Expert Systems with Applications 38 (2011) 15094–15102

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Financial ratio selection for business crisis prediction

Fengyi Lin [a], Deron Liang [b,*], Enchia Chen [c]

[a] Department of Business Management, National Taipei University of Technology, Taipei, Taiwan
[b] Software Research Center and Department of Computer Science and Information Engineering, National Central University, Jongli City, Taoyuan County 320, Taiwan
[c] Department of Computer Science, National Taiwan Ocean University, Taiwan

ARTICLE INFO

Keywords:
Financial predictors (variables)
Feature selection
Financial prediction
SVM

ABSTRACT

Recent research has used financial ratios to establish the diagnosis models for business crises. This research explores a broader coverage of financial features, namely the recommended financial ratios from TEJ (Taiwan Economic Journal) database in addition to those financial ratios studied in prior literature. The aim of this research is to discover potentially useful but previously unaware financial features for better prediction accuracy. In this study, we had applied data mining techniques to identify five useful financial ratios, which two of them, tax rates and continuous four quarterly EPS are previously unaware to the research community. Our empirical experiment indicates that our proposed feature set outperforms those models proposed by prior scholars in terms of the prediction accuracy.

1. Introduction

Financial prediction is a challenging problem that generates extensive studies over the past decades. Recent outbreak of corporate financial crises worldwide has intensified the need to reform the existing financial architecture. It is generally believed that symptoms and alarms can be observed prior to a business encounters financial difficulty or crisis. The overall objective of business crisis prediction is to build models that can extract knowledge of risk evaluation from past observations and to evaluate business crisis risk of companies with a much broader scope. Eichengreen (1999) identifies the policies of the new international financial architecture as crisis prevention, crisis prediction and crisis management. Financial indicators have been consulted by researchers as a major basis for predicting financial distress and business crises while other common methodologies include peer group analysis, comprehensive risk assessment systems, and statistical and econometric models (Ozkan-Gunay & Ozkan, 2007).

There are two major factors influencing financial distressed prediction as shown in Fig. 1. First, using different financial features to prediction may cause different prediction results. Most of the features emphasize finance ratios, such as adequacy of long term capital, current ratio, inventory turnover, EPS and debt coverage stability, fixed asset turnover, profit growth rate, revenue per share, net profit growth rate before tax and after tax, etc. (Min, Lee, & Han, 2006; Shin, Lee, & Kim, 2005). Altman (1968) selected 5 financial ratios such as Sales to total assets; Beaver (1966) adopted 6 ratios including debt ratio. Ohlson (1980) utilized nine different features. However, the single financial feature used to discern the firms would show some variability, because different predicting directions and capabilities with regard to finance ratios, along with conflicting results, lead to widely different predictions.

If we could obtain the integrated combination of all significant predicted variables, it is of great help to reduce the quantity of variables necessary. In this paper, we examine the financial data offered by Taiwan Economic Journal (TEJ), the authoritative financial data bank covering extensive financial data sets of all listed companies traded in Taiwan Stock Exchange (TWSE) since 1980. We select all 74 financial ratios, referred as the TEJ feature set, and combine this set with those 21 financial ratios recommended by previous research (shown in Table 2). We call this set of 21 ratios as the literature feature set (as opposed to the TEJ feature set). In addition, we explore if there are some financial ratios that have not been mentioned in prior research but with great potential to increase the prediction accuracy.

The second factor that has significant influence on the model prediction accuracy is the classifier used in building the prediction model. Since 1960s, there had been numerous scholars conducting research into business crisis prediction. Scholars applied statistical methods such as Multiple Discriminate Analysis (MDA) (Altman, 1968; Beaver, 1966; Chuvakhin & Gertmenian, 2003) and Logit (Ohlson, 1980; Zmijewski, 1984). As computer technology is widely used in the business prediction, it is easy to apply complex algorithms in analyzing huge data sets. Therefore, aside from the aforementioned classification methods, new algorithms such as the Decision Tree (DT) (Tam & Kiang, 1992), Neural Network (Lee, Han, & Kwon, 1996; Ozkan-Gunay & Ozkan, 2007; Tam & Kiang, 1992) and Support Vector Machine (SVM) (Chandra, Ravi,

* Corresponding author.
E-mail addresses: fengyi@ntut.edu.tw (F. Lin), drliang@csie.ncu.edu.tw (D. Liang), M97570010@mail.ntou.edu.tw (E. Chen).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.05.035

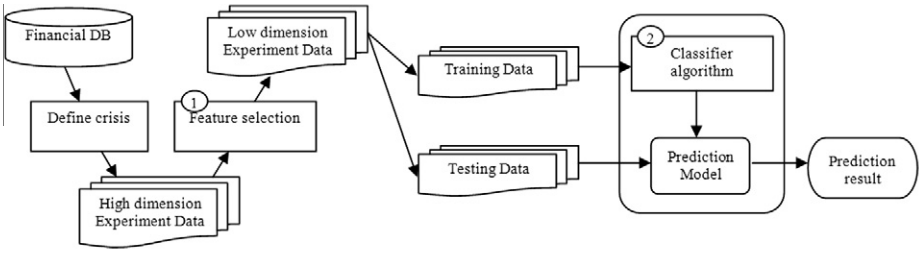

Fig. 1. Two major factors influencing financial distressed prediction.

Table 2
Selected financial ratio for financial failure prediction.

