of the GAAP; (2) control over the time of occurrence and recognition of transactions; (3) elastic treatment of economic items not stated in the GAAP; and (4) adjustment to discretionary accruals.

When an enterprise fails to reach its expected objective, which would pose a risk of default of debt contracts or have an adverse impact on returns to management, the management may be motivated to make adjustment to earnings or other profit indicators by interfering with the preparation of financial statements attributable to their self-interest [6]. In the case of adjustment to information of financial statements for a specific purpose, a greater adjustment amount represents a greater deviation of information presented in financial statements from actual conditions; at this moment, financial information loses the feature of faithful representation. Big-bath charges, commonly known as “take a big bath”, is a common earnings management practice, pursuant to which, losses that should be recognized, but not recognized in previous years, or losses that may be incurred in subsequent periods, are recognized in one accounting period under a centralized method by manipulating manipulative accruals, in order to transfer profits among different accounting periods and accordingly achieve the purpose of adjusting profits. Where a listed company applies big-bath charges, a CPA shall not issue an audit report with an unqualified opinion.

Davidson [7] defined earnings management (EM) as “the process where the managers of companies take various measures to achieve expected earnings on the premise of not violating the generally accepted accounting principles”. Healy and Palepu [8] called the behavior interfering with the preparation of financial statements as earnings management, including that, among others, managers could judge and establish transactions applying discretionary accruals, in order to distort the financial operations and reports of the company and further mislead stakeholders’ awareness of the company’s operating performance, or change the contracts and agreements heavily affected by financial figures. In recent years, Doyle et al. [9] proposed a similar view by defining earnings management as “management attempts to obtain expected interests through accounting income through a certain method or procedure”. Most managers would select to implement earnings management through accounting principles or controlling the time point of related transactions [1].

Through years of development to date, research on earnings management can be classified into three major categories: accrual-based earnings management, real earnings management [10], and classification shifting earnings management, as proposed by McVay [11]. Regarding the purpose of accrual-based earnings management, the management, at their own discretion, allots accrued profits at liberty on the premise of not violating accounting principles, thus, earnings management could be implemented under a flexible method; however, such practice would affect the information presented in financial statements. For the purpose of real earnings management, an enterprise manipulates their reporting of earnings based on real economic activities. Accordingly, information presented in corporate financial statements would deviate from the actual conditions; through such practice, earnings management would significantly reduce the reliability of information presented in financial statements [12,13]. Where an earnings threshold fails to be reached through manipulation of accruals due to an excessive gap between actual earnings and the objective, the management of an enterprise would apply real earnings management instead [10]. Regarding the purpose of accounts transfer, accounting accounts in financial statements are reclassified mainly through deviation in the assessment of specific accounts by investors and stakeholders. However, due to inherent subjective judgment on classification and the unaffected total amount, such practice would not have material impact on the information presented in financial statements.

Among earnings management methods, accrual-based earnings management does not violate the provisions of GAAP, and management may, at their own discretion, make adjustments; therefore, it is the most common earnings management method. Accordingly, accrual-based earnings management has been discussed in many researches [4–6,14–18]. Earnings management, manipulation through accruals featuring reverse in a subsequent period, and no impact on actual economic activities are the most common earnings management practices [14]. Accruals can be divided into discretionary accruals (DA) and non-discretionary accruals. Just like a bad debt ratio of estimated accounts receivable, through manipulation of discretionary accruals, the management of an enterprise may adjust

information of earnings without violating GAAP. Non-discretionary accruals are generated from the normal operating activities of an enterprise and fall within the non-maneuverable items of accruals.

The models commonly applied for measuring earnings management level include the Jones Model [17], Modified Jones Model proposed by Dechow et al. [4] through modifying the Jones Model, and the Performance Model, as proposed by Kothari et al. [18] and based on the Modified Jones Model by incorporating the rate of return for consideration.

Jones [17] considered that non-discretionary accruals are not fixed amounts, and would change subject to external environmental factors. Therefore, non-discretionary accruals are estimated using a time series model, which considers corporate operating conditions and amount of depreciation, then discretionary accruals equal to the total accruals less non-discretionary accruals are obtained. In its estimation equation, a change in sales revenue is used for measuring corporate operating conditions, and controlling changes in items receivable and payable relating to operation in non-discretionary accruals, while total assets is used for controlling change in depreciation expenses, and the computational equation is as shown in Eq. (1):

$$\frac{TA_{it}}{A_{it-1}} = \alpha_{it}\frac{1}{A_{it-1}} + \beta_{1it}\frac{\Delta REV_{it}}{A_{it-1}} + \beta_{2it}\frac{PPE}{A_{it-1}} + \varepsilon_{it} \quad (1)$$

Where,

$TA_{it}-1$: is the total accrual of i enterprise in the t-th year.
$A_{it}-1$: is the total asset of i enterprise at the beginning of the t-th year.
REV_{it}: is the revenue of i enterprise in the t-th year. ΔREV refers to change in revenue.
PPE_{it}: is the total fixed asset of i enterprise in the t-th year.
ε_{it}: is the residual of the estimation equation.

Dechow et al. [4] was of the view that, in the Jones Model, the possibility of management manipulating total accrual items through the time point of recognition of charge sales is ignored, thus, only the impact of operating revenue on accruals is considered. Therefore, a modified model is proposed, in which, the change in charge sales was deducted from operating revenue in Eq. (1); in addition, in order to avoid the impact of enterprise scale on the estimation effect, the total assets at the beginning of the period are used as the base for deflation. The proposed modified estimation equation is shown in Eq. (2):

$$\frac{TA_{it}}{A_{it-1}} = \alpha_{it}\frac{1}{A_{it-1}} + \beta_{1it}\frac{\Delta REV_{it}-\Delta REC_{it}}{A_{it-1}} + \beta_{2it}\frac{PPE}{A_{it-1}} + \varepsilon_{it} \quad (2)$$

Where,

ΔREC_{it}: refers to a change in charge sales of the current year.

