CHEMICAL ENGINEERING TRANSACTIONS
VOL. 71, 2018
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Xiantang Zhang, Songrong Qian, Jianmin Xu

ISBN 978-88-95608-68-6; **ISSN** 2283-9216
DOI: 10.3303/CET1871007

Modeling and Application for Air Quality Evaluation Based on Improved Grey Clustering and Neural Networks Classification

Jie Liu[a], Xiaodong Wang[b,*], Ping Huang[c]

[a]Faculty of Public Security and Emergency Management, Kunming University of Science and Technology, Kunming 650093, China
[b]Faculty of Land and Resources Engineering, Kunming University of Science and Technology, Kunming 650093, China
[c]College of Environment and Resources, Fuzhou University, Fuzhou 130021, China
angiaoongwxd@163.com

According to the air quality monitoring date of the six urban evaluation sites arranged in six main administrative districts in Beijing during the four seasons from 2013 to 2014, an evaluation system based on air quality indexes of Chinese was formulated. And, next, based on concentration limits of six main pollutants including PM_{10}, $PM_{2.5}$, O_3, CO, SO_2 and NO_2, analysis of the improved grey clustering method and the neural networks classification method were used for evaluation by using the detailed evaluation procedures. An improved grey clustering method based on the whitening function of the exponential types was adopted to improve the calculation precision of the fuzzy factors. It could be seen that the method adopted could more fully reflect the comprehensive effects of all the pollutants on the air quality. Meanwhile, two classification methods based on neural networks were adopted to classify the pollutants into different levels by MATLAB tools. By these methods the comprehensive influence of the pollutants on ambient air could be determined accurately and quickly. The evaluation results demonstrated that the improved grey clustering method which modified weight and extend the whitening function scope could greatly improve information utilization. By considering the weights of whitening values and standard values in different grey classes, the cluster result was more comparable. The neural networks classification method effectively improved the network generalization by the massive data of training samples and testing samples constructed by LINSPACE function in MATLAB tools. The evaluation results obtained by BP and RBF neural network learning showed high accuracy. Among them, the evaluation results of RBF neural network classification and grey clustering modified were more similar that the air quality level of six main administrative districts in Beijing was generally in grade 2. According to the evaluation purposes of the ambient air quality and the practical evaluation demands, it was possible to combine these two methods from different perspectives to gain a more reasonable evaluation result. Thus, it could be concluded that the evaluation of the pollutant contents in the air was of great practical value for eventual heightening of the new ambient air quality standard.

1. Introduction

Ambient air quality evaluation is the basis for scientific decision and integrative management of air pollution control, so there have been many previous researches on the evaluation methods (Thomas et al., 2016; Han and Wang, 2006). AQI (Air Quality Index) is the main air quality evaluation method at present, which is based on the effect on human health and mainly reflects the effect of the individual pollutant. For analysis of comprehensive pollutant contents, the comprehensive assessment based on classification or membership degree of pollution level should be developed. There is still no unified comprehensive evaluation method at present. Generally, there is complicated non-linear relationship between evaluation factors and pollution levels in complex evaluation problems. Hereby, the improving or modifying of the comprehensive evaluation method should be developed on the basis of reasonable algorithms. To this point, some comprehensive evaluation index methods were proposed (Chen et al., 2012; Reddy et al., 2004). But limiting by policies, regulations and historical background of Chinese society, the previous evaluation systems established have not already solved the current evaluation works due to their insufficient such as the incomplete of the evaluation index or

Please cite this article as: Liu J., Wang X., Huang P., 2018, Modeling and application for air quality evaluation based on improved grey clustering and neural networks classification, Chemical Engineering Transactions, 71, 37-42 DOI:10.3303/CET1871007

the disunity of the standard values. Therefore, establishing a complete and accurate evaluation method base on Chinese new standards is helpful to analyze the comprehensive feature of the current air quality Beijing. Moreover, there is practical significance on the improving of the ambient air quality of Beijing. According to the monitoring data of pollutants and their limitation standards, pollution index which effected air quality was researched for evaluation. Two models of grey clustering modified and neural networks classification were then adopted and their evaluation results were compared to provide an effective method for comprehensive evaluation of ambient air quality (Chen et al., 2018).

