Corn futures price forecast based on Arima time series and support vector machine

Wu Shengchao[a], Shao Fengjing[b,*], Sun Rencheng[c]

School of Qingdao University, Shandong 266001, China.

[a]17853252521@163.com, [b,*]sfj@qdu.edu.cn, [c]qdusunstar@163.com

Keywords: Arima time series, Corn futures price

Abstract: This paper applies arima time series and support vector machine combination to predict corn futures prices. At present, when the price forecast is made on the Chinese futures market, the data is unstable. Since linear prediction or nonlinear prediction alone cannot flexibly deal with corn futures price forecasting problems, the combination forecasting of Arima time series and support vector machine has a prominent advantage in dealing with price forecasting problems.

1. Introduction

With the deepening of the reform of the grain circulation system, the market will play a fundamental role in the process of arranging food resources and forming food prices. While playing the role of the market, it is of vital importance to strengthen and improve macroeconomic regulation and control in order to ensure the basic stability of food supply and prices. Therefore, it is of great theoretical significance and more practical to study the impact of food prices on market supply and demand. significance. Then, how to predict the trend of futures prices, help investors avoid risks and increase profits has become a major research direction in the futures field. Since the emergence of the futures trading market, many scholars have conducted a lot of research on futures prices. The basic principle is to use the past and present futures prices to predict future futures price trends, and provide a reference for investors. In the past, people mainly used regression and other means to establish models to make predictions. Since futures prices are highly nonlinear and have high redundancy due to factors such as national policies, price indices, and trading varieties, they are used. The traditional regression algorithm accurately predicts the complex data of futures prices, which leads to the problem that the prediction accuracy is low and the forecast of the delivery price cannot be met. In recent years, due to the breakthrough development of computer hardware and superb computing power[1], complex neural networks based on artificial intelligence have begun to make breakthroughs in prediction and classification[2], and have achieved good results in practice. However, the neural network has higher requirements on computing hardware. Moreover, in the actual application process, the learning algorithm of BP neural network is based on gradient descent, so it has slow training speed, easy to fall into local minimum value, poor global search ability, etc. Disadvantages, the neural network can't achieve better results in the field of prediction. In contrast, although the Arima time series model and the SVM support vector machine model have their own advantages and disadvantages, they have advantages for linear and nonlinear model processing, and the advantages between the two are complementary. Forecasting corn futures prices, using Arima time series model and SVM support vector machine model to achieve a good prediction effect.

2. Corn futures price method based on Arima time series and SVM combination

2.1 Arima time series

The ARIMA model, called the Autoregressive Integrated Moving Average Model, was proposed by Box and Jenkins in the early 1970s for a well-known Time-series Approach prediction method. The so-called ARIMA model refers to a model that converts a non-stationary time series into a stationary time series and then returns the dependent variable only to its lag value and the present

  DOI: 10.25236/icscbd.2018.007

and lag values of the random error term. The ARIMA model includes the moving average process (MA), the autoregressive process (AR), the autoregressive moving average process (ARMA), and the ARIMA process depending on whether the original sequence is stationary or not[3]. The data sequence formed by the predicted object over time is regarded as a random sequence, and a certain mathematical model is used to approximate the sequence. Once this model is identified, it can predict future values from past and present values of the time series.

2.2 Support Vector Machine

Support Vector Machine (SVM) is a statistical learning theory learning method based on the principle of structural risk minimization. It has strong generalization ability and overcomes the over-fitting in neural network, slow convergence and easy to fall into local pole. Shortcomings such as value have broad application prospects in the field of economic forecasting[4].

The basic model of Support Vector Regression (SVR) is to define the classifier with the largest separation in the feature space. It is a two-class classification model. When kernel techniques are used, support vector machines can be used for nonlinear classification.

Support vector machine is the most practical content in statistical learning theory. It is transformed from a special type of hyperplane, the so-called optimal classification hyperplane, in the case of linear separability, and then the problem is transformed into a convex quadratic programming problem.

