Variable Selection Method Affects SVM Approach in Bankruptcy Prediction

Chih-Hung Wu [1] **Wen-Chang Fang** [2] **Yeong-Jia Goo** [3]

[1] Department of Business Administration, Takming College, Taipei, Taiwan

[2,3] Department of Business Administration, National Taipei University, Taipei, Taiwan

ABSTRACT

This paper examined bankruptcy predictive accuracy of five statistics models--discriminant analysis logistic regression, probit regression, neural networks, support vector machine (SVM), and genetic-based SVM (GA-SVM) that influenced by variable selection. Empirical results indicate that the SVM-based models are very promising models for predicting financial failure, in terms of both best predictive accuracy and generalization ability. In addition, variable selection had the lowest influence of predictive accuracy in the GA-SVM model with optimal values of parameters.

***Keywords**:* Variable Selection, Bankruptcy Prediction, Support vector machine.

1. Introduction

Predicting corporate failure has been an important research topic in accounting and finance for the last three decades (Lee, Han, and Kwon, 1996). Previous application of neural networks in finance and accounting, notably in bankruptcy prediction, are limited to back-propagation neural networks (Yang, Platt, and Platt, 1999). Recently, new algorithms in learning machine, Support vector machines (SVMs), were developed by Boser, Guyon, and Vapnik (1992) to provide better solutions to decision boundary than could be obtained using the traditional neural network. Since the new model was proposed (Boser, Guyon, and Vapnik, 1992; Cortes and Vapnik, 1995), SVM has been successfully applied to numerous applications, including the handwriting recognition, particle identification (e.g. muons), digital images identification, text categorization, bioinformatics, function approximation and regression, and database marketing. Although SVMs have become more widely used to time series forecasting and dynamically reconstruct of chaotic systems. However, few articles have been devoted to the study of analyzing the power of variable selection to influence SVM-based models on problems of finance prediction. Consequently, this study analyzed the bankruptcy predictive accuracy of five various statistics models--discriminant analysis, logistic regression, probit regression, neural networks, support vector machines (SVM), and genetic-based SVM (GA-SVM) that influenced by variable selection.

2. Overview of methodologies for predicting bankruptcy

The corporate distress literature includes several diverse methodologies for discriminating between failed and non-failed firms, following Beaver's univariate comparison of financial ratios in 1966. Extensive studies in this area have applied statistical and AI approaches over the last three decades. The well-known multivariate models used in this area include multiple discriminate analysis (MDA) (Altman, 1968; Altman, Haldeman, and Narayanan, 1977), regression modeling (Edmister, 1972), logit analysis (Ohlson, 1980; Platt and Platt, 1990), and probit analysis (Zmijewski, 1984). Most recently, AI approaches, such as neural network approaches have shown promise as classification tools (Odom and Sharda, 1990; Berry and Treigueiros, 1991; Coakley and Brown, 1991; Raghupathi, Schkade, and Raju, 1991; Lee, Han, and Kwon, 1996; Yang, Platt, and Platt, 1999). Apart from abovementioned methodologies, SVM is herein extended to model the financial distress classification problem.

3. Research Design

3.1 Research Data

Financial-statement data of the failed and non-failed firms were obtained from the database of the Taiwan Economic Journal (TEJ), covering in cases of three years prior to failure

and one year after failing. “Failure” is defined as the inability of a firm to pay its financial obligations as they mature. This study defined the firms in financial distress as those whose listed securities have been classified as the category of alter-trading-method. According to the definition of Beaver (1966), the “first year before failure” is defined as that year included in the most recent financial statement prior to the year in which the firm is reported to have failed. The data sample consists of firms in Taiwan that failed in the period from 1998 to 2002. The failed firms were selected from the lists of bankruptcy companies by the Taiwan Stock Exchange (TSE) and the database of TEJ. The size of matched sample was 88 firms, including 22 failed firms and 66 non-failed firms. In the simulated sample, the total sample size was 22 companies, including 15 failed firms and 66 non-failed firms. The holdout sample comprises of all corporations listed on the TSE and OTC market from 2001 to 2002. The sample size for 2000 was 538 firms, including 373 firms on the TSE and OTC market in 2000. The sample size for 2001 was 534 firms, including 356 firms on the TSE and OTC market.

