Proceedings of the 2012 2nd International Conference on Computer and Information Application (ICCIA 2012)

An Air Quality Predictive Model of Licang of Qingdao City Based on BP Neural Network

Ruobo Xin, Zhifang Jiang, Ning Li
School of Computer Science & Technology
Shandong University
Jinan, 250101, China
chinaxin123@163.com

Lujian Hou
Jinan Academy of Environmental sciences
Jinan, 250014, China
112750122@qq.com

***Abstract*—In order to obtain high precision results of urban air quality forecast, we propose a short-term predictive model of air quality in this paper, which is on the basis of the ambient air quality monitoring data and relevant meteorological data of a monitoring site in Licang district of Qingdao city in recent three years. The predictive model is based on BP neural network and used to predict the ambient air quality in the next some day or within a certain period of hours. In the design of the predictive model, we apply LM algorithm, Simulated Annealing algorithm and Early Stopping algorithm into BP network, and use a reasonable method to extract the historical data of two years as the training samples, which are the main reasons why the prediction results are better both in speed and in accuracy. And when predicting within a certain period of hours, we also adopt an average and equivalent idea to reduce the error accuracy, which brings us good results.**

***Keywords-** prediction; air quality; BP neural network*

I. INTRODUCTION

With the development of the world's industry, global environmental problems, especially urban air pollution problems, has become increasingly prominent, and aroused more and more general concern of the society. People began efforts to explore new methods of urban air quality prediction in view of ambient air quality monitoring. Ambient air quality was difficult to predict due to the wind speed, wind direction, temperature, humidity, atmospheric pressure, rainfall, solar radiation intensity, geographical conditions, human activities, and many other factors, and its predictive accuracy is also difficult to meet people's needs. Error back-propagation neural network, BP (Back Propagation) neural network, has a strong self-learning and generalization ability as a new numerical prediction tools. In mathematics theory, it is shown to have any complex nonlinear mapping function, and has begun to be used for air quality prediction.

Qing Chen etc. (2008) introduced learning rate and momentum factor into BP neural network, and predicted the concentration of pollutants in Chaoyang district, Beijing. [1]. Aizhi Wang etc. (2009) used Early Stopping algorithm in BP neural network, and predicted PM10 air pollution index of 19 districts in Shanghai in MATLAB environment [2], Cuiling Zhu etc. (2007) considered the emissions of pollution sources and meteorological factors, and created a predictive model with BP neural network under different weather conditions [3], Xiujie Zhou etc. (2004) studied air pollution index based on BP neural network with meteorological factors [4]. Qinghua Zhang etc. (2010) predicted and compared PM10 air pollution index of Xi'an city with different methods such as Momentum BP algorithm, LM algorithm and SVG algorithm [5]. Xia Qin etc. (2007) used Bayesian Regularization method and Early Stopping algorithm in BP neural network, and predicted the hourly concentration of $PM_{2.5}$ of monitoring site of Marylebone Road in London, UK. [6]. Zhifang Jiang etc. (2010) proposed a method of air quality prediction based on BP neural network of samples self-organization clustering [7].

II. THEORY OF BP NEURAL NETWORK

BP neural network is a multi-layer feed-forward network being trained with error back propagation algorithm. It is proposed by a team of scientists headed by D.E. Rumelhart and J.L. McClelland, which has three parts of an input layer, a hidden layer and an output layer. The simple three-layer forward BP network topology is shown in Figure 1. The network has only one hidden layer, and the number of neurons of the input layer, hidden layer and output layer are respectively n, p and q. Its input vector (also called input signal) is $X(x_1, \ldots, x_i, \ldots, x_n)$. Its output vector (also called output signal) is $Y(y_1, \ldots, y_k, \ldots, y_q)$. The connection weights between the i-th node of input layer and the j-th node of hidden layer is W_{ji}, and the one between the j-th node of hidden layer and the k-th node of output layer is V_{kj} .

