International Journal of Security and Its Applications
Vol. 10, No. 7 (2016) pp.269-280
http://dx.doi.org/10.14257/ijsia.2016.10.7.24

Research on Stock Price Prediction Model based on GA Optimized SVM Parameters

Liang bang-long[1], Lin jie[1] and Yuan Guanghui[2]

[1] *School of Economics and Management, Tongji University, Shanghai 200092, China*

[2] *School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China*

Abstract

This paper construct the predicted model based on support vector machine (SVM) for the Shanghai Composite Index, acquired the model parameters using genetic algorithm optimization was carried out, combined with k-fold cross method. Experiments based on the start date to February 2011 total 4948 trading day data, 10 fold cross circulation experiments of GA optimization; get the most accurate model parameter of SVM. At last, the regression model is used to predict, and the relative error of regression prediction is 0.11, and the accuracy of regression prediction is higher. In conclusion, this model can be used to predict the Shanghai Composite Index.

Keywords*: SVM; genetic algorithm; K fold cross experiment; regression prediction*

1. Introduction

The stock market forecast is an important branch of financial and economic forecast. It is mainly reflected in the stock market a variety of information of collection, collation, comprehensive work [1], according to the stock market historical data, the status quo and summing up of the relevant rules, the use of scientific method, the stock market in the future development prospects were measured. Stock forecasts are more in the face of the following problems. With information overload, numerous technical indicator of the stock market, stock market is influenced by political and economic factors influence and internal rules are very complex and the multi-level, multi event, nonlinear and time variable dynamic nature of making some of the traditional analysis, such as Dow analysis method, chart analysis method, the column chart analysis method, point and figure chart analysis, often fails to predict the stock market dynamic, but merely as a means of analysis of historical data.

Since opening since accumulated stock market direct data quantity slue, data mining technology has obvious advantage in massive data processing and its advantage is from a large database of extracting information of interest, in the implementation of data mining process of support vector machine algorithm, genetic algorithm can properly applied to the prediction of stock market. In the literature, scholars proposed prediction model based on improved neural network stock price [2], but effective against stock prediction, lack of generality; referred to support vector machine (SVM) method to predict, the process is more complicated, the first feature extraction and then to predict [3]; some literature is using data mining methods, from the point of view of association rule study stock market as a whole and interested in stock dynamics [4-5]. Most of the follow-up studies mentioned in the literature are based on the model of the two methods, the research object and the indicators have a slight difference.

In this paper, we choose the genetic algorithm to optimize the parameters of support vector machine; it is because the sensitivity of the support vector machine model to the parameters will bring a great impact on the results of regression prediction [4].

ISSN: 1738-9976 IJSIA

Considering the joint parameters optional value range is wide, the exhaustive method to search the optimal value of infeasible [5], we must adopt an effective search strategy. Compared with other heuristic algorithms, genetic algorithm has a good global search capability. Secondly, the stock market is a very complex nonlinear system, the traditional data analysis methods are generally not on the stock to make is reasonable and effective in predicting, and support vector machine model in dealing with the nonlinear prediction has obvious advantage [6-7]. Therefore, the genetic algorithm optimization vector machine parameter optimization, building support vector machine model to achieve the return on the stock forecast.

2. Algorithm Analysis

2.1. The GA Algorithm Principle

Selection, crossover and mutation are the three steps of genetic algorithm operation. The three steps is the implementation of genetic algorithm core search ability[8]. Each operator has a different effect, in which the selection of the usefulness of the operator is analog to the nature of the mechanism of survival of the fittest, the usefulness of the crossover operator is analogy of genetic breeding and hybrid mechanism, mutation mechanism is analogy of genetic mutations. The three operators can be used as a genetic tool control process.

1) ***Selection:*** selection operator is from the parent population screened individuals with high fitness value to form a new group can be understood as individual according to its own survival ability to copy itself into the next generation of process, reflecting the law of nature "survival of the fittest, survival of the fittest".

2) ***Crossover:*** the crossover operator is operating in two steps. In the selected operator selection bit string which match the pool, a position on by paired; then according to the (random) method choice intersection, a cross point values of exchange sites on group.

