Economic Modelling 46 (2015) 1–10

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.elsevier.com/locate/ecmod

Detecting biotechnology industry's earnings management using Bayesian network, principal component analysis, back propagation neural network, and decision tree

Fu-Hsiang Chen, Der-Jang Chi *, Yi-Cheng Wang

Department of Accounting, Chinese Culture University, No.55, Hwa-Kang Road, Taipei 11114, Taiwan

ARTICLE INFO

Article history:
Accepted 24 December 2014
Available online 16 January 2015

JEL classification:
G3
C8
M4
M1

Keywords:
Data mining
Bayesian network (BN)
Back propagation neural network (BPN)
Principal component analysis (PCA)
C5.0 decision tree
Accrual earnings management

ABSTRACT

The characteristic of long value chain, high-risk, high cost of research and development are belong to high knowledge based content in the biotech medical industry, and the reliability of biotechnology industry's financial statements and the earnings management behavior conducted by the management in their accrual manipulation have been a critical issue. In recent years, some studies have used the data mining technique to detect earnings management, with which the accuracy has therefore risen. As such, this study attempts to diagnose the detecting biotechnology industry earnings management by integrating suitable computing models, we first screened the earnings management variables with the principal component analysis (PCA) and Bayesian network (BN), followed by further constructing the integrated model with the back propagation neural network (BPN) and C5.0 (decision tree) to detect if a company's earnings were seriously manipulated. The empirical results show that combining the BN screening method with C5.0 decision tree has the best performance with an accuracy rate of 98.51%. From the rules set in the final additional testing of the study, it is also found that an enterprise's prior period discretionary accruals play an important role in affecting the serious degree of accrual earnings management.

1. Introduction

Nowadays biotechnology is not only a burgeoning industry but also a newest target of investors in the universe. More and more people tend to rely on biotechnology for extending life expectancy or maintaining youth; the potential of the biotechnology industry has greatly improved. Biotechnology industry is characterized of a complicated system, a long value chain, specialized divisions of labor, and a prolonged timeline of product development. The biotechnology company's performance is more difficult to accurately evaluate from traditional financial reports (Kessel and Frank, 2007). Earnings are commonly deemed to be the status of an enterprise's past business performance. Given the fact that the stakeholders of an enterprise (usually including investors, creditors, analysts and customers) cannot be directly aware of the enterprise's operating performance, most of them regard corporate earnings as an important index. As a result, earnings management has turned out to be the major impetus for the management.

Nevertheless, if earnings management becomes a norm, financial statement users are very likely to have bias in their judgment on financial statements, which could further lead them to make wrong decisions. Earnings management refers to some method or procedure used by the management in order to have accounting earnings to attain expected goals (Doyle et al., 2013). As such, Schipper (1989) defined earnings management to be the process where the management forcefully intervenes the preparation of financial statements based on economic consideration and motive, and the right conferred by the generally accepted accounting principle to achieve the earnings goal within a lawful range. On the other hand, Healy and Wahlen (1999) considered earnings management to be the management's attempt to affect the content of financial statements with their power, so as to further mislead stakeholders in their understanding of the actual corporate operating performance. In other words, earnings management could be either legal or illegal. The earnings management beyond a certain boundary could possibly turn out to be the management's corruption.

As pointed out by Watts and Zimmer, the bonus plan hypothesis (Watts and Zimmerman, 1990) is the impetus for management personnel to conduct earnings management, in which the company implementing the bonus plan by using the employee bonus mechanism based on accounting earnings as the bonus basis may lead to moral risk. The reason for it is that the reward of the management is related to net income, i.e.,

* Corresponding author at: No.55, Hwa-Kang Road, Department of Accounting, Chinese Culture University, 11114Taipei, Taiwan, Tel.: +886 2 28610511x35525, +886 918084917 (mobile); fax: +886 2 28614177.
E-mail address: derjang@yahoo.com.tw (D.-J. Chi).

http://dx.doi.org/10.1016/j.econmod.2014.12.035
0264-9993/

with the motive of self-interest, the management may increase the earnings reported in the current period through their accounting policy selection. Healy (1985) indicated that, if a company's earnings level is too low in the current year and the level cannot be elevated through any kind of manipulation, the management may bring forward the future incidents unfavorable for earnings to the current year. In so doing, the earnings in the future could substantially increase, which is the so-called "big bath" theory. According to Healy and Wahlen (1999), the earnings management motives come in capital market motive, contract motive and law and regulation motive. Also, in most cases, earnings manipulation is generally made from selection of the accounting method, accrual management and control of the time of transaction occurrence.

In the past, most earnings management related studies explored the correlation of earnings management with other items. For instance, Schipper (1989) pointed out that a lack of sufficient communication and information asymmetry could result in earnings management, and that is the correlation between earnings management and information transparency, or the studies which explore the correlation between earnings management and auditing quality (Cohen et al., 2008) and the correlation between earnings quality and corporate social responsibilities (Pyo and Lee, 2013). Currently, fewer studies have been conducted for detection of earnings management level. In the past, Jones Model (Jones, 1991), Modified Jones Model (Dechow et al., 1995) and Kothari et al. Model (Kotharia et al., 2005) have commonly been used as the models to evaluate if an enterprise conducts earnings management. Only very few studies have been conducted to detect earnings management levels, and most of the studies have adopted conventional statistical methods, such as regression, univariate statistical methods, multiple discriminant analysis (MDA), and logit and probit analyses in investigation. These conventional statistical methods, however, have some restrictive assumptions such as linearity, normality, and independence of predictor or input variables. Considering that the violation of these assumptions occurs frequently within financial data, the methods have intrinsic limitations in terms of effectiveness and validity. According to Höglund (2012), the earnings management model is actually not a linear model. Hence, the study has adopted the data mining method to detect earnings management.

Data mining is a process to transform data into knowledge, which is one of the most active ways in research, development and application in the field of data processing. As implied by its name, data mining is to find implied, regular and potentially useful information and knowledge which can be finally comprehensive from the massive, incomplete and fuzzy information (Gupta and Modise, 2012). In other words, data mining may present some kind of models left in the remaining data, in which the models can be collected together and defined to be a data mining model. The advantage to exploring earnings management with the data mining method is to set up a non-linear model which does not require hypotheses as what is required by the conventional method (Höglund, 2012).