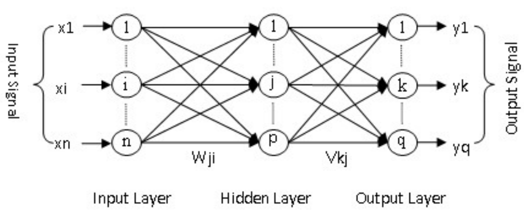

Figure 1. The simple three-layer forward BP network topology.

The basic principle of the back-propagation algorithm is: Each sample through the network training, include the input x_i (i >=1 and i<=n) and the desired output t_k (k>=1 and k<=q). The input signal (i.e. the input) x_i acts on the output node through the hidden layer node, and generates an output signal (i.e. the output) y_k after non-linear transformation in network training. Then the deviation e_k between y_k and the desired output value t_k spreads backward through the network, Shortly afterward, the connection weights W_{ji} between the nodes of input layer and hidden layer, the connection weights V_{kj} between the nodes of hidden layer and output layer, and thresholds are adjusted. This makes errors along the gradient direction decline. After learning repeatedly, the network retrieves the best parameters (including the weights and the thresholds) corresponding to the minimum error, and the network training come to an end. At this time, the network after being trained is able to handle the input information, and output the smallest error results [8].

III. PREDICTIVE ANALYSIS OF AMBIENT AIR QUALITY

A. Data sources and processing

This paper selected the air quality monitoring data and the related meteorological data of a monitoring site in Licang district, Qingdao from January 1, 2010 to June 30, 2012 for the study. The air quality monitoring data consists of the density value of three pollutants, including $SO_2(mg/m^3)$, $NO_2(mg/m^3)$ and $PM_{10}(mg/m^3)$, and the meteorological data consists of wind speed denoted as WS, wind direction denoted as WD, temperature denoted as $TEMP$, humidity denoted as $HUMD$ and atmosphere pressure denoted as $PRESS$. All data, including day data and hour data in two forms, are from the Qingdao Municipal Environmental Protection Bureau. In practice, the data is processed into two formats as shown in Table I and Table II.

TABLE I. AIR QUALITY DAY DATA

Date	SO_2	NO_2	PM_{10}	WS	WD	TEMP	HUMD	PRESS
2010/1/1	0.144	0.064	0.120	5.3	9	-0.4	42	1019
2010/1/2	0.132	0.077	0.217	6.1	7	2.1	68.2	1013
……	…	…	…	…	…	…	…	…
2011/1/1	0.008	0.052	0.012	8.7	13	-4.9	42.6	1024
……	…	…	…	…	…	…	…	…
2012/6/30	0.083	0.054	0.032	4.8	9	23.1	81.4	999

TABLE II. AIR QUALITY HOUR DATA

Datetime	SO_2	NO_2	PM_{10}	WS	WD	TEMP	HUMD	PRESS
2010/1/1 00:00:00	0.17	0.076	0.113	5.3	7	-2	44.7	1020
2010/1/1 01:00:00	0.131	0.072	0.107	3.9	7	-1.4	49.5	1019
……	…	…	…	…	…	…	…	…
2011/1/1 00:00:00	0.006	0.053	0.011	6.8	14	-6.6	36.5	1024
……	…	…	…	…	…	…	…	…
2012/6/30 23:00:00	0.03	0.113	0.115	0.8	3	22.6	79.9	1003

B. BP Model

The network model established in this paper uses a simple three-layer structure. The neurons of its input layer, hidden layer and output layer are respectively set to 6, 12 and 1, as shown in Figure 1 where n = 6, p = 12 and q = 1. Its activation function is set to 1/(1+e^(-x)) , and its weights and thresholds is set to decimal randomly distributed between (-1, 1). Simple BP neural network exists some shortcomings and deficiencies, such as slow convergence, easily trapped into local minima, and often overtraining, therefore it usually needs to be modified in applications. In order to get better results, the paper uses LMBP algorithm, Simulated Annealing algorithm and Early Stopping algorithm.

