Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2009, Article ID 830572, 8 pages
doi:10.1155/2009/830572

Research Article

Research of Financial Early-Warning Model on Evolutionary Support Vector Machines Based on Genetic Algorithms

Zuoquan Zhang,[1] Fan Lang,[1] and Qin Zhao[2]

[1] *School of Sciences, Beijing Jiaotong University, 100044 Beijing, China*
[2] *School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Zuoquan Zhang, zqzhang@bjtu.edu.cn

Received 7 September 2009; Accepted 12 October 2009

Recommended by Guang Zhang

A support vector machine is a new learning machine; it is based on the statistics learning theory and attracts the attention of all researchers. Recently, the support vector machines (SVMs) have been applied to the problem of financial early-warning prediction (Rose, 1999). The SVMs-based method has been compared with other statistical methods and has shown good results. But the parameters of the kernel function which influence the result and performance of support vector machines have not been decided. Based on genetic algorithms, this paper proposes a new scientific method to automatically select the parameters of SVMs for financial early-warning model. The results demonstrate that the method is a powerful and flexible way to solve financial early-warning problem.

1. Introduction

The development of the financial early-warning prediction model has long been regarded as an important and widely studied issue in the academic and business community. Statistical methods and data mining techniques have been used for developing more accurate financial early-warning models. The statistical methods include regression, logistic models, and factor analysis. The data mining techniques include decision trees, neural networks (NNs), fuzzy logic, genetic algorithm (GA), and support vector machines (SVMs) etc [1]. However, the application of statistics was limited in the real world because of the strict assumptions.

Recently, SVM, which was developed byVapnik (Vapnik (1995)), is one of the methods that is receiving increasing attention with remarkable results. In financial applications, time series prediction such as stock price indexing and classification such as credit rating and

financial warning are main areas with SVMs [2]. However, as SVMs are applied for pattern classification problems, it is important to select the parameters of SVMs.

This paper applies the proposed evolutionary support vector machine based on genetic algorithms model to the financial early-warning problem using a real data set from the companies which come into market in China.

2. Theoretical Background

2.1. Genetic Algorithm (GA)

A GA is a flexible optimization technique inspired by evolutionary notions and natural selection. A GA is based on an iterative and parallel procedure that consists of a population of individuals (each one representing an attempted solution to the problem) which is improved in each iteration by means of crossover and mutation, generating new individuals or attempted solutions which are then tested by [3].

There are three main questions that have become in relevant topics in GA design research: (1) encoding; (2) operators; (3) control parameters. The GA starts to work by selecting a sample (randomly or by means of any other procedure) of potential solutions to the problem to be solved—previously the problem has to be formulated in chromosomes notation. In a second step the fitness value of every chromosome (potential solution)—in accordance with an objective function that classifies the solutions from the best to the worst-is computed [4]. The third step applies the reproduction operator to the initial set of potential solutions. The individuals with higher fitness values are more largely reproduced. One of the most common methods which is used in this paper is the "roulette wheel." This method is equivalent to a fitness-proportionate selection method for population large enough. There are two essential actions in the GA procedure: (1) the creation of attempted solutions or ideas to solve the problem through recombination and mutation; (2) the elimination of errors or bad solutions (after testing them) by selecting the better adapted ones or the closer to the truth.

2.2. Support Vector Machines (SVMs)

Since SVMs were introduced from statistical learning theory by Vapnik, a number of studies have been announced concerning its theory and applications [5]. A simple description of the SVMs algorithm is provided as follows.

Given a training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \in (X, Y)^l$ with input vectors $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)}) \in R^n$ and target labels $y_i \in \{-1, 1\}$, the support vector machines (SVMs) classifier, according to Vapnik's theory, finds an optimal separating hyper plane which satisfies the following conditions:

$$H = \{x \in R^n, (\omega \cdot x) + b = 0\}, \quad \omega \in R^n, \ b \in R. \tag{2.1}$$

with the decision function $f(x) = \operatorname{sign}((\omega \cdot x) + b)$.