Recently, back propagation neural network have been extensively used in finance (Wang et al., 2011). Using the concepts, the original data are first decomposed into multiple layers by the wavelet transform. Each layer has a low-frequency and a high-frequency signal component. Then a back propagation (BP) neural network model is established by the low-frequency signal of each layer for predicting the future value. Recently, it has been found that its applications in a wide variety of fields include forecasting economic growth (Feng and Zhang, 2014), stock market prediction (Zhang and Wu, 2009) and stock index (Wang et al., 2011). Moreover the non-parametric prediction method known as decision tree (DT) has been used in an attempt to bypass the above mentioned assumptions in MDA and logit (Kim and Upneja, 2014).

DT models for earnings management prediction are set up based on the following advantages. First, a DT does not require any statistical assumptions concerning the data in a training sample. Second, a DT model can handle incomplete and qualitative data. Third, a DT is useful for exploring data to find the relationship between a large number of candidate input variables and the target variable. Finally, the DT model provides a meaningful way of representing acquired knowledge and is easily understood because it yields human comprehensible binary 'if-then' rules (Kim and Upneja, 2014; Hajaizadeh et al., 2010).

In order to help corporate stakeholders, such as investors, creditors, analysts, and customers better understand the degree of biotechnology industry's accrual earnings management, avoid them to suffer a great loss in the stock market as a result of manager's earnings management, 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. Therefore, we attempt to investigate the effectiveness of back propagation neural network (BPN) and DT approach in conducting the earnings management prediction tasks and to predict the characteristics of earnings management, so decision-makers can understand the rules of earnings management. Regarding the above purpose, this paper proposes a novel hybrid model for earnings management prediction by integrating the Bayesian network (BN), principal component analysis (PCA), BPN and DT techniques. BN and PCA methods are used for variable selection in order to obtain the significant independent variables, while BPN and DT are used to generate meaningful rules for earnings management. In order to evaluate the performance of the proposed framework, comparative experiments are conducted other than considering Type I and Type II errors.

2. Literature review

Following the years of development, earnings management has been classified into the three categories of accrual earnings management put forth by Schipper (1989), real earnings management (Cohen and Zarowin, 2010) and classification shifting proposed by McVay (2006). The accrual earnings management is that, without violating the generally accepted accounting principle (GAAP), the manager can use discretionary accruals to freely give decision. In this way, the management may be flexible in doing their earnings management. It, however, will affect the figures shown on financial statements. According to Roychowdhury (2006), real earnings management is the way to help a company break away from the normal financial statement operation rule. Its purpose is to have a company's shareholders consider the achievement of the earnings goal through price discount, increase of sales income by providing a longer credit period or cut-down of the sales cost by mass production or reduction of discretionary expenses.

The disadvantage of real earnings management is that earlier execution is required and an enterprise's real value and real economic activities will be affected. Classification change refers to financial statements' vertical moves, with which the errors are classified according to how investors value an accounting title, so as to elevate an enterprise's valuation and further mislead financial statement users' judgment. Since the figures shown on financial statements are not affected, and classification per se is a subjective matter, it is not likely to raise any doubts from auditing personnel. Given that accrual earnings management can be determined by the management within the range allowed by the GAAP, it can be flexibly adjusted and easily implemented. Furthermore, most prior studies were conducted to explore accrual earnings management (Schipper, 1989; Healy and Wahlen, 1999; Healy, 1985; Jones, 1991; Dechow et al., 1995, 2012; Kotharia et al., 2005; Höglund, 2012), so the study has been conducted to mainly explore detection of accrual earnings management.

2.1. Accrual earnings management

Dechow et al. (2012) considered accrual earnings management to have the reversal characteristic and not to involve real economic activities. The accounting accruals are divided into discretionary accruals and non-discretionary accruals. With discretionary accruals, managers may freely give decision while not violating the GAAP, e.g., set aside the bad debt ratio for account receivables. On the other hand, non-discretionary accruals are mainly related to an enterprise's normal business activities.

They are formed in an economic environment without deliberate man-made manipulation. Since there is difference between the accounting earnings under the accrual basis and the earnings under the cash basis, managers can often use the flexibility of discretionary accruals to manipulate accounting earnings according to the set purpose.

2.2. Motive of earnings management

According to the positive accounting theory put forth by Watts and Zimmerman (1986), management's earnings management motives are categorized into the following three types of hypotheses: bonus plan hypothesis; debt/equity hypothesis; and size hypothesis. Healy and Wahlen (1999) also divided the motives into the following three categories: 1. capital market motive; 2. contract motive; and 3. law and regulation motive. As for the methods adopted by enterprises for earnings management, a majority of studies divide them into the following three types: 1. discretionary accrual management; 2. accounting method selection; and 3. control of the time of transaction occurrence.

2.3. Methods to measure accrual earnings management

The more common accrual earnings management models adopted in prior studies are described as below:

1. Jones Model: Jones (1991) classified accruals into discretionary accruals and non-discretionary accruals. Non-discretionary accruals can be affected by external economic factors. By using the amount of sales change and the total amount of factory equipment depreciating assets as variables, the model is set up as per the Formula (1) below:

$$\frac{TA_{i,t}}{A_{i,t-1}} = \beta_0\left(\frac{1}{A_{i,t-1}}\right) + \beta_1\frac{\left(\Delta REV_{i,t}\right)}{A_{i,t-1}} + \beta_2\left(\frac{PPE_{i,t}}{A_{i,t-1}}\right) + \varepsilon_{i,t}. \quad (1)$$

The variables are defined as follows:

$TA_{i,t}$ the total accruals of sample company i in the tth year.
$A_{i,t-1}$ the total assets of sample company i in the $t-1$th year.
$\Delta REV_{i,t}$ the change of sample company i's sales in the tth year..
$PPE_{i,t}$ the total depreciating fixed assets of sample company i in the tth year
$\varepsilon_{i,t}$ residual terms.

