A Haze Prediction Algorithm Based on PCA-BP Neural Network

Dong Li(✉), Shudong Liu, Rong Liu, Cheng Li, and Yunjie Zhang

School of Computer and Information Engineering,
Tianjin Chengjian University, Tianjin, China
328848806@qq.com

Abstract. In this paper, we proposed a novel haze forecast model, which combine the principal component analysis with back-propagation neural network to solve the air quality problem in China. Comprehensive variables are obtained by dimension reduction on many predictive factors through principal component analysis (PCA). Then we use the comprehensive variables as the input of back-propagation (BP) neural network. Through this process, the correlation among the original predictors can be eliminated and the structure of neural network can be simplified. The simulation results show that the average prediction error of the prediction model by using principal component analysis combined with BP neural network is less than 10%, far lower than the results of the traditional prediction method which only use a single index, $PM_{2.5}$ daily average concentration, to judge whether it is haze day.

Keywords: Haze forecast · PCA · BP neural network
Comprehensive variables

1 Introduction

The solving of the air pollution problem is imminent with the intensification of urbanization in China, especially the air quality in the Beijing-Tianjin-Hebei region. According to statistics, the number of heavy pollution days up to 29 all the year round in Tianjin and the average concentration of $PM_{2.5}$, PM_{10} and NO_2 is above the national standard. Among them, the average concentration of $PM_{2.5}$ is 69 μg/m^3, which exceeds 0.97 times to the normal; the average concentration of PM_{10} was 103 μg/m^3, which was 0.47 times higher.

Because of the vast territory, complex terrain and capricious climate in China, it is very important to monitor and forecast the weather conditions and make corresponding countermeasures. In recent years, domestic and foreign scholars have put a lot of energy in haze prediction research and proposed a variety of prediction methods and models and also achieved a series of research results. Hou used the cubic exponential smoothing model to analysis and forecast haze weather [1]. According to the characteristics of BP artificial neural network, Ai proposed a haze weather forecasting system based on the BP artificial neural network, which can approximate any nonlinear function [2]. Haze forecast method of selective ensemble based on glowworm swarm optimization

C. Li and S. Mao (Eds.): WiCON 2017, LNICST 230, pp. 451–460, 2018.
https://doi.org/10.1007/978-3-319-90802-1_40

algorithm is proposed by Ni [3]. Fu presented a method to predict the haze based on multiple linear regression analysis, whose sample is online update [4]. Miao establish the haze prediction model and diffusion model with the six ambient air parameters of Changchun City in 2013 October and November [5]. These parameters are analyzed statistically under the environment of MATLAB and pay close attention to $PM_{2.5}$ values, which is defined as a good reference to evaluate and analysis of the air quality [6]. However, the performance of the model is seriously reduced when historical data is different from the predicted data in variation, which is not suitable for such a sudden and strong situation.

Many researchers in the field of haze weather prediction have been seeking a method of high reliability and accurate prediction results at present, however, the researchers are continuing to improve the performance of existing forecasting methods due to the high nonlinearity and complexity of the prediction. Various methods and models need further verification in practice [7]. Therefore, in this paper, the PCA is used to process the original multi-dimensional data. Then, input the extracted principal components to the BP neural network, which have powerful nonlinear function mapping capabilities. Finally, the neural network are trained through a lot of historical data. Through this way, either the dimensions of input variables can be reduced or correlativity among input variables can be eliminated, thus both convergence and stability of neural network can be improved.

The sections of this paper are arranged as follows:

1. The Sect. 1 introduces the research background and research significance of this paper and summarizes the model of prediction models of haze forecast in recent years.
2. The Sect. 2 introduces the fundamentals and steps of principal component analysis method.
3. The Sect. 3 introduces the fundamentals of BP neural network.
4. The Sect. 4 raised a model of haze forecast based on the combination of principal component analysis with BP neural network, which is used to evaluate the air quality of Tianjin.
5. The Sect. 5 summarizes the contents of the whole paper and look forward to the improvement of this model of haze forecast.

2 Principal Component Analysis

The principal component analysis method (PCA) is a statistical analysis method which transforms many variables into a few integers by linearly changing the factors. It greatly reduces the computational workload in the analysis process and eliminates the correlation between the original factors $x_1,x_2,\ldots,x_p$. Through the PCA, the original factors can be mapped into a set of integrated variables Z_m with less numbers than original predictor.

