Open Journal of Statistics, 2018, 8, 660-669
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Oil Price Forecasting Based on EMD and BP_AdaBoost Neural Network

Huifang Qu, Guoqiang Tang*, Qiying Lao

College of Science, Guilin University of Technology, Guilin, China
Email: *tanggq@qlut.edu.cn

How to cite this paper: Qu, H.F., Tang, G.Q. and Lao, Q.Y. (2018) Oil Price Forecasting Based on EMD and BP_AdaBoost Neural Network. *Open Journal of Statistics*, **8**, 660-669.
https://doi.org/10.4236/ojs.2018.84043

Received: July 5, 2018
Accepted: July 31, 2018
Published: August 3, 2018

Abstract

Empirical mode decomposition (EMD) and BP_AdaBoost neural network are used in this paper to model the oil price. Based on the benefits of these two methods, we predict the oil price by using them. To a certain extent, it effectively improves the accuracy of short-term price forecasting. Forecast results of this model are compared with the results of the ARIMA model, BP neural network and EMD-BP combined model. The experimental result shows that the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and Theil inequality (U) of EMD and BP_AdaBoost model are lower than other models, and the combined model has better prediction accuracy.

Keywords

Empirical Mode Decomposition (EMD), BP_AdaBoost Model, Oil Price

1. Introduction

Crude oil is part of the most important financial instruments in the commodity market. Predicting the price fluctuations and trends of the crude oil market accurately is very significant for the country, enterprises, financiers and investors. However, crude oil price fluctuations usually present non-stationary, complex, non-linear, long-term memory characteristics. And the crude oil price forecast is a major difficulty in commodity research. With the development of the crude oil market, it is particularly important to use appropriate decomposition methods and establish appropriate time-series prediction models to forecast oil prices.

In recent years, people have been paying more and more attention to the application of multi-scale decomposition methods in non-stationary financial time series. The multi-scale decomposition methods are mainly wavelet analysis methods and empirical mode decomposition methods. Wavelet analysis can perform multi-scale analysis on signals in the time domain and frequency domain, and gradually refine the original sequence into sub-sequences of different frequencies [1]. The neural network method based on wavelet transform has been extensively used in financial time-series analysis, but wavelet analysis still has some defects, which cannot be adaptively decomposed, because wavelet transform is generated on the basis of Fourier transform. The essence is still the window-adjustable Fourier transform, and there are still limitations of the Fourier transform. Therefore, wavelet analysis cannot perform multi-scale analysis well, and it also generates false harmonics when simulating the original signal.

Huang N. E. proposed an empirical mode decomposition (EMD) method in 1998; it is in a position to smooth nonlinear, non-stationary raw time series data signals while maintaining the originality as much as possible in the decomposition [2]. At present, many scholars use EMD methods and artificial neural network combined forecasting methods to predict financial time series. Yu *et al.* combined EMD with a fuzzy neural network (FNN) to predict international crude oil futures prices [3]. Zhang *et al.* used the EMD method to analyze the basic characteristics of the oil price series on different time scales, and pointed out that it can be predicted by using decomposition models combined with SVM and other prediction models [4]. Islam used the EMD method to decompose the financial time-series, and compared it with wavelet decomposition. The results show the effect of EMD decomposition is better [5]. Tao constructed an innovative multi-period oil futures price forecasting model based on EMD-based FNN model [6]. Wei proposed the ANFIS algorithm based on EMD decomposition and FNN model to predict Taiwan TAIEX and HSI stock indexes [7].

Scholars use BP neural network combined with EMD method to make predictions. However, BP neural network has the effects of local minimum value, slow convergence rate and poor generalization ability of the model. The AdaBoost algorithm can improve the prediction accuracy of the set weak predictor, and solves many problems that the weak predictor does not predict well. Therefore, in order to make up for the limitation of BP neural network weight initialization and sample data, and improve the prediction accuracy of BP neural network and EMD method, this paper proposes a BP_AdaBoost model time-series prediction method based on EMD method, and applies the model into crude oil. The empirical results are shown that the prediction accuracy of the model are preferable to the ARIMA model, BP neural network and EMD-BP combined model.

