Research Article

A Hybrid Approach of Stepwise Regression, Logistic Regression, Support Vector Machine, and Decision Tree for Forecasting Fraudulent Financial Statements

Suduan Chen,¹ Yeong-Jia James Goo,² and Zone-De Shen²

¹ Department of Accounting Information, National Taipei University of Business, 321 Jinan Road, Section 1, Taipei 10051, Taiwan

² Department of Business Administration, National Taipei University, No. 67, Section 3, Ming-shen East Road, Taipei 10478, Taiwan

Correspondence should be addressed to Suduan Chen; sduanchen@yahoo.com.tw

Received 13 May 2014; Revised 22 August 2014; Accepted 23 August 2014; Published 11 September 2014

Academic Editor: Shifei Ding

Copyright © 2014 Suduan Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As the fraudulent financial statement of an enterprise is increasingly serious with each passing day, establishing a valid forecasting fraudulent financial statement model of an enterprise has become an important question for academic research and financial practice. After screening the important variables using the stepwise regression, the study also matches the logistic regression, support vector machine, and decision tree to construct the classification models to make a comparison. The study adopts financial and nonfinancial variables to assist in establishment of the forecasting fraudulent financial statement model. Research objects are the companies to which the fraudulent and nonfraudulent financial statement happened between years 1998 to 2012. The findings are that financial and nonfinancial information are effectively used to distinguish the fraudulent financial statement, and decision tree C5.0 has the best classification effect 85.71%.

1. Introduction

The financial statement is the main basis of decision-making by investors, creditors, and other accounting information demanders and concurrently also the concrete expression of management performance, financial condition, and possessing social responsibility of the listed and OTC companies, but the fraudulent financial statement (FFS) has the trend of becoming increasingly serious in recent years [1–8].

This behavior not only makes the investing public subject to vast amount of loss but also, more seriously, influences the capital market order. Because the fraudulent case is increasingly serious with each passing day, the United States Congress passed Sarbanes-Oxley Act in 2002 and mainly hope by which to improve the accuracy and reliability of the financial statement of a company and disclosure to make the auditors able to forecast the omen of the FFS before the FFS of an enterprise occurs. When one checks corporations' financial statements due to fraud which led to a significant

misstatement, there are fairly strict norms for audit staff in Taiwan [9].

The FFS can be regarded as a typical classification problem [10]. The classification problem carries out a computation mainly in light of the variable attribute numerical value of some given classification data to acquire the relevant classification rule of every classification and bring the unknown classification data into the rule to acquire the final classification result. Many authors apply the logistic regression to make a fraudulent classification and acquire the result in the FFS issue in the past [3, 6, 7, 11–13].

Data mining is an analytical tool used to handle a complicated data analysis. It discovers previously unknown information from mass data and aims for data to make an induction from the structured model as reference amount in making a decision with many different functions, such as classification, association, clustering, and forecasting [4, 5, 8, 14]. "Classification" function is used the most often therein, and its result can serve as the decision basis and prediction. However, whether every application of data mining in the

FFS is superior to the traditional classification model is controversial.

The purpose of this study is to expect that a better method of forecasting fraudulent financial statement can be presented to forecast the omen of the fraudulent financial statement and to reduce damage to the investors and auditors. The study will adopt the logistic regression and the support vector machine (SVM) as well as the decision tree (DT) C5.0 in data mining as the basis and match the stepwise regression to separately establish classification model to make a comparison. In conclusion, the study first aims at the "fraudulent financial statement" issue to make an arrangement for and carry out an exploration of relevant literature to ensure the research variable and sample adopted by the study. We then take the logistic regression, SVM, and DT C5.0 as the bases to establish the FFS classification model. Finally, we present the conclusions and suggestions of the study.

2. Literature Review

2.1. Fraudulent Definition. The FFS is a kind of intentional or illegal behavior, the result of which directly causes the seriously misleading financial statement or financial disclosure [2, 15]. Pursuant to the provision of SAS NO.99, a kind of fraudulent pattern is dishonest financial report, and it means a kind of intentional erroneous narration, neglecting amount or disclosure, which makes the misunderstood financial statement [6].

2.2. Research Method. The classification problem carries out a computation mainly in light of the variable attribute numerical value of some given classification data to acquire the relevant classification rule of every classification and bring the unknown classification data into the rule to acquire the final classification result. Many authors apply the logistic regression to make a fraudulent classification in the FFS issue in the past [3, 11, 12, 15–17]. However, the traditional statistic method has limitation of having to accord with specific assumption in data.

