Epidemic prediction based on entropy-improved factor analysis and WOA-optimized BP network algorithm

Jianan Zhang, Hongyi Duan, Bingsong Tong

Jianan Zhang, Hongyi Duan, Bingsong Tong, "Epidemic prediction based on entropy-improved factor analysis and WOA-optimized BP network algorithm," Proc. SPIE 12645, International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2023), 126453R (23 May 2023); doi: 10.1117/12.2681323

SPIE.

Event: International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2023), 2023, Hangzhou, China

Epidemic prediction based on entropy-improved factor analysis and WOA-optimized BP network algorithm

Jianan Zhang*[1], Hongyi Duan*[2] and Bingsong Tong[2]
[1]Shanghai University of Finance and Economics, ShangHai Yangpu District 200433, CHINA
[2]Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an Xianning Road 710049, CHINA
3628750954@qq.com (J A Zhang)
2213611582@stu.xjtu.edu.cn(H Y Duan)
*These two authors contributed to the work equally and should be regarded as co-first authors.

ABSTRACT

The New Coronavirus epidemic has had a huge impact on the economy, politics, and culture worldwide. However, it is very difficult to obtain accurate data on the New Crown epidemic due to various uncertainties, such as the difficulty of detection. In this paper, we use objective and real Baidu search indexes as the basic data set, and use factor analysis with the improvement of entropy method to reduce the dimensionality of Baidu search index data to solve the problem of fixed parameters caused by its excessive dimensionality. After that, the WOA algorithm is used to optimize the parameters of the conventional BP neural network, thus making the fit and accuracy greatly improved, which is of great practical significance for the prediction of epidemic data.

Key Words. Entropy-based factor analysis; BP neural network; WOA algorithm.

1. INTRODUCTION AND REVIEW

On December 7, 2022, the Chinese government released the "New Ten" policy. With the liberalization, the difficulty in estimating the testing data of the epidemic has become a major problem due to the reduction of nucleic acid testing, etc. However, understanding the data of the epidemic is crucial to grasp the general trend of the epidemic, allocate medical resources, and play a vital role in the follow-up of epidemic prevention.

However estimating data for epidemics is a very difficult problem, with a high degree of complexity due to various factors.Many Chinese scholars have studied this. Professor Wangyu Xia used multiple linear regression to fit the data of the epidemic[1]and got good results. Professor Yan proposed a class of infectious disease dynamics models based on time lag dynamics systems, in which time lag processes were introduced to describe the incubation period and treatment cycle of viruses. The parameters of the model are accurately inverted by the published epidemic data; and the epidemic trend is accurately predicted [2].Professor Jianqiang Ren's three-step prediction model for the New Coronary Pneumonia epidemic based on machine learning, which introduced machine learning algorithms such as neural networks, random forests, long and short-term memory networks and sequence-to-sequence to predict the New Coronary Pneumonia epidemic, and achieved reliable results [3]. Professor Qiyun Wang proposed a combined COVID-19 prediction model based on the CEEMDAN-HURST algorithm for the new cases of COVID-19, which can effectively solve the problems of low prediction efficiency and low prediction accuracy commonly found in nonlinear time series prediction models[4]. Many other scholars have also proposed corresponding prediction methods, but considering the reasons for the mutation of novel coronaviruses, the infectivity of virulent strains is greatly enhanced, which also affects the adaptability of the aforementioned prediction methods.

In this regard, this paper combines the Baidu search index for a total of 92 days from September 1, 2022 to December 1, 2022 as the data object, takes into account the number of search indicators, the large hierarchy, contains different information and other characteristics, establishes factor analysis, and uses the entropy value method for improvement to obtain the processed data. Then, considering the powerful nonlinear mapping capability, self-learning and self-adaptive capability, generalization capability, fault tolerance capability of BP neural network[5]and the extremely strong tuning

International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2023), edited by Xin Feng, Aniruddha Bhattacharjya, Proc. of SPIE Vol. 12645, 126453R · © 2023 SPIE · 0277-786X · doi: 10.1117/12.2681323

capability of WOA algorithm itself, a BP neural network algorithm optimized based on WOA algorithm is established to fit and combine the data for prediction. Finally, the sensitivity analysis of the parameters and the error analysis of the whole algorithm are then performed.

