Research on Forecasting Strategy of Environmental Data Based on Neural Network

Ying Chen[a,*], Han Zheng[b] and Fengyu Yang[c]

School of Software, Nanchang Hangkong University, Fenghe Nan Street, Nanchang, P.R. China

[a]c_y2008@163.com, [b]674526380@qq.com, [c]13732976937@163.com

**Corresponding author*

Keywords: Polluted gas, Data forecast, Neural Networks

Abstract: In this paper, NO_2, the most common polluting gas in the air, is selected as experimental subject, and an improved prediction algorithm is proposed. First, the gas data is preprocessed and optimized by Chauville's method, so that the gas data become more reliable. Then the improved prediction network model is created, and the input data is input into the network model to predict the results. Finally, the network model is established through multiple random sampling learning in large samples. The simulation results show that the improved neural network-based data prediction algorithm is superior to the original method in term of accuracy and stability.

1. Introduction

The concentration of polluted gas is an important indicator to evaluate the environmental air quality. Hence, the prediction of concentration of polluted gas is helpful to the relevant government departments to adjust the plan according to the quality of air, so that people can be harmed as little as possible by the polluted gas. However, the daily variation of the concentration of pollutant gas in a region is very complicated, because which is affected by many uncertain factors, such as the change in the quantity of pollution sources, the change in the emission of pollution sources, and the changes in meteorological conditions. Therefore, how to predict air quality become research hot spot.

Many scholars at home and abroad have made a lot of achievements in the study of prediction methods of pollutant gas concentration. Gu et al. [1] made scenario simulation analysis on various influencing factors of greenhouse gas CO_2, and made scenario simulation prediction by using the basic method that the emission equals the activity data multiplied by the activity factors. Chen et al. [2] forecast the gas concentration based on the method of multiple regression, and use the least square method to estimate related parameters, then calculate and predict the gas concentration. Han et al. [3] used back propagation (BP) neural network to forecast the gas concentration in the warehouse, they used the data of the first ten days to model and forecast the gas data directly. Xu [4] uses BP neural network to predict SO_2, NO_2 and other polluted gases. Vaz et al. [5] proposed prediction method based on the analysis of rough set data, their method needs to deal with the data by discrete normalization, which reduces the amount of data and forecasts the concentration with relatively few data.

In this paper, an optimized neural network method is proposed, the original data will be pre-processed before importing the existing data into the neural network for training, and eliminate some abnormal data.

2. Artificial Neural Network

Artificial neural network (ANNs), also known as intelligent neural network, which uses advanced technology to simulate the human brain's neural system to learn data. Since the advent of neural networks in the last century, neural networks have been applied to signal recognition, pattern determination and other fields. Neural networks can predict power load, stock prices and so on, and which have a good forecasting effect. Because the content of air pollution is very complex, there are many factors affecting it, not only the amount of waste generated by human activities, but also related to temperature and other factors, so it is usually impossible to use a definite linear function to describe [6]. The advantages of neural network in nonlinear prediction can be well applied to the prediction of environmental data.

3. Optimization of Neural Network Prediction Algorithm

3.1. Data Pre-processing

The paper uses the Chauville's method to pre-process the polluted data. The Chauville's method is a method of equal confidence probability, its idea is to determine a confidence limit, the data value is greater than the error of the limit will be considered to be an outlier. Chauville's method stipulates that the confidence probability is 1-1/2*n, for example, if the probability of occurrence of an error is less than 0.5 times in n measurements, the error is considered impossible to occur, and it is considered as an abnormal value. In the paper, the abnormal value is scaled properly. We set times interval between 0.8 and 1.2. According to its confidence probability, we can calculate the Chauville's coefficient, as shown in Figure 1.

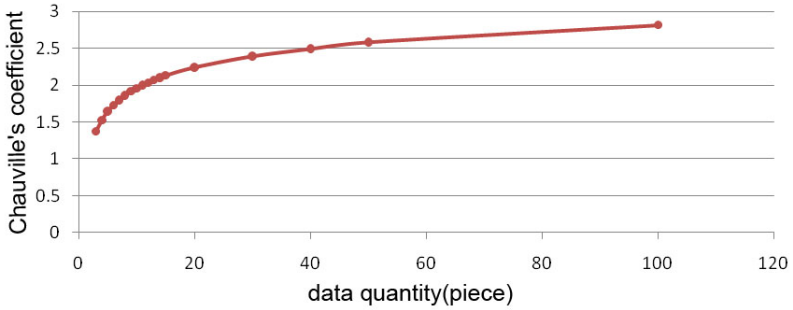

Figure 1 Chauville's coefficient

The Chauville's coefficient approximate calculation formula is shown in Equation 1.