LM (Levenberg and Marquardt) algorithm [9] is a very effective method in numerical optimization theory. And the LMBP algorithm applying this method obtains both the speed advantage of Newton Method and the convergence characteristics of Steepest Descent Method, and greatly improves the convergence rate. Simulated Annealing algorithm [10] [11] is an optimization algorithm proposed by Kirkpatrick in 1983. Although this method can avoid the network falling into local minimum, but it leads to very slow convergence, which costs greatly and is rarely used alone. Whereas the method can be used simultaneously with LM algorithm, that receives very good results. Early Stopping algorithm [12] is a training method based on self-validation, which introduces the validation set besides the training set. In training, the validation error of the network using this method varies from beginning of being smaller to latter of being larger. When the validation error becomes larger and reaches a certain level, the network will be considered as "over-fitting", and its training will be stopped.

To speed up the convergence rate of the BP network training, we conduct a normalization process on the network input data, normalize the data into interval [0.001, 0.999], and anti-normalize the output data. As shown in Equation (1) and (2).

$$x_i' = \frac{x_i - x_{i\min}}{x_{i\max} - x_{i\min}}(0.999 - 0.001) + 0.001 \quad (1)$$

$$y_i' = \frac{y_i - 0.001}{0.999 - 0.001}(y_{i\max} - y_{i\min}) + y_{i\min} \quad (2)$$

Equation (1) and (2) are respectively for the normalization and anti-normalization formula. Where, x_i is an I-dimensional component value of the input vector X, x_i' is the component value of normalization of x_i, $x_{i\min}$ is the minimum value of the I-dimensional component values of the input vector X in training sample space, $x_{i\max}$ is the maximum value of the I-dimensional component values of the input vector X in training sample space, y_i is an I-dimensional component value of the input vector Y, y_i' is the component value of normalization of y_i, $y_{i\min}$ is the minimum value of the I-dimensional component values of the input vector Y in training sample space, $y_{i\max}$ is the maximum value of the I-dimensional component values of the input vector Y in training sample space.

C. Predictive analysis of air quality

The contents of air quality forecast in China mainly include pollution index (API), the primary pollutants, air quality levels and air quality conditions. And all can be easily obtained by calculating with concentration of SO_2, NO_2 and PM_{10}. So the air quality prediction is actually the concentration value prediction of SO_2, NO_2 and PM_{10}.

Predictive model of our study includes prediction for day and prediction for time period of hours. The former one is used to predict air quality of someday in the future. The latter one is used to predict air quality of a continuous period of several hours in the future. Because the former one is the special case of the latter, we describe the predictive process of the former one.

As shown in Figure 2, t_1 and t_2 are two time (unit: day) on the timeline, where t_1 has occurred and t_2 has not occurred. Then the predictive process of the pollutant concentration can be simply described as follows: The concentration value DST'_{t2} of pollutants in the t_2 time can be predicted with the one DST_{t1} in the time t_1 and the weather forecast information in the time t_2. In other words, the input vector is $X(DST_{t1}, WS_{t2}, WD_{t2}, TEMP_{t2}, DUMP_{t2}, PRESS_{t2})$, the output vector is $Y(DST'_{t2})$. The training sample space is composed of the pollutant concentration value and meteorological data, of which the correspondence relationship and time interval between the input vector and the expected output vector of the sample is as same as the one in prediction. In order to get a better training effect, we select the data during the most recent period T_0 and a period of time ($T_1 + T_2$) within the same period in the history of last year as a training sample space, as shown in Figure 2, in which the unit of time is day, and $T_0 = T_1 = T_2 = 60$.

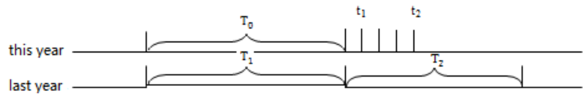

Figure 2. Predictive relation and sample space of the model.

In the prediction for time period of hours, as shown in Figure 3 where t_1 is current time, we take the length of a period of time for future (moment t_2 to t_2 + delt, and the delt <= 24) as a virtual time unit U (U is equal to day, and 1U = delt hours). And then we will get a new U-unit timeline by dividing an hour-unit timeline with t_2 as datum, and the length of time delt as step. So, the sample data with time unit of hour can be transformed to the one with time unit of U by using the arithmetic average method, in addition to wind speed and wind direction by using the vector average method. And we can predict air quality of a period of time for future by using U-unit data of being transformed, as same as the prediction for day. This method is easier than the one with hour as basic predictive unit, and the results using this method are more accurate.