The steps of principal component analysis are as follows:

1. Data standardization. At first, the original variables were normalized to eliminate its extreme variation and different dimension. Assuming that there are n sets of data, each group of data has p variables which makes up n × p order matrix:

$$X_{n*p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \vdots & \vdots & & \\ x_{n1} & x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

Generate the standard matrix Z by formula (2):

$$Z_{ij} = (x_{ij} - \frac{1}{n}\sum_{i=1}^{n} x_{ij}) \Big/ \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}, i = 1, 2, \cdots, n; j = 1, 2, \cdots p \tag{2}$$

2. Establish the correlation coefficient matrix:

$$R = Z^T Z \big/ (n-1). \tag{3}$$

solve the eigen equation of the sample correlation matrix $\left|R - \lambda I_p\right| = 0$, we can obtain the eigen value and vector a_i.

3. Calculate principal component contribution rate and determine principal component:

$$\alpha_i = \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i. \tag{4}$$

Principal component contribution rate is used to reflect the amount of information what original variable factor contains. Normally, the 1st, 2nd,…,nth principal components are fetched, which are respectively matched with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$, whose accumulative contribution rates are up to 85–95%. In other word, the value of m in $F_1, F_2, \ldots, F_m$ is determined by accumulative contribution rate of variance, G(m).

$$G(m) = \sum_{i=1}^{m} \lambda_i \Big/ \sum_{k=1}^{p} \lambda_k. \tag{5}$$

4. Calculate the main composition score:

$$l_{ij} = \lambda_i a_{ij}. \tag{6}$$

$$F_i = a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{pi}x_p, i = 1, 2, \cdots, m. \tag{7}$$

$$PCA = \sum_{i=1}^{m} \alpha_i F_i \Big/ G(m). \tag{8}$$

Comprehensive score *PCA* as input variable of BP neural network is the main basis for judging whether it is haze day.

3 BP Neural Network

The model adopted multilayer perceptron neural network of BP algorithm is called Back Propagation Neural Network, which is consisted of an input layer, single or multiple hidden layers and an output layer. There are two stages for learning process of BP Neural Network. In the first stage, output signals can transmit along the formed network to the output layer via hidden layers, during which the weight value of neurons of hidden layers is constant; the conditions of neurons of each layer can only influenced by the neurons from last layer. In the second stage, the difference between the output value from output layer and the expected value will be regarded as an error signal and transmitted back layer by layer, during which the connection weight value between each layer will be modified, which can ensure the error values drop into the allowed error range. Fig. 1 is BP neural network structure.

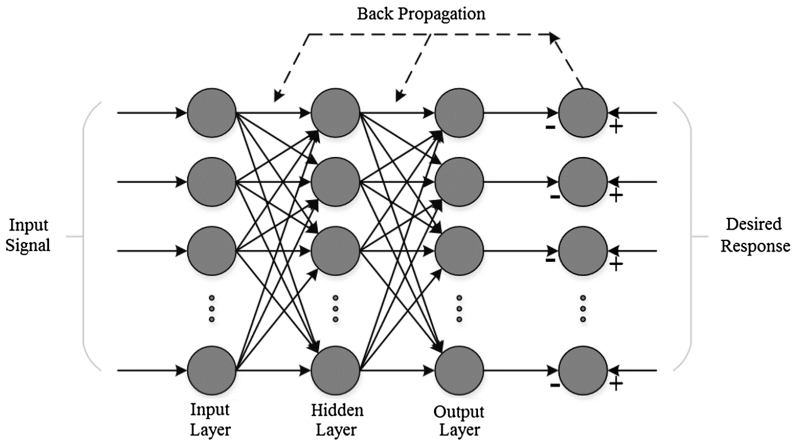

Fig. 1. BP neural network structure.

The output of each layer neuron is shown in formula (9), which show that the mapping between n dimensional space to m dimension space. The connection weight value between each layer: $\omega_{ij}, \theta_j, \omega'_{jk}, \theta'_k, \omega''_{kl}, \theta''_l$.

$$\begin{cases} x'_j = f(\sum_{i=0}^{n-1} \omega_{ij} x_i - \theta_j), j = 0, 1, \cdots, n_1 - 1 \\ x''_k = f(\sum_{j=0}^{n_1-1} \omega'_{jk} x'_j - \theta'_k), k = 0, 1, \cdots, n_2 - 1 \\ y_l = f(\sum_{k=0}^{n_2-1} \omega''_{kl} x''_k - \theta''_k), l = 0, 1, \cdots, m - 1 \end{cases} \tag{9}$$

The steps of BP neural network are as follows:

1. Weight initialization.
2. Enter each learning sample.
3. Calculate the output of each layer: x'_j, x''_k, y_l.
4. Find the back propagation error of each layer and record the values of $x_k^{''(p)}$,$x_j^{'(p)}$ and $x_i^{(p)}$ according to formula (10).