2. SELECTION OF DATA METRICS

In this article, we reasonably adopt the following assumptions

1. People's willingness to search Baidu will not fluctuate dramatically in a short period of time.

2. Data in this paper ignore the indicators that don't reach the search frequency of more than 2000

3. Baidu Search index has certain accuracy, and the indicators on its ranking are real and reliable

4. The data on the number of confirmed cases is reliable.

Considering the search terms related to novel coronaviruses pneumonia, since the frequency of search terms can reflect their importance in prediction to a certain extent, the top ten indicators were selected based on the above 92-day heat total ranking, which are: fever, pneumonia, headache, cough, ibuprofen, lianhua qingwen capsule, diarrhea, on-line consultation, epidemic, and taste, as shown in the following Table1.[6] and Figure 1. [7]

Table 1. Ten significant index

item	*index*
fever(x_1)	309982
pneumonia(x_2)	198682
headache(x_3)	208796
cough(x_4)	405098
ibuprofen(x_5)	908688
lianhua qingwen capsule(x_6)	575753
diarrhea(x_7)	156859
on-line consultation(x_8)	93435
epidemic(x_9)	953354
taste(x_9)	12312

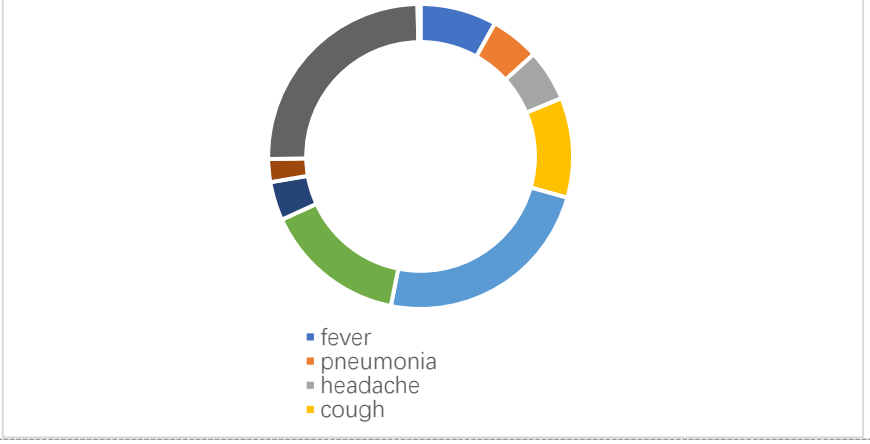

Figure 1. Ring chart of the ten significant index

3. DIMENSIONALITY REDUCTION OF INDICATORS: FACTOR ANALYSIS AND ENTROPY IMPROVEMENT.

Factor analysis was proposed by the British psychologist Spearman in 1904, and it explores the basic structure in the observed data by studying the internal dependencies between numerous variables and representing the basic data structure with a few dummy variables, making it possible to reflect the main information of the original numerous variables [8]. The basic arithmetic algorithm is as follows.

3.1 Assumptions and basic properties

Suppose there are n variables (n is 10 in this paper), $x_i = a_i + b_{i1}F_1 + b_{i2}F_2 + b_{i3}F_3 + \cdots + b_{im}F_m + e_i$

Among above there is $m < n$. The above equation can be expressed as a matrix form

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{12} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n2} & b_{n2} & \cdots & b_{nm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad (1)$$

and recorded it as $x - a = AF + e$, among them $F_1, F_2, F_3 \cdots F_m$ are public factors. They are unobservable variables, and their coefficients are called load factors. And e is a special factor that cannot be included by the first m public factors, and satisfies $E(F) = 0, E(e) = 0, Cov(F) = I_m$

3.2 Properties of factor analysis models

Through the decomposition of the covariance matrix of the original variable x, we have

$$Cov(x) = AA^T + diag(a_1^2, a_2^2, a_3^2, \cdots, a_m^2) \quad (2)$$

Thus, the proportion of common factor sharing components can be judged by the magnitude of the eigenvalue

The loading matrix is not unique. The factor loading b_{ij} is the correlation coefficient between the i-th variable and the j-th common factor, which reflects the correlation importance of the two common factors, and the larger the absolute value, the higher the correlation closeness.