3.2 The Procedure of Evaluating Models

The procedure of evaluating bankruptcy prediction models was depicted in Fig. 1. Experiments were performed to examine two kinds of analysis: (1) all variables were included in building bankruptcy prediction model (e.g. non-variable selection), (2) all variables were screened by MWW test. Besides the accuracy of the predictions of bankruptcy, Type I and Type II errors were analyzed among these experiments. Type I error was defined as the probability that a firm predicted not to fail will in fact fail, while the Type II error was defined as the probability that a firm predicted to fail will not in fact fail (Blum, 1974). The SVM model is applied with fix values of parameters. The GA-SVM model that proposed by Wu et al. (2007) is implemented with the optimal values of parameters. The bankruptcy models in this investigation employed 19 financial variables, selected in previous research on financial distress, as input variables. These variables were organized into four groups, according to whether they related to liquidity, profitability, asset management or financial structure. The input variables of all the models are the same. The hit ratio of classification is the indicator used to evaluate the predictive accuracy of model.

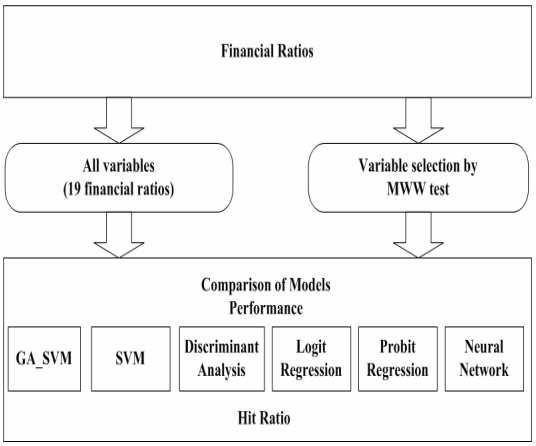

Figure 1 The procedure of evaluating models

3.3 The SVM Model

When data sets are noisy and exhibit a large overlap between data classes, error variables $\varepsilon_i > 0$ are introduced to allow the output of the outlier to be locally corrected, constraining the range of the Lagrange multiplier α_i from 0 to C. C is a constant penalty function designed to prevent outliers from affecting the optimal hyperplane. Hence, the non-linear objective function is maximize

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \left(k(x_i, x_j) \right)$$

Subject to $0 \le \alpha_i \le C, \quad i = 1, ..., l$ (1)

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (2)$$

The optimal weight w^* and bias are determined by solving the quadratic programming problem. $w^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i$, $b^* = y_i - w^{*T} x_i$ .The optimal decision function is $f(x) = sign\left(\sum_{i=1}^{l} y_i \alpha_i^* k(x, x_i) + b^* \right)$.

In machine learning theories, popular kernel functions, such as the Gaussian kernel

function, have been found to provide good generalization capabilities (Kecman, 2001; Campbell, 2002). Accordingly, the Gaussian kernel function is employed as the kernel function in this work.

4. Data Analysis

To examine the accuracy that influenced by variable selection, bankruptcy models used financial variables, selected by the Mann-Whitney-Wilcoxon (MWW) test, as input variables. Table 1 presents the results. The result of normality test revealed that most of financial ratios were not normally distributed as has been stated in previous research. Therefore, a non-parametric test (e.g. MWW test) herein used to select important variables for modeling the financial crisis models.

