ISSN 2616-5775 Vol. 4, Issue 5: 8-15, DOI: 10.25236/AJCIS.2021.040502

AQI prediction based on CEEMD-WOA-Elman neural network

Kexin Yan[1], Jiaxun Liang[2]

[1]*School of Mathematics and Statistics, Ningbo University, Ningbo, Zhejiang, 315211, China*
[2]*School of Quality and Technical Supervision, Hebei University, Baoding, Hebei, 071002, China*

Abstract: *Aiming at the problem of predicting AQI, this paper proposes a hybrid prediction model. The specific method is to use Complementary Ensemble Empirical Mode Decomposition (CEEMD) to preprocess the non-stationary sequence, and then use the Elman neural network optimized by WOA to predict. The construction of CEEMD-WOA-Elman model mainly includes four parts: preprocessing, optimization, prediction, and evaluation. In order to evaluate the effectiveness and generalization ability of the model, we introduced four evaluation indicators to comprehensively evaluate the prediction model proposed in this paper. The analysis results show that compared with other models, the hybrid prediction model proposed in this paper has higher prediction accuracy and the predicted results obtained are more excellent.*

Keywords: *complementary ensemble empirical mode decomposition, whale optimization algorithm, Elman neural network, air quality index prediction, CEEMD-WOA-Elman hybrid model*

1. Introduction

The prediction of AQI is of great significance. It can not only avoid major property losses and improve the health of citizens, but also help the government to take timely mandatory measures to prevent pollution. At present, there are three main types of models used for air quality prediction: statistical model, numerical model and neural network model. Multiple linear regression [1], gray model [2], exponential smoothing method [3] , etc. are often used in statistical models, but these models are often not so accurate in predicting nonlinear problems due to various factors; CMAQ [4], WRF-CMAQ [5], WRF-Chem [6], etc. are often used in numerical models. Theoretically, as long as the data are accurate, the predicted value predicted by these three methods can be highly similar to the actual value, but the atmospheric changes are nonlinear and non-stationary, so the prediction accuracy is often low, and the calculation amount of numerical model is very large, so the prediction is time-consuming; At present, back propagation (BP), convolution neural network, wavelet neural network and other neural networks are widely used for prediction. Artificial neural network has very strong adaptive ability and is suitable for studying air quality prediction, but as the order of the system increases, the convergence speed of network learning becomes slower. The model proposed in this paper can solve the above mentioned model defects to a certain extent.

In this article, we propose the CEEMD-WOA-Elman hybrid model to predict AQI. Firstly, we use CEEMD to preprocess the original AQI sequence and decompose it into intrinsic mode function components and residual components on different time scales. Then, each component is input into Elman neural network optimized by WOA, and the predicted value of each component is processed to obtain the final AQI predicted value. In order to evaluate the accuracy of this model, we compare this model with single BP neural network, single Elman neural network, EEMD-BP model, EEMD-Elman model, CEEMD-BP model, CEEMD-Elman model, and then get the conclusion.

2. Proposed methodology

2.1 CEEMD

Traditional signal processing methods include wavelet analysis, short-time Fourier transform, bi-linear transform, etc., but these methods are often limited by the W. Heisenberg uncertainty criterion, so they cannot handle non-stationary signals well. In the past decade, many scholars have made a lot of explorations. At present, the more famous one is Complementary Ensemble Empirical Mode

ISSN 2616-5775 Vol. 4, Issue 5: 8-15, DOI: 10.25236/AJCIS.2021.040502

Decomposition (CEEMD), which is evolved from Empirical Mode Decomposition (EMD) and Ensemble Empirical Mode Decomposition (EEMD). However, EMD has the problem of modal aliasing and EEMD cannot completely offset the increased white noise, which adds a part of the noise invisibly, and needs to complete a large enough integration average, which is a very time-consuming process.