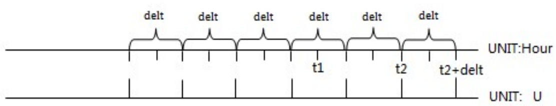

Figure 3. Custom time unit data of prediction for time period of hours.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Predictive Result

With prediction for day and time period, we get the statistics of air quality prediction of the monitoring site in Licang district, Qingdao from July 2011 to June 2012. The following results are obtained with the same set of the model parameters. The maximum number of iterations of the sample training is 3000, and the time spent on predicting air quality for one time is less than 30 seconds.

1) To predict air quality for the next day with current day's data.

TABLE III. PREDICTION ACCURACY OVER THE PAST 12 MONTHS

	Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Avg**
API AARD	0.30	0.21	0.23	0.27	0.29	0.32	0.22	0.22	0.21	0.19	0.18	0.10	0.23
Below 30%	0.74	0.71	0.70	0.71	0.50	0.52	0.71	0.76	0.81	0.70	0.77	0.97	0.72
API AAD	14.33	11.75	14.59	16.33	15.12	26.49	17.49	13.53	15.42	14.75	16.56	7.28	15.30
Below 10	0.42	0.48	0.33	0.29	0.33	0.29	0.48	0.48	0.45	0.50	0.52	0.73	0.44
Below 20	0.77	0.84	0.67	0.71	0.34	0.45	0.61	0.76	0.84	0.70	0.61	0.93	0.69

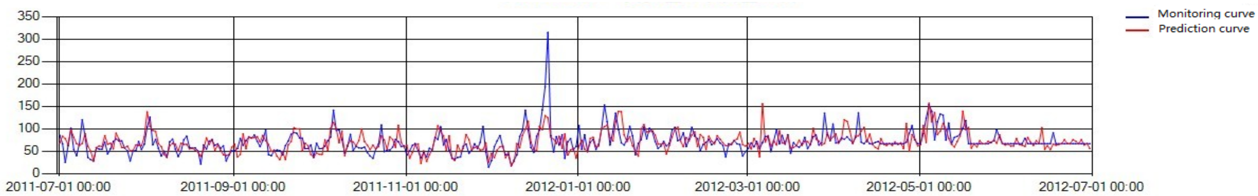

Figure 4. Curves for API predictive value and the true value over the past 12 months.

2) To predict air quality for 6:00-18:00 of current day with 6 hours' data from 18:00 on the day before

TABLE IV. PREDICTION ACCURACY OVER THE PAST 12 MONTHS

	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Avg
API AARD	0.29	0.23	0.20	0.33	0.26	0.22	0.24	0.17	0.19	0.30	0.20	0.33	0.25
Below 30%	0.61	0.71	0.70	0.65	0.70	0.74	0.74	0.83	0.87	0.57	0.81	0.57	0.71
API AAD	15.74	12.21	12.42	40.26	23.24	36.96	18.46	11.45	14.71	20.76	16.80	15.63	19.89
Below 10	0.29	0.52	0.60	0.26	0.30	0.29	0.26	0.59	0.39	0.17	0.39	0.37	0.37
Below 20	0.71	0.81	0.77	0.48	0.57	0.45	0.65	0.83	0.74	0.50	0.65	0.77	0.66

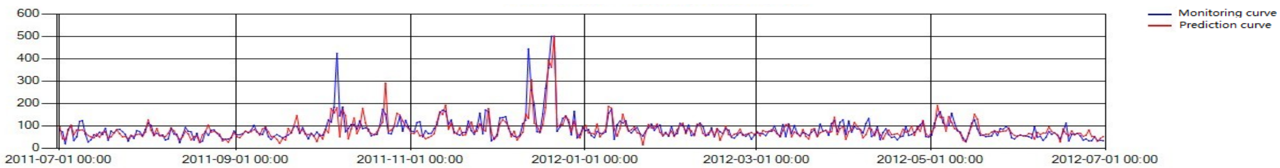

Figure 5. Curves for API predictive value and the true value over the past 12 months.

