Available online at http://www.mecs-press.net/ijeme

Research on Financial Distress Early-warning of Listed Companies Based on GA-SVM[①]

LI Yun-fei[a], ZHANG Qian[b]

[a] School of Economics and Management, Xi'an University of Technology, 710048, Xi'an, China
[a,b] School of management, Xi'an Polytechnic University, 710048, Xi'an, China

Abstract

Based on financial management and enterprises of early-warning theory, this paper constructs a financial distress early-warning model using GA-SVM. First, it uses listed companies appearing in Shanghai Stock Exchange and Shenzhen Stock Exchange in 2007-2009 as sample books. Defining ST listed companies which have abnormity of finance status as signature of the listed company's financial crisis. Then it uses the data in the financial statements known to the public as the input feature vector and combine genetic algorithm and support vector machine. Use Taking an empirical research with the financial distress early-warning model. Test results of the demonstration study shows the model has a superiority in predicting financial distress.

Index Terms: Financial Distress; Early-Warning; Genetic Algorithm; Support Vector Machine

1. Introduction

Financial distress early-warning system is a financial analysis system which can forecast financial distress for enterprises or enterprise groups by setting up some sensitivity financial indicators and observing its changes. Financial distress is the most integrative and significant representation of enterprise crisis. As continual perfecting of China's stock market mechanism and enterprise bankruptcy system, financial distress has made enterprises suffer great loss and directly affect the survival and development. Furthermore, a new financial crisis will probably occur if large number of enterprises gets into financial distress at the same time. Thus, how to effectively forecast financial distress of enterprises has become an urgent need research issues. On the current, study on financial crisis early-warning has become a hot academic research problem. In the early study, regression analysis, mathematical programming, fuzzy analysis, k-Nearest Neighbor (KNN) and expert system are used to predict financial distress. Since these are linear method in substance and their effectiveness depends

on restrictive assumptions, its applicability and effectiveness are limited. Recently a lot of research shows that Artificial Intelligence (AI) method is particularly outstanding in dealing with nonlinear complex system problems. In particular, Artificial Neural Networks (ANN) is the most widespread used method. For example, Back（1996）[1],Zhang（1999）[2],Yang（1999）[3] and Pendharkar（2005）[4] use different ANN model to predict finical distress and compare with the traditional method. The result shows that the effect of ANN is better than the traditional linear method. Although many theories and empirical studies have proved that ANN is more effective than the traditional method, the traditional neural network method is based on large training set. If the training set is small, it is difficult to determine the structure of the neural network and may have the local minimum problem which will reduce the efficiency. So the performance of this method needs to be improved when having small training set. Support vector machine (SVM) is based on statistical learning theory and structural risk minimization principle. Its main idea is mapping nonlinear problem in the original space to a linear problem in high-dimensional feature space, then finds the hyperplane and maximizes the distance between hyperplane and support vector in high-dimensional feature space. This ensures that the classification performance and generalization ability of SVM to be significantly better than ANN.

2. Principles and Methods

2.1. Support vector machine (SVM)

Support vector machine (SVM) is based on structural risk minimization principle. In other words, it gets maximum generalization through minimizing the upper bound of generalization error risk. The main idea of SVM is to achieve the maximum generalization ability by constructing the optimal hyperplane which has the largest distance between the different types of sample sets in sample space or feature space[5] . In essence, SVM is an algorithm method. When the sample is linearly separable, SVM solves the maximal-margin solution in the sample space. When the sample is linearly nonseparable, SVM is mapping sample set to high dimensional space through proper kernel function, thus the sample set is linear separable in high dimensional space.

In linear separable condition, let (x_i, y_i) , $i = 1,2,...,n$, $x \in R^d$, $y \in \{-1,+1\}$ be the sample set. $g(x) = w \cdot x + b$ is the general form of linear classification function in d-dimensional space. The hyperplane equation is

$$w \cdot x + b = 0 \tag{1}$$

Normalize the classification function and make two samples meet $|g(x)| \geq 1$. Thus, the classes margin is $2/\|w\|$ .So maximizing the classes margin equals minimizing $\|w\|$ or $\|w\|^2$. If a hyperplane can correctly classify all the samples, it must meet:

$$y_i[(w \cdot x) + b - 1] \geq 0 , i = 1,2,...,n \tag{2}$$

A hyperplane which meet the above conditions and let $\|w\|^2$ minimize is called optimal hyperplane.