Therefore, CEEMD algorithm is derived, which is a new data decomposition technology proposed by Yeh et al. [7] in 2010 to improve the EEMD algorithm. They added positive and negative pairs of random Gaussian white noise to the original signal to eliminate the residual white noise in the signal reconstruction. At the same time, they can reduce the number of noise addition and improve the operation effect and efficiency. The specific steps are as follows:

Step 1: Add a pair of random white noises with the same amplitude and opposite signs to the original signal AQI sequence $x(t)$:

$$\begin{cases} x_i^+(t) = x(t) + \mu\sigma_i^+(t) \\ x_i^-(t) = x(t) + \mu\sigma_i^-(t) \end{cases}, \ i = 1, 2, \cdots, M. \tag{1}$$

Where μ is the amplitude, M is the number of iterations, $x_i^+(t)$ is the positive noise, and $x_i^-(t)$ is the negative noise.

Step 2: EMD decomposition is performed on each positive and negative noise to obtain IMF and residual components.

Step 3: Repeat steps (1), (2), but add different white noise each time, and finally get $2M$ groups of IMF and remainder.

Step 4: Calculate the integrated average of the decomposition results

2.2 Elman neural network

Elman neural network is a typical global feed forward local recurrent neural network proposed by J. L. Elman [8] in 1990 for speech processing. Compared with the general neural network, the basic Elman neural network is special in that it adds a context layer to the hidden layer of the feedforward neural network, as shown in Figure 1.

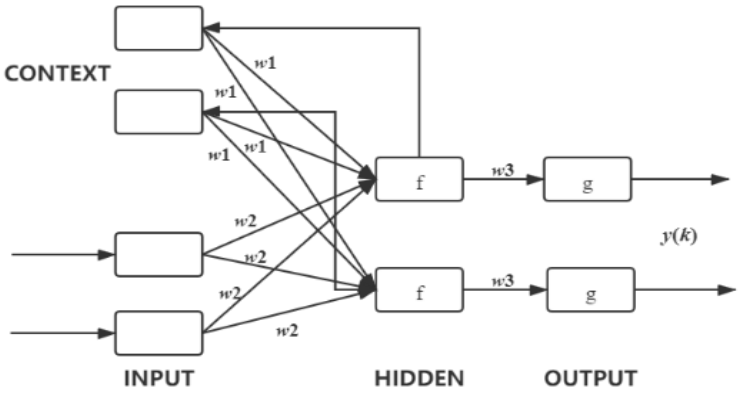

Figure 1: The schematic diagram of Elman neural network.

The units of the input layer transmit signals or data. The transfer function of the hidden layer unit can be linear or nonlinear. The context layer is used to memorize the past output value of the hidden layer unit and automatically connect to the input of the hidden layer at the next moment, which can be regarded as a one-step delay operator. This self-link mode makes it very sensitive to the data of the historical state, thus enabling the system to have the function of dynamic memory, and achieving the purpose of dynamic modeling. The output of the output layer is the result of linear weighting. Historical practice has proved that the Elman neural network has superior performance, for example, it can approximate any nonlinear mapping with arbitrary accuracy, and is better than the general neural network in terms of computing power and network stability.

ISSN 2616-5775 Vol. 4, Issue 5: 8-15, DOI: 10.25236/AJCIS.2021.040502

2.3 Whale Optimization Algorithm (WOA)

Whale optimization algorithm is a meta-heuristic search optimization algorithm which simulates the foraging mode of humpback whales with the characteristics of simple structure and few adjusting parameters [9]. The specific steps are as follows:

(1) Encircling prey: By identifying specific directions and distances from different prey and then surrounding them. The mathematical definition of this behavior is as follows:

$$\vec{D} = \left| \vec{C} \cdot \vec{X^*}(t) - \vec{X}(t) \right|, \ \vec{X}(t+1) = \vec{X^*}(t) - \vec{A} \cdot \vec{D}, \tag{2}$$

Where t indicates the current iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors, X^* is the position vector of the best solution obtained so far, and $\vec{X}$ is the position vector. X^* will be updated in each iteration if there is a better solution. The vectors $\vec{A}$ and $\vec{C}$ are calculated as follows:

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a}, \ \vec{C} = 2 \cdot \vec{r}. \tag{3}$$

Where $\vec{a}$ is linearly decreased from 2 to 0 over the course of iterations and $\vec{r}$ is a random vector in [0,1].

