ICIC Express Letters
Part B: Applications
Volume **7**, Number **3**, March **2016**

ICIC International ©2016 ISSN 2185-2766
pp. 591–597

CONSTRUCTION OF FINANCIAL DISTRESS PREDICTION MODELS USING STEPWISE REGRESSION AND DATA MINING TECHNIQUES

Yung-Ming Hsieh

Department of Accounting
Soochow University
No. 56, Kuei-Yang Street, Section 1, Taipei 10048, Taiwan
armin@scu.edu.tw

Received August 2015; accepted October 2015

Abstract. *This study combines stepwise regression (SR) and data mining techniques to construct more effective models for predicting financial distress. The stepwise regression and data mining techniques, including the artificial neural network (ANN) combined with the Bayesian belief network (BBN), decision tree, and support vector machine (SVM), are applied to establishing a two-stage financial distress prediction model with relatively higher accuracy. The research variables include both financial variables and non-financial variables. The empirical results of this study show that the accuracy rate of the financial distress prediction model (ANN + SVM model) which applied the ANN variable selection at the first stage and the SVM at the second stage can be up to 86.30% and the overall accuracy (financial distress + non-financial distress) can be up to 85.41%.*

Keywords: Data mining, Financial distress prediction model, Stepwise regression, Artificial neural network, Bayesian belief network, Support vector machine, Decision tree

1. **Introduction.** Since 1997, economies around the world have been affected in one way or another by the financial crisis. The 1997-1998 Asian Financial Crisis severely affected the economies of many Asian countries. The crisis also had a profoundly negative impact on the finances of the international companies in these countries. In 2001, companies such as Enron and WorldCom as well as other well-known enterprises experienced financial distress and financial statement fraud. In 2004, financial statement fraud and tunneling were also experienced by Taiwan listed companies, e.g., BDCOM and Infodisc Technology. The 2008 subprime crisis in the United States also triggered the 2008 financial tsunami, with an ensuing credit crisis, financial meltdown and global economic downturn.

If auditors, accountants, and senior management of these companies could have detected risks in advance, have been alerted to early warning signs or asked the right questions, perhaps remedial measures could have been taken sooner. It would then have been possible to pre-empt the crisis, prevent the huge losses that occurred and implement good financial risk management. It seems to be self-evident that the effective prediction model of financial distress would play an increasingly important role in the financial world. At the very least, such a model could effectively help business continuity in operations.

Since the 1980s, scholars have been applying artificial intelligence (AI) in various fields, one of which is to determine and predict financial distress. In recent years, many scholars have been applying data mining techniques to predict financial distress with the aim of dramatically increasing prediction accuracy. The aim of this study is to combine the stepwise regression and a number of data mining techniques to establish a more accurate model for financial distress prediction.

The reference [1] clarifies that an enterprise could be defined as having failed under the following conditions: enterprise bankruptcy, corporate debt defaults, bank overdrafts or failure to pay preferred dividends. It has been pointed out [2] that the failure of an enterprise could be defined as the state of loss, as the actual returns of the investment were far lower than the return rate of similar investments in the past or at the present time. The definition of financial distress in this study follows the TEJ's (Taiwan Economic Journal) definition of financial distress. The TEJ maintains that when any of the following occurs in the company, the company can be judged to be in financial distress: bankruptcy, reorganization, a run of checks being returned, a bailout request, a takeover, CPAs' qualified opinions, a negative net worth, delisting from the public capital market, and downtime due to tight finances.

Of the conventional statistical methods available, multivariate analysis methods (including Logit regression), discriminant analysis and Probit regression are the most commonly used. However, these statistical methods are restricted by their many assumptions and their accuracy in predicting financial distress is relatively low. As early as the 1960s, [1] used financial ration analysis, such as the measurement of profitability, liquidity, and debt repayment capabilities to predict the possibility of financial distress. [2] uses financial variables to predict financial distress using the multivariate discriminant analysis method by selecting the five most descriptive financial ratios: working capital/total assets, retained earnings/total assets, earnings before interest and taxes/total assets, market value of equity/book value of total debt, and net sales/total assets.

Since the 1980s, many AI methods (here referring to data mining methods) have been applied to establishing a financial distress prediction model, such as: a decision tree [3-5], an artificial neural network [6-9], a Bayesian belief network [5,10] and the support vector machine [6,9,11,12]. Even though these studies increase the accuracy of predicting financial distress, the models used indicate varying degrees of accuracy, indicating that the established models and statistical methods being used are still not accurate and can be improved.