As a result, the machine learning way which does not require any statistic assumption about data portfolio rises abruptly. Many scholars recently try to adopt the machine learning way as the classification machine to conduct a research. The empirical result also points out that it possesses an excellent classification effect. Chen et al. [13] applied the neural network and SVM to forecast network invasion, and the research result indicates that the SVM has excellent classification ability. Huang et al. [18] applied the neural network and SVM to explore the classification model of credit evaluation. Shin et al. [19] conducted a relevant research of bankruptcy prediction. Yeh et al. [4] apply it in prediction of enterprise failure. On the other hand, Kotsiantis et al. [3] and Kirkos et al. [10] apply DT C5.0 in the relevant research to acquire the excellent classification result. Thus, the study will adopt the foresaid logistic regression, SVM, and DT C5.0 as the classifier construction classification model.

2.3. Variable Selection. As for variable selection via relevant literature exploration, some authors adopt the financial variable as the research variable [3, 10], others adopt the nonfinancial variable as the research variable [12, 16, 17], and still others adopt both the financial variable and nonfinancial variable as the research variable [15, 20].

Because financial statement data often have cheating suspicion, if we purely consider the financial variables, the possibility of erroneous classification may increase. Therefore, the study not only adopts the financial variable as the research variable, but also adds the nonfinancial variable to construct the fraudulent financial prediction model.

3. Methodology

The purpose of this study is to present a two-stage research model which integrates the financial variable and nonfinancial variable to establish the fraudulent early warning model of an enterprise. The procedure of the study is to aim at the data to make a stepwise regression analysis, to acquire the result of the important variable of the TTF after screening, and then to take such variable as the input variable of the logistic regression and SVM. Finally, the study makes a comparison and an analysis to acquire a better FFS classification result.

3.1. Stepwise Regression. The study selects a variable of the maximum classification ability in accordance with forward selection and incorporates the predictor into the model by stepwise increase. During each process, P value of the statistic test is used to screen the variables. If P value is less than or equal to 0.05, then the variable enters the regression model, and the selected variable is the independent variable of the regression model.

3.2. Logistic Regression. The logistic regression resembles the linear regression, while the response variable and explanatory variable of the general linear regression are usually the continuous variable, but the response variable explored by the logistic regression is the discrete variable; that is, it handles the qualitative variable of the two-dimensional independent variable problem (e.g., yes or no and success or failure). The model utilizes cumulative probability density function to convert real number value of the explanatory variable to probability value between 0 and 1. The elementary assumption is different from the analytic assumption of another multivariate analysis. The influence of the explanatory variable on the response variable is to fluctuate in the index form, which means that the logistic regression does not need to conform to the normal distribution assumption. In other words, it can handle the population of the nonnormal distribution and the problem of the nonlinear model and the nonmeasuring variable.

The general logistic regression model is as follows:

$$Y^* = \beta x + \varepsilon$$

$$Y = \begin{cases} 1: Y^* > 0 \\ 0: Y^* \leq 0 \end{cases}, \quad (1)$$

where Y : response variable of actual observation, $Y = 1$: a financial crisis event occurs, $Y = 0$: no financial crisis event occurs, Y^* : latent variable without observation, x : matrix of explanatory variable, β : matrix of explanatory variable parameter, and ε : error of explanatory variable.

3.3. Support Vector Machine (SVM). The operation model of the SVM projects the initial input vector to eigenspace of the high dimension with linear and nonlinear core function and utilizes the separating hyperplane to distinguish two or many materials of different classes. The SVM utilizes the hyperplane classifier to classify the materials.

3.3.1. Linear Divisibility. When the plain formed by the training sample data is linear, which consider the training vector: $x_i = (x_i^{(1)}, \dots, x_i^{(n)}) \in R^n$ belongs to two classes $y_i \in \{-1, +1\}$. In order to definitely distinguish the training vector class, it is necessary to find out the optimal partition hyperplane able to separate the materials.