3) ***Variation:*** Mutation is similar to gene mutation in the chromosome in nature, the emergence of gene mutation makes species can appear in the ancestors of the characteristics, natural selection can make full of vitality of the new species continue. The GA algorithm in some way on a mutation operator with a smaller probability of random changes (such as 0->1 or 1->0), we can see that the crossover and mutation operators are the main reasons for the evolution of biological diversity.

Genetic algorithm in essence is a kind of stochastic optimization algorithm, but it is different compared to the simple random search, but increased to chromosomal position on the evaluation and chromosomal genes, more effectively use existing information to guide the search process, to achieve the goal of improving quality.

2.2. SVM Regression Prediction Algorithm

Support vector machine was first proposed by Vapnik[6], it is similar to the radial basis function network or multilayer perceptron network, support vector machine is generally used for pattern classification or nonlinear regression process. Its main principle is to establish a classification hyper plane as the decision surface, the positive and negative appear maximization edge isolation.

Regression prediction is defined as[8]: known training set $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(X\times Y)^l$, Where x_i is input data vector set, y_i is the output data, Finding the function f makes $y=f(x)$ as much as possible to satisfy the condition. The regression problem is divided into two kinds of linear and nonlinear, SVM can be used to solve nonlinear regression analysis, the method is mainly through the determination of nonlinear mapping Φ. The

input data vector set X is mapped to the D dimensional feature space F, and the linear mapping is carried out in this high dimensional space [8]. The fitting function is obtained as follows:

$$f(x,\omega)=\omega\cdot\Phi(x)+b$$

Kernel function can be used to determine nonlinear regression function:

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x,x_i)+b \tag{1}$$

The problem is transformed to select the appropriate kernel function $K(x,x')$ and the appropriate parameter C to construct and solve the optimization problem:

$$\min \quad Q(\alpha)=\frac{1}{2}\sum_{i=1}^{j}\sum_{j=1}^{l}y_iy_i\alpha_i\alpha_iK(x_i,x_j)-\sum_{j=1}^{l}\alpha_j \tag{2}$$

$$s.t. \quad \sum_{i=1}^{l}y_i\alpha_i=0, \quad 0\le\alpha_i\le C, i=1,\cdots,l$$

Get optimal solution, : $\alpha^*=(\alpha_1^*,\cdots,\alpha_l^*)^T$ then select a positive component of α^*, $0<\alpha_j^*<C$,and calculate the threshold value:

$$b^*=y_i-\sum_{i=1}^{l}y_i\alpha_i^*K(x_i-x_j) \tag{3}$$

In this paper, we use the kernel function $K(x,x_i)=\exp(-\gamma\|x-x_i\|^2),\gamma>0$ based on the radial basis function to carry out regression prediction.

Based on support vector machine (SVM) machine learning methods, with a small sample, high performance, the advantages of strong generalization ability to effectively overcome learning and the curse of dimensionality, through proper design can meet the stock price forecasting the timing characteristics, this paper stock prediction algorithm is the best choice. Through the analysis of the trend of stock price in recent years, found that stock price random fluctuation is often accompanied by some indexes of the rules change.

Algorithm commonly used support vector machine (SVM) model to forecast the stock, to calculate the corresponding characteristics of daily stock market parameters value, then these data to support vector hangar need text form is saved to file, called last support vector machine training function, the text file data as training samples for training [9]. On support vector machine model training is required prior to the forecast, in training and need to constantly adjust support vector machine key parameters, makes the vector machine can get better training results, this leads to a problem is how to select the parameters, there is no optimal parameter to carry out regression prediction. To build a better stock forecasting model by using support vector machine, the choice of penalty parameter c and kernel function parameter g (radial basis function γ) can be used to make the established model to be applied in the actual forecast well.