The rest of this paper is organized as follows. Section 2 describes the research methods and the source of data samples with variables selected, and then graphically shows the research design and procedure. Section 3 illustrates the results of variables selection for different financial distress prediction models and compares and discusses their prediction accuracy. Specifically, the section also explains the ten-fold cross validation that is used in comparing the performances of different prediction models. Finally, Section 4 concludes the meaning of those results and the contribution of this study to the related academic literature and practice.

2. Methodology.

2.1. Research method.

As well as stepwise regression (SR), this study utilizes several data mining techniques such as an artificial neural network (ANN), the Bayesian belief network (BNN), a decision tree (DT), and a support vector machine (SVM).

The artificial neural network (ANN) is a parallel computing model based on the human neural structure. It is "information processing technology based on the research and inspiration of the brain and nervous system", and is generally known as the parallel distributed processing model or the connectionist model [13]. It has a high-speed computation function, memory, learning, and noise filtering capabilities; therefore, it can be used to solve a number of complex classification and prediction problems. This study uses another group of artificial neural networks to test generalization with unseen samples, and observes whether the values are close to the required values. These samples are known as the testing pattern.

The Bayesian belief network (BNN) model has a non-cyclic pattern structure that uses nodes to represent various possible occurrences or visible events, and expresses their relationships by bonds. Its construction emulates the process of logical thinking. The process of constructing a Bayesian belief network involves defining and considering all the possible nodes. Once this has happened, the causal relationships of the nodes by bonds are established while the impact of changing states of nodes on other nodes by conditional probabilities is expressed. The use of the Bayesian belief network allows the prediction of the final result by computing probabilities. The deduction process of the Bayesian belief network depends mainly on the acquisition of new information in order to adjust the probability values of the states of various nodes according to the Bayes' theorem. The Bayesian belief network adjusts the network from the overall standpoint. New information provides input for the inference process, which can immediately reflect and determine the possible probabilities of all events. It is a directed acyclic graph consisting of a series of arrows between nodes. The network includes a number of decision-making variables that are connected in a single direction to form the parent-child relationships. The nodes represent the decision-making variables, while the arrows denote the variables' interdependent relationships. The variables can be discrete and continuous [14]. If the direction of the arrow is from A to B, B originated from A. In this case, A is the parent node, while B is the child node and the arrow represents their causal relationship and strength. If each node x contains different parent nodes ($Parents(x)$), then the conditional probability value of all parent nodes and node x state combinations is Equation (1) to obtain the node x conditional probability tables. The probability combination of n attributes $(x_1, x_2, \ldots, x_n)$ is Equation (2).

$$P(x|Parents(x)) \tag{1}$$

$$P(x) = P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|P_{ai}) \tag{2}$$

In data mining, the decision tree (DT) is regarded as a rule tree structure. The decision tree comprises the variables involved in the decision making, and an analysis of the capabilities of these variables can give a prediction of the results. The decision tree is often used in establishing a prediction model and shows which attribute is the candidate for the next node according to the method of entropy heuristic. This is different from the selection of a branch, as is the case with other data mining tools. The decision tree consists of a number of internal nodes and leaf nodes, where each internal node represents the category of a certain attribute, while the branch under it represents the possible value or set of a number of possible values. Decision tree algorithms include: regression tree, classification tree, CHAID, CART, etc.

The support vector machine (SVM) is a machine learning-method based on statistical learning theory and SRM (structural risk minimization). The typical SVM is a two-class classifier and from the input training data, identifies two or more classes of data through the learning mechanism. It distinguishes these by the optimal separating hyperplane to maximize the margin of the two classes of data [15]. When there are complex classification problems in processing using SVM, [16] argues that a slack variable which allows error should be added. Therefore, the input data limitation equation is as shown in Equation (3). However, when the classification of training data is wrong, ξ_i will be greater than 0. Therefore, when seeking the hyperplane, the addition of values should be minimized.

$$\text{subject to } y_i\left(w^T x_i - b\right) - 1 + \xi_i \geq 0; \quad \xi_i \geq 0\, \forall i \tag{3}$$

However, in Equation (4), the trade-off value C is added to compensate for the tolerated error. On the other hand, the minimized target function is obtained to generate the maximized margin in order to facilitate the optimal classification and solution.