Then, estimates of parameters of $\widehat{\alpha}, \widehat{\beta_1}$ and $\widehat{\beta_2}$ are obtained through regression analysis, and non-discretionary accruals are calculated using Eq. (3).

$$NDA_{it} = \widehat{\alpha}_{it}\frac{1}{A_{it-1}} + \widehat{\beta}_{1it}\frac{\Delta REV_{it}-\Delta REC_{it}}{A_{it-1}} + \widehat{\beta}_{2it}\frac{PPE}{A_{it-1}} \quad (3)$$

Discretionary accruals equal to total accruals and less non-discretionary accruals, as shown in Eq. (4):

$$DA_{it} = \frac{TA_{it}}{A_{it-1}} - NDA_{it} \quad (4)$$

The Performance Model was proposed by Kothari et al. [18] by modifying the Modified Jones Model, which was mainly because it has been found that doubt over model settings would be incurred when estimating discretionary accruals using the Jones Model or Modified Jones Model when an enterprise is subject to extreme financial performance. Therefore, return on assets has been incorporated in the estimation equation to control the change caused by it. Accordingly, calculations of total accruals and non-discretionary accruals are conducted after incorporating return on assets of the current year in Eq. (2) and Eq. (3), respectively.

While the above three models have their respective good effects in measuring earnings management level, the statistical analysis methods applied in subsequent derivative researches mainly include traditional regression analysis, univariate analysis, multiple discriminant analysis (MDA), and logistic regression analysis. Moreover, such traditional statistical models must be subject to specific restrictions or hypotheses, such as linearity, normality, and independent input variables; however, if broader aspects are considered, it will require that more complicated and diversified variables are incorporated and analyzed. Violation of above hypothetical conditions is a

common occurrence, and would further affect the effectiveness of the analysis model. Accordingly, Höglund [16] proposed the view that an earnings management model is not a simple linear model; therefore, the data mining technique was applied in this study. In short, the data mining technique is a procedure transferring complicated data into knowledge; in addition to implicit rules and information to be identified in different procedures, a data mining model could be established by integrating various models, in order to integrate incomplete and fuzzy information [19].

Elastic Net (EN), as proposed by Zou and Hastie [20], mainly modifies the least absolute shrinkage and selection operator (LASSO), as proposed by Tibshirani [21]. In addition to reserving the OLS loss function concept, the total absolute coefficient and total coefficient square are incorporated as penalty functions. As compared with the LASSO model, EN would not be subject to the restriction that sample size must be greater than the number of explanatory variables (also known as independent variables) in respect of feature selection. Therefore, EN could identify a complete cluster of key variables from a cluster of highly correlated and significant explanatory variables. Due to its superior variable selection ability, EN has been repeatedly discussed and applied in the feature selection procedures of various researches [22–26].

The decision tree (DT) is a classification model established using the inductive learning approach. The difference between it and traditional models, including MDA and Logit, lies in the fact that it is free from statistical hypothesis restriction, and tree-shaped judgment rules could be formed according to results after treating discrete and continuous variables (Viaene et al. [27]; Jan 2018). Furthermore, as stated by Chen et al. [5] the main strengths of DT also include its ability to treat incomplete data and explore the potential relations among massive and complicated input and output variables, in addition to not being subject to any statistical hypothesis of the sample data. Thus, C5.0 is one of the most commonly used DT algorithms at present [27].

Although traditional statistical methods are questioned by researchers, they also have their advantages; and machine learning methods have great advantages recommended by many researchers. Therefore, it is worth to carry on the research by combining traditional statistical methods and machine learning methods-- EN and DT.

This study focused on accrual-based earnings management and attempts to establish an innovative and high-accuracy model for detecting earnings management using hybrid machine learning methods and traditional statistical methods integrating elastic net, decision tree C5.0, stepwise regression, and logistic regression (logit regression).

2. MATERIALS AND METHODS

Data used in this study are derived from the Taiwan Economic Journal (TEJ) and applied machine learning methods include stepwise regression, elastic net, logistic regression (logit regression), and decision tree C5.0. Stepwise regression and logistic regression are commonly used traditional statistical methods, while elastic net and C5.0 are newer methods. In Stage I, important variables are selected by stepwise regression and elastic net; in Stage II, classification models are established using logistic regression (logit regression) and C5.0.

2.1 Stepwise Regression

Stepwise regression is a linear regression-type modeling method, as well as a statistical method commonly used by researchers. Its main concept is to implement stepwise and one-by-one judgment regarding whether each independent variable has significant influence on the dependent variables. F-testing is conducted upon inputting each independent variable, and t-testing is conducted to select the independent variables one by one. Independent variables with significant influence would be introduced, while independent variables without significant influence would be removed, in order to ensure that each of the identified independent variables has significant influence on dependent variables. This process is repeated until no newer variables are introduced. Stepwise regression ensures that the independent variables reserved in the model are important and not subject to serious multi-collinearity.

2.2 Elastic Net

Elastic net, as proposed by Zou and Hastie [20], mainly modifies the least absolute shrinkage and selection operator (LASSO), as proposed by Tibshirani [21], and its computation is shown in Eq. (5).

$\hat{\beta} = argmin_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)j=1p\beta j2)$ (5)

Where,

$\lambda > 0, 0 < a < 1$

Eq. (5) includes two penalty functions, total absolute coefficient (L1-Norm) and total coefficient square (L2-Norm), $\sum_{j=1}^{p}|\beta_j|$ and $\sum_{j=1}^{p}\beta_j^2$), respectively. On one hand, the biggest restriction on the LASSO algorithm lies in that n variables at most could be selected if the number of variables is greater than sample size (p > n) [28]. As compared with the restriction on LASSO, EN could select more n variables in the same case. On the other hand, where there are highly correlated and significant variables in data, LASSO would select one of a cluster, while EN would select a complete cluster; selection is mainly based on the estimated coefficient of $\hat{\beta}$ throughout the process; where the value is > 0, the variable would be selected; otherwise, the variable would be removed.