(2) Hunting behavior: Its spatial position is updated as follows：

$$\vec{X}(t+1) = \vec{D'} \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X^*}(t). \tag{4}$$

Where $\vec{D'} = \left| \vec{X^*}(t) - \vec{X}(t) \right|$ is the distance between the whale and the prey, b is a constant used to define the shape of the spiral, $\vec{X^*}(t)$ is the current best position vector, and l is a random number located in the interval $[-1,1]$. In the WOA algorithm, each chooses to swim towards the prey in a de-spiral shape or to shrink and surround the prey with a 50% probability, as shown below：

$$\vec{X}(t+1) = \begin{cases} \vec{X^*}(t) - \vec{A} \cdot \vec{D} & , \text{if } \beta < 5. \\ \vec{D'} \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X^*}(t), & \text{if } \beta \geq 5. \end{cases} \tag{5}$$

Where β is a random number located in the interval $[-1,1]$.

The mathematical model of searching for prey is as follows:

$$D = \left| \vec{C} \cdot \vec{X_{rand}} - \vec{X} \right|, \ \vec{X}(t+1) = \vec{X_{rand}} - \vec{A} \cdot \vec{D}, \tag{6}$$

Where $\vec{X_{rand}}$ is a random position vector (a random whale) chosen from the current population.

2.4 CEEMD-WOA-Elman

2.4.1 Build CEEMD-WOA-Elman model

The flow chart in Figure 2 shows the specific idea of model building.

Step 1: Use the CEEMD method to decompose the original data of the AQI into multiple IMF components and residual components with different frequencies.

Step 2: Initialize the relevant parameters of the WOA and Elman models, where the fitness function is selected as the mean square error of the whole training set and the test set. The smaller the fitness function is, the more accurate the training is and the better the prediction accuracy of the model is.

Step 3: Divide multiple IMF components and residual components into training set samples and test set samples, and use them as input variables of the WOA-Elman model. WOA will optimize the weights and thresholds of the Elman neural network and obtain the optimal values after predicting the test set samples, so as to obtain the prediction samples of each component.

Step 4: Add the predicted values of each component to get the predicted results.

ISSN 2616-5775 Vol. 4, Issue 5: 8-15, DOI: 10.25236/AJCIS.2021.040502

Step 5: Compare the prediction results with the test data, and work out the descriptive indicators of each data.

Step 6: Compare the prediction performance of different prediction models, analyze and summarize conclusions.

