A New $PM_{2.5}$ Air Pollution Forecasting Model Based on Data Mining and BP Neural Network Model

Anna Li[2] and Xiao Xu[1, *]
[1] Shanghai Engineering Research Center of Hadal Science and Technology, Research Center for Ocean Mapping and Applications, College of Marine Sciences, Shanghai Ocean University, Shanghai, China
[2] University of Science and Technology of China, Hefei, 230026, Anhui, China
[*] Corresponding author

***Abstract*—A new $PM_{2.5}$ air pollution forecasting model based on data mining and BP neural network model was established in this paper. This model combined data mining and BP neural network algorithm with data of mass concentration of $PM_{2.5}$ and meteorological data obtained from the Ministry of Original data Environmental Protection in China and Anhui Meteorological Data Service Center. The test results showed that the new $PM_{2.5}$ air pollution forecasting model had higher prediction accuracy than before.**

***Keywords*—$PM_{2.5}$ air pollution forecasting; BP neural network; Data mining; Meteorological data**

I. INTRODUCTION

Recently, haze has attracted the attention of experts in many fields since it can significantly reduce visibility and has an adverse effect on human health. Fine particulate matter ($PM_{2.5}$: particles with aerodynamic diameter $\leq$ 2.5 μm) has been strongly related to adverse health problems by numerous studies conducted mostly in developed countries, also it can directly changing the radiation balance, thus, contributing to climate effects.

The Yangtze River Delta (YRD) is one of the world's fastest-developing economic zones, which located in the center of the vibrant economic area of Eastern China. However, annual $PM_{2.5}$ in YRD was still approximately two times higher than the World Health Organization (WHO) Air Quality Guidelines (annual average: 10 μgm^{-3}). Hefei, as a representative city in the YRD, is located in the upper reaches of the Yangtze River. The recent rapid increase in agricultural activity and urbanization in Hefei has produced great influence on the atmospheric aerosol issues in China. Therefore, it is important to make reasonable air pollution forecasting prediction.

Data mining is a way of using statistical techniques, artificial intelligence technology, machine learning technology and database technology. Many scholars have already used different methods of data mining to predict $PM_{2.5}$ concentrations.

In recent years, there are multiple algorithms such as multiple linear regression, fuzzy logic, support vector machine, artificial neural network, Kalman filtering and hidden Markoff model. It has been shown that the prediction of $PM_{2.5}$ air pollution by machine learning algorithm can increase prediction accuracy because of the strong expressive power for nonlinear process from the machine learning model. In order to get higher prediction accuracy, the selection of key input variables was taken into consideration. The artificial neural network technology has great superiority in solving uncertainty, more input and complicated nonlinear problems, also it has good ability of forecasting.

Although it has high accuracy in mass concentration of $PM_{2.5}$ prediction, there are still some limitations in this field. For instance, the mechanism of machine learning has not cleared enough to explain the whole prediction process. Machine learning requires multiple historical data. Besides, the prediction method of machine learning is mainly limited to the prediction analysis of single site.

This paper presents a new $PM_{2.5}$ air pollution forecasting model based both on data mining and BP neural network model.

II. BACK PROPAGATION NEURAL NETWORK ALGORITHM

Back propagation (BP) neural network structure is shown in Figure 1, including an input layer, a hidden layer and an output layer. Each layer contains several neurons, between layers and layers of the nodes are connected by the corresponding weights. The number of hidden layer nodes formula as following:

$$\mathrm{nNode} = \sqrt{M + N + 1} + \mathrm{a} \tag{1}$$

where *nNode* represents the number of neurons, *n* represents the number of input neurons, and *m* represents the number of output neurons, where *a* represents a constant between 1 and 10.

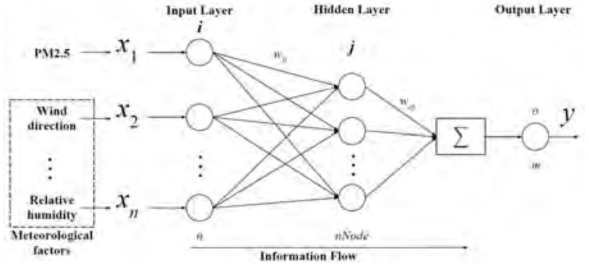


FIGURE I. STRUCTURE OF SINGLE HIDDEN LAYER BP NEURAL NETWORK

The input data vector is (x, x_2,…, x_n)^T, and output is a *y* data. The input and output of the hidden layer neurons are shown in equations (2) and (3):

$$net_{ji} = \sum_{i=1}^{n} w_{ji} * x_{n-i} + \alpha_j \qquad (2)$$

$$x_j = f_j(net_{ji}) \qquad (3)$$

The input and output of the output layer neurons are shown in equations (4) and (5)

$$net_{ji} = \sum_{i=1}^{n} w_{oj} * x_j + \alpha_o \qquad (4)$$

$$\mathrm{y} = f_o(net_{oj}) \qquad (5)$$

ω_{ji} and ω_{oj} are the weights of the connection between the input and the hidden layer respectively, the connection between the hidden layer and the output layer with the value range [-1, 1].