2. Fundamental Principles

2.1. Empirical Mode Decomposition

Huang N.E proposed the concept of the Intrinsic Mode Function (IMF) to represent the indigenous features of the signal. At any time, a signal can be the sum of a finite number of IMFs. Huang pointed out that the part obtained by decomposition must meet the following two conditions to be the IMFs: 1) The number of extreme points and the number of zero-crossing points are at most one difference; 2) The mean value of the upper and lower envelopes formed by the points of maximum and minimum is equal to zero. The specific decomposition steps are as follows [8]:

Step 1: Determine the local maximum point and the local minimum point of the original sequence $X(t)$, and then interpolate it with the cubic spline function to obtain the upper envelope sequence value $U_1(t)$ and the lower envelope sequence value $L_1(t)$ respectively.

Step 2: Calculate the mean of the upper envelope $U_1(t)$ and the lower envelope $L_1(t)$ at each moment to obtain the instantaneous average $m_1(t)$, $m_1(t)=(U_1(t)+L_1(t))/2$.

Step 3: The original sequence $X(t)$ subtracts $m_1(t)$ can obtain the difference sequence $h_1(t)$, $h_1(t)=X(t)-m_1(t)$. If $h_1(t)$ meet the two assumptions of the IMF, $h_1(t)$ is an intrinsic mode function. Then make $c_1(t)=h_1(t)$. If $h_1(t)$ is not met the two assumptions, think of $h_1(t)$ as $X(t)$. Repeat the above steps until the empirical model function is met the definition of intrinsic mode function.

Step 4: With the original sequence original $X(t)$ subtract $c_1(t)$, it can obtain the residual sequence $r_1(t)$, $r_1(t)=X(t)-c_1(t)$. Then, make $r_1(t)$ as the original sequence. Repeat steps 1 - 4 until the obtained residual sequence $r_n(t)$ is a monotonic function and cannot be extracted. At this time, the original sequence can be expressed as

$$X(t)=\sum_{i=1}^{n}c_i(t)+r_n(t) \tag{1}$$

Among them, the number of IMFs is *n*. The residual is $r_n(t)$ which represents the long-term trend of the original sequence; $c_i(t)$ represents the IMF component, $c_1(t),c_2(t),\cdots,c_n(t)$ represents the part of the original sequence with different frequencies from high to low.

2.2. Basic Principles of BP_AdaBoost Model

The AdaBoost algorithm is just an iterative algorithm. The core idea of the algorithm is to process the same test sample data, obtain multiple weak predictors, and then obtain the weight of different weak predictors through training, and finally combine the outputs of multiple weak predictors to form strong predictor. The weak predictor in the BP_AdaBoost model is a BP neural network. Depending on the prediction result of each weak predictor, changing the weight of the training sample. And train the weak predictor of BP neural network. Finally, the output of the BP neural network weak predictor is combined to form a strong predictor. The specific algorithm steps are as follows [9]:

Step 1: Initialize the distribution weight of the samples and the BP neural network. Select m training samples in the sample data, and initialize the distribution weight of the training sample $D_t(i)=1/m$. The number of input layer nodes and output layers in the BP neural network are determined by the sample input feature dimension and the output result dimension, respectively, and the weight and threshold of the BP network are initialized.

Step 2: Preprocess the data. The data is normalized so that the reprocessed data can be read by the BP neural network weak predictor.

Step 3: Weak predictor prediction. When the t-th BP weak predictor is trained through the training samples, the prediction error ε_t of the prediction sequence $g(t)$ can be obtained according to the BP neural network output, and the formula is

$$\varepsilon_t = \sum_{i=1}^{m} D_t(i) \quad \text{when } g(t) \neq y \tag{2}$$

$g(t)$ is forecast results for the network, y is expected value.

Step 4: Calculate the weight of the prediction sequence $g(t)$, and use the sum of the prediction error in Equation (2) to calculate the weight of $g(t)$. The formula is:

$$w_t = \frac{1}{2} * \ln\left[(1-\varepsilon_t)/\varepsilon_t\right] \tag{3}$$

Step 5: Update the sample weight. The next round of sampling data weight is adjusted by the predicted sequence weight w_t, and the mathematical expression is

$$D_{t+1}(i) = \frac{D_t(i)}{B_t} * \exp\left[-w_t y_i g_t(x_i)\right] \tag{4}$$

Step 6: Output the strong predictor. After training T time, T weak prediction functions are obtained, then the strong prediction function is:

$$h(x) = \sum_{t=1}^{T} \frac{w_t}{\sum_{t=1}^{T} w_t} * f(g_t, w_t) \tag{5}$$

3. Oil Price Prediction Model Based on EMD and BP_AdaBoost Model

Using algorithms to predict the crude oil prices. Figure 1 is based on the EMD method and BP_AdaBoost model oil price forecasting process.