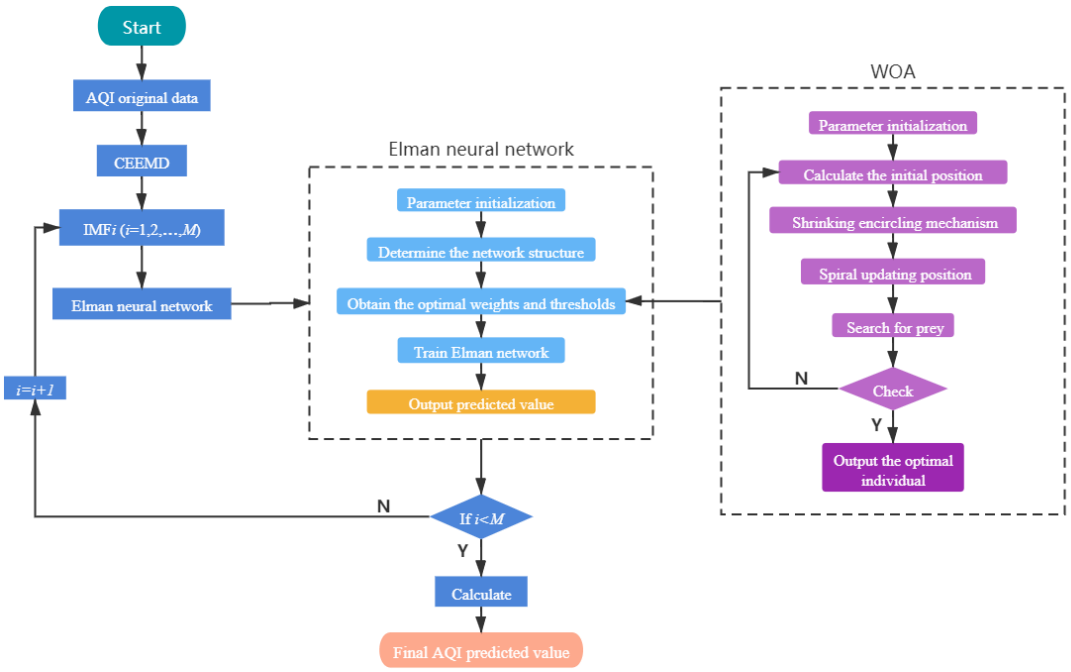

Figure 2: Flow chart of CEEMD-WOA-Elman model establishment.

2.4.2 Selection of evaluation indexes

In order to ensure a more accurate observation data processing effect, we selected mean absolute error (MAE), mean square error (MSE), mean square error root (RMSE), mean absolute percentage error (MAPE) as the evaluation index.

2.4.3 The results and analysis of CEEMD-WOA-Elman prediction model

A total of 1200 valid sample data of Shijiazhuang City's air quality index from January 1, 2018 to April 14, 2021 were selected as the data for analysis. The descriptive statistical characteristics of these data are shown in Table 1.

Table 1: Basic statistical characteristics of sample data.

Index	Number of Samples	Mean	Standard Error	Median	Standard Deviation	Variance	Kurtosis	Skewness	J-Btest
AQI	1200	110.94	1.75	98	60.63	3675.99	5.17	1.69	1

It can be seen from Table 1 that the standard deviation of the sample data is 60.63, so the volatility of the sample is relatively large, and the corresponding skewness and kurtosis are 1.69 and 5.17 respectively. The J-Btest test result is h=1, so this data series does not conform to the normal distribution. Therefore, we can find that the AQI sequence is non-stationary.

The air quality index data of Shijiazhuang from January 1, 2018 to April 14, 2021 are shown in Figure 4. It can be seen that there is no obvious regularity in this data series, and there are large fluctuations, so it is somewhat difficult to predict. Therefore, we use the CEEMD to convert non-stationary signals into multiple stationary sequences, which can make the prediction more accurate.

The decomposition results are shown in Figure 3. After CEEMD decomposition, we obtain total of 9 components, and the non-stationary and non-linear properties of the components are reduced. Then we transform the study of the original AQI sequence into the study of each component.