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{j}^{l} \xi_j; \quad C > 0 \tag{4}$$

The Lagrange Multiplier Method is applied to converting the above two equations into a second order equation, as shown in Equation (5). Finally, under KKT (Karush-Kuhn-Tucker conditions), the optimal solution of the target equation is achieved.

$$L_D = \frac{1}{2}\|w\|^2 + C\sum_{i}^{l}\xi_i - \sum_{i=1}^{N}\alpha_i\left\{y_i\left(w^T x_i - b\right) - 1 + \xi_i\right\} - \sum_{i=1}^{N}\mu_i\xi_i \tag{5}$$

2.2. **Sample.** All the samples used in this study were collected from the TEJ database. The research period covers from 2007 to 2014. For this study, the samples used including 30 companies in financial distress and 90 normal companies in non-financial distress (financial distress: non-financial distress = 1:3). All selected companies are listed in TWSE or GTSM in Taiwan capital markets.

Variables in this study included: (1) The dependent variable: the occurrence of financial distress. Values assigned were 0 for financially non-distressed companies and 1 for financially distressed companies; (2) The independent variable: 14 main variables were selected to measure the possibility of presence of financial distress. These variables and their descriptions are shown in Table 1.

Table 1. Summary of selected independent variables

No.	*Variable*	*Description/Formula*
X01	Current ratio	Current assets/current liabilities
X02	Quick ratio	Quick assets/current liabilities
X03	Debt ratio	Total liabilities/total assets
X04	Long-term funds appropriate rate	(Total stockholders' equity + long term liabilities)/total fixed assets
X05	Stockholders' equity growth rate	(Current period stockholders' equity – previous period stockholders' equity)/previous period stockholders' equity
X06	Gross profit growth rate	(Current period gross profit – previous period gross profit)/previous period gross profit
X07	Cash flow ratio	Cash flow from operating activities/current liabilities
X08	Total assets turnover	Net sales/average total assets
X09	Inventory turnover	Cost of goods sold/average inventory
X10	Fixed assets turnover	Net sales/total fixed assets
X11	Gross profit rate	Gross profit/net sales
X12	Sales revenue per employee	Sales revenue/number of employees
X13	Operating income per employee	Operating income/number of employees
X14	Audited by BIG4 (the big four CPA firms)	1 for companies audited by BIG4; otherwise, 0

2.3. **Research design and procedure.** The financial distress prediction model is established in two stages and different prediction models are also established in two stages in this study. In the first stage, stepwise regression (SR) and ANN respectively are used to screen the variables; in the second stage, BBN, SVM, and DT-CHAID are used for building prediction models and the models' accuracy/performance are finally evaluated to identify the best classification model. The research design and procedure are shown in Figure 1.

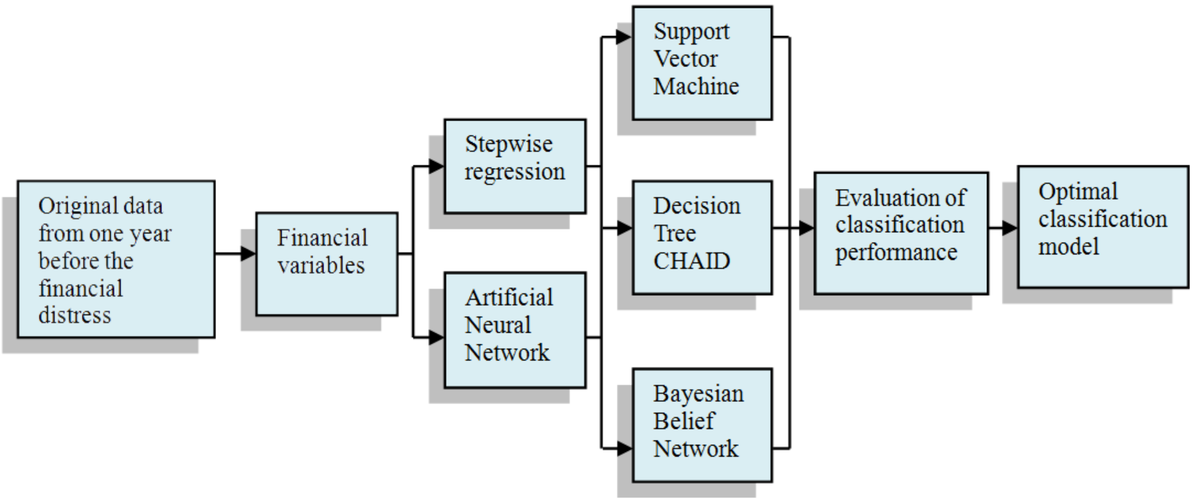
Original data from one year before the financial distress
Financial variables
Stepwise regression
Artificial Neural Network
Support Vector Machine
Decision Tree CHAID
Bayesian Belief Network
Evaluation of classification performance
Optimal classification model