2. Modified Jones Model: Dechow et al. (1995) Modified Jones Model before putting forth the Modified Jones Model. According to Dechow et al. (1995) sales income should not necessarily be total non-discretionary accruals, because the fact that an administrator would early recognize credit sales or delay the recognition might be neglected. As such, the original amount of sales change has been revised to be the amount of change in accounts receivable. The model is as per the Formula (2) below:

$$\frac{TA_{i,t}}{A_{i,t-1}} = \beta_0\left(\frac{1}{A_{i,t-1}}\right) + \beta_1\frac{\left(\Delta REV_{i,t}-\Delta REC_{i,t}\right)}{A_{i,t-1}} + \beta_2\left(\frac{PPE_{i,t}}{A_{i,t-1}}\right) + \varepsilon_{i,t} \quad (2)$$

$\Delta REC_{i,t}$ the change of sample company i's accounts receivable in the tth year.

The definitions of other items are same as above formula.

Dechow et al. (2012) provide a new approach to test for accrual-based earnings management, and uses the inherent property of accrual accounting for which any accrual-based earnings management in one period must reverse in another period, and there are priors concerning the timing of the reversal. By incorporating these priors, the test power and specification for earnings management can be significantly improved. The results indicate that the test incorporating reversals could increase test power by around 40% and provide a robust solution to mitigating model misspecification arising from correlated omitted variables. Hence, the study adopted the formula of non-discretionary accruals recommended by Dechow et al. (2012) to measure accrual earnings management.

3. Study methodology

Some scholars indicated that feature selection would help remove interference features and reduce the dimensionality of data sets by deleting unsuitable attributes, which would therefore further improve the performance of data mining algorithms (Hall and Holmes, 2003). As such, the study used Bayesian network and principal component analysis to screen variables in the first stage. For model establishment in the second stage, the study used the selected variables to set up the models of back propagation neural networks and C5.0 decision tree, and assess the four models' capacity of detecting accrual earnings management. For the third stage, out of the four models, the model with optimal capacity was picked up to establish the decision tree model, followed by finding out the relationship between the selected variables and accrual earnings management and assessing the prediction accuracy of the decision tree model and its accrued rules. Fig. 1 below shows the study flow chart.

3.1. Methods to measure the proxy variables of accrual earnings management

Since the degree of accrual earnings management cannot be directly measured, the study adopted discretionary accruals as the proxy variables of accrual earnings management and used them to measure the degree of an enterprise's earnings management. Non-discretionary accruals may change with the change of corporate operation, which is the part that the management cannot judge. Discretionary accruals are mostly calculated by first counting up the total accruals, followed by deducting non-discretionary accruals to come up with the discretionary accruals. Prior studies usually used the following two methods to estimate the total accruals: the cash flow statement method and balance sheet method. As indicated by Collins and Hribar (2002), the deviation of total accruals estimated by the cash flow statement method is smaller. Hence, the study adopted the cash flow statement method proposed by Collins and Hribar (2002) to assess total accruals.

The balance sheet method is shown as per Formula (3) below:

$$TAC_{i,t} = \Delta CA_{i,t} - \Delta CL_{i,t} - \Delta Cash_{i,t} + \Delta STDEBT_{i,t} - DEPTN_{i,t}. \quad (3)$$

The variables are defined as follows:

$TAC_{i,t}$ the total accruals of company i in the tth year.
$\Delta CA_{i,t}$ the change of company i's current assets in the tth year.
$\Delta CL_{i,t}$ the change of company i's current liabilities in the tth year.
$\Delta Cash_{i,t}$ the change of company i's cash in the tth year.
$\Delta STDEBT_{i,t}$ the change of company i's long-term debt due within one year in the tth year.
$DEPTN_{i,t}$ the depreciation expense of company i in the tth year.

The cash flow statement method is shown as per Formula (4) below:

$$TAC_{i,t} = EBXI_{i,t} - CFO_{i,t}. \quad (4)$$

The variables are defined as follows:

$TAC_{i,t}$ the total accruals of company i in the tth year.

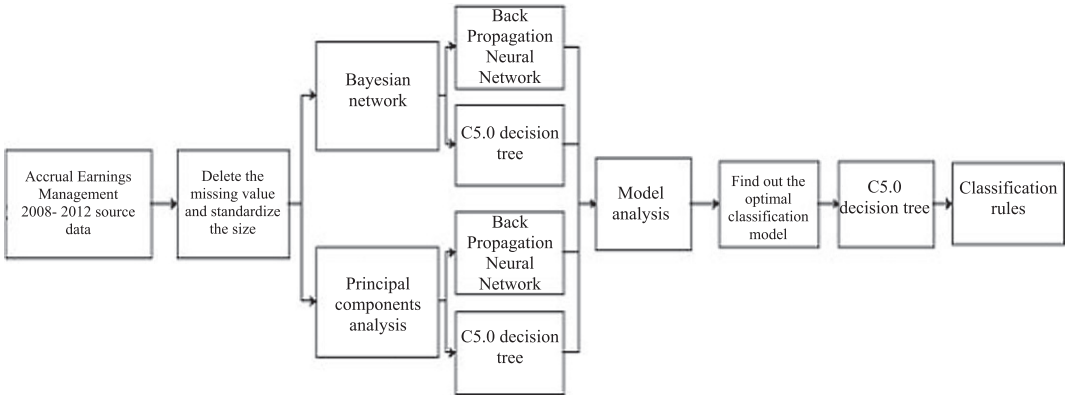

Fig. 1. Research flow chart.

$EBXI_{i,t}$ the continuing operations' income of company *i*'s in the *t*th year.
$CFO_{i,t}$ the cash flow of company *i*'s operating activities in the *t*th year.

By using the cash flow statement method, the study counted up the total accruals, followed by calculating non-discretionary accruals with the latest earnings management detection method recommended by Dechow et al. (2012), which is shown as Formula (5) below:

$$WC\_ACC_{i,t} = \left(\Delta CA_{i,t} - \Delta CL_{i,t} - \Delta Cash_{i,t} + \Delta STD_{i,t}\right)/A_{i,t-1}. \quad (5)$$

The variables are defined as follows:

$WC\_ACC_{i,t}$ the non-cash working capital accruals of company *i* in the *t*th year.
$\Delta CA_{i,t}$ the change of company *i*'s current assets in the *t*th year.
$\Delta CL_{i,t}$ the change of company *i*'s current liabilities in the *t*th year.
$\Delta Cash_{i,t}$ the change of company *i*'s cash in the *t*th year.
$\Delta STD_{i,t}$ the change of company *i*'s short-term debt in the *t*th year.
$A_{i,t-1}$ the total assets of company *i* in the $t-1$th year.