Given a training set on a feature space T=$\{(\overrightarrow{x_1}, y_1), (\overrightarrow{x_2}, y_2), \dots, (\overrightarrow{x_N}, y_N)\}$, $\overrightarrow{x_i}=(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})$,T∈x=$R^n$,$y_i \in y = \{+1, -1\}$,i=1,2,…,N .where $\overrightarrow{x_i}$ is the ith instance and y_i is the token of $\overrightarrow{x_i}$:

if $y_i = +1$ then $\overrightarrow{x_i}$ is to be Positive example

if $y_i = -1$ then $\overrightarrow{x_i}$ is to be Negative example

Given a linearly separable training data set T, assume that the separated hyperplane obtained by interval maximization learning is $\vec{w}^*.\vec{x} + b^* = 0$. Defining classification decision function, $f(\vec{x}) = sign(\vec{w}^*.\vec{x} + b^*)$. This classification decision function is also called a linear separable support vector machine. For linear separable support vector machines, the distance between a sample distance and the hyperplane can usually be separated to indicate the reliability of classification prediction: the farther a sample is from the separation hyperplane, the more reliable the classification of the sample; the closer the sample is to the hyperplane The classification of the sample is less convincing.

The goal of the SVM is to solve the separation hyperplane that can correctly divide the training data set and has the largest geometric interval[5]. The maximum geometric spacing here is also known as the hard interval maximization.

Based on the relationship between geometric spacing and function spacing, the solution problem is:

$$\max_{\vec{w},b} \frac{\hat{\gamma}}{||\vec{w}||} \tag{1}$$

s.t. $y_i(\vec{w}.\overrightarrow{x_i} + b) \geq \hat{\gamma}$,i=1,2,…,N

Construct and solve optimization problems:

$$\max_{\vec{w},b} \frac{1}{2}||\vec{w}||_2^2 \tag{2}$$

s.t. $y_i(\vec{w}.\overrightarrow{x_i} + b) - 1 \geq 0$,i=1,2,…,N

Define the Lagrangian function:

$$L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}||\vec{w}||_2^2 - \sum_{i=1}^{N} \alpha_i y_i(\vec{w}.\overrightarrow{x_i} + b) + \sum_{i=1}^{N} \alpha_i \tag{3}$$

$\vec{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$ is Lagrange multiplier vector

Find the optimal solution : $\vec{w}^*, b^*$

This gives the separation hyperplane: $\vec{w}^*.\vec{x} + b^* = 0$,Classification decision function is

$f(\vec{x}) = sign(\vec{w}^*.\vec{x} + b^*)$. Support vector machines can be used not only for classification problems, but also for regression problems. Given training data set: T=$\{(\overrightarrow{x_1}, y_1), (\overrightarrow{x_2}, y_2), \dots, (\overrightarrow{x_N}, y_N)\}$, $\overrightarrow{x_i}=(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})$T∈x=$R^n$,$y_i \in y = \{+1, -1\}$,i=1,2,…,N.

For the sample $(\overrightarrow{x_i}, y_i)$, the loss is usually calculated from the difference between the model output $f(\overrightarrow{x_i})$ and the true value y_i, and the loss is only if $f(\overrightarrow{x_i})= y_i$ zero.

The basic idea of support vector regression is to allow up to ε deviation between $f(\overrightarrow{x_i})$ and y_i We only think that the prediction is correct when $|f(\overrightarrow{x_i})- y_i|>\varepsilon$.