If the hyperplane $w \cdot x + b$ can separate the training sample, it is shown as

$$w \cdot x_i + b > 0, \quad \text{if } y_i = 1, \quad (2)$$

$$w \cdot x_i + b < 0, \quad \text{if } y_i = -1. \quad (3)$$

Adjust w and b properly; (2) and (3) can be rewritten as

$$\begin{aligned} w \cdot x_i + b &\geq 1, & \text{if } y_i = 1, \\ w \cdot x_i + b &\leq -1, & \text{if } y_i = -1. \end{aligned} \quad (4)$$

or as

$$y_i (w \cdot x_i + b) \geq 1, \quad \forall i \in \{1, \dots, n\}. \quad (5)$$

Pursuant to the statistics theory, the best interface not only separates two classes of samples correctly, but also maximizes the classification margin. The class margin of the interface $w \cdot x + b$ is shown as

$$d(w, b) = \min_{\{x_i | y_i = 1\}} \frac{w \cdot x_i + b}{|w|} - \max_{\{x_i | y_i = -1\}} \frac{w \cdot x_i + b}{|w|}. \quad (6)$$

Equation (7) can be acquired from (4):

$$d(w, b) = \frac{1}{|w|} - \frac{-1}{|w|} = \frac{2}{|w|}. \quad (7)$$

So the problem of the maximization class margin $d(w, b)$ transforms to minimization $|w|^2/2$ under constraint condition (5). Pursuant to Lagrange relaxation, the foresaid problem must accord with the hypothesis of (8) and (9). In the foresaid condition, the minimization is shown as (10):

$$\alpha_i \geq 0, \quad (8)$$

$$\sum_i \alpha_i y_i = 0, \quad (9)$$

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j, \quad i = 1, \dots, n. \quad (10)$$

Every α_i corresponds to a training sample x_i , and the training sample of its corresponding $\alpha_i > 0$ is called the support vector. Classification function acquired finally is shown as

$$\begin{aligned} f(x) &= \text{sgn}(w \cdot x_i + b) \\ &= \text{sgn} \left(\sum_{i=1}^{N_s} \alpha_i y_i x_i \cdot x + b \right), \end{aligned} \quad (11)$$

where N_s is the number of the support vector.

3.3.2. Linear Indivisibility. If the training sample is linearly indivisible, (4) can be rewritten as

$$\begin{aligned} w \cdot x_i + b &\geq 1 - \xi_i, & \text{if } y_i = 1 \\ w \cdot x_i + b &\geq \xi_i - 1, & \text{if } y_i = -1. \end{aligned} \quad (12)$$

where $\xi_i \geq 0, i = 1, \dots, n$.

If x_i is classified mistakenly, then $\xi_i > 1$. Thus, the mistaken classification is less than $\sum_i \xi_i$. Add a given parameter value in the objective function. Consider reasonably the maximum class margin and the minimum mistaken class sample; that is, seeking the minimum of $|w|^2/2 + C(\sum_i \xi_i)$ can acquire the SVM under linear indivisibility. Pursuant to Lagrange relaxation, the foresaid problem must accord with the hypothesis of (13) and (14). In the foresaid condition, the minimization is shown as (15):

$$0 \leq \alpha_i \leq C, \quad (13)$$

$$\sum_i \alpha_i y_i = 0, \quad (14)$$

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j, \quad i = 1, 2, 3, \dots, n. \quad (15)$$

3.4. Decision Tree (DT). The Decision Tree (DT) is the simplest in the inductive learning method [21]. It belongs to the data mining tool and can handle the continuous and noncontinuous variable. It establishes the tree structure diagram mainly by the given classification fact and induces some principles therein. The principles are mutually exclusive, and the DT generated can also make an out-of-sample prediction. The DT algorithms used most frequently include CART, CHAID, and C5.0 [22]. C5.0 [23] improves from ID3 [23]. Thanks to ID3 use limitation, it cannot handle the continuous numerical value materials; thus, Quinlan conducts a research for improvement, and C5.0 is developed to handle the continuous and the noncontinuous numerical value.

The DT C5.0 is mainly separated into two parts. The first part is classification criterion, which is calculated pursuant to the gain ratio. Construct the DT completely as shown in (2). Information gained in (16) is used to calculate the pretest and posttest gain of the data set and is defined as "pretest information" minus "postinformation" from (17). The entropy in (16) is used to calculate impurity, which is called randomness. In other words, it is used to calculate

TABLE 1: Results of stepwise regression variable screening.