ISSN 2616-5775 Vol. 4, Issue 5: 8-15, DOI: 10.25236/AJCIS.2021.040502

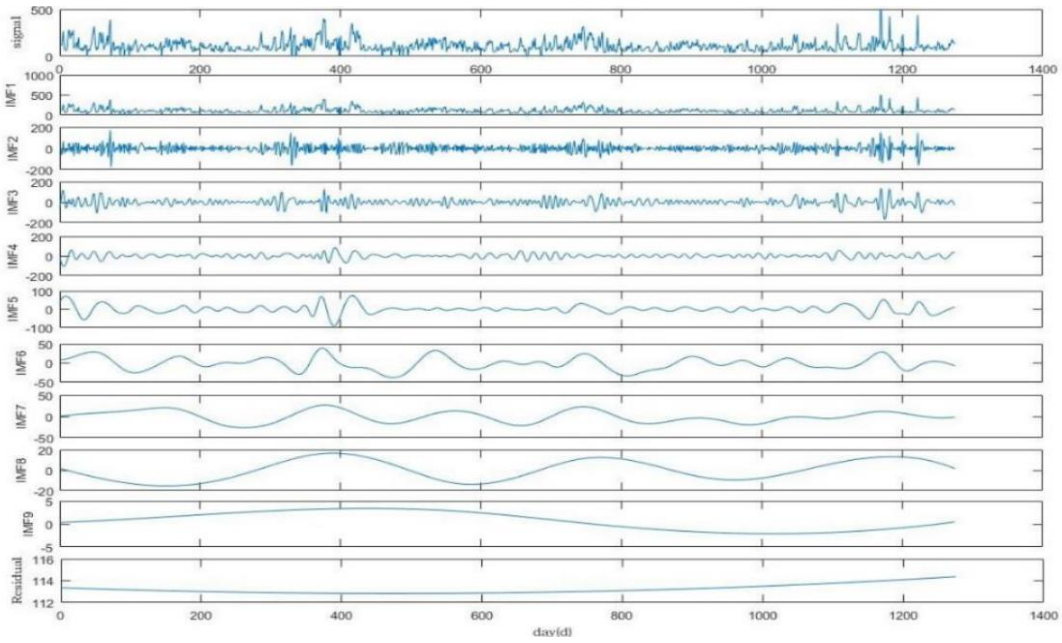

Figure 3: CEEMD decomposition diagram of Shijiazhuang AQI sequence.

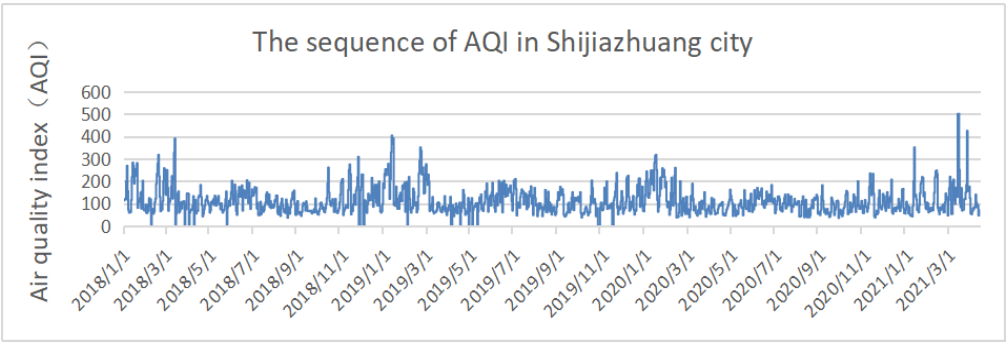

Figure 4: The diagram of AQI sequence.

2.4.4 Prediction results of each component

We use a total of 1,200 data points of Shijiazhuang air quality index as the effective samples for the study, among which we use a total of 1050 data of samples from January 1, 2018 to November 15, 2020 as the training set samples and 150 data from November 16, 2020 to April 14, 2021 as the test set sample for the test of the predicted results. The prediction results of each component in CEEMD-WOA-Elman modeling training are shown in Figure 5. Figure 6 is a comparison diagram between the actual curve and the prediction curve of each model. After processing each component, the final AQI prediction sequence is obtained and compared with the actual AQI sequence, as shown in Figure 8.

In addition, we have also drawn the evolution curve of WOA in the optimization process, as shown in Figure 7, where the abscissa represents the number of iterations, and the ordinate represents the MSE.