Describe SVR problems in mathematical language:

$$\max_{\vec{w},b} \frac{1}{2}\left|\left|\vec{w}\right|\right|_2^2 + C\sum_{i=1}^{N} L_\varepsilon(f(\overrightarrow{x_i}) - y_i) \tag{4}$$

Support vector machine regression prediction is obtained by minimizing the empirical risk on the training set. The loss function used has the form of square error and absolute value error. Where C≥0 is the penalty constant, and the larger the value, the higher the degree of fitting. L_ε is the ε insensitive loss function, ε is the maximum error allowed by the regression[6], and the number of support vectors and the generalization ability are controlled. The larger the value, the smaller the support vector, L_ε is defined as:

$$z = f(\vec{x}) - y \tag{5}$$

$$L_\varepsilon(z) = \begin{cases} 0 & |z| \le \varepsilon \\ |z| - \varepsilon & |z| > \varepsilon \end{cases} \tag{6}$$

Furthermore, in order to make the same model obtained on the training set have better generalization ability, it is necessary to consider not only the minimization of empirical risk, but also to reduce the complexity of the model. Introducing the slack variables ξ_i, ξ_j, the new optimization problem is:

$$\max_{\vec{w},b,\xi_i,\xi_j} \frac{1}{2}\left|\left|\vec{w}\right|\right|_2^2 + C\sum_{i=1}^{N} L_\varepsilon(\xi_i+\xi_j) \tag{7}$$

Assume that the final solution is $\vec{\alpha}^* = (\alpha_1{}^*, \alpha_2{}^*, \dots, \alpha_N{}^*)^T$, $\hat{\vec{\alpha}}^* = (\widehat{\alpha_1}^*, \widehat{\alpha_2}^*, \dots, \widehat{\alpha_N}^*)^T$, find a certain component of $\vec{\alpha}^*$ (C>$\alpha_j{}^*$>0), then:

$$b^* = y_i + \varepsilon - \sum_{i=1}^{N}(\widehat{\alpha_i}^* - \widehat{\alpha_j}^*)\overrightarrow{x_i}^T\overrightarrow{x_j} \tag{8}$$

$$f(\overrightarrow{x_i}) = \sum_{i=1}^{N}(\widehat{\alpha_i}^* - \widehat{\alpha_j}^*)\overrightarrow{x_i}^T\overrightarrow{x_j} + b^* \tag{9}$$

Further, if you consider using the kernel technique, given the kernel function $K(\overrightarrow{x_i}, \vec{x})$, the SVR can be expressed as:

$$f(\overrightarrow{x_i}) = \sum_{i=1}^{N}(\widehat{\alpha_i} - \alpha_j)K(\overrightarrow{x_i}, \vec{x}) + b \tag{10}$$

However, SVR is sensitive to missing data. There is no general solution to nonlinear problems. It is necessary to carefully select kernel functions for processing, and the computational complexity is high.

3. Corn futures price forecast based on Arima model and SVM model

3.1 Selection of samples and characteristic indicators

Since the overall price of food is reflected in the trend of increasing year by year, the sampling time of the sample data should not be too far from the actual time. Therefore, the corn trading data from January 1, 2017 to December 29, 2017 is selected. The date is not continuous) as sample data for the study. At the same time, referring to the indicators of corn market trading, and through the random forest algorithm for feature selection, select the corn opening price, the highest price, the lowest price, the closing price, the trading volume, and the transaction amount as the characteristic

indicators of the research sample data.

3.2 Data preprocessing

Noise is stored in the raw data of the sample data. Due to the presence of noise, machine learning is more difficult, and the prediction result may cause large errors. The processing of noise will depend on actual needs. In the course of this study, the main source of noise was found to be data missing and the data was 0. For the data of this anomaly, we used the mean prediction estimation method. The data of the first three days and the last three days of the abnormal data are averaged, and the abnormal data is replaced to reduce the impact of abnormal data on corn futures price forecast.

3.3 Analysis and prediction of data

Figure 1 is the time trend chart of corn futures opening price (kpj), highest price (zgj), lowest price (zdj), closing price (spj), trading volume (cjl), closing amount (cjje), settlement price (jsj) .