When being training, the output error of neurons in output layer is spreading into hidden layer and input layer, and then revised weight of each layer. It makes the mean square error E of network system to minimize. Finally, forecasting model is set up by the weights of record network after training. The training steps are as following:

a) Step1: Input learning mode including the input vector and the desired output vector of the training samples.

b) Step2: Compute output of hidden layer unit

c) Step3: Compute error of output layer unit.

d) Step4: Determine whether the algorithm achieves convergence conditions, if so, stop the process and output the optimal individual; otherwise, do it sequentially.

e) Step5:Adjust the weights between different layers

f) Step6: If the learning mode ends, then go to 8 steps otherwise go to 1 step.

g) Step7: Update learning model

h) Step8: Enter the test set.

III. PM2.5 FORECASTING

A. *Using the Data Mining Technology to Obtain Main Factors Influencing PM2.5 Forecasting*

Firstly, we have to figure out whether these factors can be used to classify algorithm, and data mining technology is utilized to analyze. This paper takes Hefei area as an example, collecting the $PM_{2.5}$ concentration data and meteorological data during the winter time in 2017-2018(Nov., Dec., Jan., Feb.) to data mining. Time series of $PM_{2.5}$ concentrations during the winter is shown in Figure 2.

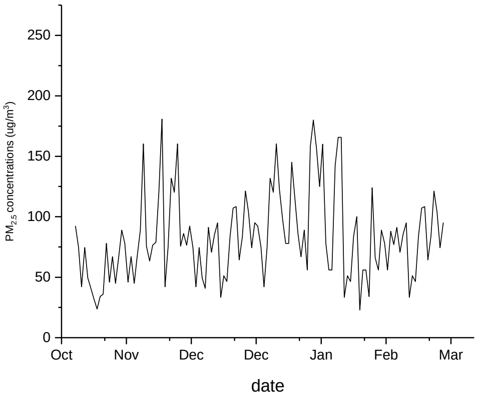
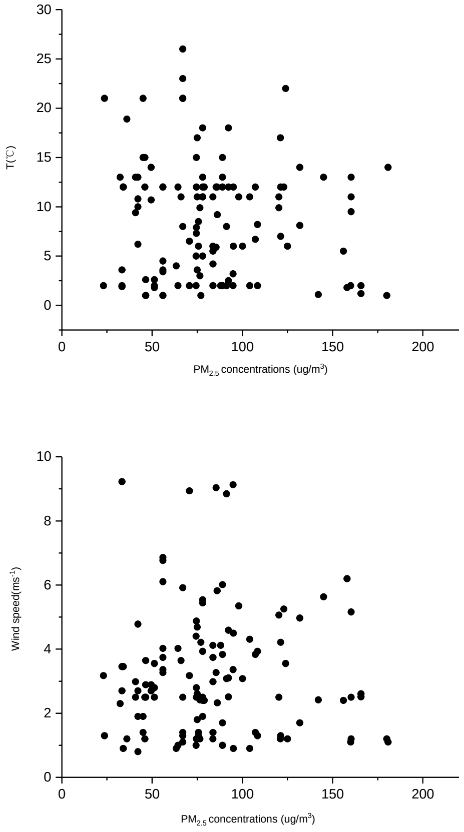


FIGURE II. TIME SERIES OF $PM_{2.5}$ CONCENTRATIONS THROUGHOUT THE WINTER

The relationship between $PM_{2.5}$ and temperature(℃), wind speed(ms^{-1}), wind direction(°) and RH(%) is shown in figure 3.



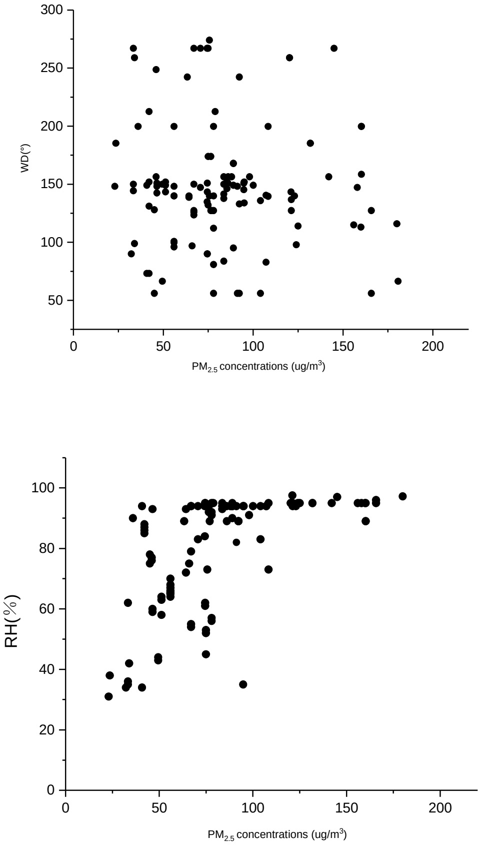


FIGURE III. THE DISTRIBUTION RELATIONSHIP BETWEEN $PM_{2.5}$ AND TEMPERATURE, WIND SPEED, WIND DIRECTION AND RH

Using the data mining technology, we can conclude relative humidity is the main effect of $PM_{2.5}$ dilution. It can be easily seen from Fig. 3 that there is a positive correlation between $PM_{2.5}$ and RH. The mass concentration of $PM_{2.5}$ increases with the increasing humidity. Generally speaking, the temperature, wind speed and wind direction had no significantly correlation with $PM_{2.5}$.