$$\omega = 1 + 0.4 \times ln(n) \tag{1}$$

If the absolute value of the difference between the measured value and the average value is greater than the product of the standard deviation of the entire data and the Chauville's coefficient, as shown in Equation 2.

$$| x_i - \bar{x} | > S_x \times \omega \tag{2}$$

3.2. Construction of Neural Network Model

The complexity of the human brain is not only due to the number of neurons, but also the neural structure of the human brain is very complex, and the artificial neural network simulation of the human brain only depends on changing the number of neurons in each layer and the number of hidden layers, so the number of neurons is not the more the better, only increase the number of neurons may cause excessive fitting. Of course, if the number of neurons is too small, the fitting degree will not be enough. In this paper, the appropriate number of hidden layers and neurons in the neural network model will be chosen according to the original data. Some scholars have proposed some common methods to determine the number of neurons, such as (input + output) / 2 (input is the number of input variables, output is the number of output variables) [7]. It cannot be applied to every situation accurately, therefore, this paper set the appropriate number of neural network layers and the number of neurons in each layer through experiment.

3.3. Cross Validation of Prediction Results

Since random sampling, weight threshold correction and other processes may lead to some changes in each prediction result when the data is used for the prediction of neural network modelling, the accuracy and stability of the prediction result will not be very good if the prediction result is taken as the final prediction value directly. In this paper, each time the neural network model is built and predicted, 25 groups of data are randomly selected as test data, and the remaining 340 groups of data are used as training data. The process is to retain cross-validation, and a prediction value is obtained by predicting the input variables with the obtained neural network. In order to obtain a reliable and stable prediction value, the neural network model established by random sampling of 20 iterations is used for prediction, and the average value of the predicted values obtained by each model will be taken as the final prediction result.

4. Predictive Data Analysis

4.1. Sample Data

In this paper, the source of pollution gas is the air pollution gas data of Tucson, a city in southeast Arizona that is famous for its tourism industry [8]. It is predicted by using the past data to predict the future data and the predicted value is compared with the actual value of the existing predicted day to detect and evaluate the pros and cons of the prediction method of this paper.

Considering the formulation of the input and output of the model, the daily concentration of NO_2 for the whole year 2010 is set as the predicted daily concentration data of 365 group, that is, the training expectation of the neural network model. The daily pollutant concentrations within the first ten days, the ten days in the previous year and the ten days in the previous two years corresponding to the prediction day are taken as the 30 training input values of the neural network model, and the model of 30 inputs and 1 output are established. 30 input data for one day: (PPB is an abbreviation for part per billion, which is a gas concentration unit) is shown in Figure 2.

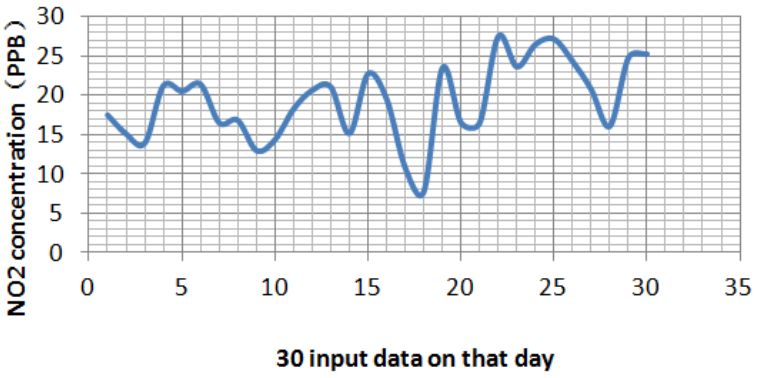

Figure 2 Input data for one day

4.2. Neural Network Model Establishment

A suitable neural network layers and the number of neurons in each layer are needed to drawn through experimental test. Due to the experiment has 30 input variables and one output variable, then the appropriate number of neurons should be around 15. NO_2 concentration data are taken as of experiment subject, 12, 13, 14, 15, 16, and 17 respectively as neural network layer neuron number test. Compared with the data training, the higher the degree of fitting, the closer the *R* is to 1. The test results are shown in Figure 3.