To calculate total accruals and non-cash operating capital accruals through the preceding steps, followed by deducting non-cash operating capital accruals from total accruals to acquire the proxy variable discretionary accruals of the study's earnings management. The formula is shown as Eq. (6) below:

$$DA_{i,t} = TAC_{i,t}/A_{i,t-1} - WC\_ACC_{i,t}. \quad (6)$$

The variables are defined as follows:

$DA_{i,t}$ the discretionary accruals of company *i* in the *t*th year.
$TAC_{i,t}$ the total accruals of company *i* in the *t*th year.
$WC\_ACC_{i,t}$ the non-cash working capital accruals of company *i* in the *t*th year.

3.2. Back propagation neural networks (BPN)

Neural networks refer to the information processing system simulating bio-neural networks. They use a large number of connected artificial neurons to simulate the capacity of bio-neural networks (Claveria and Torra, 2014; Özkan, 2013). Since neural networks are equipped with the functions of high-speed calculation and information noise filtering, they are capable of solving many sophisticated classification and prediction issues. The BPN adopted by the study is in a forwarding structure and a kind of supervisory learning network which is suitable to be applied to diagnosis and prediction issues, and is the most commonly used neural network.

The BPN is constituted by multilayer neurons and, as shown in Fig. 2, it has three layers of input layer, hidden layer and out layer.

The input layer is BPN's processing unit, which is used to receive variables. The hidden layer is constituted by neutrons and its major purpose is to increase complication of neural networks, so it can simulate complicated linear relations. The out layer generates the post-processing prediction results.

The basic theorem of BPN is to minimize the error function with the concept of the gradient steepest descent method. In general, the learning is processed in a way of a training at a time until completion of all the learning training examples, which is called a learning epoch. A network can be used to repetitively learn the training examples until the network learning attains the convergence effect.

The algorithm flowchart (Fig. 3) and steps of BPN are as follows:

1. select network structure and parameters including each layer's number of units, learning speed, inertia factors, calculation frequency and tolerance error, among others.
2. generate initial weighted value and threshold value with a random number.
3. standardize all the used example data according to the range of the neutron conversion function.
4. calculate output of hidden layer and out layer.
5. calculate the gap between out layer and hidden layer.
6. calculate connected weighted value modification volume and threshold value modification volume among respective layers.
7. renew the connected weighted value and threshold value among respective layers.
8. go back to step 4 and repeat the calculation until the error value is smaller than the tolerance error value, or when the calculation reaches the preset frequency, the network will stop calculation.

3.3. C5.0 decision tree

Decision tree is a technique commonly used in data mining for classification and prediction. It can classify huge data according to the division rule to find out valid data, so as to achieve the ideal results (Kim and Upneja, 2014; Marsala and Petturiti, 2015; Parvin et al., 2015). With the algorithm, the tree-like model could be set up by utilizing the induction method to give prediction analysis of discrete or continuous attributes. In order to classify the input data, each node of a decision tree is a judgment formula. The judgment formula is used to judge a variable by finding out if the input data is equal to an attribute value, so each node could

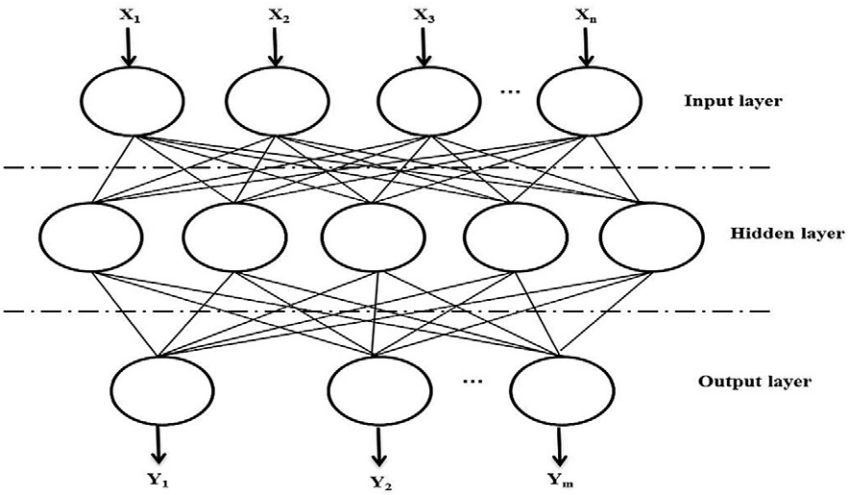

Fig. 2. Back propagation neural network structure.

divide the input data into several categories. In so doing, a tree-like structure may start taking shape.

The C5.0 adopted by the study is to deem each kind of data to be the same category before further calculating the information of a variety of attributes, followed by picking up the optimal data from respective attributes and classifying as well as comparing these data.

3.4. Selection of samples and variables

By aiming at Taiwan's biotechnology industry companies listed in the TSEC/OTC market from 2008 to 2012, the study selected its samples on the yearly basis from the database of Taiwan Economic Journal (TEJ). Out of the dependent and independent variable samples influencing accrual earnings management, those that lack any numerical values were deleted. As a result, 498 valid samples were selected. The sample selection process is shown in Table 1 below.

The study selected 25 variables which may influence earnings management from prior earnings management studies. Table 2 below shows the variables used by the study.

4. Empirical results and analysis

The study further screened the selected 25 variables in the hope of choosing the variables which could have greater influence on accrual earnings management, followed by proceeding with the second stage of model establishment with the selected variables for accuracy testing. In