B. Gathering Impact Factors Data as the Training Sample for Neural Network Training

We input Hefei $PM_{2.5}$ data and meteorological data in 2017-2018 as training sample, using Matlab neural network toolbox to complete training.

C. Using the Test Sample to Obtain the Evaluation of Forecasting Model

We input Hefei $PM_{2.5}$ data and meteorological data during the winter of 2017 (Nov., Dec., Jan., Feb.) as test sample, to evaluate the model's precision.

IV. ANALYSIS OF TEST RESULT

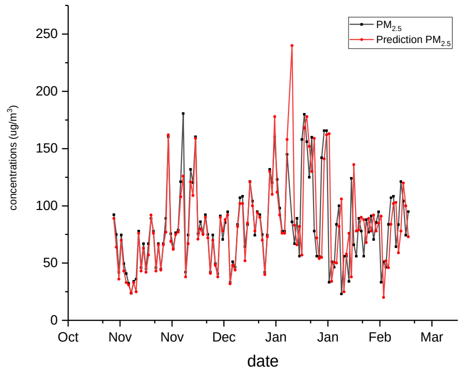


FIGURE IV. THE RESULTS OF TRAINING DATA AND PREDICT DATA

After the neural network training, we use data during the winter of 2016-2017 to test. The experimental results are shown in Figure 4. As can be seen from figure 4, the changing trend of the forecasting data is consistent with the actual data.

In addition, in order to evaluate the forecast model, following statistical parameters were calculated: MAE – mean absolute error; RMSE – root mean square error; IA – index of agreement; R^2 – coefficient of determination; R – correlation coefficient. the best configuration was tested as following:

MAE [μg/m^3]: 0.92472;

RMSE [μg/m^3]: 1.2756;

IA: 0.9452;

R^2: 0.9188;

R: 0.9315.

V. CONCLUSIONS

In this paper, we established a new $PM_{2.5}$ air pollution forecasting model based on data mining and BP neural network model. This model combines with BP neural network and data mining technology. The test results show that the forecasting model established in this paper has high prediction accuracy. Moreover, for some sites lack of meteorological recorded data, our study showed that neural models can provide accurate $PM_{2.5}$ forecasting.

However, the prediction of $PM_{2.5}$ concentration is affected by a great many factors. Hence, how to effectively model and predict still needs further study. In the future study, we need obtain more historical data of longer periods.

ACKNOWLEDGMENT

This work was supported by the Shanghai Ocean University Research Initiation Funds for the Post-Doctorial Scholarship (Grant No. A2-0203-00-100311), the Shanghai Ocean University Funds for Ocean Mapping and Applications (Grant No. A1-2801-18-300211), the State Key Program of National Natural Science of China 'Structural Reliability Analysis on the Spherical Hull of Deepsea Manned Submersibles' (Grant No.

51439004), the scientific innovation program project ‘Key technology research and experimental validation of deep manned submersible’ by the Shanghai Committee of Science and Technology (Grant No. 15DZ1207000)

REFERENCES

[1] Nebot and F. Mugica, “Small-Particle Pollution Modeling Using Fuzzy Approaches”, M.S. Obaidat et al. (eds.), Simulation and Modeling Methodologies, Technologies and Applications, Advances in Intelligent Systems and Computing, Springer International Publ. pp. 239-252, 2014.

[2] K. Polat and S.S. Durduran, “Usage of output-dependent data scaling in modeling and prediction of air pollution daily concentration values (PM_{10}) in the city of Konya”, Neural Computing and Applications, 21, pp. 2153-2162, 2012.

[3] S.F. Mihalache, M. Popescu, and M. Oprea, “Particulate Matter Prediction using ANFIS Modelling Techniques”, Proc. of 19th International Conference on System Theory, Control and Computing (ICSTCC), October 14-16, Cheile Gradistei, Romania, pp. 895-900, 2015.

[4] Lin Wang, Yi Zeng, T. C. Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. Expert Systems with Applications, 2015, 42(2): 855-863.

[5] Chan CK, Yao X (2008) Air pollution in mega cities in China. Atmos Environ 42(1):1–42

[6] Fleming SW (2007) Artificial neural network forecasting of nonlinear Markov processes. Can J Phys, 85(3):279–294(16)

[7] Pai TY, Ho CL, Chen SW, Lo HM, Sung PJ, Lin SW, Lai W-J, Tseng S-C, Ciou S-P, Kuo J-L et al (2011) Using seven types of GM (1.1) model to forecast hourly particulate matter concentration in Banciao City of Taiwan. Water Air Soil Pollut 217:25–33

[8] Domańska and M. Wojtylak, “Application of fuzzy time series models for forecasting pollution concentrations”, Expert Systems with Applications, 39, pp. 7673-7679, 2012.