To find the optimal hyper plane: $(\omega \cdot x) + b = 0$, the norm of the vector needs to be minimized, on the other hand, the margin $1/\|\omega\|$ should be maximized between two classes.

The solution of the primal problem is obtained after constructing the Lagrange. From the conditions of optimality, one obtains a quadratic programming problem with

Lagrange multipliers α_i's. A multiplier α_i exists for each training data instance. Data instances corresponding to nonzero α_i's are called support vectors [6].

On the other hand, the above primal problem can be converted into the following dual problem with objective function and constraints:

$$\begin{aligned} \text{Min}: \quad & \frac{1}{2}\sum_{i,j=1}^{k}\alpha_i\alpha_j y_i y_j (x_i x_j) - \sum_{i=1}^{k}\alpha_i \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, 2, \ldots, k, \quad \sum_{i=1}^{k}\alpha_i y_i = 0 \end{aligned} \tag{2.2}$$

with the decision function

$$f(x) = \text{sign}\left(\sum_{i=1}^{k}\alpha_i \cdot y_i (x \cdot x_i) + b\right). \tag{2.3}$$

Most of classification problems are, however, linearly nonseparable in the real world. In the nonlinear case, we first mapped the data to a high-dimensional space, using a mapping, $\phi : R^d \rightarrow H$. Then instead of the form of dot products, "kernel function" K is issued such that $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. We will find the optimal hyper plane: $(\omega \cdot \phi(x)) + b = 0$ with the decision function $f(x) = \text{sign}(\sum \alpha_i \cdot y(i)\phi(x)\phi(x_i) + b)$.

In this paper, RBF kernel functions are used as follows: $K(x, y) = e^{-\|x-y\|^{1/\sigma^2}}$. Using the dual problem, the quadratic programming problems can be rewritten as

$$\begin{aligned} \text{Min}: \quad & \frac{1}{2}\sum_{i,j=1}^{k}\alpha_i\alpha_j y_i y_j K(x_i x_j) - \sum_{i=1}^{k}\alpha_i \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, k, \quad \sum_{i=1}^{k}\alpha_i y_i = 0. \end{aligned} \tag{2.4}$$

3. The Financial-Warning Model of Chinese Companies

3.1. Evolutionary Support Vector Machines Based on Genetic Algorithms

As SVMs are applied for pattern classification problems, it is important to select the parameters of SVMs [7]. This paper applies genetic algorithms to define the parameters of SVMs. The steps of evolutionary support vector machines based on genetic algorithms are given as follows.

Step 1. Define the string (or chromosome) and encode parameters of SVMs into chromosomes. In this paper, the radial basis function (RBF) is used as the kernel function for financial warning prediction. There are two parameters while using RBF kernels: C and δ^2. In this study, C and δ^2 are encoded as binary strings and optimized by GA. In addition, the length of the GA chromosome used in this paper is 18. The first 9 bits represent C and the remaining 9 bits represent δ^2.

Step 2. Define population size and generate binary-coded initial population of chromosomes randomly. The initial random population size is 40.

Step 3. Define probability of crossover (Pc) and probability of mutation (Pm) and do the operation of GA (selection, crossover and mutation).

Generate offspring population by performing crossover and mutation on parent pairs. There are different selection methods to perform reproduction in the GA to choose the individuals that will create offspring for the next generation [8]. One of the most common method and the one used in this paper is the "roulette wheel."

Step 4. Decode the chromosomes to obtain the corresponding parameters of SVMs.

Step 5. Apply the corresponding parameters to the SVMs model to compute the output o_k. Each new chromosome is evaluated by sending it to the SVMs model.

Step 6. Evaluate fitness of the chromosomes using o_k (fitness function: predictive accuracy) and judge whether stopping condition is true, if true end; if false, turn to Step 3. The fitness of an individual of the population is based on the performance of SVMs.