As for the statistical significance of the variance contribution of the common factor F_j, it can be used to measure the relative importance of F_j by calculating the sum of squares of the elements in each column $S_i = \sum_{i=1}^{n} a_{ij}^2$, called the sum of variance contribution of F_{ij} to all x_{ij}.

3.3 Estimate the factor loading matrix

Conventional methods include principal component analysis, principal factor method, maximum likelihood method, etc. While each method has its own characteristics and responds to different information [9]. In this paper, we innovate the entropy value method to obtain a weighted load matrix with better performance for the above three methods. The details are as follows: According to the definition of entropy, the probability (proportion) of the i-th scheme under the j-th indicator of this indicator is

$$P_{ij} = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}}, i = 1, 2, \cdots, m; j = 1, 2, \cdots, n \tag{3}$$

Therefore, the information entropy of the system is

$$e_j = -k \sum_{j=1}^{n} P_{ij} I_n P_{ij} \tag{4}$$

Define the coefficient of variability as $g_i = 1 - e_j$, and for the ith metric, the larger its value, the greater its effect on the load matrix

3.4 Factor rotation

In this paper, based on several tests on the data, the optimal results were obtained using the variance maximization method, which starts from each column of the simplified factor loading matrix such that the squared difference of the loadings related to each factor is maximized. The results are analyzed as follows: the above ten indicators can be planned to the following three major factors in Table 2.

Table 2. The three factors

Factor 1(F_1)	Factor 2(F_2)	Factor 3(F_3)
fever pneumonia ibuprofen diarrhea	lianhua qingwen capsule on-line consultation epidemic	headache taste cough

The specific weighting indicators are as follows.

$$\begin{aligned} F_1 &= 0.167x_1 + 1.456x_2 - 0.987x_5 + 2.987x_7 \\ F_2 &= 1.789x_6 - 3.219x_7 + 0.878x_9 \\ F_3 &= 1.789x_3 + 0.686x_4 - 1.301x_{10} \end{aligned} \tag{5}$$

Thus, in this paper, we obtained the dimensionality-reduced data and used the above three factors to fit and predict the new crown epidemic data, introducing a BP neural network optimized with the WOA algorithm.

4. ORIGINAL BP NEURAL NETWORK

BP neural networks are often used for function approximation, classification, pattern recognition, and data compression and prediction due to their simple structure and few tuning parameters [5]. A typical BP network structure is generally shown in the figure, which has the ability to provide negative feedback from the output layer to the input layer.

In this paper, we can set the parameters of BP neural network as follows: y_j are the inputs of the j-th node in the input layer; ω_{ij} is the weight between the i-th node in the hidden layer and the j-th node in the input layer; θ_i is the threshold of the i-th node in the hidden layer; ω_{ki} is the weight between the k-th node in the output layer and the i-th node in the hidden layer; a_k is the threshold of the k-th node in the output layer; f, g are the excitation functions of the nodes in the hidden layer and the output layer.

We divide the whole process into forward learning and backward learning, and the output of the former is formulated as

$$o_k = \sum_{i=1}^{n} \omega_{ki} g(\sum_{j=1}^{\mu} \omega_{ij} x_j + \theta_i) + a_k \tag{6}$$

(there g takes the sigmod excitation function and x_j is the sample iteration update position)

The latter is to calculate the error between the output value and the expected value of the output layer, based on the negative gradient descent principle to update the network weights and thresholds of each layer from backward to forward.

5. OPTIMIZATION USING WOA ALGORITHM

According to the study shows that the threshold value of the above weights(in subsection 4) is greatly influenced by the initial value, determining the appropriate parameters affects the accuracy of the prediction, using the WOA algorithm is concise and easy to implement, and the objective function requirements are lenient and less parameter control [10]. In this paper, the WOA algorithm is used for optimization.

