International Journal of Innovative
Computing, Information and Control
Volume **9**, Number **6**, June **2013**

 ISSN 1349-4198
pp. **2245–2260**

A HYBRID SYSTEM BY THE INTEGRATION OF CASE-BASED REASONING WITH SUPPORT VECTOR MACHINE FOR PREDICTION OF FINANCIAL CRISIS

Pei-Chann Chang[1,*], Chiung-Hua Huang[2,3] and Chi-Yang Tsai[2]

[1]Department of Information Management
[2]Department of Industrial Engineering and Management
Yuan-Ze University
No. 135, Yuan-Tung Road, Chungli, Taoyuan 32023, Taiwan
[*]Corresponding author: iepchang@saturn.yzu.edu.tw; s968906@mail.yzu.edu.tw
iecytsai@saturn.yzu.edu.tw

[3]Department of Industrial Engineering and Management
Ta-Hwa University of Science and Technology
No. 1, Ta-Hwa Road, Hsinchu 30740, Taiwan
hch@tust.edu.tw

Received March 2012; revised July 2012

Abstract. *The prediction of business crises is an important academic topic of which many have used artificial intelligence methods to build an early warning system for this purpose. The objective of this study is to enhance the accuracy in predicting business crises by proposing an innovative model that combines financial variables with a system that integrates Case Based Reasoning (CBR) model with a Support Vector Machine (SVM) technique. This study is divided into three major steps: First, Stepwise Regression Analysis (SRA) is applied to the input set in selection of the most important factors; second, Case Based Reasoning (CBR), a clustering method, is employed to separate the case library into smaller clusters; and lastly, a Support Vector Machine (SVM) model is established and prediction results are being generated. In comparison with other methods, the proposed CBR-SVM model outperforms other prediction models as the prediction accuracy of business crises are being enhanced while it simultaneously produces valuable information for business owners and investors.*

Keywords: Case based reasoning, Support vector machine, Early financial warning system, Crisis prediction

1. **Introduction.** In recent years, many companies have filed for Chapter 11 bankruptcy, both domestically and overseas. Banks, brokerage firms, and other enterprises, in particular, have been greatly devastated by the subprime bust and the financial tsunami, all of which have caused great impact on the global financial market. There are many ways to access financial information for public companies. The Taiwan Stock Exchange (TSEC) posts financial reports and important announcements of public companies on the Market Observation Post System, which is updated on a regular basis. In addition, all listed companies are required to submit quarterly financial information to Taiwan's Financial Supervisory Commission (FSC). The information required for each submission includes five major performance indicators – financial structure, solvency, operating ability, profitability and cash flow, as well as twenty other financial ratios for investors. However, it is not easy to interpret information found in financial ratios/financial statements, nor is it a simple task to identify potential crises of a company. That being the case, this research is seeking to build a warning system for corporate financial distress. Using the techniques

proposed hereafter, translate financial statements published by companies into a real time and efficient predictor for its operations. Through this predictor, investors, creditors, bond holders and regulatory agencies would be able to detect any potential issues enterprise's underlying finance and governance and take appropriate responsive actions at an early stage. Looking at the current European debt crisis, if financial information of the EU counties had been analyzed or looked into at an earlier stage, crisis response would have been better prepared for and planned ahead.

Topics around financial crises and corporate bankruptcies have been the favorite subject for many academic researchers over the years [1]. However, early prediction methods that were generally based on statistics analysis need to establish relationships among numerous variables and tend to use complex mathematics and regression models, making it extremely time-consuming. In recent studies, Artificial Intelligence (AI) has gained popularity [2,3] as it is able to adjust model parameters in response to changes in the external environment with the help of fast computer arithmetic. Meanwhile, it is also able to reduce operating and data processing time for users to extract valuable information and make decisions more quickly. Artificial neural network-based heuristic algorithms, or other relevant prediction theories, are among the most widely used AI methods. Despite that this method produces faster, more accurate prediction results, the artificial neural network, or heuristic algorithm, is more difficult to understand than the traditional statistical methods. Support Vector Machine (SVM) [4] is a classification tool based on statistical learning theory that specializes on minimizing errors through experiences while maximizing generalization performance. In recent years, Case Based Reasoning (CBR), a clustering technique which develops behavioral learning theory which analyzes and makes predictions based on past experiences and data, has become one of the newer Data Mining prediction methods [5].

This study combines the CBR and a Support Vector Machine (SVM) method to group highly similar data and develops a financial crisis prediction model that is able to take suitable precautionary measures to avoid issues relating to information asymmetry, and that aims to reduce investment risks and potential losses. The steps taken in this study are as follows: 1) Select key financial indicators, 2) Develop effective financial crisis prediction model, and 3) Practical applications.

2. **Literature Review.**

Many academic researches have used a wide variety of approaches, from adjusting financial reports and using statistical methods, to searching for effective financial variables and ratios that can predict financial crisis. Beaver's study, which proposed the use of proxy variables to predict business failure [6], is among the earliest in this field. Altman was the first to use financial ratios to predict business crises [7]. Various methods were being used by researchers for establishing financial crisis prediction models; statistical methods and AI are the two major analysis tools. With respect to statistical methods, single variable analysis and multiple discriminant analysis were being frequently adopted in earlier studies. As statistic methods develops, regression analyses, like linear probability, Probit, Logit and the alike, have become more popular. Since AI was introduced in 1956 as a new analysis tool, its main purpose has been to let computers possess human-like intelligence and develop personified knowledge, arithmetic, reasoning and self-learning abilities. The field of AI includes artificial neural network, machine learning, fuzzy logic, fuzzy support vector machine [8], data mining and others. In the 1990s, data mining techniques had been used by many scholars in predicting business financial crisis, while its development has proved that it can provide more accurate predictions than that produced by traditional statistical analysis methods.