Parameters using cross validation can be optimal in some sense, can effectively avoid learning and less learning state, eventually for the test set predictions obtained ideal accuracy [11]. An example shows that the parameters selected by cross validation of the training SVM model than randomly selected training parameters obtained by SVM model is more effective in prediction and classification. The training method is generally cross training, the upcoming sample is divided into N parts, each time to take one of them as the forecast data, and other N-1 data as training samples for support vector machine training

learning [12]. The advantage of this method is that it avoids the over fitting of the sample noise.

3. GA-SVM Optimization of the Stock Market Forecast Model

As mentioned above, the general SVM must be trained, and the training parameters are very important for the prediction model. Using GA to optimize the selection of the penalty parameter c and kernel function parameter g, and its main realization flow chart as shown below.

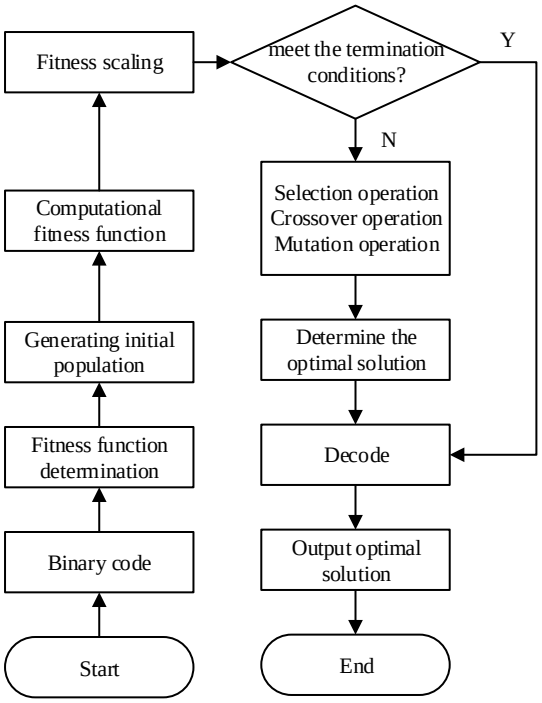

Figure 1. GA-SVM Optimization Flow Chart

The main steps are as follows:

(1) The coding and initial population generation: all of the genetic processes in the GA algorithm are carried out in the coding space, so it is necessary to convert the original problem into the content of the code space. In fact, the GA algorithm is not very high, but the coding strategy will have a great influence on the genetic operators. The most simple encoding method is binary code, in addition as the real code (using decimal method, the solution space direct operation), floating point code. According to the specific problem to choose the coding method, for the two parameters at the same time to optimize the use of the joint coding method to carry out the [13], the first definition of chromosome length $2L$, And in accordance with the experience of the two parameters of each range, will be discrete and binary coding, the length of the chromosome into two parameters, that is, as $X = x_1 x_2 \ldots x_L$, $Y = y_1 y_2 \ldots y_L$, the formation of chromosomes $XY = x_1 x_2 \ldots x_L y_1 y_2 \ldots y_L$,in this model $X = c, Y = g$. At the same time, it can be concluded that the mapping function of each parameter as $c = f_c \cdot X, g = f_g \cdot Y$, the encoding and decoding process can be completed quickly according to the mapping.

Fitness function is constructed: generally used in the process of model design method are data sets a part as the training set to design the model, a part as a test set to test the model. The accuracy of test results is model evaluation standard [14]. In order to describe

the relationship between the accuracy and the predictive goal more clearly, the true positive rate and the true negative rate were introduced into two evaluation indexes.

True positive rate: according to the screening criteria to determine the correct number of patients with the actual number of patients with the percentage as:

$$TPR = \frac{DP}{TP} \times 100\% \tag{4}$$

DP represents the number of patients correctly detected. TP represents the actual number of patients. True negative rate: according to the screening criteria for the correct judgment of the number of patients without disease and the percentage of the number of patients:

$$TNR = \frac{DN}{TN} \times 100\% \tag{5}$$

DN represents number of detected patients, TN represents the actual number of people without disease. In stock forecast we can forecast to rise as a search for true positive, while the prediction falls as the search for a true negative, so that you can through the adjustment of sensitivity and specificity in again strengthened forecast for. For example, in practice, we tend to be more accurate in the future of the stock to buy up. To this end, it is true that the weight of the positive weight is higher, and the weight of the true negative set lower.

$$Fitness = (0.6 * TPR + 0.4 * TNR) * 100\% \tag{6}$$

(3) Genetic operations: in the first chapter of the selection, crossover, and mutation operator operation in the use of a lot of mature methods. The selection operator is based on the classical roulette method, and the crossover operator uses the random crossover method according to the coding method, It has a better set of values with respect to the intersection of 1 points and multi points, and the mutation operator is turning over to a certain position of the chromosome according to the probability of mutation.