No.	Ratio definition	Mentioned by
X_1	Current ratio	Beaver (1966), Zmijewski (1984), Martens et al. (2008)
X_2	Cash flow/total debt	Beaver (1966), Deakin (1972), Blum (1974), Zmijewski (1984), Martens et al. (2008)
X_3	Cash flow/total asset	Deakin (1972), Ohlson (1980)
X_4	Cash flow/sales	Deakin (1972), Li and Sun (2009)
X_5	Debt ratio	Beaver (1966), Deakin (1972), Ohlson (1980), Martens et al. (2008), Ding et al. (2008)
X_6	Working capital/total asset	Beaver (1966), Altman (1968)
X_7	Market value equity/total debt	Altman (1968), Martens et al. (2008), Li and Sun (2009)
X_8	Current assets/total asset	Deakin (1972)
X_9	Quick asset/total asset	Deakin (1972)
X_{10}	Sales/total asset	Altman (1968), Li and Sun (2009)
X_{11}	Current debt/sales	Deakin (1972)
X_{12}	Quick asset/sales	Deakin (1972)
X_{13}	Working capital/sales	Beaver (1966), Deakin (1972), Ohlson (1980), Martens et al. (2008)
X_{14}	Net income/total asset	Beaver (1966), Deakin (1972), Ohlson (1980), Zmijewski (1984)
X_{15}	Retained earnings/total asset	Altman (1968), Ding et al. (2008)
X_{16}	Earnings before interest and taxes/total asset	Altman (1968), Li and Sun (2009)
X_{17}	No-credit interval	Beaver (1966)
X_{18}	log(total assets/GNP price-level index)	Ohlson (1980)
X_{19}	One if total liabilities exceeds total assets, zero otherwise	Ohlson (1980)
X_{20}	One if net income was negative for the past 2 years, zero otherwise	Ohlson (1980)
X_{21}	$(\mathrm{NI}_t - \mathrm{NI}_{t-1})/(\lvert\mathrm{NI}_t\rvert + \lvert\mathrm{NI}_{t-1}\rvert)$, NI_t: Latest Net income	Ohlson (1980)

& Bose, 2009; Chen & Hsiao, 2008; Ding, Song, & Zen, 2008; Hua, Wang, Xu, Zhang, & Liang, 2007; Shin et al., 2005; Wu, Tzeng, Goo, & Fang, 2007) are used. Lately, more sophisticate Case-Based Reasoning (CBR) models are proposed that includes the CBR with multiple classifiers (Li & Sun, 2009), OR-CBR (Li, Sun, & Sun 2009) and ranking-order CBR (Li & Sun, 2008). Table 1 summarizes the classifiers used in financial prediction.

The aim of this paper is twofold. While prior scholars used popular feature set that represents certain domain knowledge, in this study we select variables not only from prior literature (or the literature feature set), but also from (Taiwan) TEJ feature set, which contains 74 extra financial ratios. The Taiwan TEJ database is the authoritative datasets that are frequently used by both academic literature and practices in Taiwan. Our approach aims to search for some financial features which might be ignored by prior financial experts but could be useful to gain better business failure prediction. Six new financial ratios are identified from the TEJ feature set together with four ratios from the literature feature set after screening via data mining techniques. This set of 10 financial ratios serve as potential candidates for the construction of prediction models. Secondly, we construct SVM models based on the selected features consisting of 10 financial ratios. Further analysis shows that an SVM model built with a feature set consists of five financial ratios, two from the TEJ feature set, yields the best performance. We also compare our model against other models based on the feature sets recommended by prior studies. Experiments indicate that our model outperforms other models in prediction accuracy. In summary, our proposed model encompassing new financial features can be expected to achieve a more accurate prediction of corporate financial distress than a model based exclusively on prior scholars' results.

The next section focuses on a theoretical overview of business crisis prediction. Section 3 introduces the proposed business crisis prediction models. Section 4 outlines the research experiment framework and design adopted by our study. The experiment results and discussion are presented in Section 5. Finally, the conclusion is provided in Section 6.

2. Literature review

Business crisis prediction is a challenging problem stimulating numerous studies over the past decades. Early studies tend to treat financial ratios measuring profitability, liquidity and solvency as significant indicators for the detection of financial difficulties. However, reliance on these financial ratios can be problematic. The order of their importance, for example, remains unclear as different studies suggest different ratios as the major indicators of potential financial problems.

2.1. Financial crises and financial features

Despite the numerous definitions of business crises, the general meaning should include some narrower definitions like bankruptcy and shut-down and some broader definitions like failure, decline and distress. According to Beaver (1966), a business crisis occurs when a firm announces its bankruptcy, bond default, over-drawn bank account or nonpayment of preferred stock dividends. As financial factors are mostly backward-looking, point-in-time measures, prediction models examining only financial features are inherently constrained. This paper accordingly would like to further explore the role of non-financial features in corporate business crisis prediction.

Table 1
Classifiers used in financial prediction studies.

Classifiers	Paper studied
MDA	Altman (1968), Beaver (1966), Chuvakhin and Gertmenian (2003)
Logit Regression	Ohlson, 1980, Tam and Kiang (1992), Zmijewski (1984), Hua et al. (2004)
Neural Network	Lee et al. (1996), Shin et al. (2005), Tam and Kiang (1992)
Decision Tree	Tam and Kiang (1992)
Support Vector Machine	Shin et al. (2005), Wu et al. (2007), Hua et al. (2007), Ding et al. (2008), Chandra et al. (2009)
Case-Based Reasoning	Jo and Han (1996), Sun and Hui (2006), Li and Sun (2009), Li et al. (2009), Li and Sun (2008)

The pioneering study of Beaver (1966) introduces a univariate approach of discriminant analysis to predict financial distress. The method was later expanded into a multivariate framework by Altman (1968). Discriminant analysis had been the primary method of business failure prediction until 1980s during which the use of logistic regression method was emphasized. The standard discriminant analysis procedures assume that the variables used to characterize the members of the groups under investigation are in multivariate normal distribution. However, in real life, deviations from the normality assumptions are more likely to take place, and this violation may result in biased results. A non-linear logistic function is preferred over multivariate discriminant analysis (MDA), and there are researchers (Altman, 1968; Günther & Grüning, 2000; Huang, Chen, Hsu, Chen, & Wu, 2004) claiming that even when all the assumptions of MDA hold, a Logit model is virtually as efficient as a linear classifier. Considerable discrepancy is observed in the prediction accuracy reached by the three methods since using different methods leads to different prediction models that adopt different financial ratios.

Major financial features selected for financial distress prediction include financial leverage, long-term and short-term capital intensiveness, return on investment, EPS and debt coverage stability, etc. Selection of these features, however, is seldom based on a theory capable of explaining why and how certain financial factors are linked to corporate bankruptcy (Günther & Grüning, 2000; Huang et al., 2004). However, the selected features could have huge impact on the financial prediction. Prior studies have frequently used financial features as shown at our summarized Table 2.

