The Use of Data Mining Techniques and Support Vector Regression for Financial Forecasting

Liqiang Hou [1], Shanlin Yang [1] and Zhiqiang Chen[1]

[1]*School of Management of Hefei University of Technology, Anhui Hefei 230009*
[1]*liqianghou1020@gmail.com*

Abstract

In recent years, data mining techniques such as neural networks, support vector Regression have been applied extensively to the task of predicting financial variables. As influenced by various factors, the volatility of stock shows a non-linear characteristic, which demonstrates that the forecasting is a non-linear problem. Support vector regression (SVR) is proven to be useful in dealing with non-linear forecasting problems in recent years. The key point in using SVR for forecasting is how to determine the appropriate parameters. An improved Artificial Neural Networks(ANN) algorithm is used to optimize the parameter set of (C, σ), which influences the performance of this model directly. By doing so, this model can deal with the nonlinearity and multi-factors of volatility, and ensure stability and accuracy of support vector machine based regression. Finally, we study a case with the satisfactory result by the SPA test which is showing that this model is more accurate than other models, which guarantees its application.

Keywords: *data mining techniques, support vector regression, parameter optimization, RBF Artificial Neural Networks*

1. Introduction

Data mining techniques (DMT) have formed a branch of applied artificial intelligence (AI), since the 1960s. During the intervening decades, important innovations in computer systems have led to the introduction of new technologies [1], for web based education. Data mining allows a search, for valuable information, in large volumes of data [2].The explosive growth in databases has created a need to develop technologies that use information and knowledge intelligently. Therefore, DMT has become an increasingly important research area [3].Of the data mining techniques developed recently, several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining, are herein reviewed. The techniques for mining knowledge from different kinds of databases, including relational, transactional, object oriented, spatial and active databases, as well as global information systems, are also examined. Potential data mining applications and some research issues are discussed.

In recent years, data mining techniques such as neural networks, support vector Regression have been applied extensively to the task of predicting financial variables [4]. Then, the development of powerful communication and trading facilities has enlarged the scope of selection for investors [5]. Traditional capital market theory has also changed and methods of financial analysis have improved [6]. Forecasting stock return or a stock index is an important financial subject that has attracted researchers' attention for many years.

In this paper, we study the intrinsic link between these two algorithms in financial field, and propose RBF optimization algorithm based on *SVR* and genetic algorithm, which use Neural Networks to choose the parameters (c,σ^2) for *SVR*, then to be used for constructing RBF networks. This algorithm effectively improves generalization and don't need a large number of experiments or empirical experiences to network structure. By means of practical study, we draw the conclusion by the SPA test that the method rendered in this paper is better as compared to *ANN* , $GM(1,1)$, *EGARCH* and $LS-SVR$.

The remainder of this study is organized as follows: Section 2 and 3 outlines the *SVR* providing the foundation for structure and parameters of RBF. Section 4 describes the *SVR* providing network structure and parameters of RBF. Section 5 describes the SPA test. Data description and main findings are then reported in Section 6. Finally, summary and conclusion are presented in the last section.

2. SVR Providing the Oretical Foundation for Structure and Parameters of RBF

RBF network from input to output mapping is nonlinear, but network output is linear in terms of the weights. The k th hidden unit's output is:

$$\phi_k(x) = \exp(-\frac{\|x-c_k\|^2}{2\sigma_k^2}) \tag{1}$$

Where $\|\bullet\|$ is Euclidean norm, x is the input vector, c_k is the center vector of hidden units, σ_k is the width of hidden units. *N* denotes the number of the hidden units, wk is the weights between the hidden units and the outputs, then the outputs of RBF networks is:

$$f(x) = \sum_{k=1}^{N} w_k \exp(-\frac{\|x-c_k\|^2}{2\sigma_k^2}) \tag{2}$$

According to Mercer Conditions, SVR adopts kernel function to map a sample vector from the original space to feature space. Gaussian kernel function used here is:

$$K(x,v_i) = \exp(-\frac{\|x-v_i\|^2}{2\sigma_i^2}) \tag{3}$$

SVR in regression form is the linear combination of the hidden units, then:

$$f(x) = \sum_{i=1}^{g} w_i K(x,v_i) + b = \sum_{i=1}^{g} w_i \otimes \exp(-\frac{\|x-v_i\|^2}{2\sigma_i^2}) + b \tag{4}$$

SVR has a similarity in structure with the RBF network, so the number g of support vector which is gotten from the training of *SVR* can be the number of the hidden units in RBF networks, support vector can be the center vector of radius function, the width selected by *SVR* can be the width of RBF.