Previous researches were based on statistical theories [9,11] to develop warning systems for financial crisis. Beaver is one of the earliest scholars to evaluate financial ratios for the purpose of measuring a firm's financial distress. He used financial ratios to build a single variable prediction model, and then examined the prediction power of those ratios using a dual classification scheme. Research samples were taken from 158 companies in the US during 1954 to 1964. Then he selected 79 companies as "failed" firms and paired those with 79 healthy firms of same industry and with similar asset size, to reduce possible distortions caused by these factors. Based on the financial reports of the "failed" firms for the 5 consecutive years prior to bankruptcy, along with the financial reports of the paired healthy firms in the same years, 30 financial ratios were analyzed and classified into 6 categories. Later, the samples were separated into two groups: one being the training sample group that minimized the probability of faulty determination and the other being the test sample that calculated the distribution of the errors. The results indicated there were three ratios with better predictability, which are (1) Cash flow/total liabilities, (2) Net income after tax/total assets, and (3) Total liabilities/total assets. These three financial ratios showed signs of stability for the healthy firms and would alter when the companies faced financial distress. This method has great influence on researches afterwards, and is commonly adopted in relevant researches, as well as in this research.

Altman is the first to adopt Multivariate Analysis in 1968, using Linear Multivariate Discriminant Analysis for predicting firm's financial failure. Research samples were taken from 66 US manufacturing companies between 1946 and 1965. Within these samples, 33 bankrupt companies were paired with 33 healthy companies of similar size. 22 financial ratios were categorized into five groups, using Stepwise Regression Analysis (SRA). Five financial ratios generated better predictability, which included (1) Working capital/total assets, (2) Retain earning/total assets, (3) Income before interest and taxes/total assets, (4) Common stock equity/total liability, and (5) Net sales/total assets, forming the Z-score formula for predicting bankruptcy. The accuracy of this model in predicting 5 years in advance of the financial distress were: 95% for one year in advance, 75% for two years in advance, 48% for three years in advance, 29% for four years in advance, and 36% for five years in advance. Evidently, this model offered better predicting power on two years before crises occurred.

Case Based Reasoning (CBR) is one of the most applicable and extensively discussed methods in the AI field and is where a lot of attention is being drawn on. CBR originates from the cognitive process by which we search for similar past or incidents to solve current issues. It was initially proposed in 1982 by Schank as Dynamic Memory with the use of Memory-based reasoning system [11].

Jo and Han used three tools: CBR, discriminant analysis and artificial neural network to predict bankruptcy of banks [12]. Their research included four steps: Training, Testing, Adjusting, and Result Predicting. It concludes that the integration of the techniques produced better results than using one technique a time. With input data of training, testing and induction, it predicts better by integrating the three tools than just using one. Ahn and Kim used a hybrid case-based reasoning and genetic programming technique to build an effective corporate bankruptcy prediction model, and had proved that CBR may improve the prediction accuracy significantly [13].

Shiu et al. [14] adopted the fuzzy decision tree to conduct CBR through dividing a large-scale database into several smaller databases and finding suitable clusters. Four phases were involved in this process: the first phase was using gradient decent technique to find each characteristic's weighting, the second phase was dividing the database into several smaller ones using calculations for similarity matrix to find the equivalent matrix, the

third phase was utilizing Fuzzy Decision Tree to find the right theorem, and the last phase was to choose the suitable database through execution of the third phase. This research tested on two subjects, being the rice taste (RT) and Boston housing (BH) respectively, and achieved 90% accuracy. Shiu et al. proposed a clustering model by applying a data matrix to finding similarly weighted data and adopted the gradient method to find the clustering groups. Unlike other traditional clustering methods (K-means [15], Fuzzy C means, SOM), this case based reasoning weighted model (CBR weighted model) was able to find weighted data automatically [16,17]. Cao et al. [18] applied Fuzzy Rough in a distributed CBR system to selecting suitable cases, and connected the Client End (smaller database) to the Server End (full database). This procedure contained four steps: the first step was computing each learning feature weights in the database, the second step was separating the database based on the feature weight results from step one, the third step was using Fuzzy Rough to select suitable Fuzzy Rule, and the last step was selecting the suitable database by using similarities between the cases.

In our previous researches, we have proven that CBR can be integrated with multiple classification models, e.g., decision tree [19] and self-organizing map (SOM) [20], to improve its performance. In this research, CBR is applied to SVM algorithm to increase the prediction accuracy of early warning system for financial crisis. Our results indicate that the proposed hybrid approach outperforms results from other models.

When compared with traditional approaches, Artificial Intelligence approaches have been proven to be less vulnerable to some restrictive assumptions. With an inductive learning mechanism, Neural networks (NN) can serve as an alternative for classification problems of which traditional statistical methods have long been applied in the past. NN have shown to have better predictive power than other statistical methods or rule-based systems in predicting probability of business failure [21-23].

Recently, support vector machine (SVM), which was initially developed by Vapnik [24], has gained popularity due to its many attractive features and excellent performance in generalizing big range of problems. SVM is being widely used as a classification tool and a new machine learning method that is based on the statistical learning theory [25-27]. Furthermore, SVM applies the structural risk minimization (SRM) principle which has been shown to be superior to traditional empirical risk minimization (ERM) principle of conventional neural networks [28]. The study concludes that SVM outperforms NN, MDA and logistic regression in predicting business failure. SVM has advantage in terms of its learning ability, performance and feasibility. Moreover, SVM is able to generate good decision rules with limited training samples and produces fewer errors, making it a frequently adopted approach to deal with classifications of small samples, non-liner and High Dimensional Recognition problems [29].

Shin, Lee and Kim have conducted research on the predictability and stability of SVM and Back-Propagation Neural Network (BPNN) [30] models. The sample of their research was taken from 2320 Korea Credit Guarantee Fund enterprises of during 1996 and 1999, including 1160 healthy firms and 1160 crisis firms. In comparing SVM with BPNN, the ratio design for research samples was divided into 5 training samples: 80%, 20%, 10%, 5% and 2.5%, with the testing sample set as 20% of the full-size sample. As the ratio of training sample decreases, the ability to make predictions was compared against the predictability of the testing sample to determine the stability of the models. The research concludes that the predictability of SVM is better than BPNN, and this is also true when the training sample ratio decreases.

SVM has been found to have a wide range of applications in the field of pattern recognition, bioinformatics, and other relevant artificial intelligence researches. Along with the introduction of Vapnik's ε-insensitive loss function, SVMs have also been extended to

make estimates on nonlinear regression problems, namely the SVR. SVM has been successfully employed to in forecasting problems in many fields such as classifications [31], financial time series (stocks index and exchange rate) forecasting [32,33], mobile agent, diseases diagnosis [34] and speech recognitions [35].