A closer examination of the above 21 constructed features reveals some interesting patterns of how domain knowledge is represented through different combinations of raw accounting variables. For example, among the 21 constructed features, 8 of them are constructed by dividing raw accounting variables by total assists, 2 of them are constructed by dividing the variables of loan-specific assets by gross loans and 4 of them are constructed by dividing total sales. These constructed features in some sense reflect preliminary domain knowledge of normalization. The goal of normalization is to eliminate the effects of some irrelevant factors in describing a company's financial condition (Zhao, Sinha, & Ge 2009). While the 21 financial features are all in relatively simple forms, they constitute important domain knowledge, which is not explicitly captured in, and cannot be automatically learned from, the raw accounting data. Without some knowledge of the financial domain, even a data mining specialist would not know how to combine different raw accounting variables in meaningful ways to construct such intermediate concepts.

In this study we selected variables from both prior literature (21 ratios as listed in Table 2) and Taiwan TEJ feature set, which contains 74 financial features (see Appendix A) to predict financial crisis. This methodology aims to search for some financial features which might be ignored by prior financial experts but could be used better business failure prediction. Fig. 2 shows how we choose the TEJ feature set and prior literature feature set.

3. Business crisis prediction model: the back ground

Substantial literature can be found on business crisis prediction. We briefly review methods used in this research, i.e., the Iterative Relief and Support Vector Machine (SVM).

3.1. Iterative relief

The Iterative Relief algorithms, as one of the first feature weighting methods that have a clearly defined objective function and can be solved through numerical analysis instead of combinatorial searching, provide a promising direction for more rigorous treatment of the feature weighting and selection problems (Sun, 2007).

3.2. SVM model

As a relatively new algorithm in machine learning, Support Vector Machine (SVM) was first developed by Boster, Guyon, and Vapnik (1992) to provide better solutions to decision boundary than could be obtained using the traditional Neural Network. The machine learning techniques automatically extract knowledge from a data set and construct different model representations to explain the data set. The SVM approach has been put into several financial applications recently, mainly in the area of time series prediction and classification (Shin et al., 2005). SVM belongs to the type of maximal margin classifier, in which the classification problem can be represented as an optimization process. Vapnik (1995) showed how training a Support Vector Machine for pattern recognition could lead to a quadratic optimization problem with bound constraints and one linear equality constraint. The basic procedure for applying SVM to a classification model can be summarized as follows (Chen & Hsiao, 2008). First, the input vector is mapped into a feature space, which is possible with a higher dimension. The mapping is either linear or non-linear, depending on the kernel function selected. Then, within the feature space, the approach proceeds to seek an optimized division, i.e., to construct a hyperplane that separates two or more classes. Using the structural risk minimization rule, the training of SVMs always seeks a globally optimized solution and avoids over-fitting. It has, therefore, the ability to deal with a large number of features. The decision

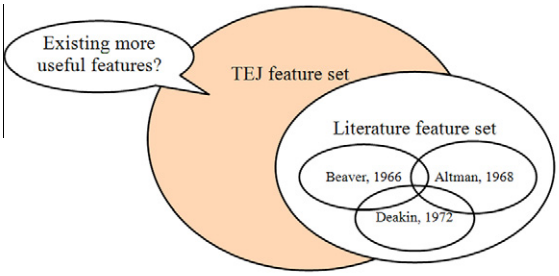

Fig. 2. TEJ feature set vs. literature feature set.

function (or hyper-plane) determined by a SVM is composed of a set of support vectors selected from the training samples.

The SVM developed by Vapnik (1995) implements the principle of *Structural Risk Minimization* by constructing an optimal separating hyper plane $w \cdot x + b = 0$. SVM uses a linear model to separate sample data through some nonlinear mapping from the input vectors into the high-dimensional feature space. Unlike most of the traditional Neural Network models which implement the *Empirical Risk Minimization Principle*, SVM seeks to minimize an upper bound of the generalization error rather than minimizing the training error. To find the optimal hyper plane $\{x \in S(w,x) + b = 0\}$, the norm of the vector w needs to be minimized while the margin between the two classes $1/||w||$ should be maximized

$$\min_{i=1,\ldots,n} |(w,x) + b| = 1. \tag{1}$$

According to Lagrange multiplier α_i, the decision function is built as follows:

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{ij=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \quad \text{subject to } 0 \leqslant \alpha_i \leqslant C, \\ \times \sum_{i=1}^{l} \alpha_i y_i = 0, \tag{2}$$

where C is the penalty parameter of the error term.

Finally, we get a nonlinear decision function in primal space for linear non-separable case

$$y(x) = sign\left(\sum_{i=1}^{l} y_i \alpha_i K(x, x_i) + b \right). \tag{3}$$

Four common kernel function types of SVM are given as follows:

$$\left. \begin{array}{l} \text{Linear kernel}: K(x_i, x_j) = x_i^T x_j, \\ \text{Polynomial kernel}: K(x_i, x_j) = \left(\gamma x_i^T x_j + r\right)^d, \\ \text{Radial basis kernel}: K(x_i, x_j) = \exp\left(-\gamma ||x_i - x_j||^2\right), \\ \text{Sigmoid kernel}: K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j + r\right), \end{array} \right\} \tag{4}$$

where d, $r \in N$ and $\gamma \in R^+$ are constant.

SVM works as a maximal margin classifier in which the classification problem can be represented as an optimization process. Support vectors are a subset of training data used to define the boundary between two classes. As suggested by Vapnik (1995), SVM can be generalized well even in high-dimensional spaces under small training sample conditions, indicating a learning ability independent of the feature space dimensionality.

The training of SVMs is equivalent to solving a linearly constrained quadratic programming, helping reach a solution that is unique, optimal and absent from local minima. It is robust to outliers. It reduces the effect of outliers by using the margin parameter C to control the misclassification error. Moreover, with Vapnik's e-insensitive loss function, SVM can model nonlinear functional relationships difficult to be modeled by other techniques (Vapnik, 1995). These characteristics make SVM a strong candidate in predicting financial distress. Therefore, our proposed model defines the bankruptcy problem as a nonlinear problem.