3. GA Providing SVR Models Parameters

The algorithm is to use genetic algorithm to optimize the *SVR* model parameters, including the parameter σ , penalty factor *C* and insensitive loss function ε , the basic steps of it are as follows:

Step1: Choose the initial population of individuals randomly;

Step2: Evaluate the fitness of each individual in that population;

Step3: Select a new generation of population from the previous generation by using selection operator;

Step4: Take the crossover and mutation operation on the current population, then take the evaluation, selection, crossover and mutation operation on the new breed, and continue.

Step5: If the fitness function value of optimal individual is large enough or the algorithm have run many generations, and the optimal fitness value of the individual can't be changed obviously, then we get the optimal value of kernel function parameter σ, penalty factor C, and insensitive loss function ε ,and we can also get the optimal classifier by the training datum.

How to construct fitness function is the key point of genetic algorithm. We use the promotion theorem of SVR in high-dimensional space to construct fitness function. We set

$$fit = 1/(T + 0.05) \tag{5}$$

Where $T = R^2/l\gamma^2$ is the testing error boundary is the radius of super-sphere which contain all the data $\gamma = 1/|w|$ is the interval value, l is the number of the samples

4. SVR Providing Network Structure and Parameters for RBF

$LS-SVR$ expands standard SVR by optimizing the square of relaxation factors and converting the constraints of inequality to equality, so the quadratic programming problem in traditional SVR becomes linear simultaneous equations, thus the calculating difficulty reduces a lot in company with the solution high efficiency and convergence speeding up.

The basic method of SVR : Define $x \in R^n$ and $y \in R$

Let R^n be the input space, by nonlinear transformation $\phi(\cdot)$, we let in the input space x map into a high dimensional characteristic space where we use the linear function to fit sample data while making sure the generalization.

In the characteristic space, the linear estimation function is defined as:

$$y = f(x, \omega) = \omega^T \phi(x) + b \tag{6}$$

Where ω is the weight and b is the skewness.

The aim function is:

$$\min_{\omega,b,\xi} J(\omega, b, \xi) = \frac{1}{2}\omega^T \omega + \frac{1}{2}C\sum_{i=1}^{n}\xi_i^2 \tag{7}$$

$$s.t. \quad y_i = \phi(x_i)\omega + b + \xi_i \qquad i = 1, \cdots, N \tag{8}$$

Where $\omega \in R^h$ is the weight vector and $\phi(\bullet)$ is non-linear mapping function, $\xi_i \in R^{N*1}$ is relaxation factor, $b \in R$ is the skewness while $C > 0$ is penalty factor.

Importing factors, $\alpha_i \in R^{N*1}$, we can easily get the function as:

$$L(\omega, b, \xi_i, \alpha_i) = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}C\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\alpha_i\left[\phi(x_i)\omega + b + \xi_i - y_i\right] \tag{9}$$

According to the KTT we get

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{N} \alpha_i \phi(x_i) = 0 \\ \dfrac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = \alpha_i - C\xi_i = 0 \\ \dfrac{\partial L}{\partial \alpha_i} = \phi(x_i) + b + \xi_i - y_i = 0 \end{cases} \tag{10}$$

$$\begin{bmatrix} 0 & E^T \\ E & \phi\phi^T + C^{-1}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{11}$$

Where E is the matrix whose elements are all 1, I is a $N \times N$ identity matrix.