Through literalistic review, data clustering has been found to be a crucial part of data mining and information retrieval. Case-based reasoning has been utilized in various fields such as product design [36] and stock pricing [37], but has never been applied to prediting financial distress models. This research is the first to employ CBR to the clustering financial data and to have it combined with Taguchi Experimental Design for selecting best operation levels, as well as from data means of Larger the Better (LTB) that generated the optimal parameters of CBR. In the study, we try to improve clustering efficiency though the use of Case-Based Reasoning clustering method to find the best clustering strategy. As a result, the prediction accuracies of our CBR-SVM and CBR-BPN models are superior to RST-SVM and RST-BPN models from Yeh's research [38].

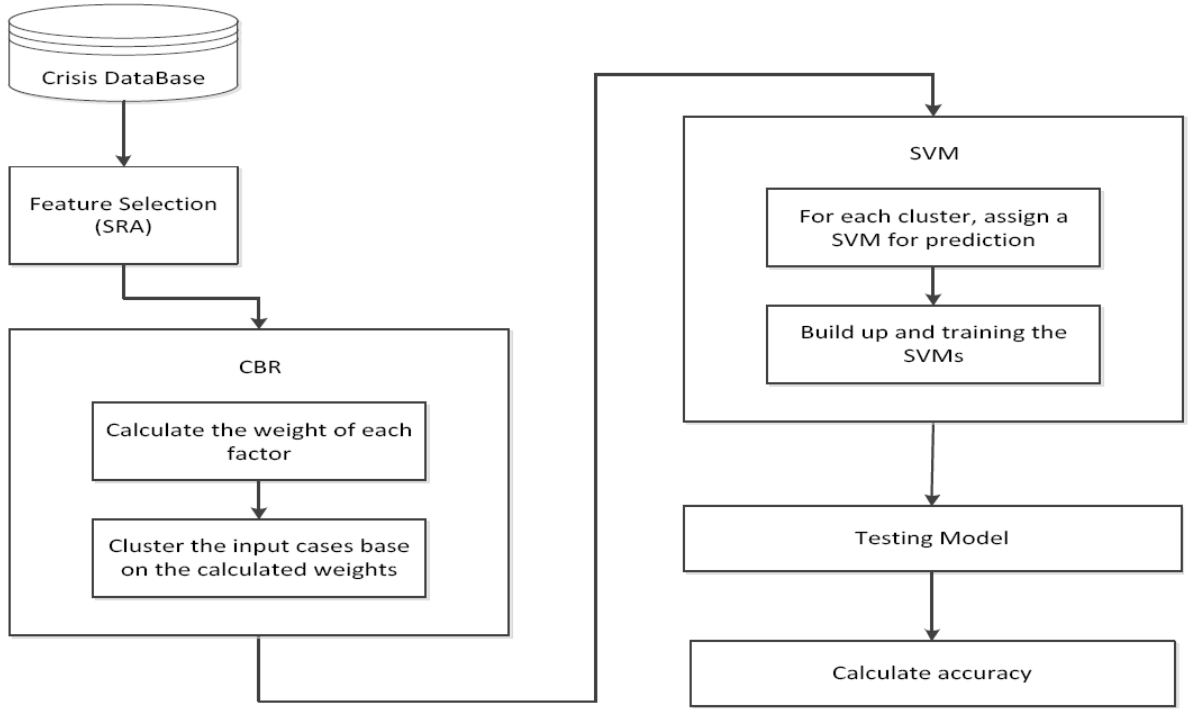

FIGURE 1. The flow chart of overall proposed framework

3. **Development of CBR-SVM System.** The objective of this research is to develop an enterprise financial distress warning model through the combination of CBR and SVM. The research flow chart is shown in Figure 1, which depicts the three major stages: (i) raw data is processed using a SRA technique to select the most significant factors and increase the similarity of the data as the prediction base, (ii) CBR is used to determine the weighting for each factor and to separate similar cases to form smaller groups so to increase the accuracy for clustering, and (iii) SVR is combined to classify different clustered cases to establish accurate predictions of enterprise financial crises. The chart for the development of CBR-SVM system is shown in Figure 2.

3.1. **Stepwise regression analysis.** Stepwise Regression Analysis (SRA) is mainly used for selecting the most effective variables or factors. Comparing with other factor screening

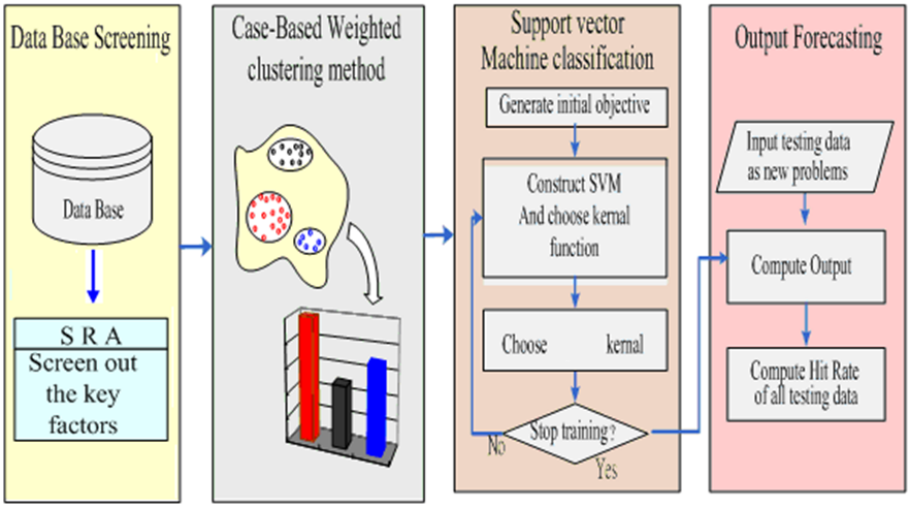

FIGURE 2. The development of CBR-SVM system

methods, SRA is simpler in terms of computation needs and can be more easily understood. The main approaches are forward selection or backward elimination to select the most suitable combination of variables for analyzing financial ratios, using Statistical F-test and Sum of Squares for Error (SSE) to determine whether the variables are statistically significant. In SRA, it begins with the first variable that comes into the model, and gradually increases the number of variables. However, once a variable is eliminated, it cannot be accepted into the model again. Before selecting variables, the first step is to determine the critical points, and the level of significance F_e (F-to-Enter) and F_r (F-to-Remove). Then, the Partial F is calculated and compare it against F_e and F_r; if $F > F_e$, then the variable is added, and if $F < F_r$, the variable is eliminated.