2. Improved grey clustering model

2.1 Whitening function

By considering the fuzziness of environmental quality classification and the greyness of environmental system in grey clustering evaluation, the whitening values of different clustering indexes are sorted by grey classes, so that the grey class that the cluster objective belongs to can be determined (Deng J., 1982). Suppose i=1,2,…,m was clustered objects, j=1,2,…,n was clustering index, k=1,2,…,p was grey class, x_{ij} (i=1,2,…,m; j=1,2,…,n) was the measured concentration values, the sample matrix of m clustered objects about n clustering indexes could be established. Suppose f_{jk} ($f_{jk}\in[0,1]$) was the whitening function of j index belonged to k grey class, y_{jk} was the standard value of f_{jk}, which was determined by classification standard of j. In order to improve the coverage and information utilization, the whitening function of exponential type was established. The grey clustering model of modified included 3 whitening functions of exponential type as Eq (1), Eq(2), Eq(3):

$$f_{j1}(x_{ij})=\begin{cases}1 & x_{ij}\in[1,y_{j1}]\\ e^{\frac{y_{j1}-x_{ij}}{y_{j1}}} & x_{ij}\in(y_{j1},+\infty)\end{cases} \tag{1}$$

$$f_{jk}(x_{ij})=\begin{cases}e^{\frac{x_{ij}-y_{j(k-1)}}{x_{ij}}} & x_{ij}\in[0,y_{j(k-1)}]\\ 1 & x_{ij}\in(y_{j(k-1)},y_j]\\ e^{\frac{y_{jk}-x_{ij}}{y_{jk}}} & x_{ij}\in(y_{jk},+\infty)\end{cases} \tag{2}$$

$$f_{jp}(x_{ij})=\begin{cases}e^{\frac{x_{ij}-y_{j(p-1)}}{x_{ij}}} & x_{ij}\in[0,y_{j(p-1)}]\\ 1 & x_{ij}\in(y_{j(p-1)},+\infty)\end{cases} \tag{3}$$

where $2\le k<p$.

2.2 Clustering coefficient

Clustering weight is the measurement of indexes to the same gray class, which is different in the same grade for the pollutants in air quality evaluation system (Xu et al., 2006). Traditionally, threshold method or zero weight was used to determine clustering weight. However, these methods have not considered that the variation ranges of standard values of pollutants in the same grade are different. So the weights of indexes in different grades should not only consider the measured concentrations of pollutants but also the standard values in different air quality grades. Hereby, the dimensionless standard values and sample values was adopted, then the standardized values with the dimensionless number was equal to 1 were used to calculate clustering weight. Clustering coefficient σ_{ik} was the membership of object i to grey class k as Eq (4):

$$\sigma_{ik}=\sum_{j=1}^{n}f_{jk}(x_{ij})\eta_{jk} \tag{4}$$

where η_{jk} was the clustering weight of level k and index j. Hereby, the clustering vector of clustered object i could be expressed as $\sigma_{ik}=(\sigma_{i1},\ \sigma_{i2},\ldots,\ \sigma_{ip})$. If $\sigma_{ik}=\max_{i\le k\le p}\{\sigma_{ik}\}$, the clustered object i would be belong to grey class k.

3. Neural network classification model

3.1 BP network

BP network is a type of multi-layered feed-forward network and its main feature is feed forward signal and back propagation error (Rava et al., 2017). Input information by layer-by-layer process was transmit to output layer and neural state at each layer only affect the next neural state. Suppose the input of BP network was $(X_1,X_2,\ldots,X_n)^T$, x_i was the input value of layer i, the number of the neurons was n, the input neuron was m, and the neurons in hidden layer was l, the connection weights between input layer i and hidden layer j was w_{ij}, the connection weights between layer j and output layer k and their threshold values were respectively w_{jk}, a_j and b_k, the expected output was y. The incentive degree of neuron was determined by its activation function. Normally, the activation function of SIGMOID type was chosen for research. In general, the learning method of BP network were in the following steps, (1) network initialization and determining the number of neurons and their connection weights between the layers of input, hidden and output, (2) calculating the output of input layer and output layer and the predicted output of the BP network, (3) updating the weights and the threshold values according to predicted output of the BP network, (4) judging whether to end the iteration and returning to step (2) if not.