The specific modeling steps are as follows:

1) Determine sample data. Suppose the sample sequence is $X = (x_1, x_2, \cdots, x_n)'$, and n is the number of sample sequences.

2) Perform a stationarity test on the sample sequence X to determine whether it is stable.

3) After decomposing by the EMD method, $t-1$ IMF components and a residual component are generated.

$$(x_1, x_2, \cdots, x_n)' \overset{EMD}{\Rightarrow} \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1t} \\ f_{21} & f_{22} & \cdots & f_{2t} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n1} & f_{n2} & \cdots & f_{nt} \end{pmatrix} \Rightarrow (F_1, F_2, \cdots F_t) \tag{6}$$

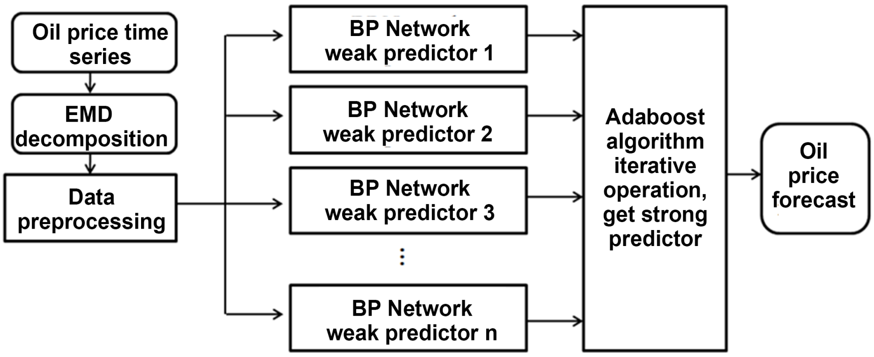
Oil price time series
EMD decomposition
Data preprocessing
BP Network weak predictor 1
BP Network weak predictor 2
BP Network weak predictor 3
BP Network weak predictor n
Adaboost algorithm iterative operation, get strong predictor
Oil price forecast

Figure 1. Oil price prediction flow chart based on EMD method and BP_AdaBoost model.

Among them, $F_i, i = 1, 2, \cdots, t-1$ is the intrinsic mode function obtained by decomposition, and F_t is the residual component.

4) Perform data preprocessing.

$$v_i = \frac{F_i - F_{\min}}{F_{\max} - F_{\min}} \tag{7}$$

$v_i, i = 1, 2, \cdots, t$ is the normalized value.

5) Identify the structural parameters of the BP_AdaBoost model. The normalized IMFs and residual component will form several BP neural network weak predictors. And depending on the principle of BP_AdaBoost algorithm. The weight of the BP weak predictor will be continuously updated. The error will be rectified repeatedly. Network training will be carried out, and finally several predictors are combined to output a strong predictor.

4. Empirical Research

4.1. Selection of Sample Data

This paper selects the daily closing price of brent crude oil from November 28, 2014 to March 18, 2018 as an experiential research object. The data is from EIA, 843 samples in total. The whole data is divided into 2 sections. Among them, 828 of data from November 28, 2014 to February 23, 2018 is chosen as the training set. The prediction model based on EMD and BP_AdaBoost model is established. The selection is taken from February 26, 2018. A total of 15 data were used as a test set on March 18, 2018. Figure 2 shows the price chart of brent crude oil. This article utilizes R language and Matlab software for programming.