Variable code	Variable classification	Variable description	Pr > ChiSq
X1	Financial	Accounts receivables/total assets	0.2401
X3	Financial	Inventory/current assets	0.0339
X10	Financial	Interest protection multiples	0.0694
X13	Financial	Debt ratio	0.0294
X15	Financial	Cash flow ratio	0.0025
X17	Financial	Accounts payable turnover	0.0295
X24	Financial	Operation profit/last year operation profit >1.1	0.0267
X29	Nonfinancial	Pledge ratio of shares of the directors and supervisors	0.0473

TABLE 2: Hit ratio of three models using the train datasets.

Research model	C5.0	Logistic	SVM
Hit ratio	93.94%	83.33%	78.79%

randomness of the data set. When randomness in the data set reaches the most disorderly state, the value will be 1.

Therefore, the less random the posttest data set is, the larger the information gain is calculated, and the more favorable it is for DT construction:

$$\text{Gain Ratio}(S, A) = \frac{\text{Information } n \text{ Gain}(S, A)}{\text{Entropy}(S, A)} \quad (16)$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v). \quad (17)$$

The second part is pruning criterion. Pursuant to the error based pruning (EBP), the DT is properly pruned to enhance the correct ratio of classification. EBP is evolved from the pessimistic error pruning (PEP), and such two pruning methods are presented by Quinlan. The main concept of the EBP is to make a judgment using the error ratio, calculate the error ratio of every node, and further judge the node which results in rise of the error ratio of the overall DT. Finally, this node is pruned properly to further enhance the correct ratio of the DT.

3.5. Definition of Type I Error and Type II Error. In order to establish the valid forecasting fraudulent financial statement, it is considerably important to measure type I type II errors of the study. Type I error is to mistakenly judge the normal financial statement company as the FFS company. This judgment does not cause investors' damage, but it carries out an erroneous audit opinion for being too conservative and further influences credit of the company audited. Type II error is that the FFS enterprise is mistaken for the normal enterprise. This classification error leads to auditing failure, auditors' investment loss, or investors' erroneous judgment.

4. Empirical Analysis

4.1. Data Collection and Variables. The research samples are the FFS enterprises from the years 1998 to 2012. 66 enterprises are selected from the listed and OTC companies of the Taiwan Economic Journal Data Bank (TEJ). The 1 by 1 pair way is adopted to match 66 normal enterprises, so there are 132 enterprises in total as research samples.

As for selection of the research variables, the study altogether selects 29 variables, including 24 financial variables and 5 nonfinancial variables (see appendix).

For consideration of the number of samples, to avoid having too few samples of the test group and to improve test accuracy, we propose to utilize 50% of the sample materials as the train sample to establish the regression classification model. The remaining 50% of the sample materials serve as the test sample to test validity of the classification model established.

In addition, to test the stability of the proposed research model, this study randomly selects three groups at a ratio of 80% from the test data as the test sample for cross-validation. The compartment and sampling of data in this research are shown in Figure 1.

4.2. Model Development. To begin with, the study aims for the financial and nonfinancial variable to screen using the stepwise regression screening method. The variables screened serve as the input variable of the logistic regression and SVM. Next, the study aims at every method to carry out the model training and test. Finally, the study compares the merit and demerit of the classification correct ratio and gives the relevant suggestions for the analytic result. The model construction is divided into three parts. The first part is the variable screening way; the second part is the classification way; the third part compares the test results of two kinds of classification models. The research process of the study is shown as Figure 2.

4.3. Important Variable Screening. While constructing the classification model, there may be quite many variables, but not every variable is important. Therefore, the variables of no account need to be eliminated to construct a simpler classification model. There are quite many variable screening

TABLE 3: C5.0 cross-validation results.

C5.0 model			Predict value		Hit ratio	Type I error	Type II error
			Non-FFS	FFS			
Actual value	CV1	Non-FFS	25	3	83.93%	10.71%	21.42%
		FFS	6	22			
	CV2	Non-FFS	25	3	87.50%	10.71%	14.28%
		FFS	4	24			
	CV3	Non-FFS	25	3	85.71%	10.71%	17.85%
		FFS	5	23			
Average			25	3	85.71%	10.71%	17.85%
			5	23			

TABLE 4: Logistic regression cross-validation results.

Logistic regression model			Predict value		Hit ratio	Type I error	Type II error
			Non-FFS	FFS			
Actual value	CV1	Non-FFS	25	3	80.36%	10.71%	28.57%
		FFS	8	20			
	CV2	Non-FFS	26	2	82.14%	7.14%	28.57%
		FFS	8	20			
	CV3	Non-FFS	25	3	80.36%	10.71%	28.57%
		FFS	8	20			
Average	Non-FFS	25	3	80.95%	9.52%	28.57%	
	FFS	8	20				

ways, among which the stepwise regression variable screening method is used most frequently [24].