ISSN 2616-5775 Vol. 4, Issue 5: 8-15, DOI: 10.25236/AJCIS.2021.040502

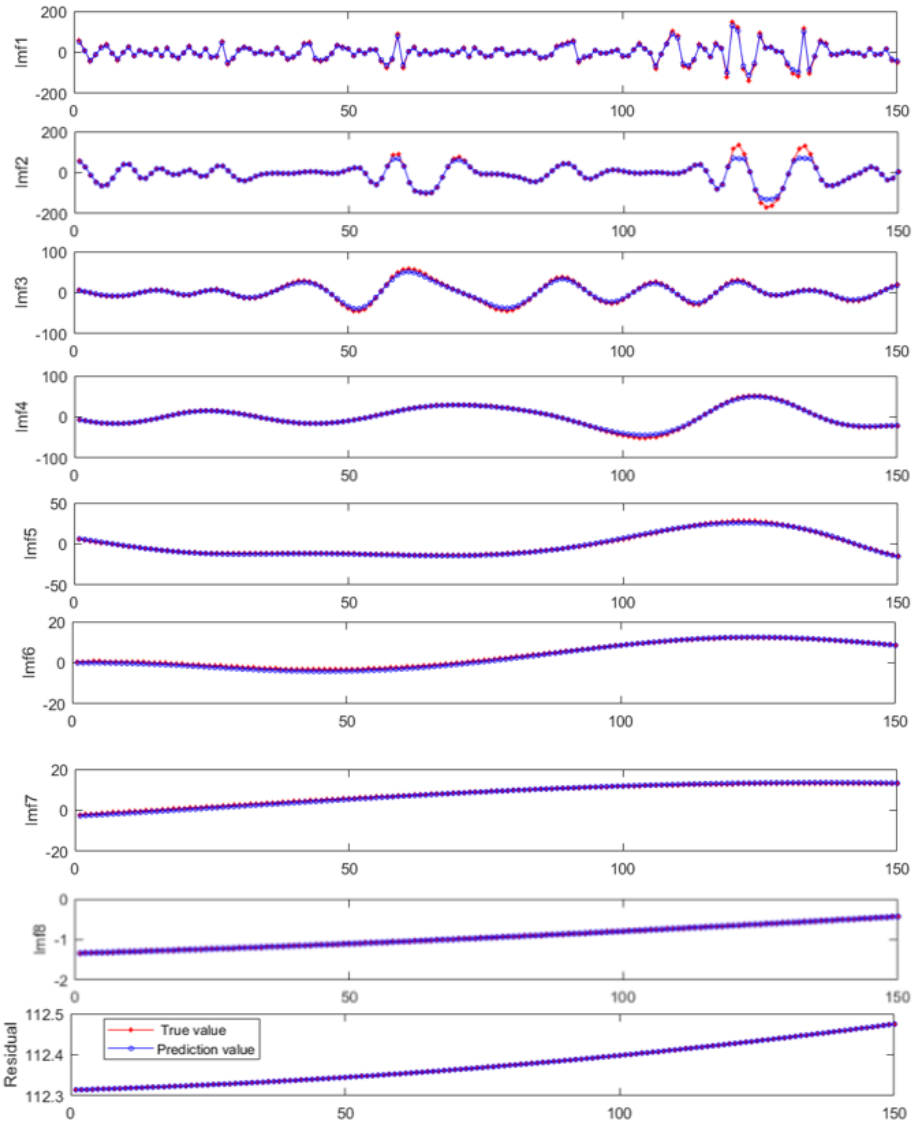

Figure 5: Comparison curves between predicted and ture values of each component.