5.1 Hunting behavior

Assuming that the position of the target prey or the optimal solution is within the search range, the current best search agent is randomly selected in the whale population, and other search agents try to update their positions to the best search agent. The corresponding formula is:

$$\vec{y}(t+1) = \vec{y}(t) - \vec{x}_1 \cdot \vec{x}_2 \quad \text{and} \quad \begin{cases} \vec{x}_2 = 2\vec{a} \cdot \vec{e}_1 - \vec{a} \\ \vec{x}_1 = \left| \vec{\alpha} \cdot \vec{y}(t) - \vec{y}(t+1) \right| \\ \vec{\alpha} = 2\vec{e}_2 \end{cases} \tag{7}$$

In the formula, t is the iteration time, $\vec{x}_2$ and $\vec{a}$ are coefficient vectors, and $\vec{e}_1, \vec{e}_2 \in [0,1]$

5.2 Bubble network attack mechanism

In the whale population, the individual closest to the prey is selected as the best search agent, other whales will approach the currently selected whale individual, and perform a spiral update of the moving trajectory in the two-dimensional coordinate system, the iteration formula is

$$\vec{y}(t+1) = \vec{D}^t e^{bl} \cos(2\pi l) + \vec{y}'(t) \quad \text{and} \quad \vec{D}^t = \left| \vec{y}'(t) - \vec{y}(t) \right| \tag{8}$$

(b is the undetermined coefficient, $l \in rand[-1,1]$)

Based on the calculation of the distance between the current position and the current best prey, the spiral equation can be established[11]

5.3 Set the probability model

The probability of having p is used to shrink the envelope mechanism or expand the search range; the probability of having 1-p is used to spiral update the position [12]. The corresponding formula is as follows

$$\vec{y}(t+1) = \begin{cases} \vec{y}'(t) - \vec{A} \cdot \vec{D} \cdots\cdots\cdots\cdots\cdots p \geq 0.5 \\ \vec{D}^t e^{bl} \cos(2\pi l) + \vec{y}'(t) \cdots p < 0.5 \end{cases} \tag{9}$$

6. FITTING AND EXPERIMENTAL RESULTS

In order to compare the effect after optimization with the WOA algorithm, this paper fitted the above three major factors with the number of new coronary pneumonia using BP neural network and WOA optimized BP neural network respectively, and the results obtained are shown in the following three Figure 2. Figure3. and Figure4.