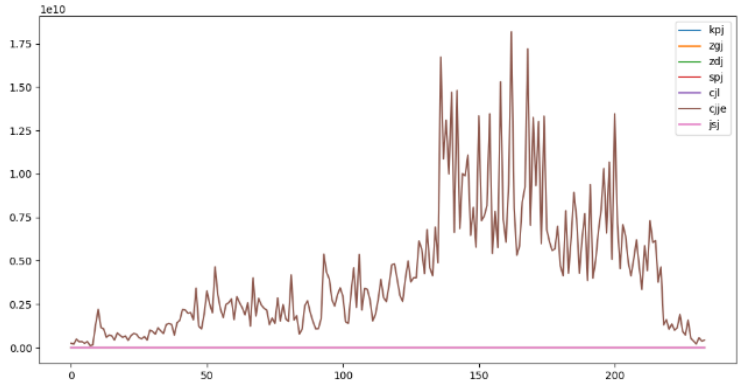

Figure 1. Corn futures chart

The trend of corn opening price (kpj), highest price (zgj), lowest price (zdj), closing price (spj), settlement price (jsj) is shown in Figure 2.

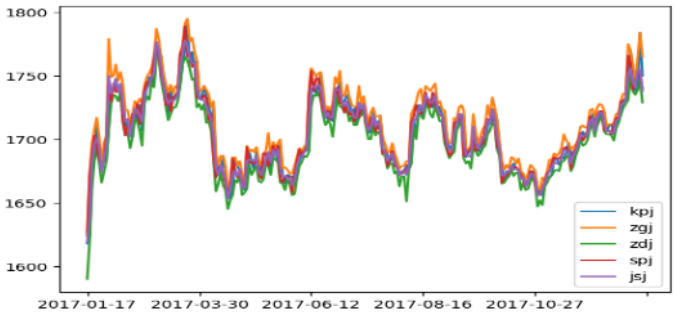

Figure 2. Overlap trend chart

The autocorrelation and partial correlation are shown in Figure 3.

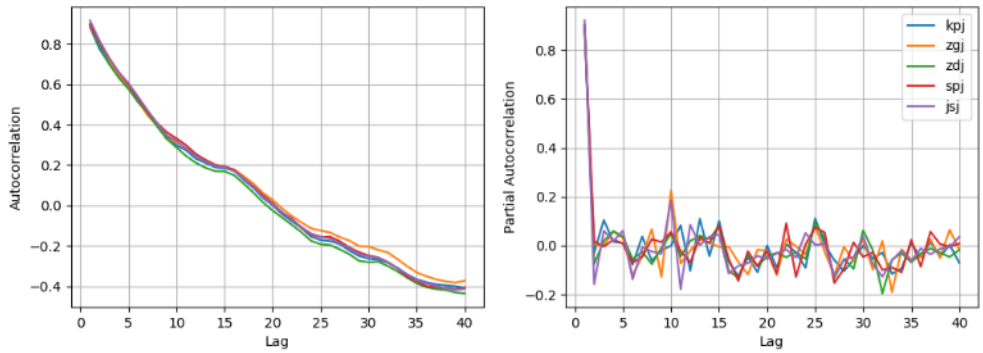

Figure 3. Autocorrelation and partial correlation diagram

Next, the Arima model is used for prediction. It is assumed that the time series Y_t is regarded as a combination of the linear autocorrelation part L_t and the nonlinear residual N_t, namely:

$$Y_t = L_t + N_t$$

Firstly, the linear part is modeled by Arima model. Let the prediction result be $\hat{L}_t$, the residual of sequence L_t is N_t, and $N_t$contains the nonlinear relationship of sequence Y_t.