FIGURE 1. Research design and procedure

3. **Results and Discussion.** After the first stage, that is of variable selection, the second stage of this study established the financial distress prediction models before comparing the prediction accuracy rates of six models ($2 \times 3 = 6$).

As there are a number of variables to be input, two methods are used to select the variables that could possibly improve the prediction accuracy.

Table 2 illustrates the selection results of a stepwise regression, where three of fourteen variables remained after the selection of the stepwise regression: X03 (debt ratio), X12 (sales revenue per employee), and X14 (audited by BIG4).

TABLE 2. Selection results of stepwise regression

Variable Name	Variable Importance
X03 (Debt ratio)	0.568
X12 (Sales revenue per employee)	0.396
X14 (Audited by BIG4)	0.036

Table 3 shows the selection results of ANN, where five of fourteen variables remained: X12 (sales revenue per employee), X03 (debt ratio), X07 (cash flow ratio), X13 (operating income per employee), and X14 (audited by BIG4).

TABLE 3. Selection results of ANN

Variable Name	Variable Importance
X12 (Sales revenue per employee)	0.33
X03 (Debt ratio)	0.314
X07 (Cash flow ratio)	0.187
X13 (Operating income per employee)	0.088
X14 (Audited by BIG4)	0.039

SPSS Clementine is used to establish the models. Stepwise regression and ANN are used to select the variables. BBN, SVM, and CHAID are used for model building and classification performance testing. The selected variables are normalized for random unrepeated sampling, with the training group and testing group in a ratio of 8:2. That is, 80% of the total samples are used for training and modeling, while 20% of the samples are used for testing and computation of accuracy [17].

After the stepwise regression in the first stage and the second state prediction model's ten-fold cross validation, it was found that the SR + SVM model's financial distress prediction accuracy is the highest at 85.67%. Regarding the overall prediction accuracy (financial distress + non-financial distress), the SR + BBN model's financial distress prediction accuracy is 84.33%, as shown in Table 4. Regarding the Type I error, the SR + SVM model is the lowest at 14.33% while for the overall error rate (Type I error + Type II error), the SR + BBN model is the lowest at 15.67%, as shown in Table 5.

Table 4. Ten-fold cross validation performance – selected by stepwise regression

Model	Financial distress	Non-financial distress	Overall accuracy
SR + BBN	85.32%	83.33%	84.33%
SR + SVM	85.67%	81.67%	83.67%
SR + CHAID	84.13%	82.49%	83.31%

Table 5. Type I error & Type II error

Model	Type I error	Type II error	Overall error rate
SR + BBN	14.68%	16.67%	15.67%
SR + SVM	14.33%	18.33%	16.33%
SR + CHAID	15.87%	17.51%	16.69%

After the first stage stepwise regression and the second state prediction model's ten-fold cross validation, it was found that, the ANN + SVM model's financial distress prediction accuracy is the highest at 86.30%, and its overall prediction accuracy (financial distress + non-financial distress) is also the highest at 85.41%, as show in Table 6. For the Type I error, the ANN + SVM model is the lowest at 13.70%; and the overall error rate (Type I error + Type II error) is also the lowest at 14.59%, as shown in Table 7.

Table 6. Ten-fold cross validation performance – selected by ANN

Model	Financial distress	Non-financial distress	Overall accuracy
ANN + BBN	84.71%	82.83%	83.77%
ANN + SVM	86.30%	84.52%	85.41%
ANN + CHAID	83.75%	80.67%	82.21%

Table 7. Type I error & Type II error

Model	Type I error	Type II error	Overall error rate
ANN + BBN	15.29%	17.17%	16.23%
ANN + SVM	13.70%	15.48%	14.59%
ANN + CHAID	16.25%	19.33%	17.79%

In summarizing the above empirical results, of all the prediction models, the ANN + SVM model's financial distress prediction accuracy is the highest at 86.30%. Its overall prediction accuracy (financial distress + non-financial distress) is also the highest at 85.41%, while the Type I error is the lowest at 13.70%, and the overall error rate is the lowest (Type I error + Type II error) at 14.59%.