4. Empirical Analysis

From the above analysis, the results show that the effect of GA-SVM in the stock market forecast is related to the parameters of SVM, in order to better create a model to determine the parameters, using K cross validation method for experimental design. K-fold cross validation principle is to divide the data set into k subsets, and the loop will k sub set a as a test set, and the rest of the k-1 data set as the training set, at last, the error of K subsets of data to calculate the mean value and K iterations of verification is division of supervised learning algorithm of the evaluation method and data set division generally use the equal and equational or randomly, its specific process as shown in Figure 2.

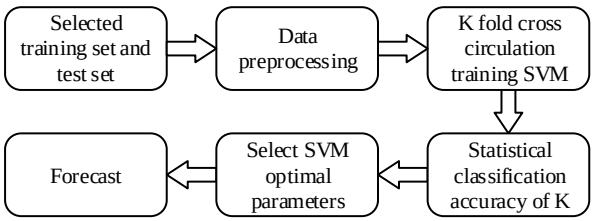

Figure 2.The Experiment Design Process

4.1. Index Selection

In the stock price's technical analysis, there are dozens of commonly used indicators, if we put these dozens of indicators all calculated into the support vector machine training, there will be the following problems:

1) ***Increase the dimension of the problem.*** If the dozens of indicators are used as a feature of the stock sample, then the entire stock sample dimension has dozens of dimensions, and the number of stock samples is limited, it is difficult to accurately classify.

2) ***Large correlation.*** Between these dozens of indicators, there is a lot of strong correlation, so we only need to select a representative index to carry out the feature selection can achieve the same effect.

Whether it is using the traditional statistical methods or the emerging data mining technology, the stock price trend forecast should be selected before a period of time of the stock market data for research. If the time is too long, it will lead to an increase in the amount of calculation and the model is not easy to converge, because the stock market is a highly dynamic process of change. Of course, this period of time can not be too short, because the data is too short to show the impact of the stock market trends. No matter how much data is selected, the design of SVM is consistent with the training process. Therefore, it is very important to select a suitable sample. The Shanghai composite index is chosen as the training and forecasting object of this experiment, because the Shanghai index is one of the important comprehensive evaluation indexes of the domestic stock market, which can accurately reflect the dynamic of the domestic stock market.

From the above analysis, the selected stock index includes as follows: opening price, closing price, the highest price, the lowest four direct evaluation object, Moving Average Convergence Divergence (MACD), relative strength index (RSI) indirect evaluation object. The sample in the next three or more than 1% of the sample marked as “+1”, down more than 1% of the sample marked as “-1”. Other samples were labeled as “0”. So we can observe the Shanghai composite index rising trend.

4.2. Experiment Design

Experiments were selected data from Dec.19, 1990 to Fen.08 2011, making SVM training and parameter optimization process by using total of 4948 trading days of the Shanghai Composite Index data. Experimental environment of the software and hardware conditions are as follows: Windows 7 operating system, 4G transport storage and 1.9GHz CPU, MATLAB2014a, embedded by the design and development of National Taiwan University professor of libsvm software packages. The software package can quickly and effectively SVM pattern recognition and regression, packaging, a lot can directly use functions also provides cross validation function, using the default parameters can be realized conventional SVM. However, the kernel function parameter g of SVM and the penalty function C have not formed a unified model. In this paper, GA is used to optimize the process of SVM parameter optimization, the main parameters of SVM are: the kernel function type: RBF, the range of nuclear parameter g as $[0,1000]$, the range of punishment coefficient C as $[0,1000]$. The main parameters of GA are: population size 40, code length 40, crossover probability 0.9, mutation probability 0.1, and maximum evolution algebra 300.