B. Analysis of experimental results

As shown in Table 1 and Table 2, API average relative deviation of total 12 months are 23% and 25%, API average relative deviation of each month of less than 30% are 72% and 71%, API average absolute deviation of total 12 months are 15.30 and 19.89, API average absolute deviation of each month of less than 20 are 69% and 66%, and predictive average accuracy of primary pollutant are 80% and 79%. With Figure 4 and Figure 5, we can conclude: The model can well predict primary pollutant and API value, the consistency of curves for API predictive value and the true value is higher, and the proportion of the excellent results is higher.

V. CONCLUSION

In this paper, BP neural network is applied to the ambient air quality prediction, and meteorological factors such as wind speed, wind direction, temperature, humidity, and atmospheric pressure are considered, which receive better result. Because the same network structure and parameters are used during the whole process, the prediction has a strong potential for promotion.

REFERENCES

[1] Q. Chen and Y.X. Shao, "The Application of Improved BP Neural Network Algorithm in Urban Air Quality Prediction: Evidence in China," Pacific-Asia Workshop on Computational Intelligence and Industrial Application (PACIIA '08), Dec. 2008, pp. 160-163.

[2] A.Z. Wang and D.B. Li, "BP Neural Network and Its Application to the Forecast of the Air Quality Based on the MATLAB," Journal of Safety and Environment, Vol. 09, No. 02, pp. 84-88, April 2007.

[3] C.L. Zhu, Z.F. Jiang and Q. Wang, "Forecast Model of Environment Air Quality Based on B-P Neural Network," Computer Engineering and Applications, Vol. 43, NO. 22, pp. 223-227, 2007.

[4] X.J. Zhou, X.H. Su and M.Y. Yuan, "Forecast of Air Pollution Index Based on BP Neural Network," Journal of HARBIN Institute of Technology, Vol. 36, No. 05, pp. 582-585, May 2004.

[5] Q.H. Zhang, R. Yao, W.N. Deng and G.R. Wang,"PM10 Pollution and Its BP Neural Network Forecast Model," Journal of XI'AN University of Science and Technology, Vol. 30, No. 01, pp. 43-62, Jan. 2010.

[6] X. Qin, L. Lei and X.L. Yao, "Methods to Improve the Generalization of BP Neural Network Applied in Air Pollution Forecast," Journal of Beijing University of Technology, Vol. 33, No.08, pp. 849-852, Aug. 2007.

[7] Z.F. Jiang, B.Q. Mao and X.X. M et al. ,"An Air Quality Forecast Model Based on the BP Neural Network of the Samples Self-Organization Clustering," 2010 Sixth International Conference on Natural Computation (ICNC 2010), Aug. 2010, pp. 1523–1527.

[8] Z.Z. Shi, Neural NetWorks. Beijing: Higher Education Press, 2009.

[9] Municipal Environmental Protection Bureau of Qingdao. http://www.qepb.gov.cn/.

[10] K.Madsen, H.B.Nielsen and O.Tingleff, Efficient Scheduling Focusing on the Duality of MPL Representatives. Methods For Non-Linear Least Squares Problems 2nd Edition. Denmark: Informatics and Mathematical Modelling Technical, 2004.

[11] B. Zhang, J.W. Ye and Y.C. Hu, "Application of Optimizing the Path by Simulated Annealing," China Journal of Highway and Transport, Vol. 17, No. 01, pp. 79-81. Jan. 2004.

[12] H.G. Chen, J.S. Wu and J.L. Wang, et al. , "Mechanism Study of Simulated Annealing Algorithm," Journal of TONGJI University(Natural Science), Vol. 32, No. 06, pp. 802-805, Jun. 2004.

[13] J.C. Li, G.J. Qin and X.S. Wen et al. , "Over-Fitting in Neural Network Learning Algorithms and Its Solving Strategies," Journal of Vibration Measurement & Diagnosis, Vol. 22, No. 4, pp. 260-264, Dec. 2002.

[14] S. Han, Y.Q. Liu and W. Yang et al. , "An Optimized Vector Average Method for Calculation of Average Wind Direction," Vol. 36, No. 05, pp. 68-72, May 2002.