Inner product of regression in non-linear function can be replaced by kernel function satisfied Mercer. Let $\Omega_{ij} = \phi\phi^T$, then

$$\Omega_{ij} = \phi(x_i)^T \phi(x_j) = K(x_i, x_j) \tag{12}$$

We then have the $LS-SVR$ regression function model

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x_j) + b \tag{13}$$

Kernel function commonly used in practice are linear kernel, polynomial kernel, and RBF kernel, we use the RBF kernel as our kernel function for its better generalization. The form is as following:

$$K(x_i, x_j) = \phi(x_i)\phi(x_j) = \exp(-\left\| x_i - x_j \right\| / 2\delta^2) \tag{14}$$

In which the regularization parameter C and kernel breadth δ is the crucial parameters of $LS-SVR$.

Then, we give the figure of the RBF-SVR model in the article.

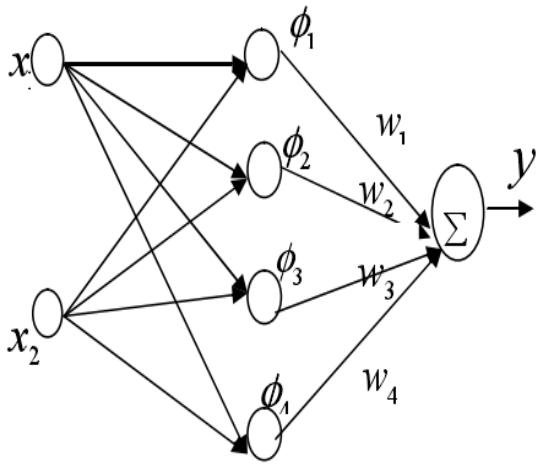

Figure 1. RBF Artificial Neural Networks Structure based on SVR

5. The SPA Test

It is generally accepted that squared daily returns provide a poor approximation of actual daily volatility. Andersen and Bollerslev [7] pointed out that more accurate estimates can be obtained by summing all squared intraday returns. If we were to apply their method directly in this paper, then we would define the *RV* measurement as

$$RV_t^{'} = \sum R_{t,d}^2 \tag{15}$$

However, this definition ignores the information contained in overnight returns. To address this problem, Hansen and Lunde [8] suggested scaling the *RV* measurement in the following way.

$$RV_t = \gamma RV_t^{'} \tag{16}$$

Where the so-called scale parameter γ is defined as

$$\gamma = \frac{\frac{1}{N}\sum_{t=1}^{N} R_t^2}{\frac{1}{N}\sum_{t=1}^{N} RV_t^{'}} \tag{17}$$

Andersen *et al*., [9] found the distribution of *RV* to be highly non-normal and skewed, but its logarithms to be approximately normal. Accordingly, they suggested that the natural logarithms of a *RV* measurement series, denoted as $\ln RV$, could be modeled by a Gaussian dynamic process.

Various forecasting criteria or loss functions can be considered in assessing the predictive accuracy of a volatility model, although, as Lopez [10] noted, it is not obvious which loss function is most appropriate for the evaluation of such models. Rather than making a single choice, we thus employ the following four accuracy statistics or loss functions as our forecasting criteria.

$$MSE = \frac{1}{N}\sum_{t=1}^{n}(RV_t - \hat{\sigma}_t^2)^2 \tag{18}$$

$$MAE = \frac{1}{N}\sum_{t=1}^{n}\left|RV_t - \hat{\sigma}_t^2\right| \tag{19}$$

$$HMSE = \frac{1}{N}\sum_{t=1}^{n}(1 - \hat{\sigma}_t^2 / RV_t)^2 \tag{20}$$

$$HMAE = \frac{1}{N}\sum_{t=1}^{n}\left|1 - \hat{\sigma}_t^2 / RV_t\right| \tag{21}$$

Where n is the number of forecasting data points; MSE and MAE are the mean square error and mean absolute error; HMSE and HMAE are the MSE and MAE, respectively, adjusted for heteroskedasticity. Different criteria serve different practical purposes. For example, in the case of Value-at-Risk applications, greater interest may lie in the accurate forecasting of a high rather than a low level of volatility, which implies that the MSE criterion is the most relevant loss function in risk management applications. Additional discussion of these criteria can be found in Ref. [11].

When a particular loss function is smaller for model A than it is for model B, it is impossible to conclude that the forecasting performance of the former is superior to that of the latter. Such a conclusion cannot be made on the basis of a single loss function and a single

sample. Recent work has focused on a testing framework that can determine whether one particular model is outperformed by another. As discussed in the Introduction, the SPA test, an extension of the White framework proposed by Hansen and Lunde [12], has been shown to possess good power properties and to be more robust than previous approaches.