$$\begin{cases} \delta_{ij}^{(p)} = \sum_{k=0}^{n_2} \delta_{jk}^{(p)} \omega'_{jk} x_j^{'(p)} (1 - x_j^{'(p)}), j = 0, 1, \cdots n_1 \\ \delta_{kl}^{(p)} = (d_l^{(p)} - y_1^{(p)}) y_1^{(p)} (1 - y_1^{(p)}), l = 0, 1, \cdots m - 1 \\ \delta_{kl}^{(p)} = \sum_{l=0}^{m-1} \delta_{kl}^{(p)} \omega''_{kl} x_k^{''(p)} (1 - x_k^{''(p)}), k = 0, 1, \cdots n_2 \end{cases} \tag{10}$$

5. Record number of samples that have been learned. If $p < P$, go to step (2) to continue the calculation, if $p = P$, go to step (6).
6. Adjust the weight of each layer according to the weight correction formula.
7. Calculate the new x'_j, x''_k, y_l, and E^T, according to the obtained new weights. If any end conditions satisfied, end the learning process, or else go to step (2) and start a new round of learning. End conditions: $\left| d_l^{(p)} - y_l^{(p)} \right| < \varepsilon$ or $E_T < \varepsilon$.

4 Simulation Analysis

The existing haze forecasting models used to be established by a single factor $PM_{2.5}$ value. This is the main drawback to reduced the prediction accuracy. On the one hand, single factor cannot clearly define the fog and haze weather and it cannot take the complexity of haze components into account.

4.1 Data Source

Daily mean concentrations of 6 kinds of pollutants which affect actually air quality in Tianjin are chosen as predictor in this paper including $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3 and CO by Tianjin Environmental Protection Bureau in 2016. Part of original monitoring data is shown in Table 1.

Table 1. Original monitoring data in Tianjin (μg/m^3).

Date	AQI	$PM_{2.5}$	PM_{10}	SO_2	CO	NO_2	O_3
1/1	220	172.4	237.4	62.7	1700	78.2	22
1/2	395	336.4	480.5	64.5	3488	138.2	7
1/3	129	97.5	109.7	34.5	1254	52.9	25
1/4	140	106	148	36.3	1533	63.2	43
1/5	53	33.8	62.3	25	875	43.5	44
1/6	74	47.8	84.4	31.5	1071	53.8	55

4.2 Principal Component Extraction

According to the steps of principal component analysis in Sect. 2, the correlation coefficient matrix between each factor which is calculated by formula (3) is shown in Table 2.

Table 2. Correlation coefficient matrix

	$Z_{PM2.5}$	Z_{PM10}	Z_{SO2}	Z_{CO}	Z_{NO2}	Z_{O3}
$Z_{PM2.5}$	1	0.864	0.672	0.746	0.778	−0.011
Z_{PM10}	0.864	1	0.595	0.644	0.695	0.084
Z_{SO2}	0.672	0.595	1	0.784	0.808	−0.325
Z_{CO}	0.746	0.644	0.784	1	0.835	−0.313
Z_{NO2}	0.778	0.695	0.808	0.835	1	−0.181
Z_{O3}	−0.011	0.084	−0.325	−0.313	−0.181	1

Table 2 show that there is the correlation between original indicators. As can be seen from Table 2, according to a single indicator, $PM_{2.5}$ daily average concentration, is unable to accurately determine whether it is haze day.

Table 3. Total variance explained.

Component	Total variance explained		
	Total	Variance %	Contribution rate %
1	4.011	66.846	66.846
2	1.164	19.408	86.254
3	0.346	5.762	92.016
4	0.215	3.58	95.596
5	0.146	2.427	98.024
6	0.119	1.976	100

According to the formula (4) to calculate the correlation coefficient matrix eigenvalues, eigenvectors and cumulative contribution rate in Table 3.

As can be seen from Table 3, the contribution rate of the first two principal components has reached 86.274%, which is more than 80%, which can represent the majority of information of the original data, so it can be determined that the number of principal components is 2. According to formulas (6) and (7), the linear combination between two principal components and each variable can be obtained:

$$\begin{cases} F_1 = 0.23Z_{PM2.5} + 0.21Z_{PM10} + 0.22Z_{SO_2} + 0.23Z_{CO} + 0.23Z_{NO_2} - 0.05Z_{o_3} \\ F_2 = 0.23Z_{PM2.5} + 0.33Z_{PM10} - 0.18Z_{SO_2} - 0.14Z_{CO} - 0.02Z_{NO_2} + 0.81Z_{o_3} \end{cases}. \tag{11}$$

Comprehensive score *PCA*:

$$PCA = (0.66846F_1 + 0.19408F_2)/0.86254. \tag{12}$$

4.3 Simulation Results

The model 1 used PCA based on the principal component analysis as input of BP neural network to predict the haze weather prediction model. The model 2 was established only by using the daily average concentration of $PM_{2.5}$ in the original sample as the input of the BP neural network is called the traditional prediction method which is widely used in many fields.