4. Experiment framework and design

As shown in Fig. 3, the financial data of all companies subjective to the experiments are all selected from the TEJ database. The selection of the experiment data is discussed in Section 4.1. The financial ratios, both from the TEJ feature set and the literature feature set, are selected based on the feature selection algorithms discussed in Section 4.2. We use SVM to construct the prediction model. The details of SVM kernel selection and parameters setting are discussed in Section 4.3.

4.1. Sample variables

In this section, we present the experiment framework and design of our proposed model. A publicly listed firm is regarded to encounter business crisis and turns into a distressed company when declared for any one of the following conditions: full-value delivery, stock transaction suspension, re-construction, bankruptcy or withdrawal from the stock market. Based on the above criteria, we selected 120 distressed and 120 non-distressed (as matched samples) companies from TEJ database range year 2000 to 2008.

TEJ financial database for general industry is divided into twelve categories: 1. Balance sheet (60 + financial accounts such as total asset, total debt, etc.). 2. Income Statement (40 + financial accounts such as operating costs, interest expense, etc.). 3. Earning distributions. 4. Cash flow statement (50 + financial accounts such as depreciation, etc.). 5. Related Party Sales. 6. Notes and supplementary. 7. Operating costs. 8. Manufacturing expenses. 9. Operating expenses. 10. Retirement pay. 11. Warrant and employee cost. 12. Financial ratios. For these 12 categories contains more than 5000 items. Those features might calculated from financial statements, or defined by experts, economy analysis, and computer science. This study employs Financial Ratios category as experiment feature set. 21 ratios in the literature feature set (see Section 2) and 74 in TEJ feature set (see Appendix A) were used as experiment variables. There is no need to delete the redundancy due to our methodology will automatically screen out the repeated variables.

4.2. Feature selection

Based on the feature recommended by prior scholars and TEJ databases, a number of variables are used to develop the diagnosis model. We adopted Iterative relief algorithm to calculate the weight of our financial variables. The I-RELIEF algorithms, as one of the first feature weighting methods that have a clearly defined objective function and can be solved through numerical analysis instead of combinatorial searching, provide a promising direction for more rigorous treatment of the feature weighting and selection problems (Sun, 2007). For those features with high correlation (correlation between two features > 0.9), our algorithm will remove one of those features which has lower weight.

As shown in Tables 3 and 4 of the features out of the 10 selected features are from prior literature feature set, and the other six T_{21}, T_{22}, T_{23}, T_{26}, T_{42} and T_{74} from the TEJ feature set, are newly adopted by our proposed models and was not be mentioned by prior scholars. We further analyze these features with their means and standard deviations as shown in Table 4.

From Table 4, it is clear that the selected features have significant differences between the distressed firms and non-distressed firms in their mean values. The lower the standard deviation, the higher stability values in features. In other words, the fluctuation will be steadier. Taking Debt ratio as example, we could tell the non-distressed firms' mean value is 40.35% and the distressed firms have mean value of 64.22%.

4.3. Selection of SVM kernels and parameters

This study conducts experiments with different kernel functions such as the linear, RBF, polynomial, and sigmoid. The selection of kernel and the corresponding parameter plays a crucial role in the prediction quality of the SVM-based models. However, there is no general guideline for this selection process. In general, the radial basis function (RBF) is suggested for SVM. The RBF kernel nonlinearly maps the samples into the high-dimensional space, so

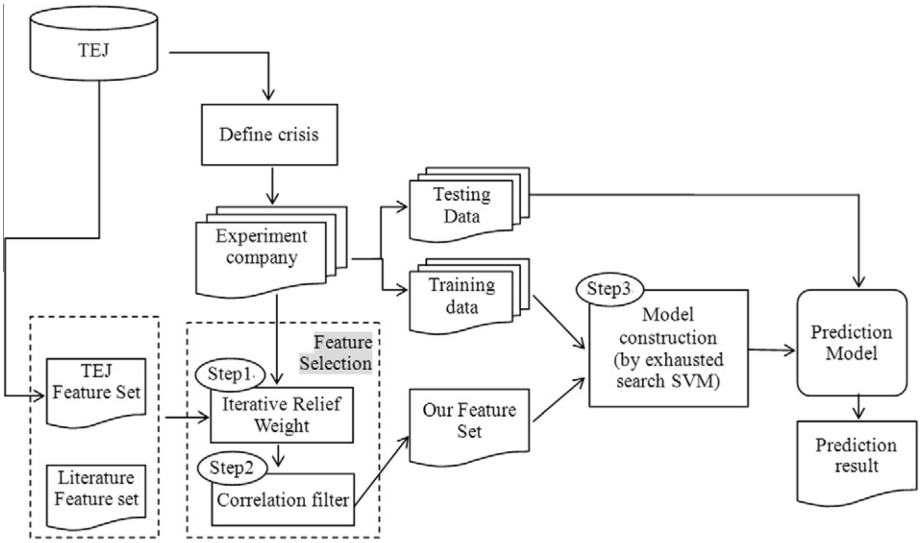

Fig. 3. Overall procedure of modeling.

Table 3
The features selected from both literature feature set and the TEJ feature set.

Features	Definition
Financial ratios selected from the literature feature set	
[X_5]	Debt ratio
[X_6]	Working capital/total asset
[X_{14}]	Net income/total asset
[X_{15}]	Retained Earnings/total asset
Financial ratios selected from the TEJ feature set	
[T_{21}]	Tax rates
[T_{22}]	Equity value per share
[T_{23}]	Continuous 4 quarterly EPS (earnings per share)
[T_{26}]	Operating earnings per share
[T_{42}]	Equity growth ratio
[T_{74}]	EPS

it can handle nonlinear problems. The linear kernel is a special case of the RBF where it has no parameter to determine except for C. on the other hand, the polynomial kernel has three parameters, i.e., C, γ, and d, to select, which causes higher complexity than RBF and Sigmoid, As suggested in literature (Ding et al., 2008; Huang et al., 2004), popular approaches in kernel selection are cross-validation via grid-search, heuristic search, and Bayesian inference. We apply grid-search with 10-fold cross-validation since we encounter a median-sized problem. Furthermore, the cross-validation procedure can prevent the prediction models from over fitting problem. In order to increase the searching efficiency, exponential increasements to the parameters pairs are used. Table 5 illustrates the grid-search approach to the RBF kernel as an example, the (C, γ) pairs are set as C ranges from 2^8, 2^9, 2^{10}, to 2^{11} and γ ranges from 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, to 2^{-1}. In Table 5, the optimal pair $C = 2^9$ and $\gamma = 2^{-3}$ is found as highest accuracy with the cross-validation rate of 85.1%.