In contrast to other evaluation techniques, the SPA test can be used to compare the performance of two or more forecasting models at the same time. Forecasts are evaluated employing a pre-specified loss function, and the ‘‘best’’ forecasting model is the one that produces the smallest expected loss. In the SPA test, the loss function relative to the benchmark model is defined as $x_{t,l}^{(0,i)} = L_{t,l}^{(0)} - L_{t,l}^{(i)}$, where $L_{t,l}^{(0)}$ is the value of loss function l at time t for benchmark model M_0, and $L_{t,l}^{(i)}$ is the value of loss function l at time t for competing model M_i, for $i = 1, \cdots, k$. The SPA test is used to compare the forecasting performance of a benchmark model against its *K* competitors. The null hypothesis that the benchmark or base model is not outperformed by any of the competing models can be expressed as $H_0 : \max_{i=1,\cdots,k} E(X_{t,l}^{(0,i)}) \le 0$. It is tested with the statistic $T_l^{SPA} = \max_{i=1,\cdots,k} (\sqrt{n}\bar{X}_{i,l} \Big/ \sqrt{\lim_{n\to\infty} Var(\sqrt{n}\bar{X}_{i,l})})$, where n is the number of forecast data points and $\bar{X}_{i,l} = \frac{1}{N}\sum_{t=1}^{n} X_{t,l}^{(0,i)}$. The estimation of $\lim_{n\to\infty} Var(\sqrt{n}\bar{X}_{i,l})$ and the p-value of T_l^{SPA} are obtained using the stationary bootstrap procedure discussed by Politis and Romano [13-15].

6. Case Study

6.1. Selection of Trained Sample Data

We choose the closing prices from 2010.7.1 to 2011.3.31 of The Shanghai Composite Index as the sample data to analyze the Shanghai Composite Index recently change situation. There are 182 valid sample data (data from Security Star Website). We choose the closing prices as the sample data series to analyze. And then, we predict the yield rate of the 40 trading days. After getting $\hat{y}_t, t = 1, 2, \cdots 40$ .As the closed point can be considered the reflection of the day's ending information, we suppose the daily closing price as k_t ,at the same time, we define the rate of The Shanghai Composite Index as y_t ,whose method of calculating is as below:

$$y_t = R_t = \ln(k_t / k_{t-1}) \tag{22}$$

The great length of the above definition is reducing the influence of skewness and kurtosis of single index. Unlike stock price, the returns can accurately reflect the fluctuate tendency in stock market.

Besides, we choose SSE 50 Index, growth rate of equity and volume of trade as the influential factors on yield. We preprocess those three kinds of data just like y_t .

That means the average growth rate per month in t day is:

$X_{1t} = \ln(x_{1,t} / x_{1,t-1})$, $X_{2t} = \ln(x_{2,t} / x_{2,t-1})$, $X_{3t} = \ln(x_{3,t} / x_{3,t-1})$. X_{2t} is the growth rate of market equity rate while X_{3t} is the growth rate of volume of trade[16-18].

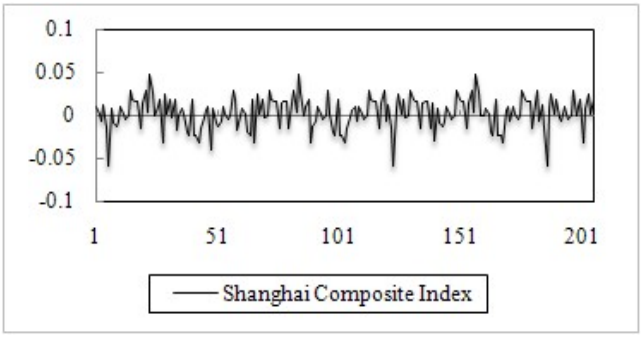

Figure 2. The Scatter Diagram of *y* Produced by Standardization of Shanghai Composite Index

6.2. *ANN – SVR* Prediction Model using trained Sample Data

(1) Generalized regression neural network select Gaussian function as the kernel function, the distribution density Spread is 0.6, the identification results is shown in Figure 3 , the mean squares error between identification results and output matrix is 0.0578.And the testing results is shown in Figure 4, the mean squares error between testing results and expected output matrix is 0.0282.