BP neural network can fit any nonlinear curve if there are more hidden layers in theory. BP network structure with two hidden layers was adopted in this paper after several contrast experiments in order to improve prediction accuracy. Each hidden layer contains 10 nodes, the first hidden layer neurons use logsig transfer function, the second hidden layer neurons use pure linear function, the training function using trainlm. Training parameters are set as follows: learning rate 0.01, the maximum number of training 5000, the target error 0.00000001, display step 50.

There are 91 sets of effective data, 70 groups as a training neural network, 21 groups as predictive comparison data from January 2016 to March 2016. Table 4 shows the forecast values, measured values and prediction errors of haze weather indicators in the first three days of April though two models. According to the $PM_{2.5}$ standards set by our country, the average daily concentration is more than 35 μg/m^3, which is regarded as the haze weather. While the PCA value is determined to be greater than 0.15 when compared with actual air quality index AQL, it is considered to be the haze weather. For example, the actual calculation of AQL on April 1 was given in Table 4 is 105, which is considered to be lightly haze day, but the $PM_{2.5}$ daily concentration predicted by the traditional prediction method is 34.19 μg/m^3, it is not haze days according to the detection standard. Based on the model 1, the error calculated according to the error formula (13) is 6.25% less than the traditional prediction method of prediction error of 29.93%, so that the forecasting accuracy is improved.

$$Errors = (MeasuredValue - ForecastValue)/MeasuredValue. \tag{13}$$

Table 4. Comparison of network forecast value and measured value.

Date	AQL	Indicator	Measured value	Haze day	Forecast value	Haze day	Errors%
04.01	105	$PM_{2.5}$	48.8 μg/m^3	√	34.19 μg/m3	×	29.93
		PCA	0.16	√	0.17	√	6.25
04.02	51	$PM_{2.5}$	17.4 μg/m^3	×	20.54 μg/m^3	×	18.06
		PCA	−0.56	×	−0.58	×	3.57
04.03	78	$PM_{2.5}$	29.7 μg/m^3	×	15.46 μg/m^3	×	47.94
		PCA	−0.32	×	−0.27	×	15.62

As Table 4 shown, it can be seen that the prediction $PM_{2.5}$ content of traditional forecasting methods has great errors in determining whether it is haze day. But the average prediction error of the prediction model that combination the principal component analysis with BP neural network is less than 10%, and the predicted value is closer to real value. The main reason is that the new synthetic variables are obtained by principal component analysis eliminate the correlation between original predictive factors, reduce influence of redundant information, greatly simplify the neural network structure and improve the prediction accuracy.

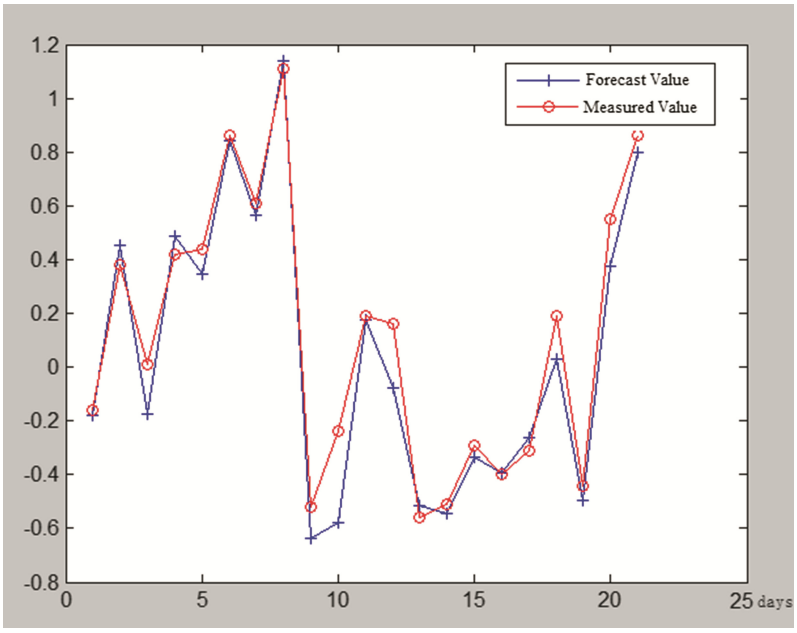

Fig. 2. The curve fitting of the comprehensive variable PCA.