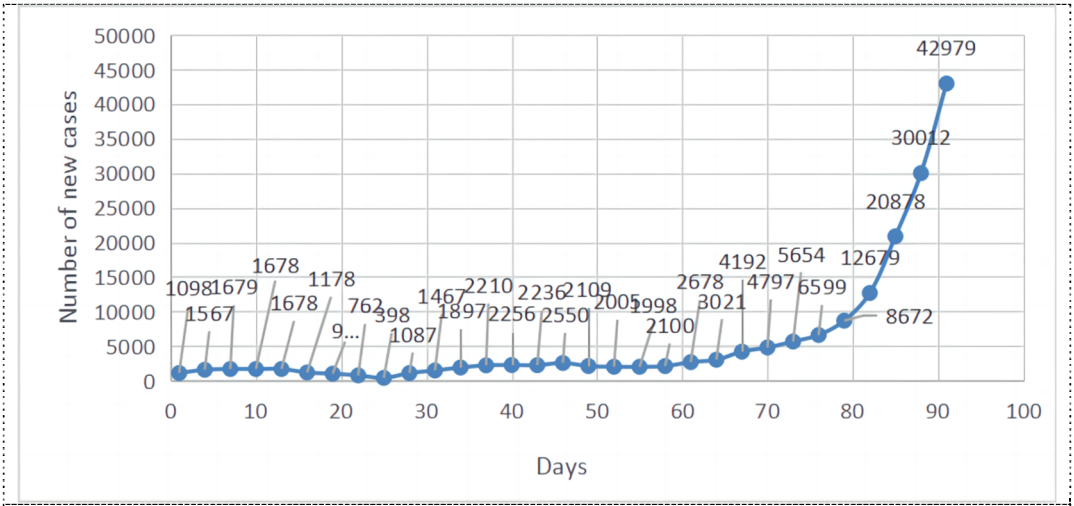

Figure 2. Real epidemic data

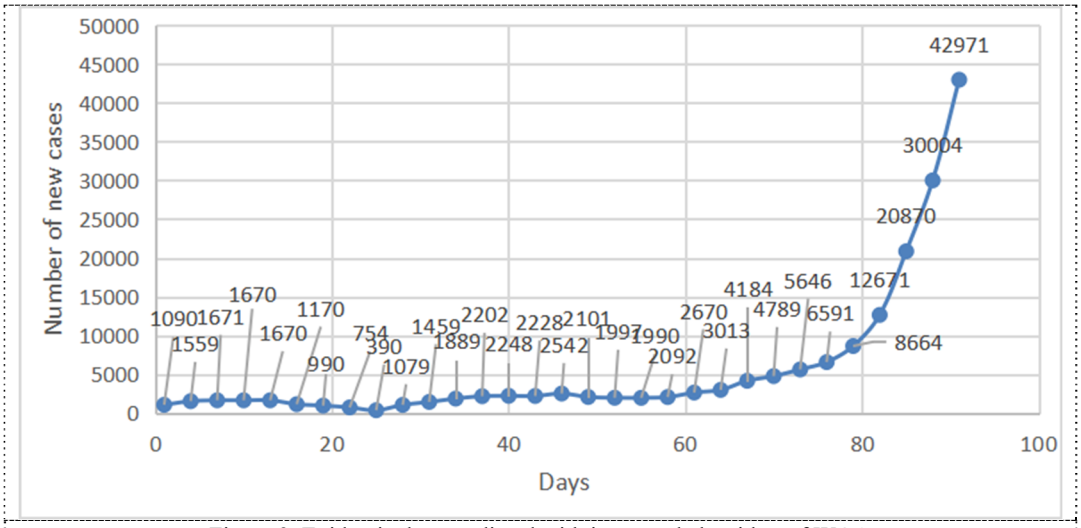

Figure 3. Epidemic data predicted with improved algorithm of WOA

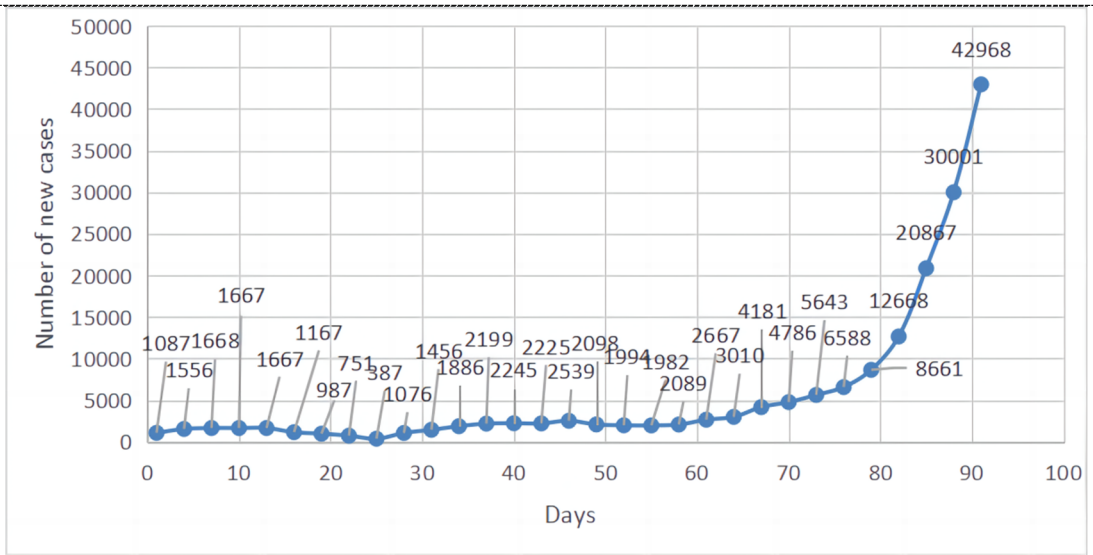

Figure 4. Epidemic data predicted based on BP neural network algorithm

From the above figure, it can be seen that the fitting effect of the BP neural network optimized by the WOA algorithm is significantly optimized for the ordinary BP neural network, but both have a feminine good fitting effect, and the error results for the BP neural network optimized based on the WOA algorithm are shown in the Figure 5. below, which shows that the model in this paper has a good fitting and prediction effect. And the fitting size is shown in Figure 6.

