ICIC Express Letters
Part B: Applications
Volume **9**, Number **2**, February **2018**

 ISSN 2185-2766
pp. 131–136

FINANCIAL DISTRESS PREDICTION USING DATA MINING TECHNIQUES

Suduan Chen[1,*] and Stephen Jhuang[2]

[1]Department of Accounting Information
National Taipei University of Business
No. 321, Sec. 1, Jinan Rd., Zhongzheng District, Taipei 100, Taiwan
*Corresponding author: suduanchen@yahoo.com.tw

[2]Ernst & Young Global Limited
9F, No. 333, Sec. 1, Keelung Rd., Taipei 11012, Taiwan
A3124196@ulive.pccu.edu.tw

Received August 2017; accepted November 2017

Abstract. *This study establishes a two-stage model for financial distress predictions by applying stepwise regression and the data mining techniques of artificial neural networks and decision trees CHAID and C5.0. Both financial and non-financial variables are factored into consideration. A total of 34 companies reporting financial distress and 68 companies not reporting financial distress from 2005 to 2015 were sampled. The research findings indicate that the SR-C5.0 model has the highest prediction accuracy of 91.65%.*
Keywords: Financial distress, Data mining, Artificial neural network (ANN), Decision tree CHAID, Decision tree C5.0

1. **Introduction.** The U.S. sub-prime mortgage crisis in the second half of 2008 developed into a global financial crisis and put many companies around the world into financial distress. Many financial institutions were taken to their knees, for example, Bear Stearns (the 5th largest investment bank in the U.S.) was being acquired by JP Morgan Chase, the collapse of Indy Mac, the bankruptcy of Lehman Brothers, and the credit downgrade of AIG due to its need for fund raising. Taiwan was also engulfed into this financial tsunami, with the Financial Supervisory Commission taking over Chinfon Commercial Bank and other financial institutions. This string of corporate events resulted in social unrest, damaged the rights of stakeholders, and undermined the health of the domestic business environment. Because good operating performances of companies are the foundation of a stable society, if managers can identify warning signs or problems ahead of time, then they will be able to take measures to prevent or rein in any crisis. In sum, effective predictions of financial distress are critical to economies.

[1] defined financial distress as high leverage, difficulty in servicing debt, defaults on debt repayment, or the initiation of bankruptcy procedures. [2] stated that financial distress is a large sum of bank overdraft that needs to be utilized, default on repayments of preferred dividends or corporate bonds, and the declaration of bankruptcy. [2] used financial ratio analysis (e.g., profitability, liquidity, and debt-servicing capability) to predict the likelihood of financial distress. Research findings in the literature have indicated that the ratio of cash flows/total liabilities serves as the best tool to differentiate companies that are in financial distress and those that are not. [3] defined financial distress as being legally bankrupt, taken over, or under restructuring according to bankruptcy laws. [3] constructed the discriminant function of $Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$, with X_1 denoting working capital/total assets, X_2 as retained earnings/total assets, X_3 as earnings before interest and taxes/total assets, X_4 as the market value of equity/book value of total debt, and X_5 as sales/total assets. However, [4] posited that Altman's

forecast model is not applicable to all time periods. A famous study by [5] used the Logit model to predict financial distress, while [6] employed the Probit model to forecast financial distress, and [7] set up a discrete time hazard model as a pre-warning tool for financial distress.

Artificial intelligence (i.e., data mining) has been used as an analytical tool in recent years to predict financial distress [8-18], as well as to improve financial distress prediction accuracy by machine learning. However, these studies only employ 1 or 2 statistical methods (such as ANN, SVM, fuzzy, or DT) of artificial intelligence to predict financial distress, such that all of them are not rigorous forecast models. Therefore, this study develops a two-stage model to predict financial distress. The first stage operates stepwise regression and artificial neural network (ANN) for variable screening. The second stage deploys decision tree CHAID and decision tree C5.0 to construct a forecast model for financial distress. The purpose is to identify the best model by comparing prediction accuracy.

Following the introduction, which describes the research background, motivation, and purposes and provides a brief organization overview, Section 2 details the statistical methods, sample data, variables, research design, and procedures. Section 3 presents results and discussions. A comparison is executed on 10-fold cross validations and a review is conducted on the accuracy of the prediction models. Section 4 states the contributions of this paper to academics and practitioners and provides suggestions for the direction of future studies.

2. **Methodology.**

2.1. **Statistical methods.**

The statistical methods used herein include stepwise regression, artificial neural network (ANN), decision tree CHAID, and decision tree C5.0.