4.2. Stationary Tests

It can be seen from Figure 2 that the entire sequence changes with time, and the original sequence has obvious non-stationary and nonlinear variation characteristics. In order to test the stationarity of the original sequence, the unit root test is performed. The results are presented in Table 1. After testing, the ADF value of the original sequence is −2.12, and the corresponding P value is 0.23, so the

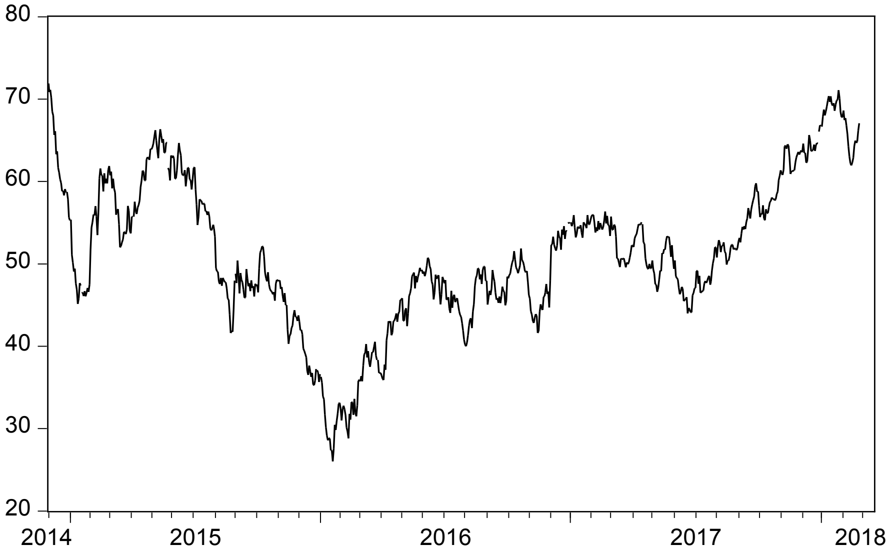

Figure 2. Crude oil price series.

Table 1. Results of stationarity test of the original sequence.

Sequence	ADF test value	1% critical value	5% critical value	10% critical value	P value	conclusion
Y	−2.1287	−3.4380	−2.8648	−2.5685	0.2335	non-stationary

rejection is not rejected under the confidence of 0.95. The null hypothesis is the case that crude oil prices are non-stationary time series.

4.3. EMD Decomposition

The EMD method is utilized to decompose the sample sequence, and 7 IMF components and one residual amount are generated. Figure 3 is the EMD decomposition result of the sample sequence, showing the frequency from the IMF1 component to the IMF7 component from high to low. The bottom is the residual component. The higher frequency IMF corresponds to the short-term trend of the crude oil price, and the lower frequency IMF corresponds to the long-term change of the crude oil price, and the residual corresponds to the trend of the crude oil price. Each intrinsic mode function obtained by EMD decomposition can represent the local features of the original time-series, so the intrinsic mode function obtained after the analysis and decomposition can well grasp the essential feature information of the original time series.

4.4. Forecast Model Parameter Settings

Firstly, in order to obtain a good prediction effect, before the BP_AdaBoost model is modeled, the IMF component and the trend term need to be reprocessed so that the value is distributed between [0, 1]. Secondly, after many attempts, we chose the parameter with the highest prediction accuracy. The training target in the BP neural network is set to 0.001, the maximum number of

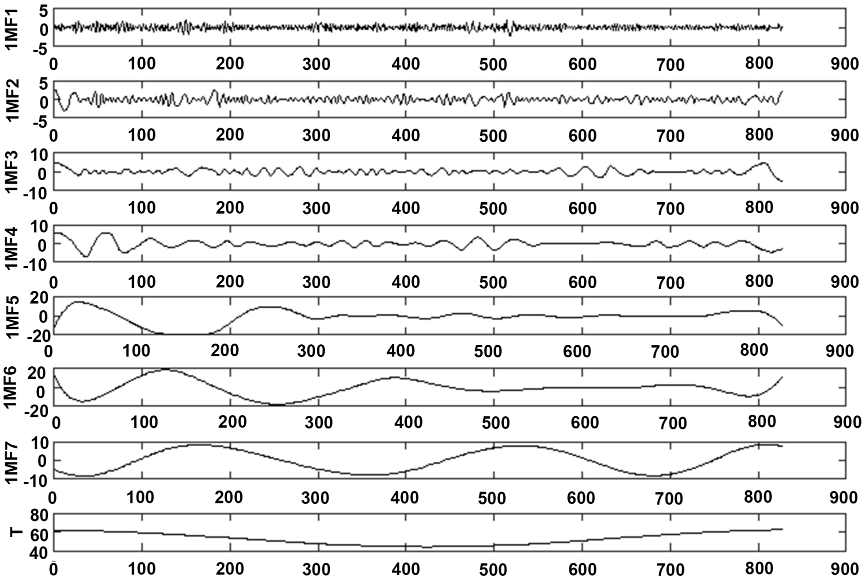

Figure 3. IMF component of the oil price series and trend items.