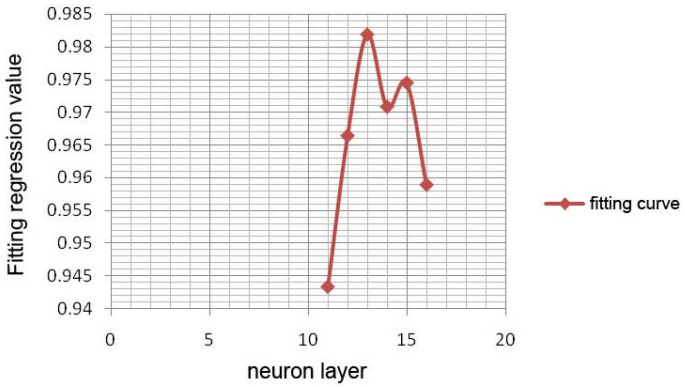

Figure 3 Monolayer hidden layer neuron number variation test chart

It is can be seen from Figure 3 that when the number of neurons is set to 13, the effect of data training is better, so the number of neurons is set to 13 for this paper. After finding the appropriate number of neurons, the number of neurons layer continue to be tested, and the layers are set to 1, 2, 3, 4, and 5 respectively for testing. The test results are shown in Figure 4.

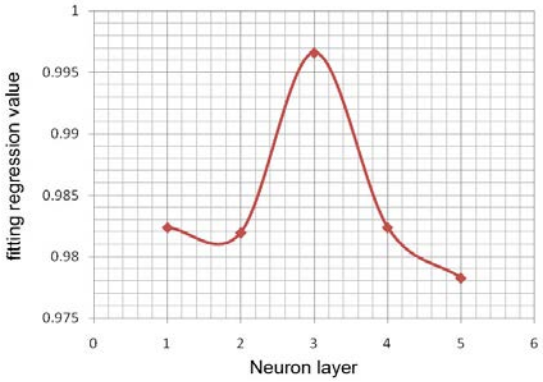

Figure 4 Hidden layer number change test chart

It is can be seen from Figure 4 that when the neuron layer is set to 3, the effect of data training is better, so the number of neurons is set to 13, and the number of neuron layers is set to 3. In this paper, the predicted experimental results of NO2 are analyzed. 30 input variables are set in the neural network, with 3 hidden layers and 13 neurons in each layer, resulting in one output. The schematic diagram of the neural network model after setting parameters is shown in figure 5

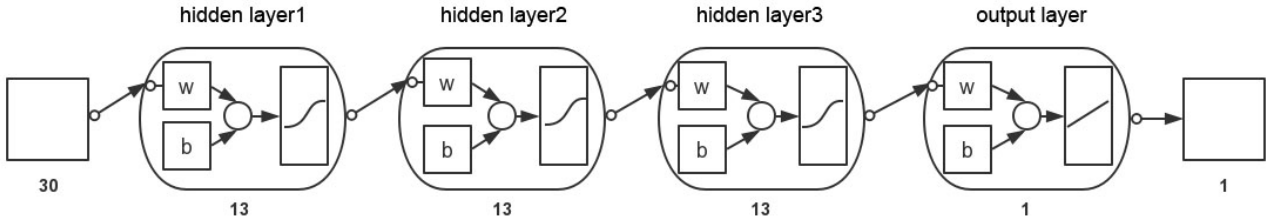

Figure 5 Model diagram of neural network

4.3. Comparison of Forecast Results

(1) Comparison of results and errors before and after data pre-processing

The data used for comparison in this paper is the actual value of the polluted gas in the air in Tucson, Arizona, USA in the past month, compared with the predicted values obtained from the previous data in the neural network modelling prediction. It is can be seen from Figure 6 that in the data of the ideal target real value after pre-processing, some values that deviate too much from the overall data have been appropriately modified. For example, the data of the third, the 25th and the 26th have changed, and the data is not pre-predicted. The third data processed is too large, and the value is reduced after pre-processing, the 25th and 26th data values are too small, and the pre-treatment is appropriately increased. Further comparative analysis shows that the results predicted by neural network modelling using pre-processed data have a better predictive effect on the overall level than the results predicted by the unpreprocessed data, such as the third. The data effects of 9, 10, 30, etc. are significantly improved, and the predicted values are closer to the true value. At the same time, according to the results of Figure 6, it can be found that the data is cross-validated in the prediction process, which is significantly more optimized than direct prediction.