We apply the same technique to each of the four kernels. After the optimal (C, γ) is found, the whole training data is trained using the SVMs with different kernels and the best parameters to generate the final models. In case of the RBF kernel, the prediction accuracy of the test data is turned out to be 86.2%, while that of the training data is 85.1%. Table 6 compares prediction performance of the SVM models using four different kernel functions. As shown in Table 6, the RBF kernel obtained the best prediction accuracy of test data (85.1%), followed by the polynomial kernel (83.7% when $d = 1$), the sigmoid kernel (83.7%), and the linear kernel (83.6%).

An Analyzing Parser is developed to process the financial statements retrieved from TEJ (Taiwan Economic Journal) databank. These data are used either as training data to construct the prediction model or as the testing data to validate the proposed model through SVM by using these optimal values. LIBSVM software (Chang & Lin, 2001) is utilized to construct the classification model and choose RBF as the kernel function.

The objective of this research is to investigate if the incorporation of financial features from TEJ feature set and literature feature set. Each of the steps is summarized as follows:

(i) Iterative Relief Weight is applied to select the features for our new model.
(ii) All the available features are ranking based on their weight and the correlation between features is then calculated. Correlation filter is used to remove the repeated features from the incorporated financial feature set. It will come out with our proposed feature set.
(iii) An SVM exhausted search is developed to code the features and to create training data based on the features determined in Steps 1 and 2. After performing the three steps as describe above, the training data are fed into the SVM tool to create the prediction models for our experiment. Finally, the testing data are prepared using the exhausted search in a manner similar to the one for training data in Step 3. Output with the highest accuracy rate will be equations.

5. Experiment results and discussion

5.1. Performance comparison of proposed feature and other scholars

To verify the efficiency and effective of our proposed diagnosis model, the prior scholars of Altman (1968), Beaver (1966), Zmijewski (1984) and Ohlson (1980) are used as benchmarks for comparison.

Model 1 is based exclusively on our selected financial features; Model 2 to Model 5 are based on feature selection result of Altman (1968), Beaver (1966), Zmijewski (1984) and Ohlson (1980), respectively. Firstly, Table 7 illustrates the statistical index of predictive accuracies on 10-fold datasets among different these models. Secondly, the comparisons of the diagnosis results between Model 1 and Model 5 are made by conducting the same SVM and I-Relief algorithm.

Table 4
Profile analysis – means and standard deviations by features.

Features		Firm type				Difference		*T*-test *p*-value ($p < 0.01$)
		Distressed firms		Non-distressed firms				
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	
X_5	Debt ratio %	64.22	17.78	40.35	16.30	−23.87	1.48	$1.32 * 10^{-22}$
X_6	Working capital/total asset	−0.02	0.23	0.19	0.20	−0.21	0.04	$8.63 * 10^{-13}$
X_{14}	Net income/total asset	−11.52	17.14	6.71	11.37	−18.23	5.77	$4.35 * 10^{-19}$
X_{15}	Retained earnings/total asset	−0.32	0.34	0.01	0.21	−0.33	0.14	$5.36 * 10^{-17}$
T_{21}	Tax rates	2.20	10.30	9.15	10.59	−6.96	−0.29	$5.03 * 10^{-07}$
T_{22}	Equity value per share	7.75	4.97	14.37	8.11	−6.62	−3.14	$5.47 * 10^{-13}$
T_{23}	Continuous 4 quarterly EPS	−3.05	2.98	1.04	3.17	−4.09	−0.20	$7.19 * 10^{-21}$
T_{26}	Operating earnings per share	−1.02	1.87	1.02	2.00	−2.04	−0.13	$1.61 * 10^{-14}$
T_{42}	Equity growth ratio	−30.55	35.55	7.31	32.07	−37.86	3.48	$6.14 * 10^{-16}$
T_{74}	EPS	−3.44	3.38	0.93	3.19	−4.37	0.19	$7.64 * 10^{-21}$

Table 5
The selection of the parameters pair (C, γ) on RBF kernel via grid-search and 10-fold cross validation.

C	γ				
	2^{-5}	2^{-4}	2^{-3}	2^{-2}	2^{-1}
2^8	83.8	84.2	84.2	84.2	82.6
2^9	84.6	83.8	*85.1*	83.8	80.9
2^{10}	84.2	83.4	84.6	82.6	80.9
2^{11}	83.8	84.6	84.2	81.3	79.7

Table 6
The performance of SVM kernels on each sub-optimal pairs (C, γ, d).

Kernel function	C	γ	d	Accuracy	
				Training	Testing
Linear	2^{11}	N/A	N/A	84.1	83.6
RBF	2^9	2^{-3}	N/A	86.2	85.1
Polynomial	2^4	2^{-3}	1	84.9	83.7
			2	84.9	80.9
			3	84.9	68.5
			4	84.9	46.1
			5	84.9	42.8
Sigmoid	2^{-2}	2^{-1}	N/A	84.3	83.7

Table 7
Statistical index of predictive accuracies on 10-fold datasets. Comparison among different models.

Statistical indices	Model 1	Model 2	Model 3	Model 4	Model 5
Minimum	75.0	62.5	62.5	78.26	58.3
Maximum	100.0	87.5	91.3	91.3	91.6
Mean	85.05	80.14	82.30	83.08	78.50
Median	83.33	86.96	86.96	79.17	82.6
S.D.	8.16	9.38	8.82	4.79	12.5

Different types of errors result in different penalty costs. As presented earlier, 120 distressed firms in the years of 2000–2008 are analyzed against 120 non-distressed counterparts. We first compare the prediction accuracy of the five models using the financial features only, one year prior to the bankruptcy of each distressed firm. This prediction is also known as the 1-year-ahead forecast (Ding et al., 2008).

As Table 7 demonstrates, 5 models represents mean values ranging from 58.3 and 78.26. The standard deviations are between 4.79 and 12.5. Model 5 seems to have lowest mean value and highest standard deviation.