Raw data must be "cleaned" to be used as input. Data "cleaning" mainly includes the following links:

1) Normalization processing: the normalization of data samples is the main means to eliminate the “large number eat small one”. In data mining, the risk must be normalized because the dimension of the data is not consistent.

Normalized mappings are as follows: $f: x \to y = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$,According to this mapping, the raw data of each attribute index can be normalized to the range of the original data. MATLAB in the mapminmax can be achieved with a function to achieve.

2) Abnormal data processing: three consecutive trading day closing price rise and drop of deviation from the value of a total of $\pm 20\%$ belonging to the abnormal fluctuations, for this kind of data commonly used to eliminate the abnormal data value, use before and after the two trading day average value instead.

After the two steps, the following figure gives the normalized data as shown in Figure 3.

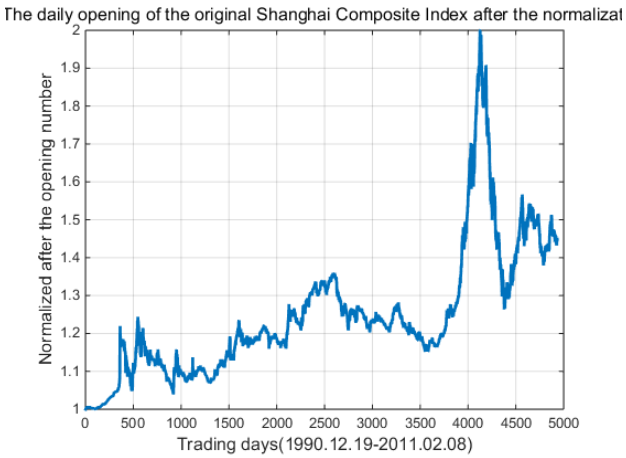

Figure 3. Normalized Shanghai Composite Index Data Chart

K-fold cross validation parameters are obtained as follows: will be "clean" data set into 10 equal sized subsets, each subset respectively as a test set, the rest of the samples as a training set, so a K fold cross validation K model should be established, finally can get k test set their identification rate. The optimal parameter of SVM is selected to optimize the parameters of SVM. Comparing the GA-SVM algorithm with the common SVM model, the superiority of the GA algorithm in the SVM algorithm is verified.

In the K fold cross validation to find better after using optimal parameters corresponding to train the SVM model, for short-term forecasting, through the correct rate and the actual data of comparison can be observed model for prediction is extended.

4.3. Results Analysis

In MATLAB2014a using libsvm software package to carry on the experiment, 10 fold cross validation in the experimental group in each group are GA-SVM optimization algorithm operation, according to the given parameter configuration and chromosome coding decoding method to get the final optimal parameters and classification accuracy rate list is as follows:

Table 1. GA-SVM Parameter Optimization Cross Validation

K	Optimum g	Optimum c	Classification accuracy
1	1.42	52.11	89.02%
2	2.05	45.78	82.14%
3	1.37	54.32	91.06%
4	3.31	50.87	92.73%
5	4.01	53.18	91.55%

6	3.94	49.39	93.37%
7	3.54	48.73	89.58%
8	3.22	51.28	92.97%
9	2.78	50.04	90.13%
10	2.39	48.96	88.12%

From the above table, it shows that the classification accuracy of the experimental group number 8 is the highest, as shown in the following figure.