Table 1 Variables removed following the MMW test

	Data Set			
	2001 TSE		2001 OTC	
Ratio	MWW	(p-value)	MWW	(p-value)
Liquidity				
Current Ratio	925.0	(0.019)*	157.0	(0.001)**
Quick Ratio	783.0	(0.007)**	217.0	(0.003)**
Cash flow Ratio	1084.5	(0.049)	**68048.5**	**(0.908)**
Profitability				
Net income to sales	278.0	(0.000)**	95.0	(0.000)**
Gross profit to sales	661.0	(0.003)**	433.5	(0.037)*
Net income to total assets	**1262.0**	**(0.126)**	102.0	(0.001)**
Net income to stockholder's equity	124.0	(0.000)**	418.0	(0.032)*
Operating income to sales	387.0	(0.000)**	**594.5**	**(0.159)**
Earning per share (EPS)	192.0	(0.000)**	86.0	(0.000)**
Growth ratio of sales	517.5	(0.000)**	**625.0**	**(0.200)**
Asset Management				
Total asset turnover	**1339.0**	**(0.180)**	266.0	(0.005)**
Fixed assets turnover	1036.0	(0.037)*	**509.0**	**(0.077)**
Inventory turnover	**1317.5**	**(0.104)**	**734.5**	**(0.412)**
Receivables turnover	**142858.0**	**(0.546)**	**706.5**	**(0.348)**
Financial Structure				
Debt ratio	141407.0	(0.000)**	67426.0	(0.006)**
Long-term liabilities to fixed assets	612.0	(0.002)**	441.0	(0.040)*
Degree of Financial Leverage (DFL)	1071.5	(0.046)*	**583.0**	**(0.145)**
Liabilities to stockholder's equity	141380.0	(0.000)**	67306.0	(0.001)**
Interest coverage ratio	445.0	(0.000)**	227.0	(0.003)**

Note: * denotes asymmetrical Sig. (2-tailed) Mann-Whitney-Wilconxon test α <0.05
** denotes asymmetrical Sig. (2-tailed) Mann-Whitney-Wilcoxon test α <0.01

The MWW test was adopted in the data preparation to determine the significant variables in the financial distress model. It calculates the sum of ranks for the larger of the two groups (either distressed firms or non-distressed firms). If the two groups are equally sized, then the MMW test is computed for the second of the two groups listed on your output.

Based on the results of the MWW test, four financial ratios (variables) were removed. These financial ratios were "net income to total assets", "total asset turnover", "inventory turnover", and "receivables turnover". Most removed ratios were in the category of Asset Management. The sample data for training were the TSE market data in 2001 and the data for predicting were the TSE market data in 2002. As result reveals, these models performed well independently of whether all variables (non-selected), or those variables selected by the MWW test, were employed. The financial ratios removed by the MMW test summarizes in Table 1. The significance level, α, was set to 0.5. As Table 1 shows, most variables that related to the category of Asset Management did not differ significantly between non-failed and failed firms. This finding implies that variables in the category of Asset Management might not able to classify the failed or non-failed firms.

Table 2 Predictive accuracies of models in TSE market

	Non-Selected			Selected by MWW		
Training	2001 TSE (538 firms), All ratios			2001 TSE (538 firms), 15 ratios*		
Predicting	2002 TSE (534 firms), All ratios			2002 TSE (534 firms), 15 ratios*		
Models	Accuracy	Type I Error	Type II Error	Accuracy	Type I Error	Type II Error
DA	0.2154	0.019	0.766	0.4513	0.021	0.528
Logit	0.4195	0.019	0.562	0.8558	0.019	0.125
Probit	0.4307	0.019	0.551	0.8446	0.019	0.137
NN	0.9739	0.026	0.000	0.9738	0.026	0.000
SVM	0.8820	0.026	0.092	0.9569	0.026	0.017
GA-SVM	0.9738	0.026	0.000	0.9738	0.026	0.000
Average	0.6492	0.023	0.329	0.8427	0.023	0.135

Note: * Four financial ratios were removed based on the Mann-Whitney-Wilcoxon test.

Based on the MWW test, seven financial ratios were removed from the financial bankruptcy model for the data in the OTC market in 2001 and 2002. These financial ratios were "cash flow Ratio", "operating income to sales", "growth ratio of sales", "fixed assets turnover", "inventory turnover", "receivables turnover", and "degree of financial leverage (DFL)". One of these removed variables related to the category of Liquidity, two to the category

of Profitability, three to the category of Asset management and one to the category of Financial Structure.