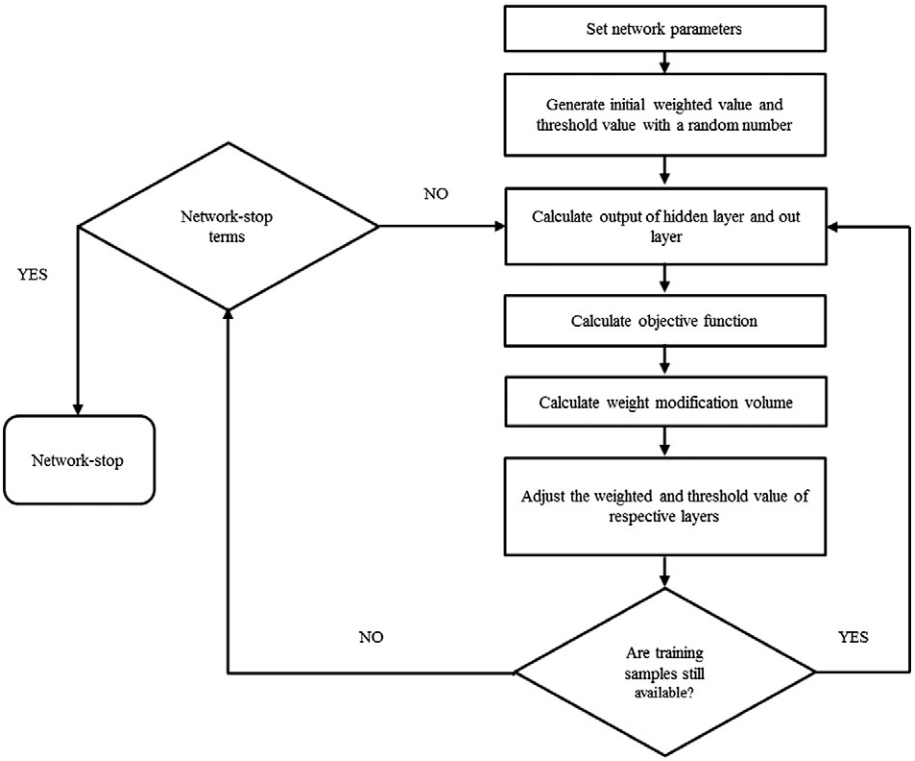

Fig. 3. Back propagation neural network algorithm flowchart.

Table 1
Sample selection process.

Sample selection process	The number of samples
2008–2012 source data	601
Samples which were left out or with incomplete data	(103)
Final number of samples	498

the end, the classification accuracy rates accrued from two screening methods pairing with two models were compared.

4.1. Variable screening

Due to numerous variables, the study used principal component analysis and Bayesian networks before establishing C5.0 decision tree and back propagation neural network models to find out important and representative variables. The screening results are described as below.

Table 2
Variables used by the study.
Source: Schipper, 1989; Healy and Wahlen, 1999; Watts and Zimmerman, 1990; Healy, 1985; Jones, 1991; Dechow et al., 1995; Kotharia et al., 2005; Höglund, 2012; Dechow et al., 2012.

Variable code	Variable name	Calculation method
X1	Director/supervisor's shareholding ratio (e.g., BOARD)	[Shares held by directors and supervisor / (outstanding shares − preferred shares issued)] * 100
X2	Corporate size (SIZE)	The natural logarithm of company's market value
X3	Debt ratio (LEV)	Total debt/stock's market value at the beginning period
X4	Does loss occur in the current year? (LOSS)	1 represents loss, while 0 represents no loss
X5	Return on assets (ROA)	(Net profit before tax and interest expense − income tax) / average total assets
X6	Operating cash flow (OCF)	Operating cash flow / total assets at the beginning period
X7	Institutional investor's shareholding ratio (ORGA)	[Shares held by institutional investors / (outstanding shares − preferred shares issued)] * 100
X8	Is the CPA one of the top 4? (BIG4)	1 represents one of the top 4; otherwise 0
X9	Return on common equity (ROE)	Net income / common equity
X10	Current ratio (Curr)	Current assets / current liabilities
X11	Debt ratio (DB)	Total liabilities / total assets
X12	Earnings per share (EPS)	(Net profit after tax − preferred stock dividend) / issued weighted average shares
X13	Net profit margin (PIS)	Net income / net sales
X14	Growth opportunity (GROWTH)	The amount of change in net sales income / prior period sales income
X15	The number of independent directors (IND)	The number
X16	Debt–equity ratio (LA)	Total debt / equity
X17	Inventory turnover (IT)	Cost of goods sold / average inventory
X18	Accounts receivable turnover (RT)	Net sales / average receivables
X19	Financial leverage (FLV)	Earnings before interest and tax / earnings before tax
X20	Acid-test ratio (QR)	(current assets − inventory − prepaid expenses) / current liabilities
X21	Prior period discretionary accruals (DA-1)	Prior period discretionary accruals
X22	Employee profitability (EN)	Net profit before tax / total number of employees
X23	Price/earnings ratio (P/E)	Stock price / earnings per share
X24	Performance threshold (Thod)	Non-discretionary accruals − prior period non-discretionary accruals
X25	Operating cash flow (CFO)	Operating cash flow

Table 3
Results of Bayesian network screening.

Variable	Bayesian network importance
X9	0.1024
X6	0.0904
X22	0.0577
X23	0.0514
X17	0.0477
X20	0.0514
X21	0.0477
X16	0.0477
X10	0.0336

4.1.1. Screening of Bayesian network (BN)

With the inference of the Bayesian theorem, Bayesian network uses the probability of applicable terms as the basis to construct a directive and non-cycled directed graph, which can give effective inference for the uncertainty of massive variables (Neapolitan, 2004; Lu and Tokinaga, 2014). Table 3 below shows the importance of the variables selected through Bayesian network, in which the importance levels are ranked to be X9, X6, X22, X23, X17, X20, X21, X16 and X10 representing return on common equity, operating cash flow, employee profitability, price/earnings ratio, inventory turnover, acid-test ratio, prior period discretionary accruals, debt–equity ratio and current ratio respectively.

4.1.2. Screening of principal component analysis (PCA)

Statistically, PCA is a data simplification technique, a linear conversion. It converts the data into a new coordinate system, so the biggest variance projected by any data can be shown on the first principal component, whereas the second biggest variance can be shown on the second principal component and so forth. Principal component analysis often reduces data dimensions while keeping the most significant characteristic of data contribution to variance. This is the way to pass retention of the low-end principal component and ignore attainment of the high-end principal component. In doing so, the low-end principal component can often keep the most important part of the data. Table 4 below shows the importance of the variables selected through PCA, in which the importance levels are ranked to be X20, X10, X3, X16, X5, X9 and X13 representing acid-test ratio, current ratio, debt ratio, debt–equity ratio, return on assets, return on common equity and net profit margin.