The variables for this research are collected through methods explained above. In view of the large number of variables involved, and in an effort to avoid the effect of certain weak variables or noise due to interactions between variable that may lead to a decreased accuracy of the model, SRA is being adopted in this research to select the most relevant variables associated with enterprise financial crises as input variables to construct the mode.

3.2. **Case-based reasoning.** After input variables are selected through SRA, CBR is applied to compute the weightings for these variables, and then cluster the cases into groups. There are two phases in this reasoning:

First Phase: Find the minimum evaluation value $E(w)$ by taking the following steps.

Step 1: Calculate Weight Distance Matrix. The initial value is produced randomly.

$$d_{pq}^{(w)} = d^{(w)}(e_p, e_q) = \left(\sum_{j=1}^{N} w_j^2 (x_{pj} - x_{qj})^2\right)^{1/2} = \left(\sum_{j=1}^{N} w_j^2 \chi_j^2\right)^{1/2}$$
$$w_j : \text{ weight}, w_j \in [0,1), j\ (1 \le j \le n) \tag{1}$$
$$N : \text{ Total number of cases}$$
$$e_p, e_q : \text{ case } p, q,$$
$$n : \text{ All of the important factors}$$

Step 2: After calculating $d_{pq}^{(w)}$ for each case first, then calculate the Similar Matrix $SM_{pq}^{(w)}$.

$$\begin{aligned} SM_{pq}^{(w)} &= \frac{1}{1+\alpha \cdot d_{pq}^{(w)}} \\ p &= 1, \cdots, N, q < p \\ \alpha &: \text{ User defined.} \end{aligned} \tag{2}$$

Step 3: Calculate $E(w)$

$$E(w) = \frac{2 \cdot \left[\sum_{pq(q<p)} \sum \left(SM_{pq}^{(w)} \right) \left(1 - SM_{pq}^{(1)} \right) + SM_{pq}^{(1)} \left(1 - SM_{pq}^{(w)} \right) \right]}{N \cdot (N-1)} \tag{3}$$

$SM_{pq}^{(1)}$: Weight equals 1 for each important variable.

Step 4: Use gradient decent technique to change the weighting of Δw_j in order to minimize the value $E(w)$.

$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j} \tag{4}$$

η: Learning rate, user defined.

$$\frac{\partial E(w)}{\partial w_j} = \frac{2 \left[\sum_{pq(q<p)} \sum \left(1 - 2 \times SM_{pq}^{(1)} \right) \times \frac{\partial SM_{pq}^{(w)}}{\partial d_{pq}^{w}} \times \frac{\partial d_{pq}^{(w)}}{\partial w_j} \right]}{N(N-1)} \tag{5}$$

$$\frac{\partial SM_{pq}^{(w)}}{\partial d_{pq}^{w}} = \frac{-\alpha}{\left(1 + \alpha \times d_{pq}^{(w)} \right)^2} \tag{6}$$

$$\frac{\partial d_{pq}^{(w)}}{\partial w_j} = \frac{w_j (x_{pj} - x_{qj})^2}{\left(\sum_{j=1}^{n} \left(w_j^2 (x_{pj} - x_{qj})^2 \right) \right)^{1/2}} \tag{7}$$

Second Phase: Find the best clusters for the case.

Step 1: Give a threshold value $\beta \in (0, 1)$.

Step 2: Let $SM = SM_{pq}^{w}$.

Step 3: Calculate $SM1 = SM$, $SM = s_{pq}$

$$s_{pq} = \max_k \left(\min \left(SM_{pk}^{w}, SM_{kq}^{w} \right) \right) \tag{8}$$

Step 4: When $SM1 \subset SM$ then go on to step 5; if not, let $SM = SM1$, and go back to step 3.

Step 5: Determine the clusters. If the quantity of Case p and Case q is equal while $s_{pq} \geq \beta$, then are put into the same group.

3.3. **Support vector machine in predicting model.** This research has obtained the classified results of each given piece of data: output value 1 represents crisis companies, and 0 represents healthy companies. By applying the clustered information obtained from CBR along with corresponding properties of each piece of data to SVM establishes the prediction model for each classified group individually. Finally, the model is examined for its accuracy and the ability to make predictions.

The purpose of SVM is to find the optimal margin hyper-plane of two classes with different distances among a clustered training data as prediction model, and decrease the computational errors caused by precision issues, thus making it more accurate in terms of clustering and enabling it to classify ungrouped data.

When the SVM classification technique is applied to predict enterprise financial crises, non-liner classification problems often occur as more financial ratios come with a lot of information and input variables as well. Therefore, kernel function mapping is needed for data conversion to make SVM classification more precise. Radial basis function (RBF) kernels are popular for classifying non-linear and high dimensional data. It is easier to operate while it possesses better predictability. The data is limited between $[0, 1]$, which can minimize the complexity and the time required for computation. As a result, RBF kernels are applied in this research at the initial stage to convert non-linear data to establish the prediction model of SVM.

$$K\left(x_i \cdot x_j\right) = e^{-\|x-y\|^2/2\sigma^2}, \text{ where } \sigma^2 \text{ is user defined.} \tag{9}$$

When non-linear data mapping are converted through the kernel function, the overlapping data on the boundary may cause errors, causing difficulties in finding the best Separating Hyperplane. To mitigate this issue, a deviation value ξ is needed to Support Hyperplane equations.

$$\begin{aligned} \left(w^T \cdot x\right) + b &> 1 - \xi_i \text{ if } y_i = +1 \\ \left(w^T \cdot x\right) + b &> -1 + \xi_i \text{ if } y_i = -1 \end{aligned} \tag{10}$$

Parameter $\xi_i \geq 0$ is the deviation value for the training data. The smaller the range of ξ_i the better. Based on the deviation of the whole mechanism, the cost of punishment for this mechanism model is defined as follows:

$$cost = C \sum_i \xi_i \tag{11}$$

C: The weighting set for the cost of in this mechanism model.