Considering the real problem, we define the predictive accuracy of the testing set as the fitness function. It is represented mathematically as follows:

$$\text{fitness} - \text{function} = \sum_{i=1}^{n} \frac{Y_i}{n}, \tag{3.1}$$

where Y_i is one, if the actual output equals the predicted value of the SVMs model, otherwise Y_i is zero.

3.2. The Selection of Input Variables

There are many financial ratios which can represent the profitability of company, and the differences between industries are obviously, such as, household appliances and pharmaceutical industry. So the horizontal comparability of many financial ratios is not reasonable. This paper focuses on the profitability of company, and then selects six ratios as the input variables: (1) Sell profit rate; (2) Assets earning ratio; (3) Net asset earning ratio; (4) Profit growth rate of main business; (5) Net profit growth rate; (6) Total profit growth rate [9].

3.3. The Selection of Output Variable

We assume that the economy environment is similar, and select ROE (Rate of Return on Common Stockholders' Equity) as the standard of selection of output variable because ROE is one of the important ratios which are used to judge the profitability [10]. The method is represented as follows. We select those companies whose ROE is greater than 0 in the year $n-1$ and the year n, and we distinguish those companies into two kinds by judging the numerical of ROE in year $n+1$: the first kind, ROE is greater than 0, and the output is $y = 1$; the second kind, ROE is equal or less than 0, and the output is $y = -1$.

Table 1: The training set.

	Variables							
Companies	Sell profit rate	Assets earning ratio	Net asset earning ratio	Profit growth rate of main business	Net profit growth rate	Total profit growth rate	$N+1$ year's ROE	Output
Chenming Paper	0.09132	0.03615	0.0415	−0.3056983	−0.429	−0.3424	0.1077	1
Foshan electrical and lighting	0.24647	0.09228	0.0909	0.29356763	0.0757	0.109821	0.1044	1
Huali Group	0.30593	0.13124	0.1806	−0.1730657	−0.148	−0.11433	0.1561	1
GreeElectric appliances	0.04546	0.04738	0.1568	−0.1580513	0.0949	0.100019	0.1617	1
Zhuhai Zhongfu	0.19179	0.05514	0.0661	−0.2462878	−0.086	−0.11713	0.1011	1
Zijiang enterprise	0.26843	0.10353	0.1346	0.46888081	0.3058	0.320772	0.1621	1
Qingdao Haier	0.23414	0.14571	0.1253	0.12413238	0.5833	0.546907	0.078	1
Fujian Nanzhi	0.12411	0.05107	0.0734	0.13528976	0.1961	0.050202	0.0619	1
ST Swan	0.03803	0.02276	0.0012	−0.7855768	−0.987	−0.69645	−0.326	−1
ST Macro	0.08378	0.01404	0.2073	−0.0473601	−0.85	−0.89666	−3.943	−1
ST Tianyi	0.05505	0.01348	0.0096	−0.2675278	−0.865	−0.83182	−0.281	−1
ST Jizhi	0.08603	0.01643	0.005	−0.3048749	−0.575	−0.73226	−0.24	−1
ST Hushan	0.16272	0.04433	0.0503	0.09753527	−0.799	−0.81446	−0.256	−1
ST Jiangzhi	0.30537	0.09123	0.0813	0.41089893	0.1747	0.220284	−0.86	−1
ST Ziyi	0.10243	0.03277	0.0014	−0.1932007	−0.986	−0.96686	−0.32	−1
Xiaxin electronic	0.04935	0.05256	0.0552	−0.428435	−0.765	−0.78969	−0.335	−1
Chunlan Gufen	0.21223	0.09518	0.0787	−0.3200746	−0.121	−0.11254	0.0405	1
Shangfeng Industrial	0.44223	0.101126	0.0598.	−0.5987261	−0.281	−0.30444	0.0327	1
Aucma	0.10466	0.03811	0.0363	0.3780161	0.6449	0.720158	0.0286	1
Xinjiang Tianhong	0.10348	0.04947	0.0451	−0.0591081	−0.155	−0.25771	0.0406	1
Jincheng Paper	0.1038	0.03736	0.051	−0.4548067	−0.406	−0.48903	0.0145	1
Wuzhong Yibiao	0.33955	0.06541	0.0631	0.26631751	0.046	0.007402	0.0125	1
Qingshan Paper	0.24708	0.06499	0.071	−0.0290705	−0.137	−0.13544	0.0088	1
Hakongtiao	0.29178	0.12776	0.1819	0.20166437	0.9176	0.578143	0.0104	1