(2) In optimizing neural networks by SVR, we set the maximum genetic generations is 30, we get the maximum value of fitness by genetic algorithm searching, the penalty factor C is 312.9, the width of Gaussian function σ is 1.25, the insensitive loss function ε is 0.0467, furthermore we get the number of support vectors is 4 by the learning of input and output datum, so we set the number of hidden units in RBF network is 4, the Gaussian function center vector are the support vectors, the width of Gaussian function is the same with the regression machine. Then the RBF is constructed, just as shown in Figure 5. Identification results are shown in Figure 6.

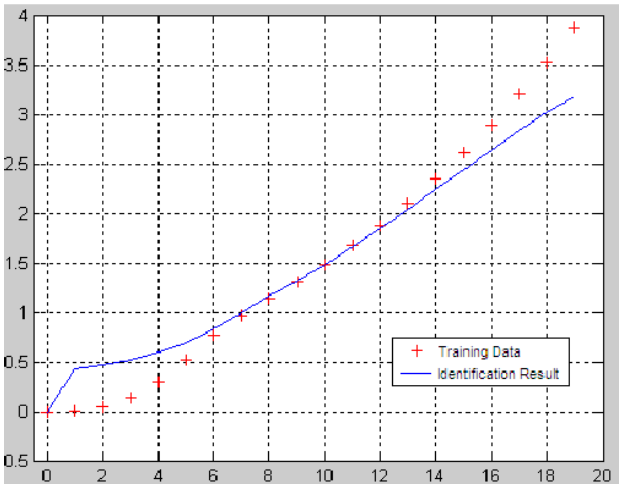

Figure 3. Identification Results of Generalized Regression Neural Network

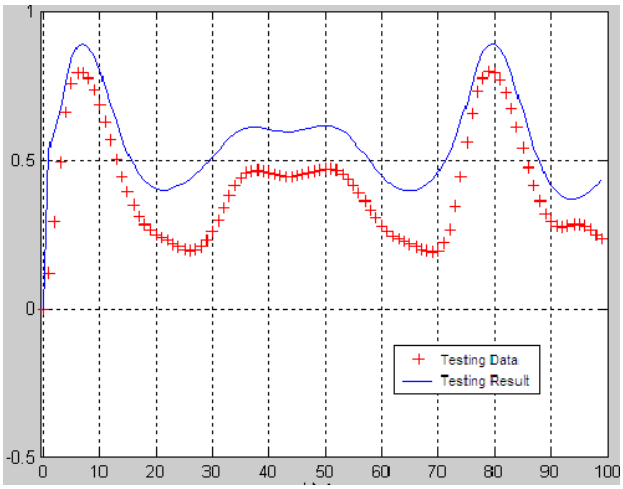

Figure 4. Testing Results of Generalized Regression Neural Network

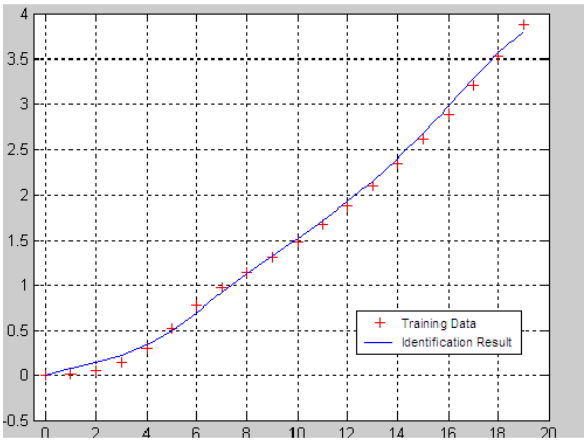

Figure 5. Identification Results of RBF Optimized by SVR

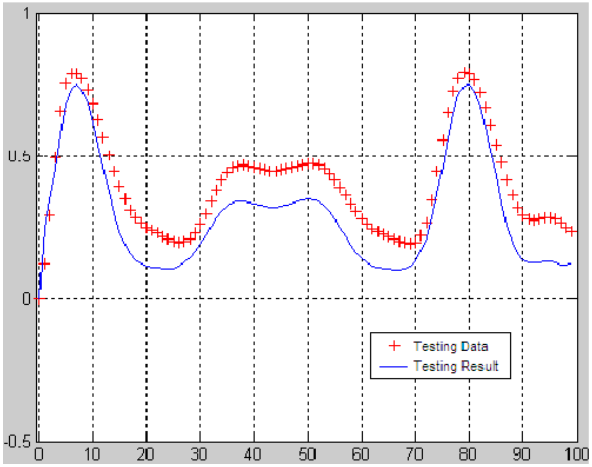

Figure 6. Testing Results of RBF Optimized by SVR

Here we implement $ANN-SVR$ model to the historical data and meanwhile to get the predicted error. The effect is showed as the below picture.

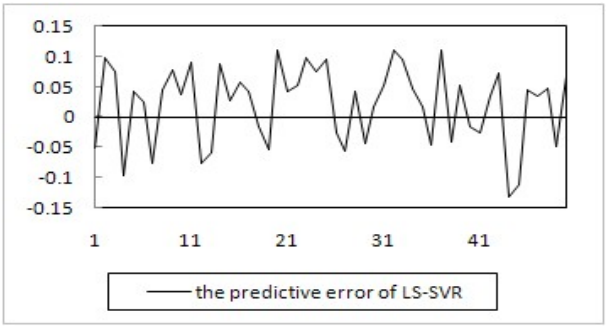

Figure 7. The Error of Prediction using ANN-SVR

6.3. SPA Test

We use Esq. (18)-(21) to calculate the loss function and the SPA value which is showed in Chart 4. In this chart will present the p-values of the SPA tests on the six models. The first column in the table lists the name of the base model (M_0) in the SPA test, and thus the seven remaining models are treated as competing models (M_i). Every number in Table 1 is a SPA p-value obtained through 10,000 times of bootstraps under a specified loss function. The larger the p-value, the less likely it is that the SPA null hypothesis ‘‘the base model is not