training is 1000, and the activation function of the hidden layer node is tansig., the training target in the BP neural network is set to 0.001, the maximum number of training is 10,000, and the activation function of the hidden layer node is tansig. Then, through training data training network, the test sample prediction sequence output is obtained, and the test sample weight value is updated according to the output result, and the BP neural network weak predictor and its corresponding weight are obtained. Finally, when using the AdaBoost algorithm for combination, this paper takes the sample with the prediction error greater than 0.001 between the actual output of the BP neural network and the real value as the sample that needs to strengthen the learning, continuously corrects the error, repeatedly trains the network, and optimizes the BP through the AdaBoost iterative algorithm. The BP_AdaBoost prediction model is calculated by the neural network, and the strong predictor is output. The BP_AdaBoost prediction model of this paper is made up of 10 BP network weak predictors.

Table 2 shows the true value of the brent crude oil price series and the predicted values of the model constructed in this paper, and gives the corresponding absolute error value and relative error value.

5. Model Comparison

5.1. Evaluation Criteria

The prediction accuracy and effectiveness of different methods are compared by the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), Theil inequality coefficient (U) and other evaluation indicators. Assuming that the original price series is Y_i and the predicted price series is $\hat{Y}_i$, the calculation formulas for these evaluation indicators are defined as follows:

Table 2. Oil price predict results.

	True value	Predict value	Absolute error	Relative error
1	67.96	67.0878	−0.8722	−0.01283
2	67.59	66.7606	−0.8294	−0.01227
3	66.08	66.1036	0.0236	0.000357
4	64.23	65.3557	1.1257	0.017526
5	64.26	64.6622	0.4022	0.006259
6	65.78	64.1719	−1.6081	−0.02445
7	65.67	63.9315	−1.7385	−0.02647
8	65.09	63.9065	−1.1835	−0.01818
9	63.87	64.0198	0.1498	0.002345
10	65.19	63.8727	−1.3173	−0.02021
11	64.53	63.9884	−0.5416	−0.00839
12	64.2	64.3648	0.1648	0.002567
13	63.61	64.6254	1.0154	0.015963
14	63.67	64.8264	1.1564	0.018162
15	64.68	65.0571	0.3771	0.00583

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2} \tag{8}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|Y_i-\hat{Y}_i\right| \tag{9}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i-\hat{Y}_i}{Y_i}\right| \tag{10}$$

$$TheilU=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2}+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i^2}} \tag{11}$$

5.2. Model Effect Comparisons

In order to check the validity of the BP_AdaBoost model constructed in this paper, it is compared with the ARIMA model, BP neural network and EMD-BP combined model. Crude oil price series prediction is carried out by the above method, and compared with the original sequence. The accuracy is evaluated. Table 3 gives a comparison of the forecast performance of different predictive models.

From the prediction results in Table 3, the model with the highest prediction accuracy in the ARIMA model, BP neural network and EMD-BP combined model is the EMD-BP combined model with root mean square error (RMSE) and mean absolute error (MAE) are 1.5226 and 1.2101 respectively. The root

Table 3. Comparison of prediction models for brent crude oil price.

Predict model	RMSE	MAE	MAPE	Theil U
ARIMA model	2.8196	2.5881	0.0401	0.2126
BP Neural Networks	2.3517	1.8090	0.0276	0.0182
EMD-BP Combined model	1.5226	1.2101	0.0186	0.1173
EMD + BP_AdaBoost Combined model	0.9823	0.8337	0.0128	0.0076

mean square error (RMSE) and mean absolute error (MAE) of the EMD-BP_AdaBoost model constructed in this paper are only 0.9823 and 0.8337, which indicate that it has higher prediction accuracy than other models. According to the training samples, the EMD + BP_AdaBoost combined model uses the AdaBoost algorithm to form a strong predictor of BP neural network weak predictors, which can improve the generalization ability. The prediction error is significantly lower than the ARIMA model, BP neural network and The EMD-BP combined model has a certain improvement in prediction accuracy and has evident reference value for crude oil price prediction.