3.4. Research Data and Experiments

The research data we employ is from the companies which come into market in China, and consists of 50 medium-size firms from 1999 to 2001. The data set is arbitrarily split into two subsets; about 50% of the data is used for a training set and 50% for a testing set. The training data for SVMs is totally used to construct the model. The testing data is used to test the results with the data that is not utilized to develop the model. The training set is shown at Table 1.

Table 2: Classification accuracies of various parameters in the first model.

	δ^2									
	1		10		30		50		80	
C	train	Test	train	test	train	test	train	test	train	test
1	86.87	87.50	80.00	70.83	66.67	70.83	66.67	70.83	66.67	70.83
10	93.33	75.50	86.67	70.83	66.67	70.83	66.67	70.83	66.67	70.83
30	93.33	75.50	86.67	70.83	66.67	70.83	66.67	70.83	66.67	75.00
50	100	79.17	86.67	79.17	66.67	70.83	66.67	70.83	66.67	75.00
90	100	79.17	93.33	87.50	66.67	70.83	66.67	70.83	66.67	75.00
100	100	79.17	93.33	79.17	66.67	70.83	66.67	79.17	66.67	70.83
150	100	79.17	86.67	79.17	66.67	70.83	66.67	79.17	66.67	70.83
200	100	79.17	86.67	75.00	66.67	70.83	66.67	70.83	66.67	70.83
250	100	79.17	86.67	79.17	66.67	70.83	66.67	70.83	66.67	70.83

Table 3: Prediction accuracy of the second model.

Training	Testing
93.33	87.50

Additionally, to evaluate the effectiveness of the proposed model, we compare two different models.

(1) The first model, with arbitrarily selected values of parameters, varies the parameters of SVMs to select optimal values for the best prediction performance.

(2) We design the second model as a new scientific method to automatically select the parameters optimized by GA.

4. Experimental Results

4.1. Classification Accuracies of Various Parameters in the First Model (Table 2)

Based on the results proposed by Tay and Cao (2001), we set an appropriate range of parameters δ^2 as follows: a range for kernel parameter is between 1 and 100 and a range for C is between 1 and 250. Test results for this study are summarized in Table 2. Each cell of Table 2 contains the accuracy of the classification techniques. The experimental result also shows that the prediction performance of SVMs is sensitive to the various kernel parameters δ^2 and the upper bound C. In Table 2, the results of SVMs show the best prediction performances when δ^2 is 10 and C is 90 on the most data set of various data set sizes. Figure 1 gives the results of SVMs with various C where δ^2 is fixed at 10.

4.2. Results of Evolutionary Support Vector Machines Based on Genetic Algorithms

Table 3 describes the prediction accuracy of the second model. In Pure SVMs, we use the best parameters from the testing set out of the results. In Table 3, the proposed model shows better performance than that of the first model.