Convert preprocessed data into a time series:

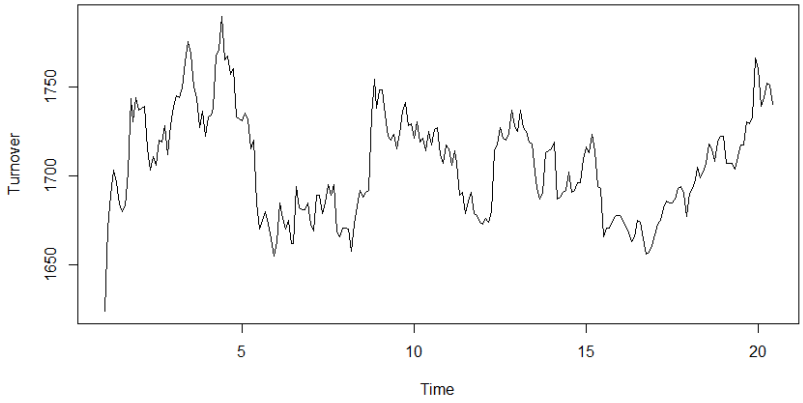

Figure 4. Time series chart of corn closing price

It can be clearly seen from the data that the transaction amount of corn has obvious seasonality. In the autumn and winter of the second half of the year, it is the harvest season of corn, the grain reserve is large and the transaction is frequent, the price fluctuates greatly, and the average price will be higher. However, the seasonality of the data has a greater impact on the forecast of corn futures prices, so the need to split and remove seasonal processing. Next, the data is subjected to seasonal processing to make the data curve more stable[7].

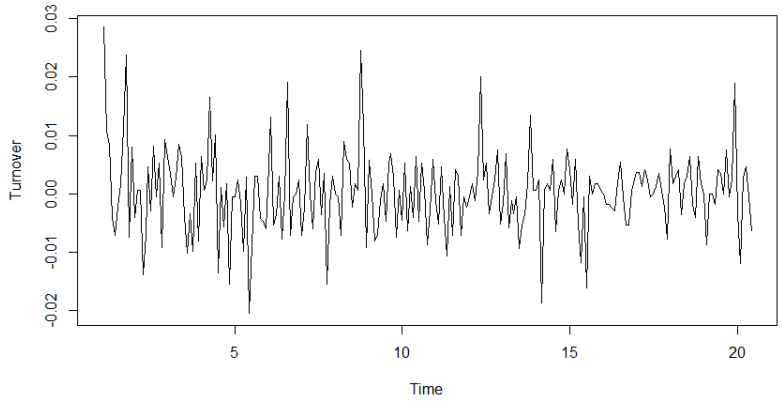

Figure 5. Time series of corn transaction amount to remove seasonal map

Since the direct analysis of time series by non-stationary time series leads to the problem of pseudo-regression, it is necessary to perform ADF detection on the model to see the stability of the model.

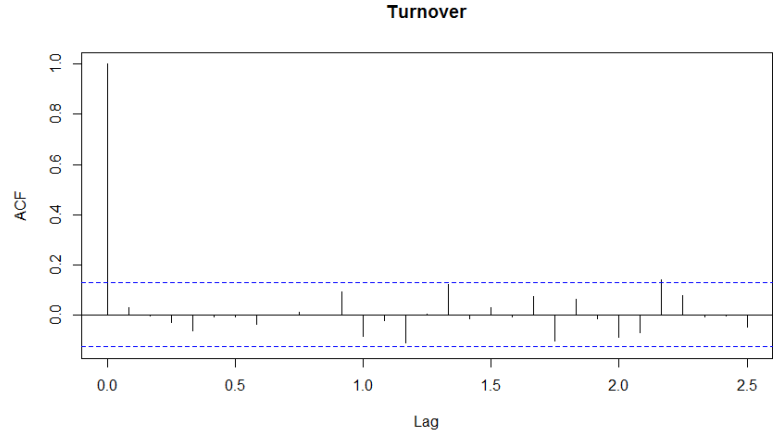

Figure 6. Time series ADF test chart

In time series analysis, the partial autocorrelation function (PACF) gives a partial correlation of the time series with its own hysteresis value, controlling all-time series values with shorter lags. It is in contrast to the autocorrelation function, which cannot control other lags. The ADF test is performed on the time series below.