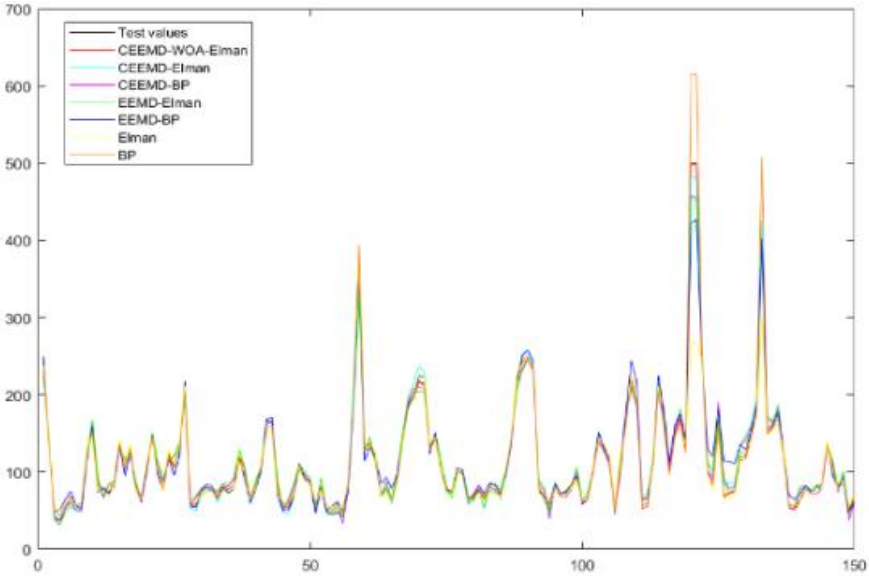

Figure 6: Comparison chart of actual curve and prediction curve of each model

ISSN 2616-5775 Vol. 4, Issue 5: 8-15, DOI: 10.25236/AJCIS.2021.040502

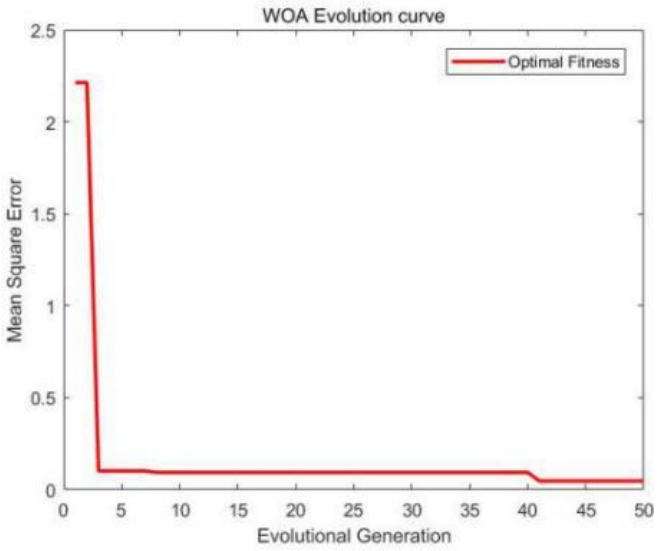

Figure 7: The evolution curve of WOA in the optimization process.

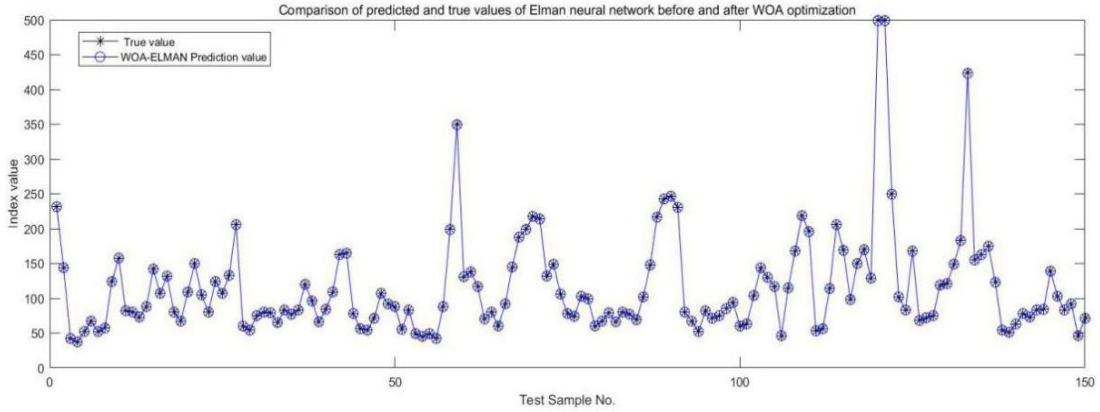

Figure 8: Comparison of predicted and true values of Elman neural network before and after WOA optimization.

3. Experimental analysis

In order to further compare the accuracy of each model in predicting AQI series, we calculated the prediction errors of each model respectively, and the specific data are shown in Table 2 below.

Table 2: Comparison of error index of each model.