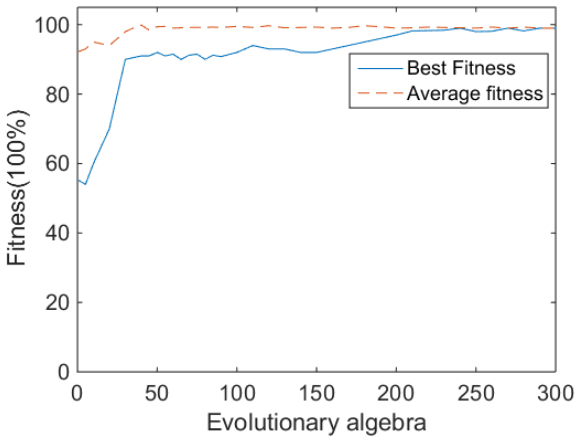

Figure 4. Fitness Curve Optimization Process

In order to verify the effect of GA search on the accuracy of SVM classification, 4 groups of parameter c and parameter g were randomly selected in the range, and the classification accuracy was listed as follows:

Table 2. Random Selection Parameters

Serial	Optimum g	Optimum c	Classification accuracy
1	10.11	78.54	38.17%
2	33.21	120.05	23.18%
3	4.65	60.12	84.06%
4	41.32	201.46	30.29%

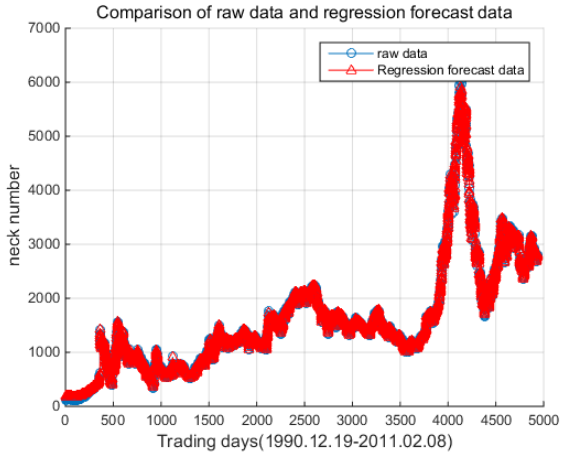

Figure 5. Comparison of Raw Data and Regression Forecast Data

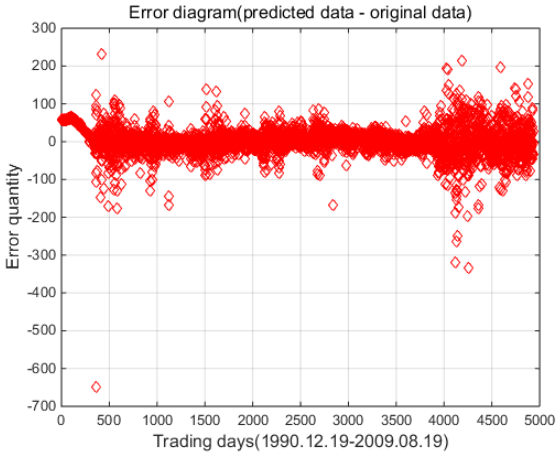

Figure 6. Absolute Error Map

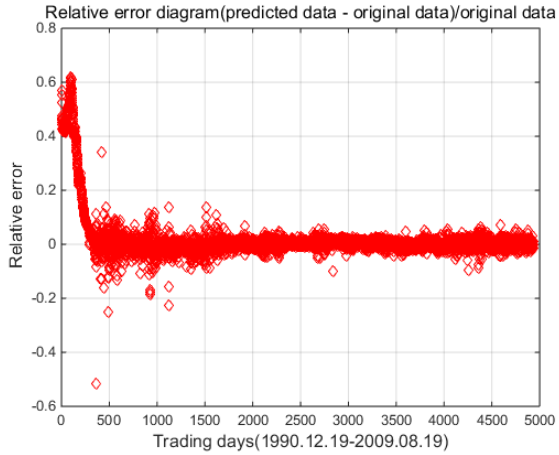

Figure 7. Relative Error Diagram

Through comparison, we can see that the random selection of parameters on the classification accuracy of the high and low uncertainty, after the GA algorithm to find the parameters of the classification accuracy is significantly higher than the results of the random method. The final determination of the parameters of the K number 8 is selected as the best parameter to carry on the follow-up forecast. At this point, the process of establishing a predictive model of GA-SVM is realized. The optimal parameters are obtained from the experimental process. The model is used to predict the short-term 20 days. The results are compared with the results of the Shanghai Composite Index.