Therefore, the study adopts the suggestions of Pudil et al. [24] to screen the variables using the stepwise regression by which to retain the research variables with more influence. The input variables of the study are screened via the stepwise regression to acquire the results as shown in Table 1, including 7 financial variables and 1 nonfinancial variable. Subsequently, the study takes these 8 variables as the new input variables to construct the classification model.

4.4. Classification Model. The prediction accuracy of the three types of models using the train datasets is displayed in Table 2.

As shown in Table 2, C5.0 has the best performance in the establishment of the prediction model and its accuracy rate is 93.94%. The traditional logistic model is the second best. The accuracy rate of the SVM model, at 78.79%, is the lowest of the three. The cross-validation results of the proposed three prediction models are shown in Tables 3 to 5.

4.4.1. Decision Tree (DT). The study constructs the DT C5.0 model, sets EBP at $\alpha = 5\%$, and adopt the binary partition principle to obtain the optimal spanning tree. The prediction results of the DT C5.0 classification model are shown as Table 3.

On average, 25 of the 28 non-FFS materials are correctly classified in the non-FFS, and three of them are incorrectly classified in the FFS. The type I error is 10.71%. On the other hand, 23 of the 28 FFS materials are correctly classified, and

the remaining five FFS materials are incorrectly classified in the non-FFS. The type II error is 17.85%.

4.4.2. Logistic Regression. Table 4 is the empirical results of the logistic classification model, which shows that 25 of 28 non-FFS materials are correctly classified and that three of them are incorrectly classified in the FFS. The overall type I error is 9.52%. In addition, 20 of the 28 FFS materials are correctly classified, and the remaining eight FFS materials are incorrectly classified in the non-FFS. The type II error is 28.57%.

4.4.3. Support Vector Machine (SVM). The operation core is set at RBF when the study constructs the SVM model. As for the parameter, the C search scope is set at 2^{-10} to 2^{10} , and γ is set at 0.1. The SVM classification results are shown as Table 5.

In this part, 26 of the 28 non-FFS materials are correctly classified, and two of them are incorrectly classified in the FFS. The type I error is 7.14%. In addition, 14 of the 28 FFS materials are correctly classified, and the remaining 14 FFS materials are incorrectly classified in the non-FFS. The type II error is 48.81%.

4.4.4. Comprehensive Comparison and Analysis. Kirkos et al. [10] pointed out that the merit and demerit of the evaluation model must also consider the type I error and type II error. The type I error means to classify the nonfraudulent companies into the fraudulent companies. Occurrence of these two type errors results from the auditing failure of the auditors. Type II error means that the auditors classify

TABLE 5: SVM cross-validation results.

SVM model			Predict value		Hit ratio	Type I error	Type II error
			Non-FFS	FFS			
Actual value	CV1	Non-FFS	26	2	73.21%	7.14%	46.42%
		FFS	13	15			
	CV2	Non-FFS	26	2	71.43%	7.14%	50.00%
		FFS	14	14			
	CV3	Non-FFS	26	2	71.43%	7.14%	50.00%
		FFS	14	14			
Average	Non-FFS	26	2	72.02%	7.14%	48.81%	
	FFS	14	14				

TABLE 6: Summary of classification results.

Model	Type I error	Type II error	Hit ratio	Ranking
Logistic regression	9.52%	28.57%	80.95%	2
SVM	7.14%	48.81%	72.02%	3
DT C5.0	10.71%	17.85%	85.71%	1

TABLE 7: Paired-samples *t* test.

Model	<i>t</i> -value	DF	Significant (two-tailed)
C5.0—logistic	−5.201	2	0.35
Logistic—SVM	−16.958	2	0.03
SVM—C5.0	9.823	2	0.10

the fraudulent companies into the nonfraudulent companies. Both types of error would cause different loss costs, and the auditors must avoid occurrence of these two errors. Comparing the results of these three models, we conclude that the classification ability of the DT C5.0 is the best, the next is the logistic regression, and the last is the SVM. The classification correct ratios of three kinds of model are summarized as shown in Table 6.