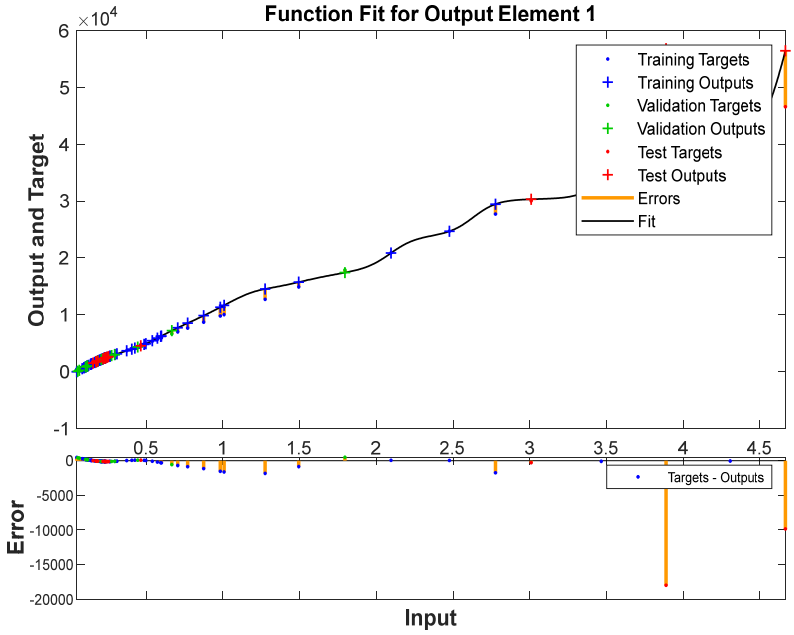

Figure 5. Fitting error graph

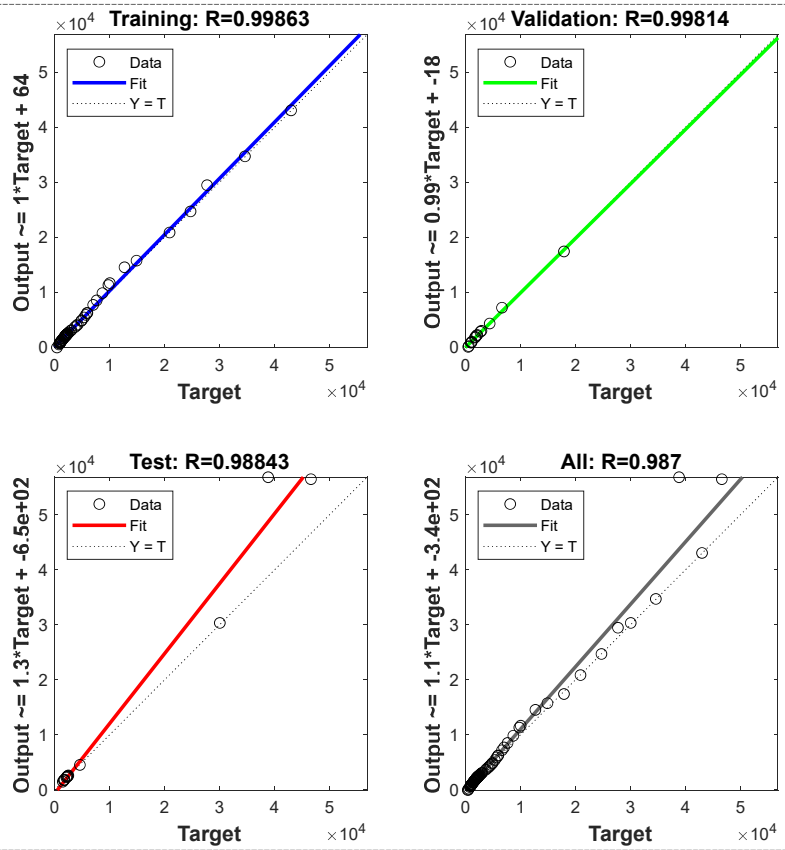

Figure 6. Fitting size graph

7. CONCLUSION

For the complex problem of predicting the number of new coronary pneumonia cases, this paper innovatively uses Baidu search index as the dataset for prediction, and uses factor analysis to reduce dimensionality with the improved-BP neural network algorithm based on WOA algorithm, which achieves a good fitting effect and is a good guide for the prediction of the current epidemic.

REFERENCES

[1] W Y Xia.(2022).Study on epidemic surveillance system based on multiple linear regression method*(Master's degree thesis, Wuhan University of Engineering)*. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFDTEMP&filename=1022718992.nh

[2] Y Yan, Y Cheng, K J Liu, X Y Luo, B X Xu, Y Jiang, J Cheng. (2020). Modeling and prediction of COVID-19 outbreaks based on a class of time-delay dynamics system. *Scientia Sinica (Mathematica)*(03),385-392. https://kns.cnki.net/kcms/detail/11.5836.O1.20200210.1444.002.html

[3] J Q Ren, Y P Cui, S J Ni.(2022).Machine learning-based methods for COVID-19 epidemic trend prediction. *Journal of Tsinghua University (Science and Technology)* doi: 10.16511/j.cnki.qhdxxb

[4] Q Y Wang, Z T Zheng. (2022). Application of the CEEMDAN-HURST algorithm in COVID-19 prediction. *Computer Engineering and Applications.* http://kns.cnki.net/kcms/detail/11.2127.TP.202211 25.1606.034.html

[5] M X Yu, G Dong, R X Xu, G L Yu. (2023). The precision temporal base source calibration prediction model based on BP neural network. *China Measurement & Test.*http://kns.cnki.net/kcms/detail/51.1714.TB.20230105.1344.001.html