The larger weighting for C indicates fewer errors made in classifying the training data, and the smaller weighting for C indicates the larger maximum boarder M. Therefore, new equations are obtained after the linear SVM model is modified as below:

$$\text{minimize} \qquad \frac{1}{2}\|w\|^2 + C \sum_i \xi_i \tag{12}$$

$$y_i \left[\left(w^{\mathrm{T}} \cdot x\right) + b\right] - 1 + \xi_i \geq 0, \quad \forall i, \quad \xi_i \geq 0 \tag{13}$$

Use the Lagrange multiplier approach, add constraints to the previous equations and reduce it to a quadratic equation:

$$L\left(w, b, \xi, \alpha, \mu\right) = \frac{1}{2}\|w\|^2 + C \sum_i \xi_i - \sum_{i=1}^{N} \alpha_i \left[y_i \left(w^T x_i + b\right) - 1 + \xi\right] - \sum_{i=1}^{N} \mu_i \xi_i \tag{14}$$

By partial differentiation of w, b and ξ, the parameter that minimizes L is found, and then the parameter is brought into Equation (13) to obtain new Karush-Kuhn-Tucker

(KKT) Conditions to modify the conversion errors:

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{15}$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{16}$$

$$\frac{\partial L}{\partial \xi} = \ 0 \rightarrow C - \alpha_i - \mu_i = 0 \tag{17}$$

$$\text{Constraint: } y_i \left[\left(w^T \cdot x\right) + b\right] - 1 + \xi_i \geq 0 \tag{18}$$

$$\text{Lagrange multiplier condition: } \alpha_i \geq 0, \mu_i \geq 0 \tag{19}$$

$$\text{Complementary Slackness: } \alpha_i \left[y_i \left(w^T x_i + b\right) - 1 + \xi_i\right] = 0, \quad \mu_i \xi_i = 0 \tag{20}$$

To serve as the basis for SVM, training data is brought to the new KKT Conditions as proof for the Support Hyperplane, and through collecting these support vectors, value b is found, after which the best Separating Hyperplane with minimum error is calculated. Finally, unclassified data are being classified, assisting investors in making accurate predictions in the fluctuating stock market.

3.4. **Performance evaluation of prediction accuracy.**

After the prediction model is established, the test data are brought into SVM to calculate the Hit Ratio of this prediction model. The equations is defined as below:

$$\text{Hit Ratio} = \frac{\sum_{i=1}^{m} x_i}{n} \times 100\% \tag{21}$$

where

m: Clusters

x_i: Number of accurate prediction for the ith cluster

n: Number of test data

A higher ratio indicates better accuracy of the model, whereas a low ratio indicates that the model needs to be adjusted and the factors used in the model need to be re-evaluated, or it may also be the case that other prediction models or financial ratios are more appropriate.

4. **Research Data.**

4.1. **Data source and research samples.**

Samples of crisis companies in this research are mainly domestic public companies, and are represented by 1, whereas healthy companies are represented by 0. Based on prospectuses and financial statements provided by the Taiwan Stock Exchange (TWSE) as well as the database [39] from the Taiwan Economic Journal (TEJ) regarding listed companies and former listed companies, a crisis company is one in which trading method has changed to full cash delivery method.

From the TEJ database, the 40 crisis company samples were selected between the period of 1999 to 2009, and were divided into 8 groups: food industry (1 firm), textile fabric industry (3 firms), electrical machinery industry (2 firms), appliance and cable industry (1 firm), iron and steel industry (2 firms), electronics industry (25 firms); construction industry (4 firms) and a miscellaneous category (2 firms). The pairing ratios used in earlier researches have always been different; in order to represent the real situation, we have adopted the most popular pairing ratio of 1 : 2, meaning that for the 40 crisis companies we have 80 healthy companies for a total sample size of 120 companies. Crisis company

samples are paired with healthy ones within the same industry, with similar capital assets, of similar nature and comparative product types for the purpose of comparison. Samples are then divided into training and test samples, with 96 samples (32 crisis firms and 64 healthy firms) between 1999 to 2007 to build the prediction model, and the rest 24 samples (8 crisis firms and 16 healthy firms) between 2009 to 2009 set up as test samples to examine the accuracy in making predictions as well as the stability of the model. The table showing the number of training and testing samples is shown in Table 1.

TABLE 1. Numbers of training and testing samples

Training and testing firms	Crises	Non-crises	Sum
Training Samples	32	64	96
Testing Samples	8	16	24
Total firms	40	80	120

4.2. **Research variables.** The categorization of financial ratios in this research is consistent with typical analysis on financial statements and prospectus for Taiwan public companies, which are divided into 6 areas including financial structure, solvency, operating ability, profitability, cash flow and growth. The selection of the 41 financial ratios used in this research are based on the following principles: first, the ratios have been used in previous relevant researches; second, the ratios were referenced from prospectuses of respective companies; and finally, several ratios were taken as averages to prevent distortion from the final data.

5. **Experimental Results.** The object of this research is to establish a prediction model for enterprise financial crises by combining CBR and SVM techniques. The model was programmed with Microsoft Visual C++ 2005, on an Intel Core 2 Duo 2.13GHz CUP with 2GB memory.

This section of the study examines the accuracy in predicting enterprise financial crises through the use of this CBR-SVM integrated model by applying data clustering techniques. First, we select the proper research objects, collect data on financial variable data, and select crucial factors as inputs for the prediction model. Next, we compute the results from each prediction and determine their accuracies, which would then be compared with those of K-Mean plus SVM methodology as well as traditional classification approaches to verify if our proposed model is superior in terms of accuracy for predicting enterprise financial crisis, thus assisting management and investors when making investment choices. The outcome indicates that this model indeed possesses better prediction accuracy.