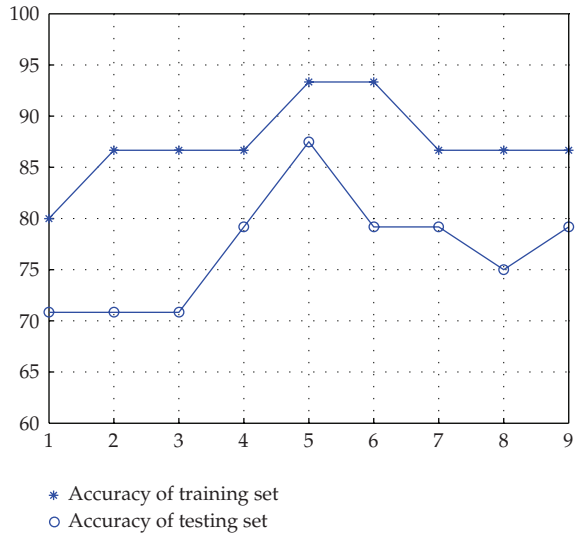

Figure 1: The results of SVMs with various C where δ^2 is fixed at 10.

The results in Table 3 show that the overall prediction performance of the second model on the testing set is consistently good. Moreover, the accuracy and the generalization using evolutionary support vector machines are better than that of the first model.

5. Conclusions

In this paper, we applied evolutionary support vector machines based on genetic algorithms to financial early-warning problem and showed its attractive prediction power compared to the pure SVMs method. In this paper we utilize genetic algorithms in order to choose optimal values of the upper bound C and the kernel parameter δ^2 that are most important in SVMs model selection. To validate the prediction performance of this evolutionary support vector machines based on genetic algorithms model, we statistically compared its prediction accuracy with the pure SVMs model, respectively. The results of empirical analysis showed that proposed model outperformed the other methods.

In a classification problem, the selection of features is important for many reasons: good generalization performance, running time requirements, and constraints imposed by the problem itself [11]. While this study used six ratios as a feature subset of SVMs model, it should be noted that the appropriate features can be problem-specific; hence it remains an interesting topic for further study to select proper features according to the types of classification problems.

Obviously, after the application of genetic algorithms, there is a significant improvement in the accuracy. That is just what we need in the selection of parameters of SVMs for financial early-warning model.

References

[1] J. Wang, "The statistical properties of the interfaces for the lattice Widom-Rowlinson model," *Applied Mathematics Letters*, vol. 19, no. 3, pp. 223–228, 2006.

[2] V. A. Kholodnyi, "Valuation and hedging of European contingent claims on power with spikes: a non-Markovian approach," *Journal of Engineering Mathematics*, vol. 49, no. 3, pp. 233–252, 2004.

[3] T. M. Liggett, *Interacting Particle Systems*, vol. 276 of *Grundlehren der Mathematischen Wissenschaften*, Springer, New York, NY, USA, 1985.

[4] V. A. Kholodnyi, "Universal contingent claims in a general market environment and multiplicative measures: examples and applications," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 62, no. 8, pp. 1437–1452, 2005.

[5] D. Lamberton and B. Lapeyre, *Introduction to Stochastic Calculus Applied to Finance*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2000.

[6] V. A. Kholodnyi, "Modeling power forward prices for power with spikes: a non-Markovian approach," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 63, no. 5–7, pp. 958–965, 2005.

[7] J. Wang, "Supercritical Ising model on the lattice fractal—the Sierpinski carpet," *Modern Physics Letters B*, vol. 20, no. 8, pp. 409–414, 2006.

[8] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York, NY, USA, 1968.

[9] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*, vol. 271 of *Grundlehren der Mathematischen Wissenschaften*, Springer, New York, NY, USA, 1985.

[10] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, pp. 637–654, 1973.

[11] Y. Higuchi, J. Murai, and J. Wang, "The Dobrushin-Hryniv theory for the two-dimensional lattice Widom-Rowlinson model," in *Stochastic Analysis on Large Scale Interacting Systems*, vol. 39 of *Advanced Studies in Pure Mathematics*, pp. 233–281, Mathematical Society of Japan, Tokyo, Japan, 2004.