As can be seen from the comparison chart regression prediction results and real data error is small, then the statistics of the relative error and absolute error is shown in Figure 6, Figure 7. We can see that the relative error is smaller, calculate the relative average error is 0.11. In conclusion GA-SVM model for the prediction of Shanghai composite index is feasible, the regression prediction accuracy is higher, relative errors illustrate the model for Shanghai composite index forecasting more reasonable and feasible.

5. Conclusions

In conclusion, Shanghai index in the domestic stock market has a role in the navigation mark; the trend represents the majority of the normal form of the trend of the stock. The

index values affect the vast majority of trading activities in the stock market. In this paper, the GA algorithm is used to search for SVM to establish the forecasting model of Shanghai Composite Index. Using K - fold cross - Test Method to model the model and improve the accuracy and robustness of the model. At the end of the paper, the Shanghai index is analyzed. By regression prediction and calculation error, it is concluded that the prediction model is accurate and can be used as a model for the prediction of Shanghai stock index.

Acknowledgments

The work of this paper is supported by National Natural Science Foundation of China (No.71302153); China Post-Doctoral Program (2014T70838); Shanghai key discipline construction project (B310).

References

[1] Z. Hongbo, "Genetic algorithm and evolutionary support vector machine", Journal of Shaoxing University, vol. 24, **(2004)**, pp. 25-28.
[2] M. Chaoqun and G. Renxiang, "Modern forecasting theory and method", Hunan University press, **(1998)**, pp. 12-16.
[3] L. Chunyan and Z. Donghua, "The forecast the stock price based on Elman neural network", Application Research of computers, vol. 10, **(2006)**, pp. 55-60.
[4] Z. Wei, "Based on the prediction of stock trend SVM algorithm optimized by GA", Jilin University, **(2010)**, pp. 102-106.
[5] Z. Shengquan, "the stock price prediction based on data mining", Huazhong University of Science and Technology, **(2009)**.
[6] Z. Shajun and X. Xianwen, "Data mining stock analysis software and CD-ROM application system design", computer software and application, vol. 17, **(2011)**, pp. 19-22.
[7] V. Vapnik and E. Lecin, "Measuring the VC-dimension of a learning machine", Neural Computation, vol. 6, **(1994)**, pp. 851-876.
[8] L. Yu, "a support vector machine based stock market trend prediction", Investment and cooperation, vol. 9, **(2013)**, pp. 44-47.
[9] Z. Yanlai, "the application of data mining in stock investment", Capital University of Economics and Business, **(2010)**, pp. 80-82.
[10] L. Yu, S. Wang and K. Lai, "Mining stock market tendency using GA-based supportvector machines", WINE, LNCS. Berlin: Springer, **(2005)**, pp. 336-345.
[11] O. Chapelle, V. Vapnik and O. Bousquet, "Choosing kernel parameters for support vector machines", Machine Learning, vol. 46, **(2001)**, pp. 131-160.
[12] W. Jinglong, Y. Shuxia and L. Chengdong, "Short term load forecasting support vector machine parameter optimization genetic algorithm method", Journal of Central South University, vol. 40, **(2009)**, pp. 180-183.
[13] L. M. Li, G. R. Wen and W. Shengchang, "Genetic algorithm of regression support vector machine parameter selection method", Computer engineering and applications, vol. 44, **(2008)**, pp. 23-26.

Authors

Liang Banglong, 1980.7, YunCheng, ShanDong, China.

Current position, grades: Ph.D. of School of Economics and Management, Tongji University, Shanghai, China.

Scientific interest: His research interest fields include Decision support systems, big data of financial and etc.

Publications: more than3papers published.

Experience: He has completed one scientific research projects.

Lin Jie, 1967.11, QuXian, SiChuan, China.

Current position, grades: Doctoral tutor, Professor of School of Economics and Management, Tongji University, Shanghai, China.

Scientific interest: His research interest fields include Financial decision-making systems, electronic commerce, etc.

Publications: more than 130 papers published and 5 textbooks edited.

Experience: He has teaching experience of 30 years, has completed nine scientific research projects.