In Model 1, we endeavor to examine the financial model known for its capability to solve classification problems in financial prediction and would like to discover any new financial features for better prediction in the future. Based on the best experiment on Model 1, X_5, X_6 and X_{15} features comes from prior scholars, and features T_{21} and T_{23} emerge from 74 TEJ financial predictors listed in Appendix A. The average accuracy of the 1-year-ahead forecast is 85.05% with Type I and Type II error rates being 10.76% and 20.6%, respectively. Type I error (misclassifying a distressed firm as a healthy one) appears more frequently than Type II error (misclassifying a healthy firm as a distressed one). These results are summarized in Table 8.

Model 2 examines Altman's (1968) features (X_6, X_7, X_{10}, X_{15} and X_{16}) to predict distressed firms with SVM. As summarized in Table 8, the average accuracy of the 1-year-ahead forecast in Model 1 is 85.05%, significantly superior to those of Model 2 (80.14%), Model 3 (82.30%), Model 4 (83.08%), and Model 5 (78.50%). Model 1 also performs better than the other four models in terms of Type I errors with an error rate of 10.76% and a Type II error rate of 20.6%. Compared to Model 2, Model 3, Model 4 and Model 5, Model 1 sustains an improved prediction performance thanks to its lower rate of Type I error. The prediction capability of various models for longer terms is discussed later.

For Model 3 and Model 4, the used of Beaver's (1966) (X_1, X_2, X_5, X_6, X_{14} and X_{17}) and Zmijewske's (1984) features (X_1, X_2 and X_{14}), are identified as the more accurate of all the adopted financial. The average accuracy for both Models reported as 82.3% and 83.08%, respectively. Compared with the other three models, Type I error occurs with a less frequency in Model 3 and type II error occurs with a less frequency in Model 4. In actual practice, the cost of misclassifying a failed firm into a healthy one (Type I error) is likely to be much greater than that of misclassifying a healthy firm into a failed one (Type II error). As indicated above, the Type I errors of Model 3 were much lower than those of Model 1, Model 2, Model 3 and Model 5. Empirical results indicate that Model 3 examining financial features can serve as a promising alternative for existing financial distress prediction models.

We further adopted Brier Score (BS) (1950) for comparison of prediction accuracy. The Brier Score (BS) is a measure of prediction accuracy well-known in meteorology and medical science. It is formulated as $[\mathrm{BS} = \frac{1}{n}\sum_i^n(\theta_i - 1)]$ where θ_i is a binary indicator for the actual realization of the default variable (1 if default, 0 if no default) and pi, is the estimated probability of default. The difference between the Brier Score and the percentage of correctly classified observations is that the former is more sensitive to the level of the estimated probabilities. The Brier Score takes the estimated probabilities directly into account. According to the results presented in Table 8, our proposed financial features (Model 1) achieves a lower average Brier Score (BS) of 14.95% after taking into consideration of all experiment results. Fig. 4 lists the average accuracy, type I error and type II error by using SVM.

Based on the outcomes in the first phase as shown in Fig. 4, the average accuracy for 1-year-ahead forecast of all five models

Table 8
Performance comparison among various SVM models.

Evaluation criterion	The proposed feature set (Model 1)	Altman (Model 2)	Beaver (Model 3)	Zmijewski (Model 4)	Ohlson (Model 5)
Type I error	10.76%	19.05%	10.68%	16.00%	21.21%
Type II error	20.60%	21.59%	25.03%	18.37%	21.89%
Brier Score (BS)	14.95%	19.86%	17.70%	16.92%	23.50%
Average accuracy	85.05%	80.14%	82.30%	83.08%	78.50%
Feature used	$[X_5][X_6][X_{15}][T_{21}][T_{23}]$	$[X_6][X_7][X_{10}][X_{15}][X_{16}]$	$[X_1][X_2][X_5][X_6][X_{14}][X_{17}]$	$[X_1][X_2][X_{14}]$	$[X_2][X_3][X_5][X_{13}][X_{14}][X_{17}][X_{18}][X_{19}][X_{20}][X_{21}]$

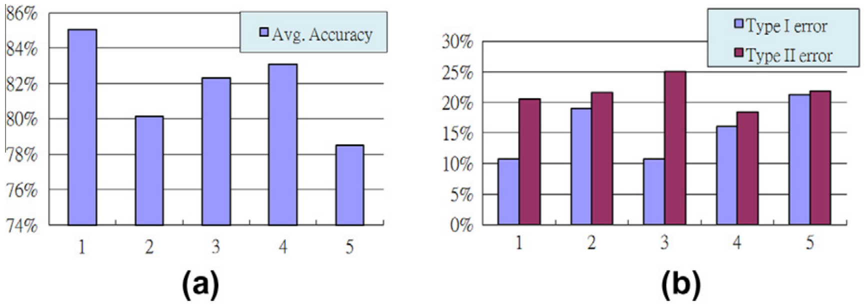

Fig. 4. Performance comparison of various models, (a) comparison average accuracy (b) Type I & Type II error.

Table 9
Statistical indices of predictive accuracies on 10-fold datasets.

Statistical indices	SVM	Logit	MDA	RBFN
Minimum	75.00	70.83	67.78	75.00
Maximum	100.0	100.00	100.00	95.83
Mean	85.05	83.82	84.23	82.99
Median	83.33	83.33	83.33	79.12
S.D.	8.16	7.99	9.80	6.69

Table 10
The best prediction accuracies of SVM, Logit, MDA and RBFN.

	SVM	Logit	MDA	RBFN
Training data (%)	87.0	83.8	84.2	83.2
Testing data (%)	85.1	83.8	84.2	83.0

falls in the range between 78.50% and 85.05%. The proposed feature set is able to predict bankruptcy one year ahead with an accuracy of 85.05%. Compared with other models, our feature set takes some features from TEJ (T_{21}: tax rates, T_{26}: continuous 4 quarterly EPS (earnings per share) into consideration and leads to an increase in average accuracy to 85.05%. This implies that tax benefits and continuous earnings per share deserve equal scrutiny in predicting financial distress. It is worth mentioning that the established model only uses five financial features and we found two value features which prior literatures could have ignored it. The data included in the above features can be obtained from publicly-available financial reports and TEJ. Therefore, combined consideration of both financial and non-financial features can be expected to greatly enhance the accuracy of a financial distress prediction model.