2.3 Logistic Regression

Logistic regression (logit regression), i.e., the Logit model, is one of the discrete choice approaches, and falls within multivariate statistical analysis. The logarithmic function used in logistic regression is the Sigmoid function. Logistic regression, which is similar to linear regression analysis, is mainly to explore the correlation between independent variables and a dependent variable. A dependent variable in linear regression is generally a continuous variable, while the dependent variables discussed in logistic regression are mainly nominal variables.

By the Logit model of detection earnings management, which is mainly converted from a log probability function. It is divided into two forms: the occurrence of earnings management (Y=1) and without earnings management (Y=0). The probabilities are p and 1-p, respectively.

When the conditional probability of earnings management is set to p(x), the Logit function can be written as Eq. (6).

$P(x) = ln\left[\frac{p}{1-p}\right] = ln\left[e^{f(X)}\right] = f(X) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ (6)

2.4 Decision Tree C5.0

C5.0 is one of supervised learning algorithms derived from DT, and was proposed by Quinlan [29] by modifying ID3.

Assuming S is the data set of s, including n different categories C (i=1,2,…,n), while si refers to the Number of Distinct Categories of each Ci; therefore, the expected information could be expressed, as shown in Eq. (7).

$$I(s_{1,}s_{2,\ldots\ldots,}s_{n,}) = -\sum_{i=1}^{n} p_i \, log_2(p_i) \quad (7)$$

Where,

P denotes the possibility that any event may fall within C. Assuming A attribute has v different values, therefore, A attribute could divide S data set into v subsets, among which, Sj refers to the set formed by events falling within aj value of A attribute. When A attribute is selected as a test attribute, it will include all subsets formed according to the nodes of S set. Assuming sij refers to the number of events in sj subset of Ci category, its entropy would be used as the expected information for dividing subsets according to A attribute, and can be expressed, as shown in Eq. (8).

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+s_{2j}+\cdots+s_{jn}}{s} I(s_{1j,}s_{2j,\ldots\ldots,}s_{jn,}) \quad (8)$$

The smaller the value of entropy, the higher the purity of its subset. For subset sj, expected information can be expressed, as shown in Eq. (9).

$$I(s_{1j,}s_{2j,\ldots\ldots,}s_{jn,}) = -\sum_{i=1}^{n} p_j \, log_2(p_j) \quad (9)$$

In measuring with entropy, Gain is also called information Gain, as shown in Eq. (10).

$$\Delta info = I(Parent) - \sum_{j=1}^{k} \frac{n(v_j)}{N} I(v_j) \quad (10)$$

2.5 Sampling and Variable Selection

2.5.1 Data sources

Samples of this study are the electronic companies listed on the Taiwan Stock Exchange, and data are derived from the Taiwan Economic Journal (TEJ), for a period of ten years from 2008 to 2017. Regarding the calculation of DA (discretionary accruals), the Performance Model, as proposed by Kothari et al. [18], was applied in this study for estimation. A stricter standard was applied for threshold value. DA values that are 0.5 standard deviation greater than the average

value would be classified into EM samples, and others would be classified into Non-EM samples. Samples were matched referring to the research model by Kotsiantis et al. [30], one EM sample was matched with three Non-EM samples (EM: Non-EM = 989: 2967 = 1: 3). The sample distributions used in this study are shown in Table 1.

Table 1. Sample distribution

Sample classification	Number of sample (Companies)
EM samples	989
Non-EM samples	2967
Total samples	3956

Table 2. Research variables and definitions

No.	Variable definition or formula (The year before the year of earnings management)
X01	Total assetsgrowth rate:△Total assets ÷ Total assets prior year
X02	ROAgrowth rate:△ROA ÷ ROA prior year
X03	ROEgrowth rate:△ROE ÷ ROE prior year
X04	Total assets turnover : Net Sales ÷ Average total assets
X05	Inventory turnover : Cost of goods sold ÷ Average inventory
X06	Accounts receivable turnover: Net sales ÷ Average accounts receivable
X07	Sales-to-equity ratio: Sales revenue ÷ Total equity
X08	Debt ratio: Total liabilities ÷ Total assets
X09	Quick ratio: Quick assets ÷ Current liabilities
X10	Current ratio: Current assets ÷ Current liabilities
X11	Operating expenses ratio: Operating expenses ÷ Operating income
X12	Operating profit margin: Operating expenses ÷ Sales revenue
X13	Profit margin: Gross margin ÷ Sales revenue
X14	Pre-taxincome-to-capital ratio: Pre-taxincome ÷ Capital
X15	Operating income-to-capital ratio: Operating income ÷ Capital
X16	Sales revenue growth rate: △Sales revenue ÷ Sales revenue prior year
X17	Operating profit margin growth rate:△Operating profit margin÷Operating profit margin prior year
X18	Net income growth rate: △Net income÷ Net income prior year
X19	Equity growth rate: △Equity÷ Equity prior year
X20	Operating cash flow ratio: Operating cash flow ÷ Current liabilities
X21	Cash reinvestment ratio: Operating cash flow ÷ (Fixedassets + Long-term investment + Other assets + Networking capital)
X22	Times interest earned: Earnings before interest and tax ÷ Interest expense
X23	Interest expenture ratio: (Interest expenture+ Interest expenture capitalized) ÷ Sales revenue
X24	Debt-to-equity ratio: Total liabilities ÷ Total equity
X25	Long-term funds appropriate rate: (Total stockholders' equity+Long term liabilities) ÷Total fixed assets
X26	Earnings per share: (Net income–Dividends of preferred stock) ÷Average common stocks outstanding
X27	The net asset value of each share: Equity÷Average common stocks outstanding
X28	Operating cash flow per share: (Cash flow from operating activities –Dividends of preferred stock) ÷Average common stocksout standing
X29	Sales revenue per share: Sales revenue÷Average common stocks outstanding
X30	Operating income per share: Operating income÷Average common stocks outstanding
X31	Pre-tax income pershare: Pre-tax income÷Average common stocks outstanding
X32	The ratio of stocks held by directors and supervisors: Number of stocks held by directors and supervisors÷Total number of common stock outstanding
X33	The ratio of pledged stocks held by directors and supervisors: The number of pledged stocks held by directors and supervisors÷Number of stocks held by directors and supervisors
X34	Family firms?: yes is 1, no is 0
X35	Audited by BIG4 (the big four CPA firms)?: Audited by BIG4 is 1, otherwise is 0
X36	Going concern doubt?: Yes is 1, no is 0
X37	Financial failure?: Yes is 1, no is 0