3.2 RBF network

Also, like multi-layered feed-forward network, RBF network is formed by the layers of input $(X_1,X_2,\ldots,X_n)^T$, hidden layer K_i and output layer R_i (Rava et al., 2017). The hidden layers formed by hidden units transfer input vectors so that the linearly non-separable in low-dimensional space can be changed into linearly separable in high-dimensional space. The input layers of RBF network are responsible for information transfer and the output are the weighted sums of hidden layer units. The distance between input vector and weight vector are taken as independent variable ($||dist||$). Normally, Gauss function was used as RBF activation function. The output of RBF network is the product of the distance between connection weight w' and input vector and their threshold values b'. The difference between BP and RBF networks is mainly in activity functions. Similar to BP network, the learning method of RBF network were in the following steps, (1) network initialization and choosing training sample as the cluster center at random, (2) grouping the input training sample by the nearest neighbor algorithm, (3) adjusting the cluster center to the RBF center, or returning to step (2) to solve the center of next round, (4) solving the variance and the weight between hidden and output layer.

4. Results and discussions

4.1 Improved grey clustering and its resolving

The clustering grey class was divided into 6 grades to reflect air quality, which was consistent with the Chinses AQI method (Liu et al., 2015). According to the monitoring data of six main pollutants obtained by twelve automatic monitoring stations in six main administrative districts in Beijing during the four seasons from 2013 to 2014, the mean concentrations of four seasons were calculated to evaluate air quality by grey clustering modified based on classification standard. Thus the sample matrix of mean concentrations in four seasons could be established. The sample data in spring were chosen for example to establish the sample matrix D_1:

$$D_1 = \begin{bmatrix} 117 & 88 & 29 & 54 & 1.528 & 69 \\ 121 & 82 & 28 & 48 & 1.618 & 60 \\ 108 & 81 & 31 & 55 & 1.519 & 65 \\ 130 & 88 & 29 & 51 & 1.623 & 51 \\ 117 & 87 & 32 & 61 & 1.539 & 52 \\ 123 & 81 & 32 & 61 & 1.539 & 52 \\ 112 & 92 & 31 & 77 & 1.568 & 47 \\ 127 & 85 & 22 & 55 & 1.522 & 42 \\ 97 & 77 & 21 & 42 & 1.581 & 74 \\ 126 & 86 & 31 & 61 & 1.599 & 48 \\ 137 & 84 & 30 & 42 & 1.688 & 56 \\ 140 & 88 & 25 & 62 & 1.596 & 62 \end{bmatrix}$$

Dongsi, Dongcheng district (site 1)
Temple of Heaven, Dongcheng district (site 2)
Guanyuan, Xicheng district (site 3)
Guanyuan, Xicheng district (site 4)
Olympic Sports Center, Chaoyang district (site 5)
Guanyuan, Chaoyang district (site 6)
Wanliu, Haidian district (site 7)
North New Area, Haidian district (site 8)
Beijing Botanical Garden, Haidian district (site 9)
Huayuan, Fengtai district (site 10)
Yungang, Fengtai district (site 11)
Gucheng, Shijingshan district (site 12)

According to the construction method of whitening function of the exponential types as Eq (1), Eq(2), Eq(3) and the classification standard of air pollutant concentration of Chinese AQI, the whitening function of index j belong to grey classes could be obtained. Then, take the sample data of the first object (site 1) in spring for

example, the whiten function values of six indexes belong to grade 1-6 were obtained:

$$f_1 = \begin{bmatrix} 0.2417 & 1 & 0.7869 & 0.3443 & 0.1507 & 0.0845 \\ 0.2193 & 0.8396 & 1 & 0.7370 & 0.4954 & 0.1592 \\ 1 & 0.509 & 0.0178 & 0 & 0 & 0 \\ 0.6964 & 1 & 0.6259 & 0.0998 & 0.0159 & 0.0001 \\ 1 & 0.7343 & 0.1983 & 0.0003 & 0 & 0 \\ 1 & 0.644 & 0.2714 & 0.1229 & 0.0598 & 0 \end{bmatrix}$$

Likewise, the whiten function value of four seasons of indexes belong to grey classes could be obtained according to the above method (and not explained here). By calculating the clustering weight of six pollutants in site 1 η_1 and according to Eq (4), the clustering coefficient of the first index of the first clustered object belong to grey classes was obtained:

$$\sigma_{11} = f_{11}(x_{11})\eta_{11} + f_{21}(x_{12})\eta_{12} + f_{31}(x_{13})\eta_{13} + f_{41}(x_{14})\eta_{14} + f_{51}(x_{15})\eta_{15} = 0.4949 .$$