Artificial neural network (ANN) is a data processing system that mimics a biological neural system, collates information from external environments or other artificial neural networks, processing such information, and then delivering outputs to external environments or other artificial neural networks. ANN boasts good predicative power, reliability, and fault tolerance when processing a massive amount of complex data. Its basic structure consists of three layers of neurons. The first layer is the input layer, which receives the information of various characteristics. The second layer is the hidden layer, which adjusts the input values and output values by assigning weightings. The third layer is the outer layer, which generates final outputs. A too small number of neurons cannot handle complex issues; while a too large number of neurons cause inefficiency or over-fitting. The concept of ANN neuron [19] is shown as Figure 1, in which n represents the number of input variables, X_i is the ith input variable, W_{ij} is the weight of the ith variable of the jth neural cell, and P_j is the combination function of the jth neural cell. If represented by activation function $f(x)$, then $Y_j = f(P_j)$ is the output value of the jth neural cell.

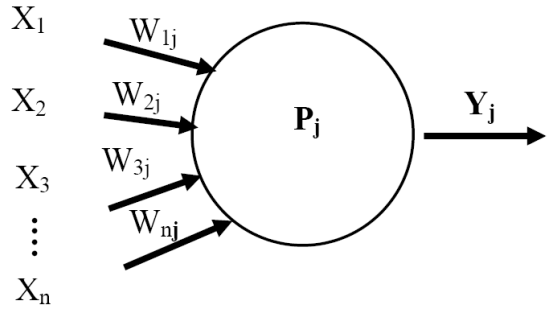

FIGURE 1. ANN neural cell

Decision trees categorize mass data according to rules and identify effective data-points for specific purposes. In order to classify each data-point, each node on the decision tree undergoes testing of the attributes of a specific variable and then determines whether a data-point meets the attribute criteria or not. Each node is able to classify inputs and form these inputs into a tree structure (DT concept diagram [19] is shown as Figure 2). The C5.0 algorithm is suitable for processing large datasets, as it entails rapid computing and consumes a smaller amount of memory resources. This algorithm, also known as a Boosting tree, adopts the Boosting method to improve forecasting accuracy. The CHAID method adjusts the Chi-square values in order to segment the sample into subgroups of the same attributes, where segmentation is conducted via searches in succession. The CHAID algorithm combines data-points according to the response to each explanatory variable and continues with the segmentation, where the purpose is to derive the minimum segmentation number on the level of each explanatory variable. Once the segmentation number for each explanatory variable is established, it is possible to refer to the explanatory variable with the highest level of statistical significance to segment the original sample into subgroups. This process continues until the segmented results show no statistical variance, or until the number of observations in the segments becomes too small to render meaningful estimates.

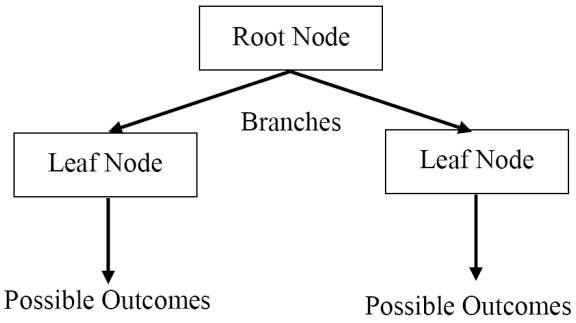

FIGURE 2. DT concept diagram

2.2. **Samples and variables.** This study uses the data of electronics companies listed on the Taiwan Stock Exchange and Taipei Exchange between 2005 $\sim$ 2015 from the Taiwan Economic Journal. A total of 102 companies are selected: 34 companies reporting financial distress and 68 companies operating under normal financial conditions (making a ratio of 1 to 2). This study examines 21 variables (consisting of 18 financials and 3 non-financials) concerning financial reporting fraud. Financial variables are X_1: Current ratio, X_2: Quick ratio, X_3: Debt ratio, X_4: Long-term funds appropriation rate, X_5: Gross profit growth rate, X_6: Operating cash flow/current liabilities, X_7: Total assets turnover, X_8: Fixed assets turnover, X_9: Inventory turnover, X_{10}: Accounts receivable turnover, X_{11}: Net income/equity, X_{12}: Gross profit rate, X_{13}: Net income growth rate, X_{14}: Debt/equity ratio, X_{15}: EBIT/I, X_{16}: EPS, X_{17}: Operating income/average equity, X_{18}: Net value growth rate. Non-financial variables are X_{19}: Sales revenue per employee, X_{20}: Operating income per employee, and X_{21}: Fixed assets per employee.