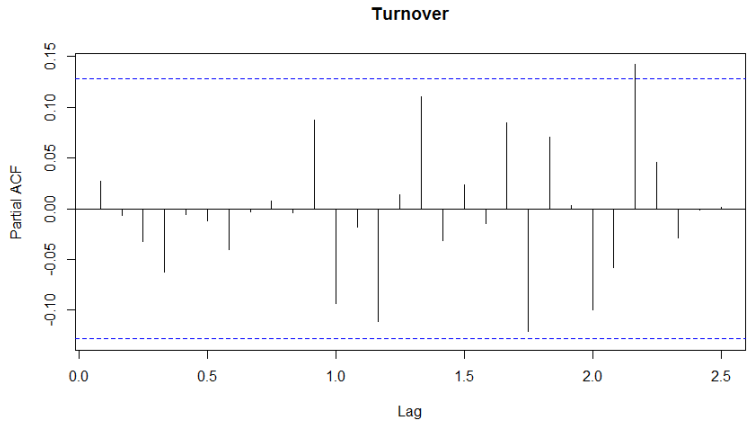

Figure 7. Time Series ADF Test Chart (PACF)

As can be seen from Figures 6 and 7, the stability of the model is general, but still within a stable range. Therefore, this Arima model has a certain stability. The prediction analysis starts below, and

the results of the prediction analysis are shown in Fig. 8.

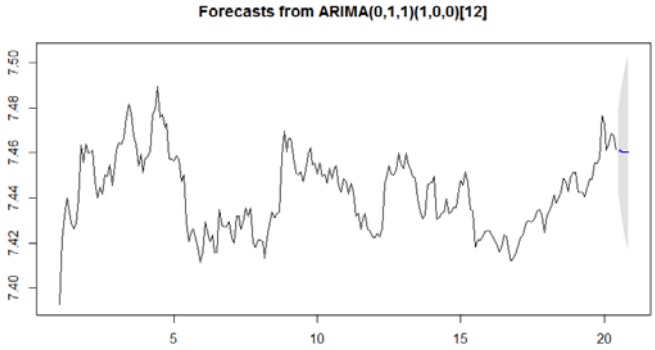

Figure 8. ARIMA predictive analysis results

Make the rolling residuals of the prediction result and the sample result 4 times, and use the first 4 columns as the input of svm and the 5th column as the output.

Next, use SVR for predictive analysis. The following table provides an estimate of the predicted results of SVR for corn futures data[8].

Table 1. Estimate of the predicted results of SVR

Estimate of the predicted results	value
Default evaluation value of linear kernel function support vector machine	0.9939249728330464
R_squared value of linear kernel function support vector machine	0.9939249728330464
Mean Square Error of Linear Kernel Function Support Vector Machine	4.733952267177894
Average kernel error of linear kernel function support vector machine	1.4669229207799146
Default evaluation of polynomial kernel functions	0.04795421254697785
R_squared value for a polynomial kernel function	0.9939249728330464
Mean square error of polynomial kernel function	4.733952267177894
Average absolute error of a polynomial kernel function	1.4669229207799146

The SVM model uses Gaussian kernel rbf support vector review prediction, as shown in Figure 9.

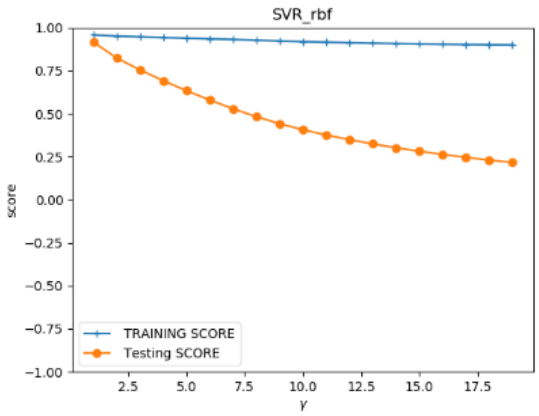

Figure 9. Gaussian kernel rbf support vector review prediction chart

3.4 Accuracy of prediction results

The predicted result of the corn futures price is subtracted from the test value, and the accuracy of the predicted result is evaluated according to the range of the obtained difference. The results of

verification according to this method are -0.25489, -0.43767, -0.37086, -1.07379, etc., all within the range of accurate prediction.