Then, the clustering weights and clustering coefficients of other indexes belong to grey classes could be calculated so that the matrix of clustering coefficient were obtained:

$$(\sigma_{ik})^1_{m\times p} = \begin{array}{c} \begin{matrix} \text{level 1} & \text{level 2} & \text{level 3} & \text{level 4} & \text{level 5} & \text{level 6} \end{matrix} \\ \begin{bmatrix} 0.4949 & 0.8530 & 0.7077 & 0.4057 & 0.2543 & 0.0798 \\ 0.5304 & 0.8642 & 0.6761 & 0.3640 & 0.2230 & 0.0679 \\ 0.5152 & 0.8747 & 0.6954 & 0.3617 & 0.2182 & 0.0669 \\ 0.4988 & 0.8408 & 0.6956 & 0.4024 & 0.2528 & 0.0810 \\ 0.4586 & 0.8540 & 0.7425 & 0.4029 & 0.2532 & 0.0876 \\ 0.5087 & 0.8809 & 0.7044 & 0.3544 & 0.2102 & 0.0631 \\ 0.4088 & 0.8434 & 0.7953 & 0.4504 & 0.2706 & 0.0941 \\ 0.4478 & 0.8464 & 0.7499 & 0.4177 & 0.2587 & 0.0945 \\ 0.5735 & 0.8897 & 0.6554 & 0.3366 & 0.1995 & 0.0544 \\ 0.4619 & 0.8520 & 0.7334 & 0.4066 & 0.2509 & 0.0840 \\ 0.5254 & 0.8611 & 0.6818 & 0.3831 & 0.2386 & 0.0775 \\ 0.4412 & 0.8589 & 0.7632 & 0.4242 & 0.2651 & 0.0957 \end{bmatrix} \end{array} .$$

Likewise, the matrix of clustering coefficients of clustered objects in other seasons could also be obtained (and not explained here). According to maximum membership principle, the grey class that the maximum of the clustering coefficient σ_{ik} belonged to was found out. From the research above it could be concluded that the air quality in six main administrative districts in Beijing during the four seasons from 2013 to 2014 was in grade 2 on the whole, some areas was in grade 3.

4.2 Neural networks classification and its realizing

According to BP and RBF networks theory, the monitoring data in above research were used for air quality classification evaluation. Network structure design, weight initialization, network training and output, error predication, were implemented by using MATLAB tools. When using neural network for classification evaluation, there would be great errors if the training sample was only the classification standard of air pollutant concentration due to the less data (Luo et al., 2004). In order to solve the problem of the insufficient samples, the training samples, testing samples and target output were constructed by interpolating the classification standard of air pollutant concentration by using LINSPACE function in MATLAB tools, so the massive training sample data, testing sample data and target output were constructed. For training sample, 500 groups of data were generated between the adjacent grades and 3000 training sample data were generated between six grades in total. Similarly, for testing sample, 100 groups of data were generated between the adjacent grades and 600 training sample data were generated between six grades in total. The output was one neuron and the generated data were set as 1, 2, 3, 4, 5 and 6 according to classification standard. The expectation target was the interpolated data of the adjacent grades and the interpolation proportion was corresponding to the training sample and testing sample. Hereby, the network output ranges of air quality grades were (0, 1], (1, 2], (2, 3], (3, 4], (4, 5], (5, 6], respectively. Normalization is the pretreatment method of sample data before neural network evaluation. By converting all sample to the data in the range of

[0, 1] or [-1, 1], the differences of order of magnitudes between dimensions could be eliminated so that the statistical distribution could be uniform.

The number of input layer neurons of BP network was depended on the air quality evaluation index (Wu, 2018; Xie et al., 2017). Based on the actual situation, the number of the output layer neurons was 6 and the output layer neurons was 1, and the reasonable number of hidden layer neurons was 5 by trial and error process. The training and evaluation of BP network were realized by the NEWFF function in MATLAB tools. The sample data were preprocessed by the methods of normalization and non-normalization for training respectively. The training times were 6 and 37, the variance and the mean square error were 0.1322 and 4.4061×10^{-5} by the method of normalization and 0.2701 and 9.0019×10^{-5} by the method of non-normalization. The results showed the method of normalization caused less errors and training times and rapid convergence rate of network. However, there was no significant difference on errors between these two pretreatments, this was because the standard limits of pollutants concentrations were mainly in the same order of magnitude and there was significant differences on the index of CO concentration only above grade 2.