5.1. **Selection factors by stepwise regression.** This research has chosen 41 financial ratios in 6 areas as research variables. Considering the large number of variables involved, we have preprocessed the collected data first, through SPSS 13 statistical software, and to avoid the unwanted interactions or noises that may occur in the training process, which would in turn reduce the model's explanation power and increase the number of possible errors. SRA has been applied to select the most relevant variables as input factors for establishing this research model, in hope that better prediction results can be achieved. Out of the financial ratios of 120 distress and healthy companies, SRA has selected variables based on the number of years prior to the crisis, respectively being one, two and three years prior. Table 2 below lists the chosen factors:

TABLE 2. Results of stepwise regression selection

Financial crisis	Financial Ratios
One year prior period	Equity-to-assets, Debt-to-equity, Receivable turnover, Asset turnover, Return on equity (ROE), Cash flows from operating activities-to-asset, Assets growth
Two years prior period	Equity-to-assets, Loan-to-equity, Receivable turnover, Return on equity (ROE), Return on fixed assets
Three years prior period	Equity-to-assets, Loan-to-equity, Receivable turnover, Return on equity (ROE), Return on fixed assets

5.2. **Setting important factors for CBR.** To generate better prediction accuracy, CBR method was employed to cluster similar cases. Microsoft Visual C++ 2005 software was used to input the selected variables into the written program set reasonable parameters and provide the right threshold value for data clustering. In order to achieve the optimal clustering result, 5 important factors were chosen based on the CBR theory for Taguchi Experimental Design [40], and with the help of statistical software MINTAB (14th Version) to find the optimal operation levels. The parameters levels are shown in Table 3.

TABLE 3. Parameters of CBR Taguchi experimental design

Parameter setting	Taguchi Experiment		
	I	II	III
a	0.6	0.8	1
Learning rate (r)	0.1	0.3	0.2
Phase-one running times	3000	1000	2000
Phase two running times	10	30	20
b	0.02	0.01	0.03

As shown in Figure 3 – Data Means of Taguchi Experimental Design for selecting the best operating levels, and from data means of Larger the Better (LTB), the optimal parameters would be 0.6 for α, 0.3 for the learning rate, 2000 phase one running times for Phase One and 20 for Phase Two, with threshold value of 0.02. These optimal parameter values are organized again in Table 4.

TABLE 4. Setup values to CBR optimal parameters

Parameter setting	Optimal values
a	0.6
Learning rate (r)	0.3
Phase-one running times	2000
Phase two running times	20
b	0.02

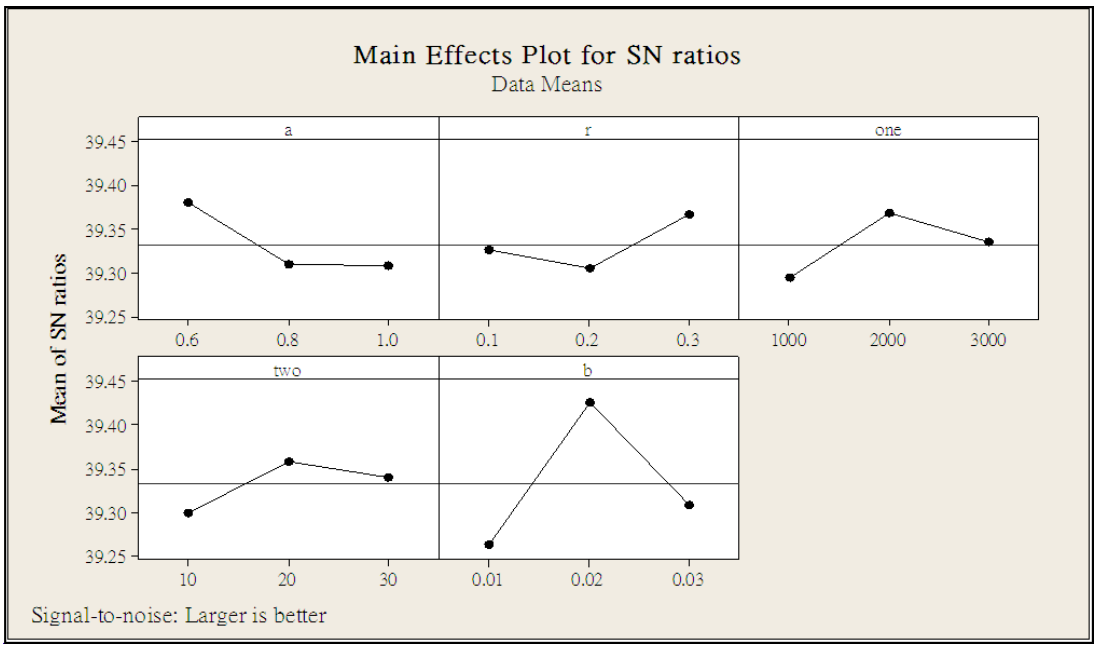

FIGURE 3. Data means of Taguchi experimental design

5.3. **Comparison between CBR-SVM and traditional clustering method.** This research has developed through the CBR-SVM methodology by combining these two methods: first, applying CBR to clustering financial data and then using SVM to classify crisis and healthy companies. As described above, the research obtained the best clusters through CBR parameters setup results to find the appropriate threshold value, and then used STATSTICA (Version 7.0) to input each clusters into SVM respectively. As summarized in Table 5, the result was that there were 4, 6, and 2 clusters for the first, second and the third year prior to the financial crisis. Then the prediction model was built, accuracy of the model was determined after computing the test data, and lastly, then compared against traditional clustering methods. Regarding the application of SVM classifiers, in consideration of the objectivity for comparison, RBF kernels were accepted as the basis for non-linear data conversion, and the value of the Gamma parameter was set as $1/k$, with k being the number of input variables; as with the Bayesian classification method, STATISTICA (Version 7.0) was also used for the classification of predictions.

TABLE 5. The optimal CBR-SVM clustering results

Financial crisis	One year prior	Two years prior	Three years prior
b	0.02		
Best number of clustering (Cases)	4	6	2

5.4. **Comparing CBR-SVM with KMEAN-SVM and SVM methods.** With regard to accuracy in predicting for financial crises with CBR-SVM, KMEAN-SVM and SVM, the proposed CBR-SVM model outperforms other prediction methods. A more detailed description regarding one, two, and three years before the financial crisis are as follows.

5.4.1. *One year prior to the financial crisis.* For the CBR-SVM method used this research, CBR clusters data into four groups with prediction accuracy being at 93.62%. For the

KMEAN-SVM method, KMEAN clusters data into six groups with prediction accuracy being at 91.59%. As for the SVM method without clustering, the prediction accuracy stands at 89.57%.

5.4.2. *Two years prior to the financial crisis.* For the CBR-SVM method used in this research, CBR clusters data into 6 groups with prediction accuracy being at 92.02%. For the KMEAN-SVM method, KMEAN clusters data into 5 groups with prediction accuracy being at 85.47%. As for the SVM method without clustering, the prediction accuracy stands at 82.05%.