Therefore, features selected with our method from union of both popular feature set and TEJ feature set can be expected to enhance the accuracy of a financial distress prediction model. In addition, the BS value of our proposed model achieves the lowest average values compare to other 4 Models. In summary, our proposed model encompassing financial features can be expected to achieve a more accurate prediction of corporate financial distress than a model based exclusively on scholars' survey results from experts.

Table 11
McNemar value (*P*-value) for comparison of performance.

	Logit Regression	MDA
MDA	0.927563	X
SVM	0.54843	0.48875

$P < 0.1$.

Table 12
The 1-year ahead to 3-year ahead forecasts of Model 1 to Model 5.

	Model 1	Model 2	Model 3	Model 4	Model5
1-year-ahead forecast	85.1%	80.1%	82.3%	83.1%	78.5%
2-year-ahead forecast	74.4%	73.6%	74.4%	71.8%	74.9%
3-year-ahead forecast	68.1%	66.7%	70.4%	72.1%	67.2%

5.2. Performance comparison of SVM model against models based on various classifiers

For benchmark purpose, we conducted Logit, MDA and RBFN models as their SVM counterparts. As these models are built with the proposed feature set as shown in Tables 8 and 9 indicates the statistical description results. The standard deviations of the models are between 6.69 and 9.8.

The prediction accuracies of the 1-year-ahead forecast are summarized in Table 10, where the RBFN, MDA and Logit models consistently fall short of their SVM counterpart models. For example, SVM yields both 87.0% and 85.06% accuracy in training data and testing data that is the highest accuracy rate than others. Namely, SVM outperforms the models of MDA, Logit and RBFN. Therefore, we conclude that our proposed model appears to be the best model in prediction accuracy among the five models, whereas, the RBFN model seems to be the least desirable model.

Moreover, we conduct McNemar test to assess the significance of the difference between results of different models. As a nonparametric test for two related samples using the chi-square, McNemar test is useful for detecting before–after measurement of the same subject. As shown in Table 11, there is no significant different between the compared models which means the feature set we had selected is not only good for SVM model but also could be used for other classifiers with similar results. As Table 10 shows, SVM accuracy rate outperform other classifiers. Therefore, the proposed model can provide managers with an easy and effective way to diagnose crises in business units.

5.3. The analysis of predictive accuracy for longer-term forecast

We further conduct additional experiment to observe the effect of the prediction capability of these models for longer terms. Table 12 shows the results of applying these five models for 1-year-ahead forecast to 3-year-ahead prediction. Model 1 sustains an accuracy of 85.1% for 1-year-ahead forecast and 68.1% for 3-year-ahead forecast. The accuracies for 1-year-ahead and 3-year-ahead forecasts read respectively 80.1% and 66.7% for Model 2 and 82.3%, and 74.4% for Model 3. Our proposed model outperforms other Models for 1-year-ahead forecasts. It is clear that the 1-year data set performs better than the three year data set. This implied that the most recent year's data plays an important role in business crisis prediction. Moreover, as the results indicate, for longer-term forecast, Model 4 takes the lead in terms of predictive accuracy, followed respectively by Model 3 and Model 1.

For long-term forecasts, Model 4 is slightly higher than Model 3 and Model 1 in term of prediction accuracy. However, Model 4 focuses only on financial ratios related to a firm's business performance ($[X_1][X_2][X_{14}]$) whereas our proposed model adds on the tax rate and continuous EPS financial features concerning firms future development. For Model 1, even the long-term prediction reach only 68.1%, the average prediction accuracy is relatively high.

Model 4 using Current ratio, Cash flow/Total debt, and Net income/Total asset as critical feature to conduct financial crisis prediction which established good prediction accuracy for 3-years prediction. Extending the data period of financial variables from one to three years reduces the accuracy rate of our proposed model. This implies that the most recent year's financial data plays a major role in financial prediction. However, the mixed effect that multi-year data has on financial prediction model s requires further study.

6. Conclusion

In this paper, we consider a set of financial features that includes 21 commonly used financial ratios proposed in prior research, called the literature feature set, and 74 financial ratios from TEJ data base, called the TEJ feature set. We apply data mining techniques to identify five financial ratios, three from the literature feature set and two from the TEJ feature set that effective in identifying financial distressed firms.

We construct SVM prediction models based on the proposed feature set (Model 1), and four feature sets proposed by Altman (1968), Beaver (1966), Zmijewski (1984) and Ohlson (1980) (Model 2 to Model 5 respectively). Our experiments show that the proposed feature set (Model 1) outperforms other feature sets recommended in previous studies in terms of the prediction accuracy. Further analysis indicates that the proposed feature set performs well in predict models that are constructed by various classifier such as MDA, Logit and Neural Network; Though the SVM model yields better results prediction in all counts.

Based on the outcomes of feature comparison, the average accuracy for 1-year-ahead forecast of all five models falls in the range between 78.5% and 85.1%. The proposed our feature set is able to

Table A1
A list of financial features in TEJ.