4.1.3. Classification of the degrees of accrual earnings management

In order to specifically sort out serious degrees of earnings management, the study used the statistical method to reasonably classify discretionary accruals. It first calculated the average value and standard deviation of total samples' discretionary accruals (DA), followed by setting the value of adding a notch of standard deviation value to the average value as the ceiling and the one of deducting a notch of standard deviation from the average value as the floor. If the value of discretionary accruals was over the ceiling value or below the floor value, it would be defined as extremely upward or downward accrual earnings management, whereas other sample observation values falling in the area between the ceiling and floor would be deemed to be minor level of accrual earnings management. By using the aforesaid method to classify the intervals, the numbers of samples and descriptive statistics, Table 5 shows that the

Table 4
Results of principal component analysis screening.

Variable	Principal component analysis importance
X20	0.934
X10	0.924
X3	0.872
X16	0.872
X5	0.871
X9	0.859
X13	0.838

Table 5
Accrual 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 < −0.12865	54	−0.16804	−0.19978
Slightly downward earnings management	−0.0254512 > DA ≥ −0.12865	198	−0.05652	−0.06357
Slightly upward earnings management	−0.0254512 < DA ≤ 0.07774	192	0.00332	0.01028
Extremely upward earnings management	DA > 0.07774	54	0.13547	0.16161
Total		498	−0.08577	−0.09146

average value of discretionary accruals calculated from the total sample observation value is −0.0254512, whereas the standard deviation is 0.103194. When the discretionary accrual value is greater than 0.07774 (the average value plus a notch of standard deviation) or smaller than −0.12865 (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. The extremely upward accrual earnings management has 54 observation values and their average value is 0.16161, whereas the extremely downward accrual earnings management has 54 observation values and their average value is −0.19978.

4.2. Model establishment

The study adopted SPSS Clementine to execute decision tree and back propagation models, with which C5.0 decision tree was first used. At the same time, 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 slight earnings management error rate, whereas type II error is the minor error in earnings management, but is classified to have a serious earnings management error rate. For the study, type I error is the error with importance. Table 6 shows the accuracy rates of BN + BPN and BN + C5.0, and Table 7 shows the accuracy rates of PCA + BPN and PCA + C5.0. As shown in Tables 6 and 7, the accuracy of either BPN or C5.0 decision tree is above 83.58% in the test group. According to Table 6, BN + C5.0 has the highest overall classification accuracy at 98.51% with the lowest type I error rate in the test group, followed by BN + BPN at 92.54% with the type I error rate at 5.97%. As shown in Table 7, PCA + BPN has the overall classification accuracy at 85.07% with the type I error rate at 14.92%, whereas PCA + C5.0's accuracy is 83.58% with the type I error rate at 16.42%.

4.3. Model evaluation and additional testing

4.3.1. Model evaluation

The study picked a best performance model from PCA and Bayesian network respectively, i.e., PCA + BPN model and BN + C5.0 model, for gain chart evaluation. Gains are defined according to the proportion of total hits that occur in each quantile; they are computed by the (number of hits in quantile / total number of hits) × 100% (Delen et al., 2013). The gain chart is to use the graph to show the improvement provided by the data mining model after comparison with random predictions and measure the "lift" score related changes. By comparing respective parts of the data set and gain scores of different models, the optimal model and the percentage of the copied model prediction benefiting cases out of the data set could be determined. With the gain chart, the prediction accuracy of several models with the same predictable attribute could be compared, while the prediction accuracy of a single result (the single value of a predictable attribute) or all the results (all values of a specified attribute) could also be evaluated. Even though the income chart also includes the same information related graph patterns as the gain chart, it could also indicate expected income increase and the income increase related to the use of each model. Simply put, the better model shall be the one whose gain curve could reach the ideal optimal accumulated gain at the earliest time before the sampling percentage becomes 100%.

Fig. 4 is the PCA + BPN accumulated gain chart, in which the blue line is the ideal optimal gain curve, the red line is the model's gain curve in the test group and the orange line is the diagonal whereas Y axle represents percentage of accumulated gain and X axle is the sampling percentage. When the sampling percentage reaches 99%, the gain is still at a level of 98.2%. Fig. 5 is the BN + C5.0 accumulated gain chart (since this model could reach the maximum gain of 100% earlier than the PCA + BPN model, the blue and red lines are mostly overlapped), in which, when the sampling is at 94%, the model will attain the maximum gain of 100%. To sum up, the gain chart can give a clearer picture about better performance of the BN + C5.0 model when evaluating the models of PCA + BPN and BN + C5.0.

4.3.2. Additional testing

For the final part of the empirical analysis, the study used the rules set coming out from the BN + C5.0 model, which has the best performance and the highest accuracy in the test group. Table 7 shows the rules set generated from C5.0, in which "1" represents the serious degree of earnings management. Fig. 6 is the decision tree chart formed from the C5.0 decision tree. As shown in Fig. 6 there are two rules for the serious degree of accrual earnings management, in which rules 1 and 2 are that it is likely to result in the serious degree of accrual earnings management when X21 (prior period discretionary accruals) is smaller or equal to −0.13 and greater than 0.76 (Table 8).

Table 6
Accuracy of BN + BPN and BN + C5.0.

BN + BPN, BN + C5.0				
	Overall accuracy rate		Overall error rate	
BPN	Training group	89.10%	Type I error	6.96%
			Type II error	3.94%
	Test group	92.54%	Type I error	5.97%
			Type II error	1.49%
C5.0	Training group	99.77%	Type I error	0.00%
			Type II error	0.23%
	Test group	98.51%	Type I error	0.00%
			Type II error	1.49%

Table 7
Accuracy of PCA + BPN and PCA + C5.0.

PCA + BPN, PCA + C5.0				
	Overall accuracy rate		Overall accuracy rate	
BPN	Training group	78.42%	Type I error	21.11%
			Type II error	0.47%
	Test group	85.07%	Type I error	14.92%
			Type II error	0.01%
C5.0	Training group	77.73%	Type I error	22.27%
			Type II error	0%
	Test group	83.58%	Type I error	16.42%
			Type II error	0%

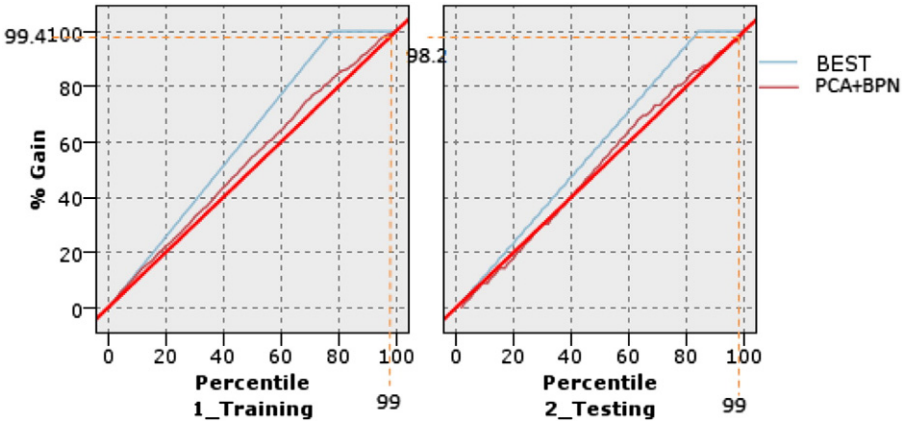

Fig. 4. PCA + BPN accumulated gain chart.