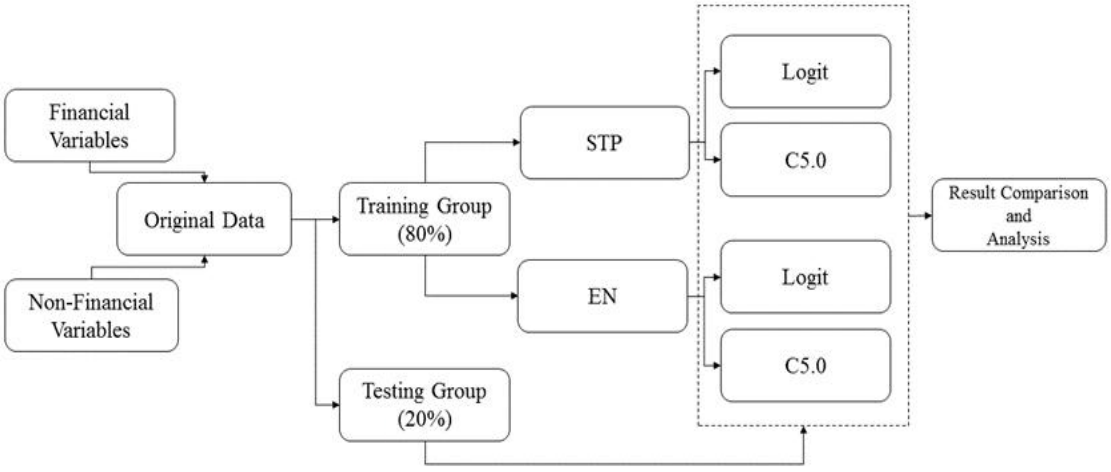

Fig. 1. Research procedure

2.5.2 Variable definitions

Dependent variable Y is used to judge whether there is earnings management; the value greater than the average value plus 0.5 standard deviation is 1, and otherwise is 0. This study summarized the independent variables used in prior literature and in practice for measuring whether there is earnings management of an enterprise. A total of 37 research variables are used, including 31 financial variables and 6 non-financial variables (corporate governance variables). The definitions of the research variables are listed in Table 2.

2.6 Research Process

In prior studies, feature selection was used to facilitate removing unsuitable attributes that would interfere with features and reduce the dimensions of the data set, thus, further improving the performance of the data mining algorithm [23,31–34]. In addition to the aforesaid data collection preparation work, in the process of variable selection in Stage I, the traditional method - stepwise regression, and a new method - elastic net, were applied in this study for selecting important variables. In this stage, machine learning and variable selection were implemented to select fewer important variables. Then, in Stage II, modeling was implemented using commonly used logistic regression and C5.0, respectively, according to the variables selected using the aforesaid two selection methods. Finally, the 4 models for detecting enterprises' earnings management established in this study were compared and analyzed to identify the model with the best detection accuracy. The research procedure is shown in Fig. 1.

3. RESULTS AND DISCUSSION

Models for detecting earnings management were established in two stages in this study. In Stage I, important variables were selected by stepwise regression and elastic net; then Stage II modeling was implemented using logistic regression and C5.0. The very rigorous ten-fold fold cross-validation was applied in this study for obtaining the detection accuracy. The results are detailed, as follows.

3.1 Stepwise Regression Selection

Stepwise regression is a feature selection method commonly used in research. Huberty [35] was of the view that stepwise regression has the following three functions: (1) select or delete variables; (2) assess the importance of variables; (3) select variables and assess their importance. The process of variable selection was by stepwise regression, and the important variables are shown in Fig. 2 and Table 3. It can be seen from Fig. 2 that stepwise regression stopped selection after completion of step 12, and four indicators were used as selection standards, namely, AIC, AICC, SBC, and Adj R-Sq, in order to reach the optimum value.