[6] Baidu Inc. (2023). Baidu Index of Searching https://index.baidu.com/baidu-index-mobile/index.html#/?qq-pf-to=pcqq.c2c

[7] Baidu Inc. (2023). Baidu Index of Information https://index.baidu.com/v2/index.html#/

[8] Z H Liu, J H Wang, F P Tong, L Sun, G Li, R Chenet al. (2022). The comprehensive evaluations of antimony resistance in different Sorbus clones based on factor analysis. *Forest Research* (06). doi: 10.13275/j.cnki.lykxyj.2022.006.017

[9] S Y Wang, Z J Zhao, X X Zhang, Z Z Yan, Y Zhang, Z M Zhang, K D Zhou, et al. (2022).Quantitative analysis of 558 patients with kidney Yang deficiency syndrome based on the factor analysis methods.
Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology. http://kns.cnki.net/kcms/detail/11.5699.R.20221121.0937.002.html

[10] Y Z Xu, Y K Fu, T Z Wu. (2023).Estimation of the lithium-ion battery SOC based on the WOA-BP neural network. *Battery Bimonthly.*https://kns.cnki.net/kcms/detail/43.1129.TM.20230103.1106.002.html

[11] Y T Yao, J J He, Y L Li, D Y Xie, Y Li.(2021).ET_0 simulation of the modified whale optimized BP neural network. *Journal of Jilin University (Engineering and Technology Edition)*, 1798-1807. doi: 10.13229/j.cnki.jdxbgxb20200545.

[12] M Xu, J J Liu, H Lei, Q Li, X X Zhang.(2022).Prediction of crystallization apparatus liquid level fluctuation of BP neural network based on whale algorithm. *Metallurgical Industry Automation.* https://kns.cnki.net/kcms/detail/11.2067.TF.20221125.0950.002.html