Table 3 Predictive accuracies of models in OTC market

	Non-Selecting			Selected by MWW		
Training	2001 OTC (373 firms), All ratios			2001 OTC (373 firms), 12 ratios*		
Predicting	2002 OTC (356 firms), All ratios			2002 OTC (356 firms), 12 ratios*		
Models	Accuracy	Type I Error	Type II Error	Accuracy	Type I Error	Type II Error
DA	0.8848	0.011	0.104	0.6517	0.008	0.340
Logit	0.9719	0.014	0.014	0.9803	0.014	0.006
Probit	0.9747	0.014	0.011	0.9803	0.014	0.006
NN	0.9860	0.014	0.000	0.9775	0.014	0.008
SVM	0.9831	0.011	0.006	0.9831	0.014	0.003
GA-SVM	0.9860	0.014	0.000	0.9831	0.014	0.003
Average	0.9644	0.013	0.023	0.9260	0.013	0.061

Note: * Seven financial ratios were removed based on the Mann-Whitney-Wilcoxon test.

5. Conclusions

This study pioneered on examining that the variable selection affects predictive accuracy of SVM-based models. Empirical results reveal that the SVM-based models are a very promising AI model for predicting bankruptcy in terms of both predictive accuracy and generalization ability. The predictive accuracies of these models are less influenced by the variable selection method. Moreover, this study demonstrated that the GA-SVM model and SVM model performed well when applied in the holdout sample, revealing the stationary of these models to forecast bankruptcy firms.

REFERENCES

[1] Altman, E. I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 1968; 3:589-609.

[2] Altman, E. I. Predicting Financial Distress of Companies: Revisiting the Z-SCORE and ZETA Models. adapted and updated from Altman (1968) and Altman (1977), 2000 (July).

[3] Beaver, W. Financial Ratios as Predictors of Failures. In Empirical Research in Accounting. Journal of Accounting Research (Supplement), 1966; 4:71-102.

[4] Berry, R. and Treigueiros, D. The Application of Neural Network Based Methods to the Extraction of Knowledge from Accounting Reports. IEEE International Joint Conference on Neural Networks, 1991; 136-146.

[5] Blum, M. Failing Company Discriminant Analysis. Journal of Accounting Research, 1974; 12 (Spring):1-25.

[6] Boster, B., Guyon, I., and Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. in Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992;144-152.

[7] Campbell, C. Kernel Methods: A Survey of Current Techniques. Neurocomputing, 2002; 48:63-84.

[8] Coakley, J. R. and Brown, C. E. Neural Network Applied to Ratio Analysis in the Analytical Review Process. The 4th International Symposium on Expert Systems in Accounting, Finance and Management, 1991; 1-35.

[9] Cortes, C., and Vapnik, V. N. Support Vector Networks. Machine Learning, 1995; 20:273-297.

[10] Edmister, R. Financial Ratios and Credit Scoring for Small Business Loans. Journal of Commercial Bank Lending, 1972.

[11] Kecman, V. Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models, MIT Press, Cambridge, Massachusetts, London, England, 2001.

[12] Lee, K. C., Han, I., and Kwon, Y. Hybrid Neural Network Models for Bankruptcy Predictions. Decision Support Systems, 1996; 18:63-72.

[13] Odom, M. and Sharda, R. A Neural Network for Bankruptcy Prediction. Proceedings of the IEEE International Conference on Neural Network. 1990; 2:pp.163-168.

[14] Ohlson, J. A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 1980; 18:109-131.

[15] Raghupathi, W., Schkade, L., and Raju B. S. A Neural Network Application for Bankruptcy Prediction. IEEE International Joint Conference on Neural Networks, 1991; 147-155.

[16] Wu, C. H., Tseng, G. H., Goo, Y. J., and Fang, W. C. A Real-Valued Genetic Algorithm to Optimize the Parameters of Support Vector Machine for Predicting Bankruptcy. Expert Systems with Applications. 32(2), March, 2007. (forthcoming)

[17] Yang, Z. R., Platt, M. B., and Platt, H. D. Probabilistic Neural Networks in Bankruptcy Prediction. Journal of Business Research, 1999; 44(2):67-74.

[18] Zmijewski, M. E. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. Journal of Accounting Research. 1984; 22:59-82 (Supplement).