5.4.3. *Three years prior to the financial crisis.* For the CBR-SVM method used in this research, CBR clusters data into 2 groups with prediction accuracy being at 84.05%. For the KMEAN-SVM method, KMEAN clusters data into 5 groups with prediction accuracy being at 78.63%. As for the SVM method without clustering, the prediction accuracy stands at 74.36%.

The above results indicate that the CBR-SVM model has better prediction accuracy than either the KMEAN-SVM or the SVM, while accuracy increases as it the timing or prediction gets closer to the occurrence of the financial crises. The result also clearly shows that clustering could increase the prediction accuracy.

5.5. **Comparison between the CBR-SVM and BAYES method.** With regard to accuracy in predicting for financial crises with CBR-SVM and BAYES, the proposed CBR-SVM model is better than the BAYES method. A more detailed description regarding one, two, and three years before the financial crisis are as follows.

5.5.1. *One year prior to the financial crisis.* For the CBR-SVM method used in this research, CBR clusters data into 4 groups with prediction accuracy being at 93.62%. As for the BAYES method without clustering, prediction accuracy stands at 90.43%.

5.5.2. *Two years prior to the financial crisis.* For the CBR-SVM method used in this research, CBR clusters data into two groups with prediction accuracy being at 92.02%. As for the BAYES method without clustering, prediction accuracy stands at 85.47%.

5.5.3. *Three years prior to the financial crisis.* For the CBR-SVM method used in this research, CRB clusters data into two groups with prediction accuracy being at 84.05%. As for the BAYES method without clustering, prediction accuracy stands at 42.74%.

As the results above indicate, the prediction result under the CBR-SVM model is more accurate than that of the traditional BAYES method. It is also obvious in this case that clustering could increase the prediction accuracy.

5.6. **Comparison between the CBR-SVM and benchmark model.** Yeh et al. [38] and this research in predicting financial crisis also drew samples from the database of Taiwan Economic Journal (TEJ) on public companies. The sample ratio between distressed and healthy companies was also 1:2, and the scope of the research concerned with data within one year prior to the occurrence of the financial crises. Through comparison between our research results from the CBR-SVM model and the CBR-BPN (Back-propagation Neural Network) model with the result from the benchmark research models presented by Yeh et al., which were (Rough Set Theory) RST-SVM and RST-BPN models, we have organized the prediction accuracies in Table 6.

The prediction accuracy of our CBR-SVM model stands at 93.62% while the accuracy for Yeh's RST-SVM model stands 6.78% less at 86.84%. The CBR-BPN model proposed in this research uses CBR to cluster data first, then uses BPN model from MATLAB (Version 7.6) to predict accuracy, ending up with an accuracy of 91.30% for this model,

TABLE 6. Comparison of the prediction accuracy of CBR-SVM/CBR-BPN with other models

Benchmark against	Hit ratios
CBR-SVM	93.62%
RST-SVM	86.84%
CBR-BPN	91.30%
RST-BPN	82.95%

which is 8.35% higher than the 82.95% accuracy rate of RST-BPN model from Yeh's research. The difference between these two models is the application of distinct clustering methods, and the result proves that CBR performance is better than RST in predicting financial crises.

From the analysis above, the prediction accuracies from this study are superior to other methods regardless of the time to the occurrence of the financial crisis. This study is based on relative characteristics of historical data to cluster similar cases from a mainframe database into several smaller ones for testing performance of CBR, which turns out to be more effective than the methods without clustering. Even if a longer period exists before the occurrence of the financial crises, prediction results under our model still remain relatively accurate, which indicates that better clustering method can indeed enhance the prediction accuracy.

6. **Conclusion.** This study is a combination between the CBR and SVM methods in establishing a model for the prediction of enterprise financial crisis. From the TEJ database, distressed companies between 1999 and 2009 were selected research samples and paired with the healthy companies in the same industry, of similar capital sizes, and of comparable product types on a pairing ratio of one to two. 41 financial ratio variables were chosen under 6 properties, including financial structure, solvency, operating ability, profitability, cash flow and growth. Through Stepwise Regression Analysis, the most effective financial ratios were discovered, and then the CBR model was used for clustering. Finally, the SVM method was adopted to determine if the companies were healthy or not, and the prediction accuracy was computed, which could be used in lowering investment risk in the future.

This study contains three major steps: the first was to use Stepwise Regression method to select optimal factors from the input set; the second was to determine the weighting of each factors for data clustering through CBR method in order to increase the clustering; and the third was to establish the SVM model to predict for financial crises and to compare the results with traditional clustering techniques and other research methods. In conclusion, the predictions from the proposed model were proven to be more accurate, and this superior prediction accuracy is a result of: 1) the use of Stepwise Regression for selecting optimal factors, 2) the use of CBR method for clustering, and 3) the effective use of SVM method.

This study has led directions for future researches, and proposes other aspects to be considered. 1) This research involved only financial ratios from the prospectus, yet many other factors were not taken into account, but could be added as inputs for future research analysis with different selection approaches. 2) This study shows clustering in advance can enhance prediction accuracy, leaving future researcher to go into deeper discussions regarding other clustering methods.

REFERENCES

[1] G. S. Ng, C. Quek and H. Jiang, FCMAC-EWS: A bank failure early warning system based on a novel localized pattern learning and semantically associative fuzzy neural network, *Expert Systems with Applications*, vol.34, pp.989-1003, 2008.

[2] W. L. Tung, C. Quek and P. Cheng, GenSo-EWS: A novel neural-fuzzy based early warning system for predicting bank failures, *Neural Networks*, vol.17, pp.567-587, 2004.

[3] L. H. Chen and H. D. Hsiao, Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study, *Expert Systems with Applications*, vol.35, pp.1145-1155, 2008.

[4] Y. E. Shao and B.-S. Hsu, Determining the contributors for a multivariate SPC chart signal using artificial neural networks and support vector machine, *International Journal of Innovative Computing, Information and Control*, vol.5, no.12(B), pp.4899-4906, 2009.

[5] P. C. Chang, Y. W. Wang and C. H. Liu, Combining SOM and GA-CBR for flow time prediction in semiconductor manufacturing factory, *Lecture Notes in Computer Science*, pp.767-775, 2006.

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

[7] E. I. Altman, Financial ratios, discriminated analysis, and the prediction of corporate bankruptcy, *Journal of Finance*, vol.23, pp.589-609, 1968.