Figures 2, 3 shows the compare results of the predicted value and measured value from 21 days after march 2016. The Y-axis of Fig. 2 is the value of the integrated score of PCA, and the Y-axis of Fig. 3 is the daily average of $PM_{2.5}$ predicted by traditional

prediction method.As shown in Fig. 2, the trend of the predicted and measured values of haze is basically the same. Although there is same error between the predictive value and the true value of the comprehensive variable, but the overall curve fit is higher when the data changes dramatically, the prediction curve can respond quickly to changes in the real curve. Figure 3 is a traditional prediction method for single haze factor $PM_{2.5}$ forecast. Contrast Figs. 2 and 3, there are much more error in the traditional prediction results. Actually, the number of iterations of the traditional prediction method reaches the maximum value after many training and contrast. Indicating that the learning accuracy can not meet the requirements.

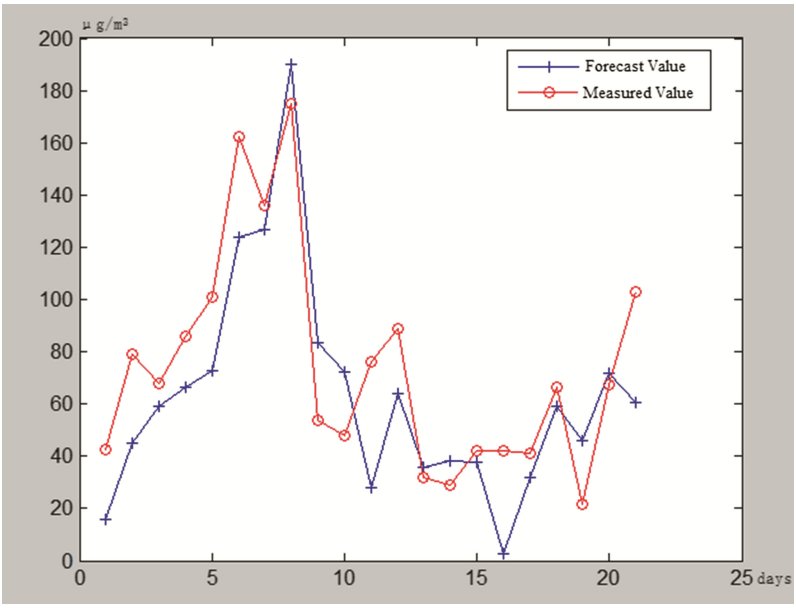

Fig. 3. The curve fitting of traditional forecasting

5 Conclusion

A haze forecast model based on the combination of PCA and the BP neural network is proposed in this paper. Principal components analysis is used to decreases the dimensions of the six predictive factors. Then, the comprehensive variables are taken as input variable for back-propagation neural network, which eliminates the correlation among the original predictors and simplifies the structure of neural network, Results show that complexity of training and training time is reduced and the prediction accuracy had improved, prediction of the fitting curve is much closer to the true value. Although the BP neural network is widely used, it still has some shortcomings, such as the convergence rate is slow, easy to fall into the local minimum etc., in the future, BP neural network can be further optimized to build a more completed prediction model.

Acknowledgments. This research was supported by the Fundamental Research Funds for the Universities in Tianjin (2016CJ11), Projects of the national "863Program" (NO. 2015BAF09B02-3).

References

1. Hou, Q., Yang, H.: Analysis and forecasting of haze weather based on the cubic exponential smoothing model. J. Environ. Prot. Sci. **38**, 73–77 (2014)
2. AI, H., Ying, S.: Study on prediction of haze based on BP neural network. J. Comput. Simul. **32**, 402–405 (2015)
3. Ni, Z., Zhang, C., Ni, L.: Haze forecast method of selective ensemble based on glowworm swarm optimization algorithm. J. PR AI **29**, 143–153 (2016)
4. Fu, Q.: Research on Haze prediction based on multivariate regression. J. Comput. Sci. **43**, 526–528 (2016)
5. Miao, Y.N.: Research on Haze ARIMA-GM Forecast and Diffusion Model Based on the Kalman Filtering (2016)
6. Li, L., Sun, Y.: Haze environment analysis and research based on equalization of PCA. J. Appl. Res. Comput. **42**, 1373–1375 (2015)
7. AI, H., Pan, H., Li, Y.: Research on optimization of PM2.5 content prediction in air haze. J. Comput. Simul. **34**, 392–395 (2017)