The comparison shows that, although the logistic classification model performs the best for type I errors, the DT C5.0 possesses the best classification effect, both for type II errors and the hit ratio. The correct classification ratio is 85.71%, followed by 80.95% for the logistic model, and 72.02% for the SVM model.

Unlike general studies using type I errors to judge the performance of prediction models, FFS studies use type II errors to determine the performance of prediction models. For the sake of prudence, we conduct the statistical test of type II errors in the abovementioned cross-validation results to confirm whether the differences in between models are significantly other than 0. The analysis results are shown in Table 7, which shows that the *t*-values of the prediction model type II error differences are −5.201 (C5.0—Logistic); −16.958 (Logistic—SVM); and 9.823 (SVM—C5.0), respectively, and all of them reach the significance level.

5. Conclusion and Suggestion

As the fraudulent financial statement (FFS) increases on the trot in recent years, the auditing failure risk of the auditors also rises thereby. Therefore, many researches focus on developing a good classification model to reduce the relevant risk. In the past, the accuracy of forecasting FFS purely using regression analysis has been relatively low. Many scholars have pointed out that prediction by data mining can improve the accuracy rate. Thus, this study adopts stepwise regression to screen the important factors of financial and nonfinancial variables. Meanwhile, it combines the above with data mining techniques to establish a more accurate FFS forecast model.

A total of eight critical variables are screened via the stepwise regression analysis, including two parts: financial variables (accounts receivables/total assets, inventory/current assets, interest protection multiples, cash flow ratio, accounts payable turnover, operation profit/last year operation profit > 1.1) and nonfinancial variables (pledge ratio of shares of the directors and supervisors).

The financial variables include operating capabilities, profitability index, debt solvency ability index, and financial structure. The nonfinancial variables include relevant variables of stock rights and scale of an enterprise's directors and supervisors. The results indicate that when auditors investigate FFS, they must focus on the alert provided by the nonfinancial information as well as the financial information.

In the classification model, the study adopts the logistic regression of the traditional classification method and the DT C5.0 and SVM of data mining to construct the classification model. The empirical result indicates that the SVM model performs the best in the type I error after comparison, and the DT C5.0 has the best classification performance in the type II error and overall classification correct ratio.

TABLE 8: Selection of the research variables.

Variable classification	Variable code	Variable description and computation
Financial variables	X1	Accounts receivables/total assets
	X2	Gross profit/total assets
	X3	Inventory/current assets
	X4	Inventory/total assets
	X5	Net profit after tax/total assets
	X6	Net profit after tax/fixed assets
	X7	Cash/total assets
	X8	Log total assets
	X9	Log total liabilities
	X10	Interest protection multiples (debt service coverage ratio, times interest earned)
	X11	Gross profit margin
	X12	Operating expense ratio
	X13	Debt ratio
	X14	Inventory turnover
	X15	Cash flow ratio
	X16	Net profit ratio before tax
	X17	Accounts payable turnover
	X18	Revenue growth rate
	X19	Debt/equity ratio
	X20	Earnings before interest, taxes, depreciation, and amortization
	X21	Current liabilities/total assets
	X22	Total assets turnover
	X23	Account receivable/last year accounts receivable >1.1
	X24	Operation profit/last year operation profit >1.1
Nonfinancial variables	X25	Shareholding ratio of the major shareholders
	X26	Shareholding ratio of directors and supervisors
	X27	Whether the chairman concurrently holds the position of CEO
	X28	Board size
	X29	Pledge ratio of shares of the directors and supervisors

FIGURE 1: Train and test subsets design.

FIGURE 2: Research model.

One of the research purposes is to anticipate accommodating the auditors with another assistant auditing tool besides the traditional analysis method, but the research about the forecasting FFS is not sufficient. Therefore, the subsequent researchers can also adopt other methods to forecast the FFS to provide a better reference. In addition, future researchers can also try to adopt different variable screening methods to enhance the classification correct ratio of the method. As for the variable, some nonfinancial variables are difficult to measure, and material acquisition is difficult, so the study does not incorporate them. Finally, as for the sample, the study focuses on the FFS scope research, and a certain number of the FFSs may not be found. Therefore, the pair companies can also be the FFS companies in the coming year, which can influence the accuracy of the study. The findings of this study can provide a reference to auditors, certified public accountants (CPAs), securities analysts, company managers, and future academic studies.