[8] Y. Wang, S. Wang and K. K. Lai, A new fuzzy support vector machine to evaluate credit risk, *IEEE Transactions on Fuzzy Systems*, vol.13, pp.820-831, 2005.

[9] E. P. Davis and D. Karim, Comparing early warning systems for banking crises, *Journal of Financial Stability*, vol.4, pp.89-120, 2008.

[10] D. Reagle and D. Salvatore, Forecasting financial crises in emerging market economies, *Open Economies Review*, vol.11, pp.247-259, 2000.

[11] C. K. Riesbeck and R. C. Schank, *Inside Case-Based Reasoning*, Erlbaum, Northvale, NJ, 1989.

[12] H. Jo and I. Han, Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction, *Expert Systems with Applications*, vol.11, pp.415-422, 1996.

[13] H. Ahn and K. J. Kim, Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, *Applied Soft Computing Journal*, vol.9, pp.599-607, 2009.

[14] S. C. K. Shiu, C. H. Sun, X. Z. Wang and D. S. Yeung, Maintaining case-based reasoning systems using fuzzy decision trees, *LNAI*, pp.285-296, 2000.

[15] P. C. Chang, J. J. Lin and W. Y. Dzan, Forecasting of manufacturing cost in mobile phone products by case-based reasoning and artificial neural network models, *Journal of Intelligent Manufacturing*, pp.1-15, 2010.

[16] S. C. K. Shiu, D. S. Yeung, C. H. Sun and X. Z. Wang, Transferring case knowledge to adaptation knowledge: An approach for case-base maintenance, *Computational Intelligence*, vol.17, pp.295-314, 2001.

[17] P. C. Chang, J. J. Lin and C. H. Liu, An attribute weight assignment and particle swarm optimization algorithm for medical database classifications, *Computer Methods and Programs in Biomedicine*, pp.382-392, 2012.

[18] G. Cao, S. C. K. Shiu and X. Wang, A fuzzy-rough approach for the maintenance of distributed case-based reasoning systems, *Soft Computing*, vol.7, pp.491-499, 2003.

[19] P. C. Chang, C. Y. Fan and W. Y. Dzan, A CBR-based fuzzy decision tree approach for database classification, *Expert Systems with Applications*, vol.37, pp.214-225, 2010.

[20] P. C. Chang, C. Y. Fan and Y. W. Wang, Evolving CBR and data segmentation by SOM for flow time prediction in semiconductor manufacturing factory, *Journal of Intelligent Manufacturing*, vol.20, pp.421-429, 2009.

[21] G. Zhang, M. Y. Hu, B. E. Patuwo and D. C. Indro, Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis, *European Journal of Operational Research*, vol.116, pp.16-32, 1999.

[22] P. Ravi Kumar and V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review, *European Journal of Operational Research*, vol.180, pp.1-28, 2007.

[23] A. F. Atiya, Bankruptcy prediction for credit risk using neural networks: A survey and new results, *IEEE Transactions on Neural Networks*, vol.12, pp.929-935, 2001.

[24] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, vol.20, pp.273-297, 1995.

[25] X. Zhou, S. Chen, B. Liu, R. Zhang, Y. Wang, P. Li, Y. Guo, H. Zhang, Z. Gao and X. Yan, Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support, *Artificial Intelligence in Medicine*, vol.48, pp.139-152, 2010.
[26] R. G. Brereton and G. R. Lloyd, Support vector machines for classification and regression, *Analyst*, vol.135, pp.230-267, 2010.
[27] H.-Y. Wu, C.-Y. Hsu, T.-F. Lee and F.-M. Fang, Improved SVM and ANN in incipient fault diagnosis of power transformers using clonal selection algorithms, *International Journal of Innovative Computing, Information and Control*, vol.5, no.7, pp.1959-1974, 2009.
[28] J. H. Min and Y. C. Lee, Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, *Expert Systems with Applications*, vol.28, pp.603-614, 2005.
[29] P. C. Petrantonakis and L. J. Hadjileontiadis, Emotion recognition from EEG using higher order crossings, *IEEE Transactions on Information Technology in Biomedicine*, vol.14, pp.186-197, 2010.
[30] K. S. Shin, T. S. Lee and H. J. Kim, An application of support vector machines in bankruptcy prediction model, *Expert Systems with Applications*, vol.28, pp.127-135, 2005.
[31] Y. W. Wang, P. C. Chang, C. Y. Fan and C. H. Huang, Database classification by integrating a case-based reasoning and support vector machine for induction, *Journal of Circuits, Systems and Computers*, vol.19, pp.31-44, 2010.
[32] P. F. Pai and C. S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting, *Omega*, vol.33, pp.497-505, 2005.
[33] N. Sapankevych and R. Sankar, Time series prediction using support vector machines: A survey, *IEEE Computational Intelligence Magazine*, vol.4, pp.24-38, 2009.
[34] E. Çomak, A. Arslan and I. Tükoğlu, A decision support system based on support vector machines for diagnosis of the heart valve diseases, *Computers in Biology and Medicine*, vol.37, pp.21-27, 2007.
[35] M. Ferras, C. C. Leung, C. Barras and J. L. Gauvain, Comparison of speaker adaptation methods as feature extraction for SVM-based speaker recognition, *IEEE Transactions on Audio, Speech and Language Processing*, vol.18, pp.1366-1378, 2010.
[36] S. Jung, T. Lim and D. Kim, A method for optimal design of high-tech products using CBR and neural network, *International Journal of Innovative Computing, Information and Control*, vol.5, no.12(B), pp.4961-4969, 2009.
[37] P.-C. Chang, C.-Y. Fan and J.-J. Lin, A case based clustering-based TSK fuzzy rule systems for stock price forecasting, *The 3rd International Conference on Innovative Computing Information and Control*, Dalian, China, pp.279, 2008.
[38] C. C. Yeh, D. J. Chi and M. F. Hsu, A hybrid approach of DEA, rough set and support vector machines for business failure prediction, *Expert Systems with Applications*, vol.37, pp.1535-1541, 2010.
[39] *Taiwan Economic Journal Database*, http://www.finasia.biz/ensite/Default.aspx?TabId=121.
[40] N. Kwak and C. H. Choi, Input feature selection for classification problems, *IEEE Transactions on Neural Networks*, vol.13, pp.143-159, 2002.