4.3.3. Discussion and findings

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

1. There were numerous predictive variables to 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 BN and PCA. Tables 3–6 prove that the BN method could effectively improve the accuracy rate of accrual earnings management prediction, regardless of the methods and variables used. The analysis presented above suggests that variable selection can enable researchers to give accrual earnings management prediction without having any special domain knowledge.
2. In summary, according to Tables 3–6, the ranking of the models in order of accuracy is as follows: BN + C5.0, BN + BPN, PCA + BPN, and PCA + C5.0. 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 be stable in terms of accuracy because it is the optimal model in all aspects of accuracy, Type I error, and predictive variables.
3. Finally, as shown in Figs. 4 and 5, compared with the other model (PCA + BPN), the proposed hybrid model (BN + C5.0) is much better in terms of accuracy and predictive variables. In addition, the proposed hybrid approach is also superior to other hybrid models.

5. Conclusion and recommendations

The study used Taiwan's biotechnology industry companies listed in the TSEC/OTC market from 2008 to 2012 as the samples to detect accrual earnings management. It first tried the principal component analysis and Bayesian network to screen the variables in the first stage, followed by further constructing and integrating the model with back propagation neural networks and C5.0 decision tree to detect if serious earnings manipulation occurred. By using back propagation neural networks and C5.0 decision tree and combining the principal component analysis and Bayesian network variable screening method, the study has explored whether an enterprise has a serious degree of accrual earnings management behavior. The empirical results show that pairing Bayesian network with C5.0 can have the best performance in exploring a serious degree of accrual earnings management, in which the test group's accuracy even tops 98.51% s and its type I error is also the lowest.

This paper detecting accrual earnings management and using a hybrid model combining PCA, BN, BPN and DT, has been developed not only to enhance classification accuracy but also elicit meaningful rules for accrual earnings management prediction. To demonstrate the proposed approach, the study used BN + C5.0, BN + BPN, PCA + BPN, and PCA + C5.0 models as benchmarks in an attempt to improve the previous model in several ways. First, the study overcame the restriction of the conventional statistic method, so the model could be more applicable to the real world. Second, the empirical results show that combining the Bayesian Network with C5.0 could best investigate the status of extreme accrual earnings management. Its accuracy rate in the test

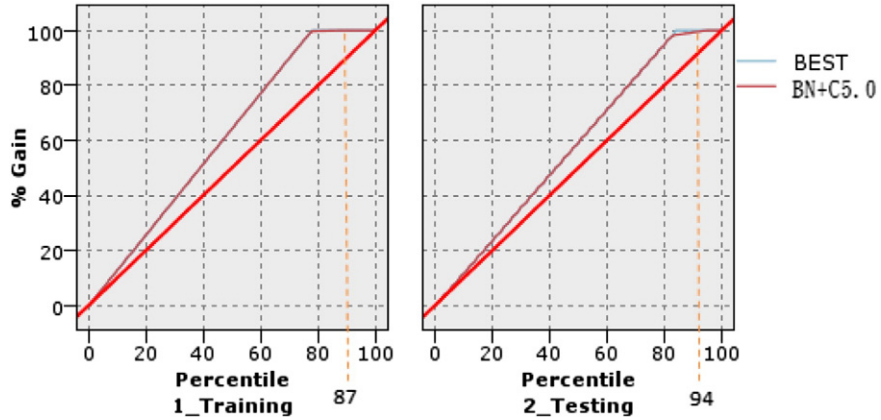

Fig. 5. BN + C5.0 accumulated gain chart.

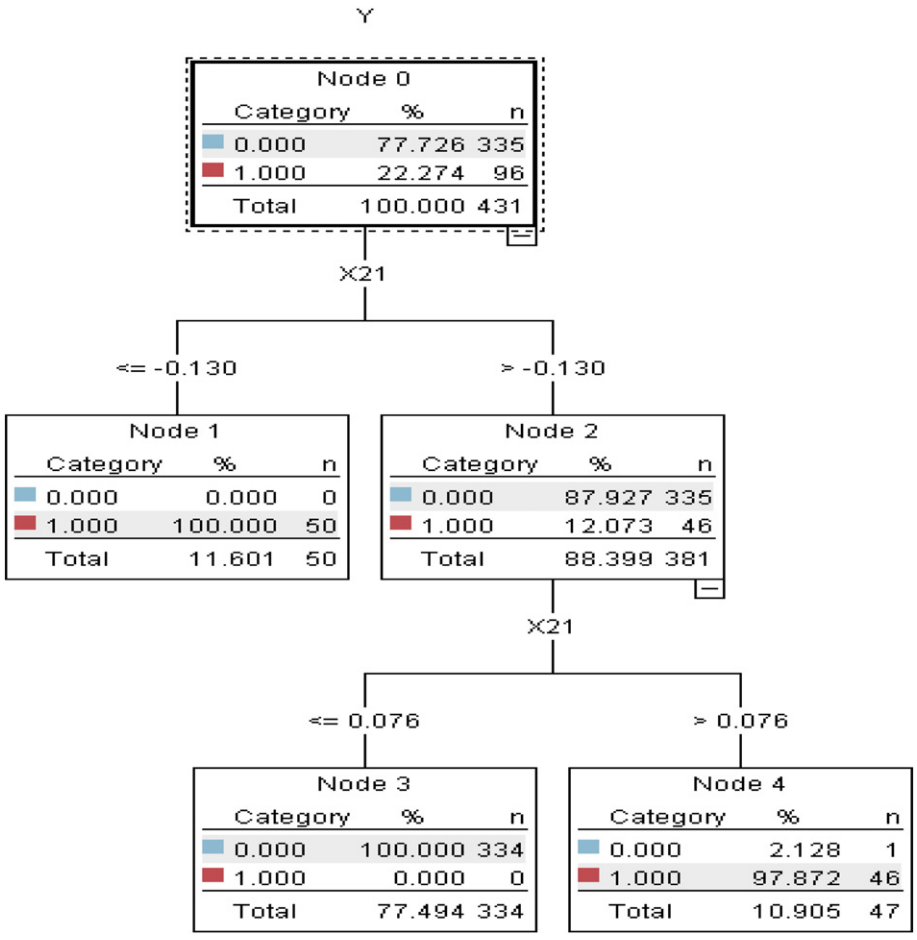

Fig. 6. C5.0 decision tree chart.