Appendix

See Table 8.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

- [1] C. Spathis, M. Doumpos, and C. Zopounidis, "Detecting falsified financial statements: a comparative study using multi-criteria analysis and multivariate statistical techniques," *The European Accounting Review*, vol. 11, pp. 509–535, 2002.
- [2] Z. Rezaee, "Causes, consequences, and deterrence of financial statement fraud," *Critical Perspectives on Accounting*, vol. 16, no. 3, pp. 277–298, 2005.
- [3] S. Kotsiantis, E. Koumanakos, D. Tzelepis, and V. Tampakas, "Forecasting fraudulent financial statements using data mining," *Transactions on Engineering Computing and Technology*, vol. 12, pp. 283–288, 2006.
- [4] C.-C. Yeh, D.-J. Chi, and M.-F. Hsu, "A hybrid approach of DEA, rough set and support vector machines for business failure prediction," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1535–1541, 2010.
- [5] W. Zhou and G. Kapoor, "Detecting evolutionary financial statement fraud," *Decision Support Systems*, vol. 50, no. 3, pp. 570–575, 2011.
- [6] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decision Support Systems*, vol. 50, no. 3, pp. 585–594, 2011.
- [7] K. A. Kamarudin, W. A. W. Ismail, and W. A. H. W. Mustapha, "Aggressive financial reporting and corporate fraud," *Procedia-Social Behavioral Sciences*, vol. 65, pp. 638–643, 2012.
- [8] P.-F. Pai, M.-F. Hsu, and M.-C. Wang, "A support vector machine-based model for detecting top management fraud," *Knowledge-Based Systems*, vol. 24, no. 2, pp. 314–321, 2011.
- [9] Accounting Research and Development Foundation, *Audit the Financial Statements of the Considerations for Fraud*, Accounting Research and Development Foundation, Taipei, Taiwan, 2013.
- [10] S. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," *Expert Systems with Applications*, vol. 32, no. 4, pp. 995–1003, 2007.
- [11] T. B. Bell and J. V. Carcello, "A decision aid for assessing the likelihood of fraudulent financial reporting," *Auditing*, vol. 19, pp. 169–178, 2000.
- [12] V. D. Sharma, "Board of director characteristics, institutional ownership, and fraud: evidence from Australia," *Auditing*, vol. 23, no. 2, pp. 105–117, 2004.
- [13] W.-H. Chen, S.-H. Hsu, and H.-P. Shen, "Application of SVM and ANN for intrusion detection," *Computers and Operations Research*, vol. 32, no. 10, pp. 2617–2634, 2005.
- [14] J. W. Seifert, "Data mining and the search for security: challenges for connecting the dots and databases," *Government Information Quarterly*, vol. 21, no. 4, pp. 461–480, 2004.
- [15] M. S. Beasley, "An empirical analysis of the relation between the board of director composition and financial statement fraud," *Accounting Review*, vol. 71, no. 4, pp. 443–465, 1996.
- [16] P. Dunn, "The impact of insider power on fraudulent financial reporting," *Journal of Management*, vol. 30, no. 3, pp. 397–412, 2004.
- [17] G. Chen, "Positive research on the financial statement fraud factors of listed companies in China," *Journal of Modern Accounting and Auditing*, vol. 2, pp. 25–34, 2006.
- [18] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study," *Decision Support Systems*, vol. 37, no. 4, pp. 543–558, 2004.

- [19] K.-S. Shin, T. S. Lee, and H.-J. Kim, "An application of support vector machines in bankruptcy prediction model," *Expert Systems with Applications*, vol. 28, no. 1, pp. 127–135, 2005.
- [20] S. L. Summers and J. T. Sweeney, "Fraudulently misstated financial statements and insider trading: an empirical analysis," *The Accounting Review*, vol. 73, no. 1, pp. 131–146, 1998.
- [21] G. Arminger, D. Enache, and T. Bonne, "Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feedforward networks," *Computational Statistics*, vol. 12, no. 2, pp. 293–310, 1997.
- [22] S. Viaene, G. Dedene, and R. A. Derrig, "Auto claim fraud detection using Bayesian learning neural networks," *Expert Systems with Applications*, vol. 29, no. 3, pp. 653–666, 2005.
- [23] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.
- [24] P. Pudil, K. Fuka, K. Beránek, and P. Dvůrák, "Potential of artificial intelligence based feature selection methods in regression models," in *Proceedings of the IEEE 3rd International Conference on Computational Intelligence and Multimedia Application*, pp. 159–163, 1999.