2.3. **Framework.** The first stage of this study uses SR (a traditional statistical method) and ANN (a data mining statistical method) to screen the important variables including the detection of financial and non-financial variables frequently used in investigating financial distress. In Stage II, this study employs CHAID and C5.0 (data mining statistical methods) to establish the models, followed by a comparison of the models' prediction accuracy, in order to identify the best model. The research design and procedures are illustrated in Figure 3.

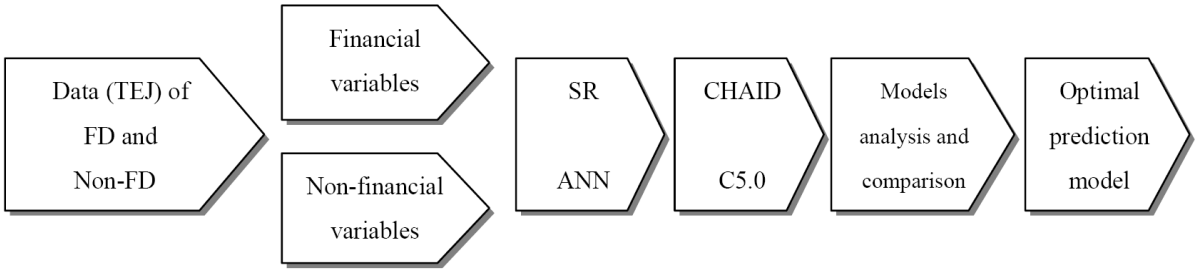

FIGURE 3. Research design and procedures

3. **Empirical Results and Analysis.** In Stage I, SR and ANN are taken as the variable screening methods. The screening results are presented below.

3.1. **Variable screening by SR.** A total of 9 variables are screened out, with the order of importance of the variables being: X_3: Debt ratio (0.23), X_{11}: Net income/equity (0.19), X_1: Current ratio (0.15), X_{16}: EPS (0.10), X_6: Operating cash flow/current liabilities (0.09), X_7: Total assets turnover (0.05), X_{21}: Fixed assets per employee (0.04), X_8: Fixed assets turnover (0.04) and X_{10}: Accounts receivable turnover (0.04).

3.2. **Variable screening by ANN.** A total of 10 variables are screened out, giving the order of variable importance as: X_{11}: Net income/equity (0.11), X_6: Operating cash flow/current liabilities (0.09), X_{21}: Fixed assets per employee (0.08), X_{20}: Operating income per employee (0.07), X_8: Fixed assets turnover (0.06), X_{10}: Accounts receivable turnover (0.05), X_{12}: Gross profit rate (0.05), and X_{18}: Net value growth rate (0.05).

After minor important variables are separated out in Stage I by SR and ANN, the research utilizes CHAID and C5.0 for Stage II modeling, because they are very suitable for predicting. In Stage II, aiming at the important variables screened out in Stage I, we employ CHAID and C5.0 to predict financial distress and modeling (SR-CHAID, SR-C5.0, ANN-CHAID, ANN-C5.0), respectively, and obtain the accurate forecast rate of each model.

3.3. **SR models.** This study had adopted rather rigorous ten-fold cross validation for an accurate forecast rate [13,18,19]. In other words, we have conducted modeling and verification for a total of 10 times first and then obtained the average value of the forecast rates in the 10 results. The ten-fold cross validation in the SR-C5.0 model has a higher prediction accuracy (FD: 92.00%; Non-FD: 91.29%; overall accuracy: 91.95%), as seen in Table 1. The SR-C5.0 model also has a lower Type I error rate (8.00%), Type II error rate (8.71%), and overall error rate (8.35%), as noted in Table 2.

TABLE 1. SR models: Ten-fold cross validation

Model	FD prediction accuracy	Non-FD prediction accuracy	Overall accuracy
SR-CHAID	88.43%	89.95%	89.19%
SR-C5.0	92.00%	91.29%	91.65%

TABLE 2. SR models: Type I error and Type II error

Model	Type I error rate	Type II error rate	Overall error rate
SR-CHAID	11.57%	10.05%	10.81%
SR-C5.0	8.00%	8.71%	8.35%

3.4. **ANN models.** Just as with the description in Subsection 3.3, a ten-fold cross validation shows that the ANN-CHAID model has higher Non-FD and overall prediction accuracy (Non-FD: 89.64%; overall accuracy: 90.13%), but a lower FD prediction accuracy (90.62%), as listed in Table 3. At the same time, the ANN-CHAID model also has a lower Type II error rate (10.36%) and overall error rate (9.87%), but suffers a higher Type I error rate (9.38%), as shown in Table 4.