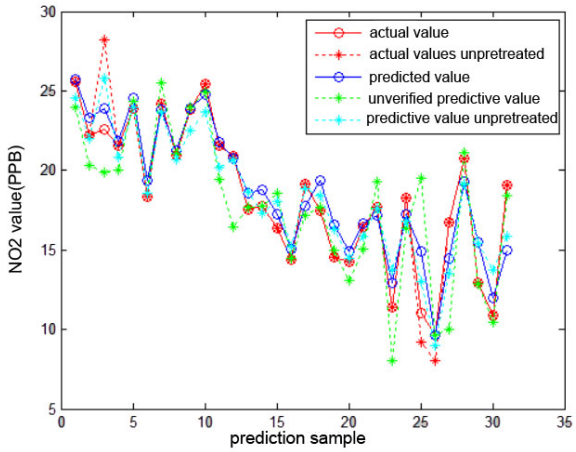

Figure 6 NO2 concentration prediction deviation figure

(2) Comparison of results and errors before and after data cross-validation

In order to compare and analyze more intuitively, this paper calculates the error between the ideal target value processed and the result predicted by the pre-processed data, and the result is shown in Figure 7.

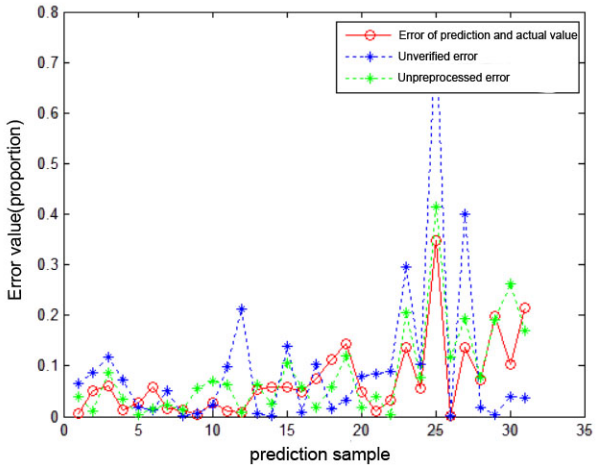

Figure 7 Comparison diagram of NO2 concentration prediction deviation ratio

It is easy to find that the error between the predicted result and the ideal value is smaller and the accuracy of the predicted result is higher after the data is pretreated. Further analysis shows that the error value after cross-validation is significantly reduced, and the accuracy and stability of prediction results are greatly improved.

5. Conclusion

By comparing the predicted results with the expected value, it is found that the predicted value is gentler than the ideal target value, and for some abrupt changes of the target value, the prediction error may be larger. Considering the fact that the concentration of a certain pollutant in the air of a city should not change abruptly within a day, the reliability of the abrupt change data, or that the abrupt change data do not appear frequently, may be taken into account here. So a relatively gentle prediction result is obtained, although the deviation of some values is large, it is acceptable for the whole. It is can see from the prediction results that most of the errors of the prediction results are controlled at about 0.1, that is, within the true value of 0.8 to 1.2 times, so the accuracy and stability of the prediction results is better. Compared with the original neural network prediction method, the prediction error is much smaller, and the accuracy and stability of the prediction results are obviously improved.

Acknowledgements

This work is supported by NSFC (Grant Nos. 61762067), Natural Science Foundation of Jiangxi Province (Grant no. 20161BAB212034), and Jiangxi Province Education Department (Grant No. GJJ160692).

References

[1] *Gu W.H., Xu R.H.. (2013) Forecast Greenhouse Gas Emissions From China's International Shipping. Journal of Wuhan University of Technology, 1, 2-4.*

[2] *Chen L.Q., Yu C.Z.. (2013) Forecasting Gas Concentration Based on Multiple Regression. Journal of Shanghai Institute of Technology (Natural Science), 1, 2-5.*

[3] *Han J., Wang G.H., Zhang X.Y.. (2011) Position based on BP network prediction of air quality. Electronic Design Engineering, 18, 1-3.*

[4] *Xu L.Y.. (2015) Research of Air quality prediction model Based on Rough Set-BP neural Network. Journal of Northeast Electric Power University, 5, 2-7.*

[5] *Vaz A.G, Elsinga B, Brito M.C, et al. (2016) An artificial neural network to assess the impact of neighbouring*

photovoltaic systems in power forecasting in Utrecht, the Netherlands. Renewable Energy, 82, 631-641.

[6] Zhi L.H., Hu P.. (2018) Viscosity prediction for six pure refrigerants using different artificial neural networks. International Journal of Refrigeration, 88, 432-440.

[7] Ren C., An N., Wang J.Z., et al. (2014) Optimal parameters selection for BP neural network based on particle swarm optimization: A case study of wind speed forecasting. Knowledge-Based Systems, 56, 226-239.

[8] Dataju. U.S. Pollution Data Pollution in the U.S. since 2000. http://dataju.cn/Dataju/web/datasetInstanceDetail/211?utm_source=qq&utm_medium=social&utm_oi=842771955634556928. 2018-07-01.