6. Conclusions

This paper aims to fully consider the non-stationary and non-linear characteristics of crude oil price data, introduces the EMD method to decompose crude oil price data, and proposes an oil price forecasting method based on EMD and BP_AdaBoost model. In this paper, the EMD multi-scale decomposition method is used to decompose the crude oil price series into 8 IMF components and a residual quantity, then normalize the data, select the BP_AdaBoost model to predict the price series, and finally obtain the prediction result of the original sequence. The prediction results of BP_AdaBoost model are compared with ARIMA model, BP neural network and EMD-BP combination model. The empirical results show that the AdaBoost iterative algorithm optimizes the combination of multiple BP neural network weak predictor outputs for oil price prediction, which effectively reduces the problem that a single BP neural network is easy to fall into local minimum, and the optimized model can improve generalization performance. As well as prediction accuracy, its prediction effect is preferable to other models.

Compared with the existing prediction models, the EMD + BP_AdaBoost combination model constructed in this paper has certain advantages:

1) The EMD method can realize adaptive decomposition, which can extract signals of different frequencies and decompose the original complex signals into simple sub-sequences without loss of information.

2) Compared with the BP neural network model based on the EMD method, the prediction model based on EMD and BP_AdaBoost has stronger generalization ability, reduces the influence of local minimum values in BP neural network, and improves the prediction accuracy. And it can better meet the needs of non-linear, time-varying crude oil price forecasting, and has a useful application prospect.

Acknowledgements

This project was funded through National Natural Science Foundation of China (61703117); Guangxi Key Laboratory of Spatial Information and Surveying and Mapping (15-140-07-33); Guangxi Key Laboratory of Spatial Information and Surveying and Mapping (16-380-25-20).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Torrence, C. and Compo, G.P. (1998) A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society*, **79**, 61-78. https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2

[2] Huang, N.E., Shen, Z., Long, S.R., *et al.* (1998) The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proceedings of the Royal Society of London A*: *Mathematical, Physical and Engineering Sciences*, **454**, 903-995. https://doi.org/10.1098/rspa.1998.0193

[3] Yu, L., Wang, S. and Lai, K.K. (2008) Forecasting Crude Oil Price with an EMD-Based Neural Network Ensemble Learning Paradigm. *Energy Economics*, **30**, 2623-2635. https://doi.org/10.1016/j.eneco.2008.05.003

[4] Zhang, Y.J. and Wei, Y.M. (2010) The Crude Oil Market and the Gold Market: Evidence for Cointegration, Causality and Price Discovery. *Resources Policy*, **35**, 168-177. https://doi.org/10.1016/j.resourpol.2010.05.003

[5] Islam, M.R., Rashedalmahfuz, M., Ahmad, S., *et al.* (2012) Multiband Prediction Model for Financial Time Series with Multivariate Empirical Mode Decomposition. *Discrete Dynamics in Nature and Society*, **2012**, 87-88. https://doi.org/10.1155/2012/593018

[6] Xiong, T., Bao, Y.K. and Hu, Z.Y. (2013) Beyond One-Step-Ahead Forecasting: Evaluation of Alternative Multi-Step-Ahead Forecasting Models for Crude Oil Prices. *Energy Economics*, **40**, 405-415. https://doi.org/10.1016/j.eneco.2013.07.028

[7] Wei, L.Y. (2016) A Hybrid ANFIS Model Based on Empirical Mode Decomposition for Stock Time Series Forecasting. *Applied Soft Computing*, **42**, 368-376. https://doi.org/10.1016/j.asoc.2016.01.027

[8] Huang, N.E., Shen, Z. and Long, S.R. (1999) A New View of Nonlinear Water Waves—The Hilbert Spectrum. *Annual Review of Fluid Mechanics*, **31**, 417–457. https://doi.org/10.1146/annurev.fluid.31.1.417

[9] Hu, D., Zheng, D. and Fu, H. (2015) Application of AdaBoost-BP Model to Dam Deformation Prediction. *Journal of China Three Gorges University*, **37**, 5-8.