Table 8
Rules set of the BN + C5.0 model.

Rules set for the BN + C5.0 model	
Rules set in the circumstance where the dependent variable is "1" (serious degree of accrual earnings management)	
Rule 1	If X21 ≤ −0.13 then 1
Rule 2	If X21 > 0.76 then 1

group is 98.51%, and its type I error is also the lowest at 0%. Finally, for the additional testing, the study also disclosed the rules generated by C5.0 against the serious degree of accrual earnings management through tables and charts. These rules show that prior period discretionary accruals have great impact on the serious degree of accrual earnings management. Conventional financial statement auditing is often restricted by the auditing time limit, cost and human resources, so it is hard for the auditor to find out earnings manipulation from colossal and complicated financial statement data. The aforesaid decision making rules can render auditing personnel some help when limited by the auditing time and cost.

References

Claveria, O., Torra, S., 2014. Forecasting tourism demand to Catalonia: neural networks vs. time series models. Econ. Model. 36, 220–228.
Cohen, D.A., Zarowin, P., 2010. Accrual-based and real earnings management activities around seasoned equity offerings. J. Account. Econ. 50 (1), 2–19.
Cohen, D.A., Dey, A., Lys, T.Z., 2008. Real and accrual-based earnings management in the pre-and post-Sarbanes-Oxley periods. Account. Rev. 83 (3), 757–787.
Collins, D.W., Hribar, P., 2002. Errors in estimating accruals: implications for empirical research. J. Account. Res. 40 (1), 105–134.
Dechow, P.M., Sloan, R.G., Sweeney, A.M., 1995. Detecting earnings management. Account. Rev. 70 (2), 193–225.
Dechow, P.M., Sloan, R.G., Hutton, A.P., Kim, J.H., 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.
Doyle, J.T., Jennings, J., Soliman, M.T., 2013. Do managers define non-GAAP earnings to meet or beat analyst forecasts? J. Account. Econ. 56 (1), 40–56.
Feng, L., Zhang, J., 2014. Application of artificial neural networks in tendency forecasting of economic growth. Econ. Model. 40, 76–80.
Gupta, R., Modise, M.P., 2012. South African stock return predictability in the context data mining: the role of financial variables and international stock returns. Econ. Model. 29 (3), 908–916.
Hajaizadeh, E., Ardakani, H., Shahrabi, J., 2010. Application of data mining techniques in stock market: a survey. J. Econ. Int. Financ. 2 (7), 109–118.
Hall, M.A., Holmes, G., 2003. Benchmarking feature selection techniques for discrete class data mining. IEEE Trans. Data Eng. 15 (3), 1–16.
Healy, P.M., 1985. The effect of bonus schemes on accounting decisions. J. Account. Econ. 7 (1/3), 85–107.
Healy, P.M., Wahlen, J.M., 1999. A review of earnings management literature and its implications for standard setting. Account. Horiz. 13 (4), 365–384.
Höglund, H., 2012. Detecting earning management with neural networks. Expert Syst. Appl. 39 (10), 9564–9570.
Jones, J.J., 1991. Earnings management during import relief investigations. J. Account. Res. 29 (2), 193–228.
Kessel, M., Frank, F., 2007. A better prescription for drug-development financing. Nat. Biotechnol. 25 (8), 859–866.
Kim, S.Y., Upneja, A., 2014. Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Econ. Model. 36, 354–362.
Kotharia, S.P., Leoneb, A.J., Wasley, C.E., 2005. Performance matched discretionary accrual measures. J. Account. Econ. 39 (1), 163–197.
Lu, J., Tokinaga, S., 2014. Estimation of state changes in system descriptions for dynamic Bayesian networks by using a genetic procedure and particle filters. Econ. Model. 39, 138–145.
Marsala, C., Petturiti, D., 2015. Rank discrimination measures for enforcing monotonicity in decision tree induction. Inf. Sci. 291, 143–171.
McVay, S.E., 2006. Earnings management using classification shifting: an examination of core earnings and special items. Account. Rev. 81 (3), 501–531.

Neapolitan, R.E., 2004. Learning Bayesian Networks. Pearson Prentice Hall, Upper Saddle River.
Özkan, F., 2013. Comparing the forecasting performance of neural network and purchasing power parity: the case of Turkey. Econ. Model. 31, 752–758.
Parvin, H., MirnabiBaboli, M., Alinejad-Rokny, H., 2015. Proposing a classifier ensemble framework based on classifier selection and decision tree. Eng. Appl. Artif. Intell. 37, 34–42.
Pyo, G., Lee, H.Y., 2013. The association between corporate social responsibility activities and earnings quality: evidence from donations and voluntary issuance of CSR reports. J. Appl. Bus. Res. 29 (3), 945–962.
Roychowdhury, S., 2006. Earnings management through real activities manipulation. J. Account. Econ. 42 (3), 335–370.
Schipper, K., 1989. Commentary on earning management. Account. Horiz. 3, 91–102.
Wang, J.Z., Wang, J.J., Zhang, Z.G., Guo, S.P., 2011. Forecasting stock indices with back propagation neural network. Expert Syst. Appl. 38 (11), 14346–14355.
Watts, R., Zimmerman, J., 1986. Positive Accounting Theory. Prentice-Hall Inc.
Watts, R.L., Zimmerman, J.L., 1990. Positive accounting theory: a ten year perspective. Account. Rev. 65, 131–156.
Zhang, Y., Wu, L., 2009. Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. Expert Syst. Appl. 36 (5), 8849–8854.