out performed by all competing models’’ can be rejected, that is, the better the forecasting performance of the base model (M_0) is relative to its alternatives (M_i). The values in bold are the largest p-values under a specific loss function. We can easily be seen from the Table 1. The precision of the $ANN-SVR$ prediction model which is established by this article is significantly better than the other models. It’s p values were close to 1.

Table 1. SPA p-values for out-of-sample Daily Volatility Forecasts

L_i	M_0	M_i			
		GM	ANN	EGARCH	LS-SVR
MSE	*ANN – SVR*	0.899	0.873	0.872	0.876
MAE		0.903	0.899	0.874	0.898
HMSE		0.922	0.867	0.859	0.861
HMAE		0.871	0.845	0.834	0.843

Notes: The values in bold are the largest p-values under a specific loss function. The larger the p-value, the less likely it is that the SPA null hypothesis ‘‘the base model is not outperformed by all competing models’’ can be rejected

7. Conclusion

As the volatility of stocks are stochastic, stable and asymmetric in information, singe predictive method turns out to be insufficient. In this paper, we combine ANN and $LS-SVR$ to predict the volatility of stocks due to their superiority, And we study a case with the satisfactory result showing that this model is more accurate than other models, which guarantees its application.

China's stock market has been undergoing 20 years during which great change has taken place. We cannot ignore that there exist certain uneconomic factors influencing the behavior

of the market, such as banker operators, massive speculative behaviors. However, after removing those uneconomic factors, the method presented in the paper is a convincible way to predict the volatility of stocks especially those short-time ones.

Acknowledgements

This research is conducted with the support of National Natural Science Foundation of China (71101041, 71071045).