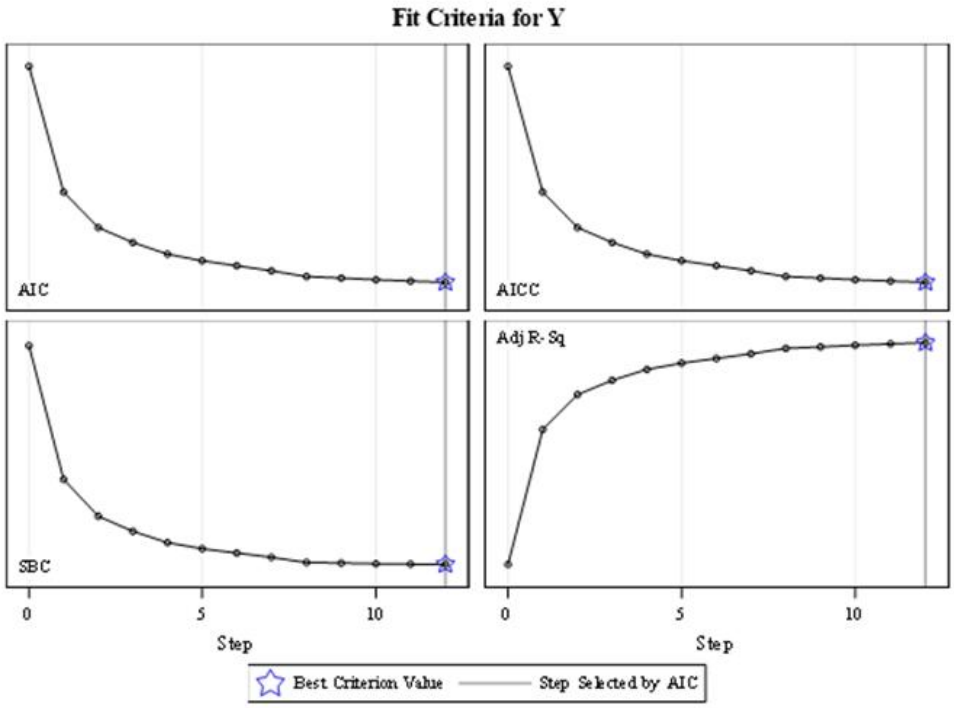

Fig. 2. Variables screening by STP

Table 3. Summary of selection by stepwise regression

Step	Effect entered	t value	Estimate	AIC	SBC
0	Intercept	-8.81	-0.098313	-5498.1012	-11269.440
1	X4	27.44	0.244733	-6350.1442	-12114.821
2	X16	10.11	0.000991	-6590.4098	-12348.425
3	X2	7.31	0.003765	-6692.3177	-12443.672
4	X12	-6.36	0.001542	-6770.5542	-12515.247
5	X28	9.84	0.007998	-6815.7503	-12553.782
6	X29	-5.91	-0.000753	-6849.4342	-12580.804
7	X1	5.53	0.001032	-6883.7370	-12608.445
8	X3	-6.12	-0.001656	-6922.7621	-12640.809
9	X10	3.83	0.000092057	-6932.9211	-12644.306
10	X37	3.63	0.045480	-6944.4659	-12649.190
11	X6	-3.32	-0.000677	-6953.5018	-12651.564
12	X19	3.20	0.000187	-6961.7487	-12653.150
				F value	141.08

As shown in Table 3, important variables selected by stepwise regression and in descending order are X4: total assets turnover, X16: sales revenue growth rate, X2: ROA growth rate, X12: operating profit margin, X28: operating cashflow per share, X29: sales revenue per share, X1: total assets growth rate, X3: ROE growth rate, X10: current ratio, X37: financial failure, X6: accounts receivable turnover, and X19: equity growth rate.

3.2 Selection by Elastic Net

The selection process of elastic net is summarized in Fig. 3. The process of variable selection goes through 18 steps in total. Changes in the coefficient are presented in the upper half of Fig. 3, while changes in SBC are presented in the lower half of the figure. It can be seen from the figure that SBC reaches the bottom upon selection of X18.

The important variables selected by elastic net and their related values are summarized in Table 4. Important variables, as selected through the 18 steps in a descending order are X4: total assets turnover, X16: sales revenue growth rate, X2: operating profit margin, X1: total assets growth rate, X12: operating profit margin, X19: equity growth rate, X36: going concern doubt, X37: financial failure, X28: operating cashflow per share, X3: ROE growth rate, X35: audited by BIG4, X32: the ratio of stocks held by directors and supervisors, X25: long-term funds appropriate rate, X10: current ratio, X27: the net asset value of each share, X6: accounts receivable turnover, X21: cash re-investment ratio, and X18: net income growth rate.

3.3 Modeling and Cross-Validation

In this stage, the variables selected by stepwise regression and elastic net were introduced in logistic regression and C5.0 for establishing the classification models. In order to compare the accuracy and stability of the models, this study applied random sampling; where 80% of data in the data set were sampled at random as the training group; 20% of data in the original data set were sampled at random as the testing group. Furthermore, in order to observe the stability of the proposed models, the very rigorous ten-fold cross-validation, as recognized by the academic community and researchers, was applied in this study for obtaining detection accuracy [31,36,37]. Pursuant to the method, modeling and verification were implemented ten times, respectively, and finally, the average accuracy of the ten results was obtained. Such average accuracy of the ten results, from the perspective of scientific research, is naturally convincing and widely accepted [36].

3.3.1 Stepwise regression-logistic regression model

The ten-fold cross-validation results of the stepwise regression-logistic regression model (STP-Logit model) are listed in Table 5, which shows that the average accuracy of the training group and testing group are 87.43% and 87.37%, respectively.