No.	Feature	No.	Feature
T_1	ROA(C) before tax, interest and depreciation	T_{38}	Net income year over year
T_2	ROA(A) after tax, before interest	T_{39}	Ordinary income year over year
T_3	ROA(B) after tax, before interest and depreciation	T_{40}	Recurring income year over year
T_4	Return on equity% – after tax	T_{41}	Total assets year over year
T_5	Interest cover	T_{42}	Total equity year over year
T_6	Return on equity% – ordinary income	T_{43}	Depreciation FA year over year
T_7	Gross margin, %	T_{44}	Return on TA year over year
T_8	Yield of accomplished sales	T_{45}	C/F adequacy ratio
T_9	Operating income %	T_{46}	Cash reinvest %
T_{10}	Pre-tax income %	T_{47}	Current ratio
T_{11}	Net income %	T_{48}	Quick ratio
T_{12}	Net non-operating income/rev.	T_{49}	D/E ratio
T_{13}	Net income%-Exc Disp	T_{50}	Debt ratio
T_{14}	Operating expenses %	T_{51}	Equity
T_{15}	Employee fee %	T_{52}	(L-T Debt + SE)/FA %
T_{16}	R&D %	T_{53}	Debt/equity %
T_{17}	Bad debt /revenue	T_{54}	Contingent debt %
T_{18}	CFO/CL %	T_{55}	TCRI
T_{19}	Inventory expenses/debt	T_{56}	Inventory &A-R
T_{20}	Interest expenses %	T_{57}	Total asset turnover
T_{21}	Tax rates	T_{58}	A/R&N/R turnover
T_{22}	Equity value per share	T_{59}	Days-A/R turnover
T_{23}	Continuous 4 quarterly EPS	T_{60}	Inventory turnover
T_{24}	Cashflow per share	T_{61}	Days-inventory turn
T_{25}	Sales per share	T_{62}	Fixed asset turnover
T_{26}	Operating earnings per share	T_{63}	Equity turnover
T_{27}	Pre_tax income per share	T_{64}	Days-A/P turnover
T_{28}	Operating income	T_{65}	Net operating cycle
T_{29}	Pre_tax income	T_{66}	Degree of operating lever
T_{30}	PER	T_{67}	Degree of financial lever
T_{31}	PBR	T_{68}	Sales per employee
T_{32}	Retention ratio	T_{69}	Operation income/employee
T_{33}	Sales year over year	T_{70}	Fixed assets/employee
T_{34}	% Gross margin growth	T_{71}	Period
T_{35}	Yield of accomplished sales year over year	T_{72}	Yield of dividend
T_{36}	Operating income year over year	T_{73}	Yield of cash
T_{37}	Pre-Tax Income year over year	T_{74}	EPS

predict bankruptcy one year ahead with an accuracy of 85.1%. Compared with other models, our feature set takes some new features from TEJ (T_{21}: tax rates, T_{26}: continuous 4 quarterly EPS (earnings per share) into account and leads to an increase in average accuracy to 85.1%. This implies that tax benefits and continuous earnings per share deserve equal scrutiny in predicting financial distress. It is worth mentioning that the established model only uses five financial features and we found two value features which prior literature could have ignored it.

In summary, our proposed model encompassing financial features can be expected to achieve a more accurate prediction of corporate financial distress than a model based exclusively on scholars' survey results from experts.

There are, on the other hand, limitations in this article that call for further researches. Our models are inevitably affected by several factors. The predictive accuracy might be further improved in the future by considering to pair sampled companies by industry or to extend the survey period. This exclusive focus on corporate governance-related factors has prevented us from considering in our present study other potentially influential non-financial features, such as market share, management style, and industry prospect. Further researches may be conducted to explore such potential non-financial indicators.

Appendix A. Appendix

See Table A1.

References

Altman, E. I. (1968). Financial ratio, discriminant analysis, and the prediction of corporate bankruptcy. *Journal of Finance, 23*, 589–609.

Beaver, W. (1966). Financial ratios as predictors of failure. Empirical research in accounting: Selected studies. *Journal of Accounting Research, 4*, 71–111.

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research, 12*(1), 1–25.

Boster, B., Guyon, I., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152).

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1–3.

Chandra, D. K., Ravi, V., & Bose, I. (2009). Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications, 36*, 4830–4837.

Chang, C. C., & Lin, C. J. (2001). *LIBSVM: A library for support vector machines*.

Chen, L., & Hsiao, H. (2008). Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study. *Expert Systems with Applications, 35*(3), 1145–1155.

Chuvakhin, N., & Gertmenian, L. W. (2003). Predicting bankruptcy in the WorldCom age. *Journal of Contemporary Business Practice, 6*(1).

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research, 10*(1), 167–179.

Ding, Y., Song, X., & Zen, Y. (2008). Forecasting financial condition of Chinese listed companies based on support vector machine. *Expert Systems with Applications, 34*(4), 3081–3089.

Eichengreen, B. (1999). Is greater private sector burden sharing impossible? *Finance and Development, 36*(3), 16–19.

Günther, T., & Grüning, M. (2000). Einsatz von Insolvenzprognoseverfahren bei der Kreditwürdigkeitsprüfung im Firmenkundenbereich. *Die Betriebswirtschaft, 60*, 39–59.

Hua, Z., Wang, Y., Xu, X., Zhang, B., & Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications, 33*(2), 434–440.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems, 37*(4), 543–558.

Jo, H., & Han, I. (1996). Integration of case-based forecasting neural network and discriminant analysis for bankruptcy prediction. *Expert Systems with applications, 11*(4), 415–422.

Lee, K. C., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems, 18*(1), 63–72.

Li, H., & Sun, J. (2008). Ranking-order case-based reasoning for financial distress prediction. *Knowledge-Based Systems, 21*(8), 868–878.

Li, H., & Sun, J. (2009). Predicting financial failure using multiple case-based reasoning combine with support vector machine. *Expert Systems with Applications, 36*(6), 10085–10096.

Li, H., Sun, J., & Sun, B. L. (2009). Financial distress prediction based on OR-CBR in the principle of *k*-nearest neighbors. *Expert Systems with Applications, 36*(1), 643–659.

Martens, D., Bruynseels, L., Baesens, B., Willekens, M., & Vanthienen, J. (2008). Predicting going concern opinion with data mining. *Decision Support Systems, 45*(4), 765–777.

Min, S. H., Lee, J., & Han, I. (2006). Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert Systems with Applications, 31*(3), 652–660.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109–131.

Ozkan-Gunay, E. N., & Ozkan, M. (2007). Prediction of bank failures in emerging financial markets: An ANN approach. *The Journal of Risk Finance, 8*(5), 465–480.

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Application, 28*(1), 127–135.

Sun, J., & Hui, X. F. (2006). Financial distress prediction based on similarity weighted voting CBR. *Lecture Notes in Artificial Intelligence, 4093*, 947–958.

Sun, Y. (2007). Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), 1035–1051.

Taiwan Economic Journal. Available at: http://www.tej.com.tw/.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science, 38*(7), 926–947.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY: Springer-Verlag.

Wu, C. H., Tzeng, G. H., Goo, Y. J., & Fang, W. C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications, 32*(2), 397–408.

Zhao, H., Sinha, A., & Ge, W. (2009). Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications, 36*(2), 2633–2644.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research, 22*, 59–82.