The solution of the optimal hyperplane can be transformed into the following constrained optimization problem. That is, find the minimum value of the following function in constraints of equation (2).

$$\varphi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}(w \cdot w) \tag{3}$$

This problem can be transformed into Lagrange function:

$$L(w,b,a) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{n} a_i \{ y_i [(w \cdot x_i) + b] - 1 \} \tag{4}$$

After solving the above problems, the optimal classification function is

$$f(x) = \text{sgn}\{(w^* \cdot x) + b^*\} = \text{sgn}\{\sum_{i=1}^{n} a_i^* y_i (x_i \cdot x) + b^*\} \tag{5}$$

in which $\text{sgn}()$ is the sign function. Equations above applies only to the condition that samples are strictly linear separable. A relaxation $\xi_i \geq$ is added in equation (2) to meet the conditions if samples are linearly nonseparable. Thus, the equation (2) becomes

$$y_i[(w \cdot x) + b - 1] + \xi_i \geq 0, i = 1, 2, \ldots, n \tag{6}$$

At the same time the objective function changed to solving the minimum of the following function.

$$\Phi(w, \xi) = \frac{1}{2}\|w\|^2 + C \left| \sum_{i=1}^{n} \xi_i \right| \tag{7}$$

That is getting the generalized optimal hyperspace by considering both least wrong sub-sample and maximum classes margin. In equation, $C > 0$ is a constant which is the degree of punishment of wrong sub-samples.

However, it is not enough to deal with nonlinear complex system only through the method above. The most prominent advantage of SVM is not to solve linearly separable problems, but to solve nonlinear problems. It achieves linear classification and gets the optimal hyperspace by transforming the nonlinear problem into a linear problem in high dimensional space. The input space is transformed into a high dimensional space by a inner product defined nonlinear transform function. And this nonlinear transformation is achieved by defining the appropriate inner product function. Different inner product function (kernel function) will have a different support vector machine and different formation of algorithms. Such as equation (8) shows

$$K(x, x_i) = \exp\{-\frac{|x - x_i|^2}{\sigma^2}\} \tag{8}$$

The main difference between the radial basis kernel function support vector machine and the traditional RBF neural network is that each center of its basis function corresponds to a support vector and support vector and output values are automatically determined by the algorithm[6].

2.2. Application of genetic algorithms

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations. Based on the above we can see that, if the parameters in the SVM model is undetermined, the value of those parameters according to specific issues directly affect the actual SVM classification results. The role of genetic algorithm in this paper is to optimize the parameters in the SVM model. Two parameters need be determined for this radial kernel function of SVM. One is Penalty coefficient C and another is variance [7] σ . The specific work processes are as follows.

Step 1, encoding. The two parameters C and σ which need to be determined are encoding respectively. Binary code is often used in encoding.

Step 2, initialization. Select a group; choose a strand or a set of individual. This initial group is an assumed solution set of problems. We evenly distribute these strands in the string space instead of use random manner, thus the corresponding values of C and σ evenly distributed in the parameter space.

Step 3, Selection, crossover and mutation operations. In selecting operations, superior individuals are selected with random walk sampling methods based on the fitness of the individual in the parent population. The generation gap value is 0.9. Then use single point crossover after the reorganization and obtain the new population after the mutation operator.

Step 4, determine optimal global convergence. If the fitness of best individual achieves the given threshold or the fitness of best individual and group fitness is no longer rise, the iteration process reaches convergence and the algorithm ends. Otherwise, the previous generation is replaced by a new generation with selection crossover and mutation and return to step 3 to continue to loop.

3. Model

3.1. Idea to construct model

Financial distress of listed companies can be understood as a problem of the division of mid-point of geometric space. Assuming there are n indicators in feature space of the sample of listed companies. The number of financial normal company is m, composed of a set X and the number of financial distress company is k, composed of a set Y. Obviously, the n indicators in feature space of the sample space of listed companies can form a Euclidean n-space. Using the n indicators for the coordinates, x_i (x_{i1}, x_{i2}, …, x_{in}) (where i =1, 2,…, m) and y_j(y_{j1}, y_{j2}, …, y_{jn}) (where j =1, 2,…, k) are scattered in the n-dimensional space. They can be summed up in two sets. One is finical normal set X and another is finical distress set Y. Intersection may exist between the two sets. The model aims at how to separate the two sets in the minimum error rate and the highest collection efficiency.