References

[1] S. Ha, S. Bae and S. Park, "Web mining for distance education", IEEE International conference on management of innovation and technology, **(2000)**, pp. 715-719.
[2] S. H. Weiss and N. Indurkhya, "Predictive Data Mining: A Practical Guide", San Francisco, CA: Morgan Kaufmann Publishers, **(1998)**.
[3] U. Fayyad, S. G. Djorgovski and N. Weir, "Automating the analysis andcataloging of sky surveys", U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining. Cambridge, MA: MIT Press, **(1996)**, pp. 471-494.
[4] H. Yuan-Sheng, D. J. Yuan Zhenzhen, "SVM short-term load forecasting based on ARMA error calibration and the adaptive particle swarm optimization", Power System Protection and Control, vol. 14, no. 39, **(2011)**, pp. 26-32.
[5] E. J. Elton and M. J. Gruber, "Modern Portfolio Theory and Investment Analysis (4thedn.)", New York: John Wiley & Sons, **(1991)**.
[6] T. Poddig and H. Rehkugler, "A world of integrated financial marketsusing artificial neural networks", Neurocomputting, vol. 10, **(1996)**, pp. 251-273.
[7] T. G. Andersen and T. Bollerslev, "Answering the skeptics: yes, standard volatility models do provide accurate forecasts", International Economic Review, vol. 39, **(1998)**, pp. 885-905.
[8] P. R. Hansen and A. Lunde, "A forecast comparison of volatility models: does anything beat a GARCH(1, 1)", Journal of Applied Econometrics, vol. 20, **(2005)**, pp. 873-889.
[9] T. G. Andersen, T. Bollerslev, F. X. Diebold and H. Ebens, "The distribution of realized stock return volatility", Journal of Financial Economics, vol. 61, **(2001)**, pp. 43-76.
[10] J. A. Lopez, "Evaluation of predictive accuracy of volatility models", Journal of Forecasting, vol. 20, **(2001)**, pp. 87-109.
[11] T. Bollerslev, R. F. Engle and D. Nelson, "ARCH models", R.F. Engle, D.L. McFadden (Eds.), Handbook of Econometrics, Elsevier Science B. V., Amsterdam, vol. 4, **(1994)**, pp. 2961-3038.
[12] P. R. Hansen and A. Lunde, "A forecast comparison of volatility models: does anything beat a GARCH(1, 1)", Journal of Applied Econometrics, vol. 20, **(2005)**, pp. 873-889.
[13] Z. Chen, S. Yang and X. Wang, "PLS-SVR Optimized by PSO Algorithm for Electricity Consumption Forecasting", Applied Mathematics and Information Sciences, vol. 7, no. 1, **(2013)**, pp. 331-338.
[14] T. Hua and X. Chi, "The prediction model of uncertainty mining in the stock market", Social Scientist, vol. 4, no. 4, **(2008)**, pp. 65-67.
[15] H. Y. Zhang and H. Lin, "Option price forecasting model by applying hybrid neural network and genetic algorithm", Journal of Industrial Engineering, vol. 123, no. 1, **(2009)**, pp. 59-87.
[16] W. H. Chen and J. Y. Shih, "A study of Taiwan's issuer credit rating systems using support vector machines", Expert Systems with Applications, vol. 30, no. 3, **(2006)**, pp. 427-435.
[17] X. L. Liu, G. Cheng and S. W. Cheng, "Intraday effects analysis of Chinese futures markets", Systems Engineering-Theory & Practice, vol. 28, no. 8, **(2008)**, pp. 65-80.
[18] H. Y. Zhang and H. Lin, "Option price forecasting model by applying hybrid neural network and genetic algorithm", Journal of Industrial Engineering, vol. 123, no. 1, **(2009)**, pp. 59-87.

Authors

Liqiang Hou received his M.Sc. degree in school of management of Hefei University of technology in 2012, respectively. Currently he is a PhD in management, Hefei University of technology. His researches interests include mathematical modeling, computer simulation and Financial Engineering.

Shanlin Yang is a Professor at the school of management of Hefei University of technology. He has published more than 100 papers and 9 books in the fields of management Information, decision theory and methods and Mathematical modeling. His research interests include computer simulation, science and technology of forecasting, decision, data processing technology, etc.

Zhiqiang Chen received the MS degree in Information Management and Information System from Hefei University of Technology in 2007, and the PhD degree in Management Science and engineering from the school of Management of Hefei University of Technology in 2012. He is currently a lecture in Hefei University of Technology. His research interests are in the areas of computer simulation, data mining, and information systems.