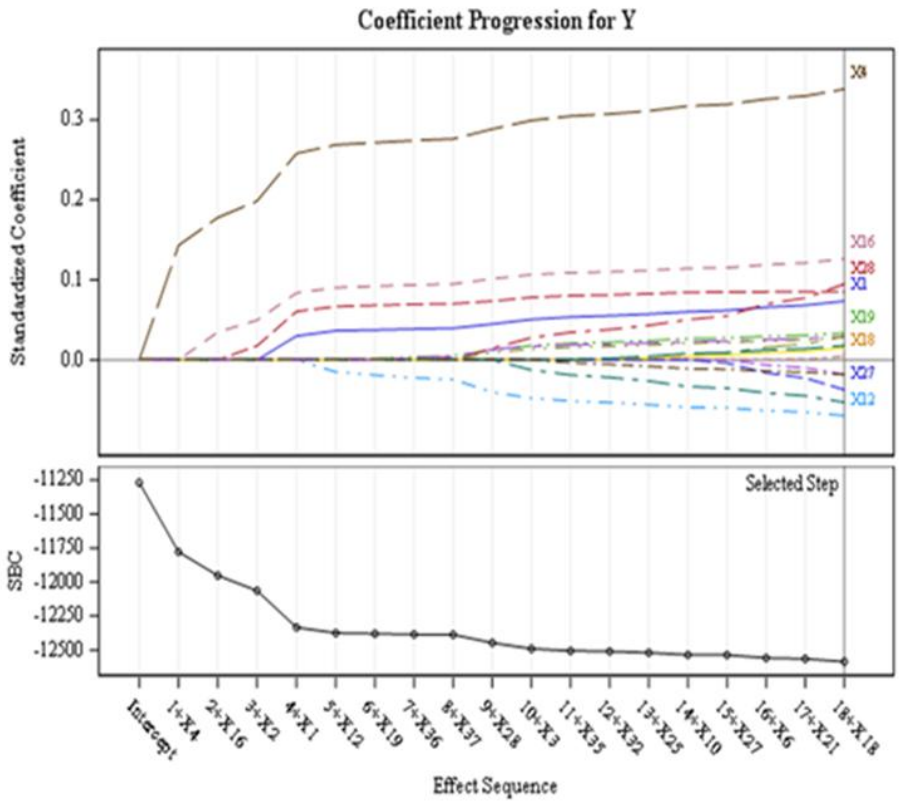

Fig. 3. Variables screening by EN

Table 4. Summary of selection by elastic net

Step	Effect entered	Estimate	AIC	SBC
0	Intercept	-0.034059	-5498.1012	-11269.440
1	X4	0.189688	-6013.4870	-11778.164
2	X16	0.000896	-6194.3505	-11952.366
3	X2	0.003251	-6312.5477	-12063.902
4	X1	0.000911	-6588.9366	-12333.629
5	X12	-0.001184	-6637.2266	-12375.258
6	X19	0.000157	-6648.1218	-12379.492
7	X36	0.127499	-6660.2589	-12384.967
8	X37	0.029183	-6668.6933	-12386.740
9	X28	0.005541	-6735.7377	-12447.123
10	X3	0.000923	-6785.6543	-12490.378
11	X35	0.023072	-6808.0479	-12506.110
12	X32	0.000463	-6818.7285	-12510.129
13	X25	0.000000780	-6834.3545	-12519.094
14	X10	0.000058543	-6857.0364	-12535.114
15	X27	0.029183	-6864.9114	-12536.328
16	X6	-0.000310	-6893.2852	-12558.040
17	X21	0.000094562	-6906.2953	-12564.389
18	X18	0.000000533	-6933.6653	-12585.097

Table 5. STP-Logit model's accuracy by the ten-fold cross-validation

No.	Training group accuracy	Testing group accuracy
1	87.74%	86.17%
2	87.31%	87.82%
3	87.14%	88.51%
4	87.35%	87.67%
5	87.33%	87.74%
6	87.31%	87.82%
7	87.67%	86.42%
8	87.57%	86.82%
9	87.50%	87.08%
10	87.36%	87.64%
Average	87.43%	87.37%

Table 6. EN-Logit model's accuracy by the ten-fold cross-validation

No.	Training group accuracy	Testing group accuracy
1	87.26%	87.66%
2	87.40%	87.12%
3	87.56%	86.48%
4	87.11%	88.33%
5	87.23%	87.80%
6	87.15%	87.50%
7	87.46%	86.28%
8	87.13%	87.59%
9	87.24%	87.17%
10	87.13%	87.60%
Average	87.27%	87.35%

3.3.2 Elastic net-logistic regression model

The ten-fold cross-validation results of the Elastic net-logistic regression model (EN-Logit model) are listed in Table 6; the average accuracy of the training group and testing group are 87.27% and 87.35%, respectively.

Through comparison of average accuracy of classification models, as established by Logit,

show that, STP-Logit is slightly superior to the EN-Logit model.

3.3.3 Stepwise regression-c5.0 model

The ten-fold cross-validation results of Stepwise regression-C5.0 model (STP-C5.0 model) are listed in Table 7, which shows that the average accuracy of the training group and testing group are 96.66% and 96.76%, respectively.

3.3.4 Elastic net-c5.0 model

The ten-fold cross-validation results of the Elastic net-C5.0 model (EN-C5.0 model) are listed in Table 8, which shows that the average accuracy of the training group and testing group are 97.14% and 97.32%, respectively. Comparison of the average accuracy of classification models, as established by C5.0, shows that the EN-C5.0 model is superior to the STP-C5.0 model in all aspects.

3.3.5 Additional discussion

In order to facilitate comparing and analyzing the results of the above classification procedures, the results of aforesaid 4 models are summarized in Table 9.

A total of 37 input variables were used in this study; therefore, it was of great importance to select important variables from a large number of variables. As input variables will affect the accuracy of models, 4 models for detecting earnings management were established by Logit and C5.0 with two groups of input variables, as selected according to the analysis results of STP and EN, respectively. As shown in Table 9, the accuracy of the C5.0 classification model is significantly higher than that of Logit, and irrespective of variable selection by STP or EN. Through further observation of the classification results of STP-C5.0 and EN-C5.0, both models show higher accuracy, at 96.76% and 97.32%, respectively. Gain value is an important basis for establishing a classification model by C5.0. In this study, the Gain values in modeling of EN-C5.0 and STP-C5.0 were charted for comparison. It can be seen from Fig. 4 that the accuracy of EN-C5.0 is indeed higher than that of STP-C5.0.

Table 7. STP-C5.0 model's accuracy by the ten-fold cross-validation

No.	Training group accuracy	Testing group accuracy
1	96.79%	96.14%
2	96.65%	96.68%
3	96.58%	96.98%
4	96.66%	96.66%
5	96.66%	97.55%
6	96.50%	97.27%
7	96.50%	97.27%
8	96.59%	96.94%
9	96.91%	95.69%
10	96.71%	96.46%
Average	96.66%	96.76%

Table 8. EN-C5.0 model's accuracy by the ten-fold cross-validation

No.	Training group accuracy	Testing group accuracy
1	97.09%	97.51%
2	97.14%	97.31%
3	97.21%	97.05%
4	97.21%	97.04%
5	97.21%	97.04%
6	97.03%	97.76%
7	97.13%	97.38%
8	97.23%	96.96%
9	97.07%	97.60%
10	97.10%	97.50%
Average	97.14%	97.32%

Table 9. The accuracy of detection models

Models	Average accuracy	Average error rate
STP-Logit	87.39%	12.61%
EN-Logit	87.35%	12.65%
STP-C5.0	96.76%	3.24%
EN-C5.0	97.32%	2.68%