3.2. Model structure design

Constructing model includes two processes, training and testing. After that processes the model has good ability to identify and can be applied to forecast specific financial distress to further verify the validity of it. Model structure is shown in Fig. 1.
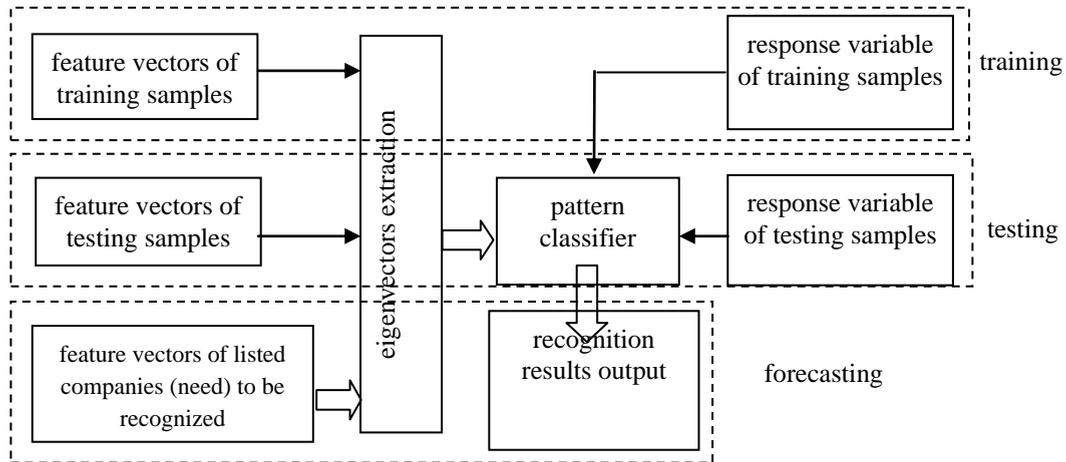
Fig. 1. Financial Distress Early-warning Model of Listed Companies

Black arrows in Figure 1 indicate the direction of data flow of the training and test samples, white arrows indicate the direction of data flow in the actual process of the financial crisis forecast. As is shown in Figure 1, the first step of the model is to establish training samples and testing samples. Then enter the training samples to train the pattern classifier. And after training, input testing samples to verify the training results. Pattern classifier can be used for financial distress early-warning if it passes the test. The model design can be proved success if the result of the pattern classifier is consistent with a greater degree with the listed company within a period of time in the future.

3.3. Sample data collection

1) Selection of sample companies
Sample companies include financial normal companies and financial distress companies. This paper chooses listed companies as experimental data subjects because the openness of the financial data of listed companies. Finical distress companies are defined as ST listed companies which have losses of two consecutive years (excluding financial and insurance companies). That is listed companies which have loss of financial distress to some extent. Financial distress refer to companies which was first listed ST companies in 2007-2009, and financial normal companies refers to other listed companies exclude companies removed by ST , ST in previous years and still remaining ST in 2007-2009.

Matched sampling method is used to select financial normal company samples. Matching criteria are as follows. 1 finical normal company must in the same or similar industry with the matched finical distress company; 2 choosing companies haven't been ST and also have good performance; 3 company's range of assets is in the level 80%-120% of the matched financial distress company, if the matched company doesn't exist with the level, select a company have differences in asset as small as possible; 4 if there doesn't exist a company meet the three criteria before, choose a better performing company of non-ST as a pair.

In accordance with the criteria of selecting financial distress samples, 60 sample ST listed companies which have losses for two consecutive years in 2007-2009 are chosen from the Shanghai and Shenzhen stock markets. Then, 60 financial normal companies are chosen in accordance with the matching criteria. Financial indicators data of the sample company's is collected and coordinated from the CSMAR developed by GTA Corporation, Shenzhen and Institute of Chinese financial research, University of Hong Kong.

2) Determine input feature vector and the output response variables

According to the characteristics of listed companies, data availability and in the light of relevant literature, this paper choose 12 candidate indicators in 5 aspects such as profitability, development capacity, shareholder profitability, solvency and operating capacity of listed companies[8,9]. Specific indicators include net operating profit rate, return on total assets ratio, net profit rate of asset, total assets growth rate, net profit ratio, earning per share, operating profit per share, liability/asset ratio, current liabilities ratio, financial leverage, DTL and circulating asset turnover rate. These indicators can fully reflect the signs of financial distress of listed companies. Therefore, the 12 indicators can be used to form the feature vectors of the financial distress early-warning model. Output response variables can be determined by financial performance of listed companies. Finance normal company samples are recorded as +1, and financial distress company samples are recorded as －1.