Table 3. ANN models: Ten-fold cross validation

Model	FD prediction accuracy	Non-FD prediction accuracy	Overall accuracy
ANN-CHAID	90.62%	89.64%	90.13%
ANN-C5.0	90.75%	89.15%	89.95%

Table 4. ANN models: Type I error and Type II error

Model	Type I error rate	Type II error rate	Overall error rate
ANN-CHAID	9.38%	10.36%	9.87%
ANN-C5.0	9.25%	10.85%	10.05%

4. **Conclusions.** The 1997 Asian financial crisis and the 2008 global financial tsunami brought many companies around the world into financial distress or bankruptcy, and the investing public and the global economy suffered great losses as a result. Fraudulent financial reporting from 2001 to 2004 by some listed companies in Taiwan, such as Procomp Tech, Infodisc Tech, and Summit Computer, wiped out the wealth of shareholders and severely damaged the confidence of the local financial market. The continued economic recession in recent years has increased the likelihood of financial distress for the corporate world, which poses a threat to the rights of shareholders and creditors, as well as to the health of the financial environment. While operating performances are eventually reflected by financial data, it is difficult for the investing public to foresee any financial crisis, and it is often too late by the time negative financial information is disclosed. It is hence important for both academics and practitioners to establish an effective prediction model for financial distress.

The first stage of the model in this study applies SR and ANN for the screening of the key variables. The second stage employs CHAID and C5.0 for the construction of the financial distress forecast model. The empirical findings show that the SR-C5.0 model has the highest prediction accuracy (FD: 92.00%, Non-FD: 91.29%, overall: 91.65%). The ranking in terms of overall accuracy is the ANN-CHAID model, the ANN-C5.0 model, and finally the SR-CHAID model. This study uses several data mining techniques to establish a rigorous model for financial distress prediction and offers results that can serve as reference to accountants, auditors, securities analysts, company management, and academic researchers.

For further research directions we offer the following: 1) employ studies on different countries or regions, which will influence the financial and non-financial variables be used; 2) utilize more statistical methods (including data mining) to predict financial stress, so as to build up a prediction model that has higher accuracy.

REFERENCES

[1] G. Andrade and S. Kaplan, How costly is financial (not economic) distress? Evidence from highly leveraged transactions that become distressed, *Journal of Finance*, vol.53, pp.1443-1493, 1998.

[2] W. H. Beaver, Financial ratios as predictors of failure, *Journal of Accounting Research*, vol.4, no.3, pp.71-111, 1966.
[3] E. I. Altman, Financial ratios discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance*, vol.23, no.4, pp.589-609, 1968.
[4] R. C. Moyer, Forecasting financial failure: A re-examination, *Financial Management*, vol.23, no.4, pp.11-17, 1977.
[5] J. A. Ohlson, Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research*, vol.18, no.1, pp.109-131, 1980.
[6] M. E. Zmijewski, Methodological issues related to the estimation of financial distress prediction models, *Journal of Accounting Research*, vol.22 (Supplement), pp.59-82, 1984.
[7] T. Shumway, Forecasting bankruptcy more accurately: A simple hazard model, *The Journal of Business*, vol.74, no.1, pp.101-124, 2001.
[8] M. D. Odom and R. Sharda, A neural network model for bankruptcy prediction, *Proc. of the IEEE International Joint Conference on Neural Network*, vol.2, pp.163-168, 1990.
[9] K. Tam and M. Kiang, Managerial applications of neural networks: The case of bank failure predictions, *Management Science*, vol.38, no.7, pp.926-947, 1992.
[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] J. Sun and H. Li, Data mining method for listed companies' financial distress prediction, *Knowledge-Based Systems*, vol.21, pp.1-5, 2008.
[12] 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.
[13] M. Y. Chen, Predicting corporate financial distress based on integration of decision tree classification and logistic regression, *Expert Systems with Applications*, vol.38, pp.11261-11272, 2011.
[14] P. Hájek, Municipal credit rating modelling by neural networks, *Decision Support Systems*, vol.51, no.1, pp.108-118, 2010.
[15] 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.3, no.3, pp.196-206, 2012.
[16] Z. Yang, W. You and G. Ji, Using partial least square and support vector machines for bankruptcy prediction, *Expert Systems with Applications*, vol.38, pp.8336-8342, 2011.
[17] N. Gordini, A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy, *Expert Systems with Applications*, vol.41, pp.6433-6445, 2014.
[18] F. J. López Iturriaga and I. P. Sanz, Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks, *Expert Systems with Applications*, vol.42, pp.2857-2869, 2015.
[19] S. Chen, Detection of fraudulent financial statements using the hybrid data mining approach, *SpringerPlus*, 2016.