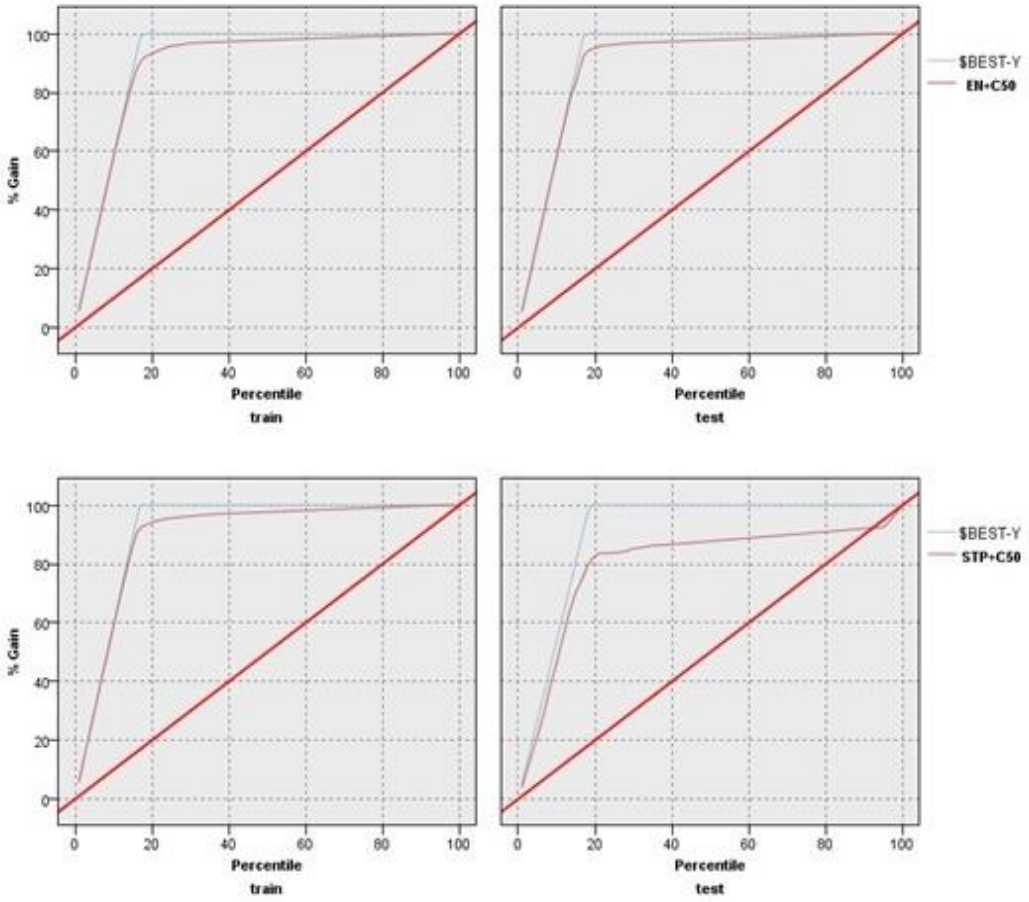

Fig. 4. Percentage of accumulated gain of STP-C5.0 and EN-C5.0

Table 10. Classified rules of EN-C5.0

Rule set of EN-C5.0 model	
Rule 1	Sales revenue growth rate >21.77 and Cash reinvestment ratio >13.16 and Accounts receivable turnover >2.99 : earnings management acts
Rule 2	Sales revenue growth rate >21.77 and Cash reinvestment ratio >13.16 and Accounts receivable turnover <2.99 and ROE growth rate <21.02 : earnings management acts

The rule set and judgment rules used in the EN-C5.0 model with the best classification performance are listed in Table 10. Rule 1 represents that the sample is classified into earnings management manipulation.

4. CONCLUSION

Earnings management would affect accounting data, in particular, the earnings reported in accounting other than the actual earnings of an

enterprise. The selection of accounting method, application of the accounting method, change in the accounting estimate, control over the time point of the application of the accounting method, and the time point of the occurrence of transactions are all typical earnings management methods. Earnings management is also a kind of window dressing for financial statements, or accounting fraud, which has the purpose of misleading or deceiving the users of the financial statements. Generally, a company with poor corporate governance may actively engage in earnings management, and various issues, including why financial statements fail to disclose a company's real financial position, and why CPAs and accounts review fail to identify defects in financial statements, have caused the academic community and practitioners to attach greater importance to earnings manipulation and corporate governance. Therefore, the research variables in this study also include several corporate governance variables.

Great contributions will be made if a rigorous and effective model for detecting enterprises' earnings management can be established. In this study, a hybrid machine learning approach were applied to detect corporate earnings management by taking electronic companies listed in Taiwan from 2008 to 2017 as the research samples. In Stage I, important variables were selected by stepwise regression and elastic net; in Stage II, effective earnings management detection models were established by logistic regression and C5.0.

Empirical results show that the model for detecting earnings management, as established by elastic net and C5.0, provides the best classification performance, with an average accuracy of 97.32%, which is the best among the four models. For the other models established in this study, the average accuracy is 96.76% for the STP-C5.0 model, 87.39% for the STP-Logit model, and 87.35% for the STP-Logit model; pursuant to which, we can know that the classification performance of C5.0 is superior to that of logistic regression. On the other hand, we can also know that the important variables (evaluation indicators), as selected by both stepwise regression and elastic net, include X1: total assets growth rate, X2: ROA growth rate, X3: ROE growth rate, X4: total assets turnover, X6: accounts receivable turnover, X10: current ratio, X12: operating profit margin, X16: sales revenue growth rate, X19: equity growth rate, X28: operating cashflow per share, X37: financial failure, and are noteworthy variables.

In consideration of the interests of substantial stockholders and short-term debt or the stock price of the company, enterprise operators may often engage in earnings manipulation or earnings management. As financial statements contain diversified and complicated information, experienced auditors may identify the key to judge through considerable time and experience, meaning general users would have difficulty examining the financial statements. Therefore, the judgment rules and earnings manipulation detection model proposed by this study can help auditors and corporate stakeholders make more accurate judgment regarding corporate financial statements within a limited time and cost.