4. Empirical model analysis

4.1. Selection of training sample and testing sample

The financial distress early-warning model of listed companies is formed by two parts, training and testing, therefore appropriate training and testing samples need to be selected. Training samples are the 60 groups data samples previously selected from listed companies (including 50% financial normal and 50% financial distress companies). Three training samples are established, one include 20 groups, one 40 groups and one 60 groups of data. Each samples contains 50% of the listed companies which are recorded as "+1" and 50% of "－1". Finally, 40 listed companies of the 60 groups are randomly selected as the sample-in test samples and another 40 company's data beyond the 60 groups as the sample-out samples.

4.2. Model training and testing

The financial distress early-warning model is trained with the previously selected data as training samples. After training is completed, use sample-in test and sample-out samples to test the model. The test results is shown in Table 1,values in the table are the correct classification rate.

Table 1. Test Results of Financial Distress Early-warning Model of Listed Companies Based on GA-SVM(%)

Training samples	20 groups	40 groups	60 groups
Sample-in correct classification rate	92.5	95.0	97.5
Sample-out correct classification rate	85.0	90.0	92.5
Average sample-in correct classification rate	95.0		
Average sample-out correct classification rate	89.2		

It can be seen from the test results in Table 1 that 3 training samples have sample-in correct classification rate of above 90%, average sample-in correct classification rate of 95%, sample-out correct classification rate of above 85% and average sample-out correct classification rate of 89.2%. The results show that the financial

distress early-warning model has strong generalization ability. In addition, it can be found from the test results that there is a corresponding increase of effectiveness of the model with the increase in the number of training samples, which shows the validity of the model is greatly influenced by quality of sample.

These empirical results show that the financial distress early-warning model of listed companies based on GA-SVM has a strong ability of pattern recognition and can accurately make judgments of the companies' financial situation according to their financial data.

5. Conclusion

Taking financial data of listed companies as the research object, this paper successfully constructs a GA-SVM financial distress early-warning model by using empirical methods. The main conclusions are as follows. First, financial data of listed companies in China is effective and has strong predictive ability. Generally speaking, listed companies' future financial situation can be predicted through data mining. Second, predictive ability of GA-SVM financial distress early-warning model is proved by sample-in and sample-out empirical analysis. The GA-SVM model can accurately provide early-warning of the financial situation of listed companies.

References

[1]   B. Back, T. Laitinen, K. Sere. Neural Networks and Genetic Algorithms for Bankruptcy Predictions[J]. Expert Systems with Applications. 1996, 11 (4): 407-413.
[2]   G. Zhang, M. Y. Hu, B. E. Patuwo, D. C. Indro. Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis[J]. European Journal of Operational Research. 1999, 116: 16-32.
[3]   Z. R. Yang, M. B. Platt, H. D. Platt. Probabilistic Neural Networks in Bankruptcy Prediction[J]. Journal of Business Research. 1999, 44: 67-74.
[4]   P. C. Pendharkar. A Threshold Varying Artificial Neural Network Approach for Classification and Its Application to Bankruptcy Prediction Problem[J]. Computers & Operations Research. 2005, 32: 2561-2582.
[5]   M. H. Shen, L. Xiao. Application of Support Vector Machine (SVM) in Pattern Recognition[J]. Telecommunication Engineering.2006,(4): 9-12.(in Chinese)
[6]   ADANKON Mathias M, CHERIET Mohamed. Optimizing resources in model selection for support vector machine[J]. Pattern Recognition. 2007, 40(3): 953-963.
[7]   Pei-Yi Hao, Jung-Hsien Chiang, Yi-Kun Tu. Hierarchically SVM classification based on support vector clustering method and its application to document categorization[J]. Expert Systems with Applications. 2007, 33(3): 627-635.
[8]   Q. S. Zhang, L. L. Luo, J. M. Liu. A Study on Financial Distress Predicting of Listed Companies Based on SVM [J].Computer Applications, 2006,(6): 105-107. (in Chinese)
[9]   Y. Q. Li, G. L. Tian. A Study on Early Warning for Listed Companies' Financial Distress[J]. Journal of Northwest University(Philosophy and Social Sciences Edition).2009,39(5):79-83(in Chinese)