The prediction results after combining SVR are basically consistent with the results of the test set, as shown in Figure 10.

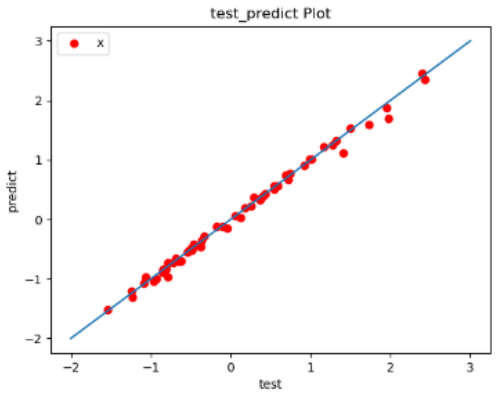

Figure 10. Prediction results and test set results (horizontal axis is the test value, vertical axis is the predicted value)

From this, the predicted results are within the range of an accurate result, and indicate that the futures price of corn is gradually rising during the forecasting stage.

4. Conclusion

The Arima time series model has the advantages of simple model, only need endogenous variables and no need for other exogenous variables, but it requires the data of the series to be relatively stable or stable after differential processing, and can only capture linearity in nature relationship. The SVM theory provides a way to avoid the complexity of high-dimensional space[9], directly use the kernel function of this space, and then directly solve the decision problem of the corresponding high-dimensional space by solving the problem in the case of linear separability. When the kernel function is known, the difficulty of solving high-dimensional space problems can be simplified, and the problem of complex and unstable data generation can be solved, but it is a problem to obtain its kernel function. Combining the two with each other, taking advantage of it and making predictions will simplify the problem and result in more accurate results.

Acknowledgments

This work was financially supported by System Model Research and Application of Big Data Analysis for Marine Disasters (41476101), National Natural Science Foundation (950,000, 2015.01-2018.12) fund.

References

[1] Wang Hai-jun.* Futures Price Forecast Based on Quadratic Optimization BP Neural Network Practice and understanding of mathematics 2008

[2] Zhang Kai Research on Oil Futures Price Forecast Based on Nonlinear Combination Method Computer Simulation 2012,7 (379-382)

[3] Price Time Series Analysis and Forecast Based on ARIMA——Taking Shanghai Aluminum 1803 Contract as an Example [J]. Ying Shaohua. Economic and Trade Practice. 2018(10)

[4] Yang Jianhui, Li Long. Option Price Forecasting Model Based on SVR [J]. Systems Engineering Theory and practice, 2011, 31 (5): 848-854.

[5] Zhang Yu, Yin Tengfei. Application Research of Support Vector Machine in Tax Forecasting

[J]. meter.

[6] Lawrence S. , Giles C. L. , Tsoi A. C.Lessons in neural network training：Overfitting may be harder thanexpected [C]Proceedings of the Fourteenth National Conference on Artificial Intelligence, Mento Park, CA：AAA1 Press ,1997.Computer Simulation, 2011, 28(9): 357 -360.

[7] Traffic flow prediction based on ARIMA and wavelet neural network combined model [J]. Cheng Yun, Cheng Xiaogang, Tan Miaomiao, Zhou Kai, Li Haibo. Computer Technology and Development. 2017(01)

[8] F Gori,D Ludovisi,P F CerritellForecast of oil price and consumption in the short term under three scenarios: parabolic,linear and chaotic behavior[J]. Energy,2007,32(7): 1291 − 1296 .

[9] Nello Cristianini, John Shawe-Taylor. Li Guozheng, Wang Meng, Zeng Huajun translation. Support Introduction to Vector Machines [M]. Beijing: Publishing House of Electronics Industry, 2004.