4. **Conclusions.**

As previous prediction models of financial distress were not well-established for high performances in accuracy or completeness, this study uses the stepwise regression and an artificial neural network (ANN) technique to select variables at the first stage, and three data mining techniques (BBN, SVM, and CHAID), to establish different financial distress prediction models that show relatively high accuracy at the second stage.

According to the empirical results, the performance of the ANN + SVM model is the best model and its overall financial distress prediction accuracy (financial distress + non-financial distress) and the overall error rate outperform other prediction models specified in this study. Our research findings can provide a well guidance or reference for further academic researches regarding related topics (e.g., more accurate in identifying financial distress companies and/or predicting enterprise bankruptcy, corporate debt defaults, and bank overdrafts or failure) which are beneficial for a majority of relevant stakeholders in capital markets. These stakeholders include stock investors (shareholders), creditors, CPAs (certified public accountants), management and internal auditors, credit rating agencies, securities/financial analysts, financial supervisory agencies/regulators, and scholars who are interested in capital markets.

REFERENCES

[1] W. H. Beaver, Financial ratios as predictors of failure, *Journal of Accounting Research*, vol.4, pp.71-111, 1966.

[2] E. I. Altman and E. Hotchkiss, *Corporate Financial Distress and Bankruptcy*, John Wiley and Sons, New York, 1993.

[3] J. Sun and H. Li, Data mining method for listed companies' financial distress prediction, *Knowledge-Based Systems*, vol.21, no.1, pp.1-5, 2008.

[4] J. Sun, M. Y. Jia and H. Li, AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies, *Expert Systems with Applications*, vol.38, no.8, pp.9305-9312, 2011.

[5] M. Salehi and F. Z. Fard, Data mining approach to prediction of going concern using classification and regression tree (CART), *Global Journal of Management and Business Research Accounting and Auditing*, vol.13, no.3, pp.1-7, 2013.

[6] M. Yang and D. W. Xiao, The selection method for hyper-parameters of support vector classification by adaptive chaotic cultural algorithm, *International Journal of Intelligent Computing and Cybernetics*, vol.3, no.3, pp.449-462, 2010.

[7] P. Hájek, Municipal credit rating modeling by neural networks, *Decision Support Systems*, vol.51, no.1, pp.108-118, 2011.

[8] T. J. Hsieh, H. F. Hsiao and W. C. Yeh, Mining financial distress trend data using penalty guided support vector machines based on hybrid of particle swarm optimization and artificial bee colony algorithm, *Neurocomputing*, vol.82, pp.196-206, 2012.

[9] Y. N. Guo, M. Yang and D. W. Xiao, The selection method for hyper-parameters of support vector classification by adaptive chaotic cultural algorithm, *International Journal of Intelligent Computing and Cybernetics*, vol.3, no.3, pp.449-462, 2010.

[10] E. Kirkos, C. Spathis, A. Nanopoulos and Y. Manolopoulos, Identifying qualified auditors' opinions: A data mining approach, *Journal of Emerging Technologies in Accounting*, vol.4, no.1, pp.183-197, 2007.

[11] A. Sengur, Multiclass least-squares support vector machines for analog modulation classification, *Expert Systems with Applications*, vol.36, no.3, pp.6681-6685, 2009.

[12] C. Xie, C. Luo and X. Yu, Financial distress prediction based on SVM and MDA methods: The case of Chinese listed companies, *Quality & Quantity*, vol.45, no.3, pp.671-686, 2011.

[13] J. A. Freeman and D. M. Skapura, Neural networks: Algorithms, applications and programming techniques, *Journal of Operational Research Society*, vol.43, 1992.

[14] A. A. Tang, N. Y. Jin and J. Han, Using bayesian belief networks for change impact analysis in architecture design, *Journal of Systems and Software*, vol.80, no.1, pp.127-148, 2007.

[15] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd Edition, Springer Science & Business Media, 2000.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[17] C. L. Huang, M. C. Chen and C. J. Wang, Credit scoring with a data mining approach based on support vector machines, *Expert Systems with Applications*, vol.33, no.4, pp.847-856, 2007.