This study could provide a reference to persons engaging in academic research relating to earnings management, as well as the management of enterprises, CPAs and accounts, and securities analyst. This study could really contribute to practice, as explained below, we offer: 1. important financial and non-financial variables as detection indicators for earnings manipulation; 2. a rigorous and effective/ high accuracy model for detecting enterprises' earnings management using hybrid machine learning methods and traditional statistical methods integrating elastic net, decision tree C5.0, stepwise regression, and logistic regression (logit regression).

ACKNOWLEDGEMENTS

The authors would like to thank the editors and the anonymous reviewers of this journal.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Healy PM, Wahlen JM. A review of earnings management literature and its implications for standard setting. Account. Horiz. 1999;13:365-384.
2. Gunny KA. The relation between earnings management using real activities manipulation and future performance: Evidence from meeting earnings benchmarks. Cont. Account. Res. 2010;27: 855-888.
3. Shahzad A. Detecting earnings management and earnings manipulation in BRIC countries; a panel data analysis for

post global financial crisis period. Int. J Account Res. 2016;4:1-10.
4. Dechow PM, Sloan RG, Sweeney AM. Detecting earnings management. Account. Rev. 1995;70:193-225.
5. Chen FH, Chi DJ, Wang YC. Detecting biotechnology industry's earnings management using Bayesian network, principal component analysis, back propagation neural network and decision tree. Econ. Modell. 2015;46:1-10.
6. Schipper K. Commentary on earnings management. Account Horiz. 1989;3:91-102.
7. Davidson S, Stickney C, Weil R. Accounting: The language of business, 7th Ed. Sun Lakes, Ariz.: T. Horton; 1987.
8. Healy P, Palepu K. Effectiveness of accounting-based debt covenants. J. Account. Econ. 1990;12:97-123.
9. Doyle JT, Jennings J, Soliman MT. Do managers define Non-GAAP earnings to meet or beat analyst forecasts? J. Account. Econ. 2013;56:40-56.
10. Cohen DA, Zarowin P. Accrual-based and real earnings management activities around seasoned equity offerings. J. Account. Econ. 2010;50:2-19.
11. McVay SE. Earnings management using classification shifting: An examination of core earnings and special items. Account. Rev. 2006;81:501-531.
12. Cohen DA, Dey A, Lys TZ. Real and accrual-based earnings management in the Pre- and Post-Sarbanes-Oxley periods. Account. Rev. 2008;83:757-787.
13. Roychowdhury S. Earnings management through real activities manipulation. J. Account. Econ. 2006;42:335-370.
14. Dechow PM, Sloan RG, Hutton AP, Kim JH. Detecting earnings management: A new approach. J. Account. Res. 2012;50: 275-334.
15. Healy PM. The effect of bonus schemes on accounting decisions. J. Account. Econ. 1985;7:85-107.
16. Höglund H. Detecting earnings management with neural networks. Expert. Syst. Appl. 2012;39:9564-9570.
17. Jones JJ. Earnings management during import relief investigations. J. Account. Res. 1991;29:193-228.
18. Kothari SP, Leoneb AJ, Wasley CE. Performance matched discretionary accrual measures. J. Account. Econ. 2005;39:163-197.
19. Gupta R, Modise MP. South African stock return predictability in the context data mining: The role of financial variables and international stock returns. Econ. Model. 2012;29:908-916.
20. Zou H, Hastie T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 2005;67:301-320.
21. Tibshirani R. Regression shrinkage and selection via the Lasso. J. Roy Stat Soc B Met. 1996;58:267-288.
22. Alberto TB, Carlo M, Jose RD. V-SVM solutions of constrained lasso and elastic net. Neurocomputing. 2018;275:1921-1931.
23. Chen S, Hou J, Xiao D. One belt, one road; initiative to stimulate trade in China: A counter-factual analysis. Sustainability. 2018;10:3242-3254.
24. Frank E, Matthias D. High-dimensional LASSO-based computational regression models: Regularization, shrinkage and selection. Mach. Learn. Knowl. Extr. 2019;1:359-383.
25. Katrutsa AM, Strijov VV. Stress test procedure for feature selection algorithms. Chemometr Intell Lab. 2015;142:172-183.
26. Katrutsa AM, Strijov VV. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Syst Appl. 2017;76:1-11.
27. Viaene S, Dedene G, Derrig RA. Auto claim fraud detection using Bayesian learning neural networks. Expert Syst Appl. 2005;29:653-666.
28. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat. 2004;32:407-499.
29. Quinlan JR. Introduction of decision trees. Mach. Learn. 1986;1:81-106.
30. Kotsiantis S, Koumanakos E, Tzelepis D, Tampakas V. Forecasting fraudulent financial statements using data mining. Int. J. Comput. 2006;3:104-110.
31. Jan CL. An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan. Sustainability. 2018;10:1-14.
32. Chen S, Goo YJ, Shen ZD. A hybrid approach of stepwise regression, logistic regression, support vector machine and decision tree for forecasting fraudulent

financial statements. Sci. World J.; 2014.
33. Goo YJ, Chi DJ, Shen ZD. Improving the prediction of going concern of Taiwanese listed companies using a hybrid of LASSO with data mining techniques. Springer Plus; 2016.
DOI: 10.1186/s40064-016-2186-5
34. Hall M, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE T Knowl Data En. 2003;15: 1437-1447.
35. Huberty CJ. Problems with stepwise methods-better alter-natives. Adv Soc Sci Meth. 1989;1:43-70.
36. Chen S. An effective going concern prediction model for the sustainability of enterprises and capital market development. Appl. Econo. 2019;51:3376-3388.
37. Yeh CC, Chi DJ, Lin YR. Going concern prediction using hybrid random forests and rough set approach. Inf. Sci. 2014;254:98-110.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/58568