Model	Error index			
	MAE	MSE	RMSE	MAPE
BP	7.7545	410.2083	20.2536	7.1496 %
Elman	7.0108	248.9889	15.7794	6.0973 %
EEMD-BP	8.4369	188.3833	13.7253	8.4956 %
EEMD-Elman	5.1409	88.7594	9.4212	4.7525 %
CEEMD-BP	7.7058	106.9241	10.3404	7.7101 %
CEEMD-Elman	4.926	53.2783	7.2992	3.604 %
CEEMD-WOA-Elman	0.14923	0.047253	0.21738	0.14954 %

From the data, we can discover the following facts that compared with a single neural network model, the MAE, MSE, RMSE and MAPE of predicted values of the BP model are about 9.59%, 39.30%, 22.09% and 14.72% higher than that of Elman network respectively, indicating that the Elman model is more sensitive to the dynamic changes in time series prediction, and the following ability of the model is better than that of the BP model. Furthermore, the MAE, MSE, RMSE and MAPE of CEEMD-WOA-Elman were 98.23%, 99.97%, 98.41% and 98.23% lower than those of EEMD-BP, respectively; 97.10%,

ISSN 2616-5775 Vol. 4, Issue 5: 8-15, DOI: 10.25236/AJCIS.2021.040502

99.95%, 97.69% and 96.85% lower than EEMD-ELMAN; 98.06%, 99.96%, 97.90% and 98.06% lower than CEEMD-BP; 96.97%, 99.91%, 97.02% and 95.85% lower than CEEMD-ELMAN.

Combined with the data and graph, it is not difficult to find that compared with the single neural network model, the four combined models (EEMD-BP model, EEMD-Elman model, CEEMD-BP model and CEEMD-Elman model) can more accurately predict the development trend of Shijiazhuang AQI, and the follow-up of the prediction is better. Furthermore, the Elman neural network performs better in AQI prediction than BP neural network. In other words, Elman neural network may be more suitable for the prediction of nonlinear and non-stationary series. By observing figure 7, it is easy to find that the CEEMD-Elman model has the best prediction effect among the four combination models, which is consistent with the actual value trend. However, it can be found that the CEEMD-WOA-Elman model has the highest prediction accuracy, and the prediction trend is almost consistent with the trend of Shijiazhuang AQI.

4. Conclusion

In view of the many shortcomings in the current AQI prediction model, this paper proposed a new hybrid model of CEEMD-WOA-Elman. It showed that compared with other models, the CEEMD-WOA-Elman model has higher prediction accuracy for the AQI sequence, no matter from the figure or from the error table, and is more suitable for forecasting air quality in Shijiazhuang. Thus, the CEEMD- WOA-Elman model has certain applicability in real life to some extent. Furthermore, the model can also be applied to short-term wind speed prediction, power load prediction and so on.

5 Reference

[1] Giuseppe Nunnari et al. Modelling SO2 concentration at a point with statistical approaches [J]. Environmental Modelling and Software, 2003, 19(10): 887-905.
[2] Guo X, Liu S. Forecasting China's SO2 emissions by the nonlinear grey Bernoulli self-memory model (NGBSM) [C]// IEEE International Conference on Grey Systems & Intelligent Services. IEEE, 2015:80-85.
[3] ZHANG XiLai, ZHAO JianHui, CAI Bo. Prediction Model with Dynamic Adjustment for Single Time Series of PM2.5 [J]. Journal of Automation, 2018, 44(10): 1790-1798.
[4] Binkowski, Francis S. Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component 1. Model description [J]. Journal of Geophysical Research: Atmospheres, 2003, 108(D6):-.
[5] Choi M W, Lee J H, Woo J W, et al. Comparison of PM2.5 Chemical Components over East Asia Simulated by the WRF-Chem and WRF/CMAQ Models: On the Models Prediction Inconsistency [J]. Atmosphere, 2019.
[6] Hagihara Y, Itahashi S, Yumimoto K , et al. Comparison of simulated and observed signals over East Asia using WRF-Chem model, A-Train data, and satellite simulators [J]. American Geophysical Union, 2011.
[7] YEH J R, SHIEH J S, HUANG N E. Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method [J]. Advances in Adaptive Data Analysis, 2010, 2(2): 135-156.
[8] Jeffrey L. Elman. Finding structure in time [J]. 1990, 14(2): 179-211.
[9] Seyedali Mirjalili and Andrew Lewis. The Whale Optimization Algorithm [J]. Advances in Engineering Software, 2016, 95: 51-67.