The number of the input layer neuron and the output layer neuron were only needed in RBF network. In according with BP network, the number of the output layer neuron was 6 and the output layer neuron was 1. An approximate RBF network was designed to automatically determine the number of the hidden layer neuron by NEWRB function in MATLAB tools. By using this method, on each iteration, one neuron would be added and not stop until the sum-squared error was stepped downward to the target error or the number of the hidden layer neuron reached the maximum. In the same way, the sample data were preprocessed by the method of normalization and non-normalization for training respectively. The results showed the training times were 7 and 59, the variances and the mean square error were 0.1228 and 4.0921×10^{-5} by the method of normalization and 0.1477 and 4.9225×10^{-5} by the method of non-normalization. The actual output, the relative error and the deviation were shown in table 1. Similar to BP network, there was no significant difference on errors between the two pretreatments, only the decrease of the training times of network after normalization.

Table 1: Outputs and errors

Evaluation grade	BP network				RBF network			
	Target output	Actual output	Relative error (%)	Error direction	Target output	Actual output	Relative error (%)	Error direction
1	1	1.0153	1.5069	-	1	0.9781	2.2390	+
2	2	2.0137	0.6803	-	2	1.9852	0.7455	+
3	3	3.0080	0.2660	-	3	3.0031	0.1032	-
4	4	4.0014	0.0350	-	4	3.9378	1.5796	+
5	5	4.9964	0.0721	+	5	4.9951	0.0981	+
6	6	5.9922	0.1302	+	6	5.9046	1.6157	+

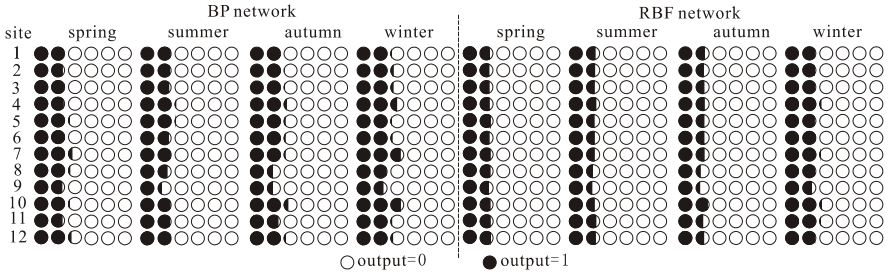

Figure 1: Sorted output of classification evaluation

The data of four seasons with normalization were classified for evaluation by the trained neural network of BP and RBF respectively (Figure 1). The results of BP network evaluation showed the air quality was between grade 2 to grade 3 in six main administrative districts in Beijing during the four seasons from 2013 to 2014. Among them, 5 evaluation sites were in grade 3 and others were in grade 2 in spring, 2 evaluation sites were in grade 3 and others were in grade 2 in summer, 8 evaluation sites were in grade 3 and others were in grade 2 in autumn, 10 evaluation sites were in grade 3 and others were in grade 2 in winter. For RBF network evaluation, only 3 evaluation sites were in grade 3 in winter, others were in grade 2. From the output results, the air quality evaluation grade of RBF network was lower than BP network, which was caused by their error direction. Generally, the positive deviation (+) was more reasonable than negative deviation (-) due to the

positive deviation was in the evaluation scope of each grade but the negative deviation was out of the scope. From the evaluation results the positive deviation of RBF was more so its evaluation results were more accurate.

5. Conclusions

By modifying the whitening function and the weight of grey clustering method, the coverage of whitening function of the exponential types was wider. Meanwhile, by considering the different of variation range of standard values in the same gray classification, the problem of zero weight was avoided. This method reflected the relationship of pollutants belong to different air quality grades. The mass training and testing sample data by interpolation of LINSPACE function in MATLAB tools were constructed for classification evaluation based on BP and RBF neural networks, the generalization of the network was improved effectively. The results showed both BP and RBF networks had good effects on air quality evaluation. The neural networks classification could research the classification of pollutants in evaluation grades and it was also an effective comprehensive evaluation method, which could also accurately and quickly decide the combined effect of pollutants on ambient air. Moreover, the evaluation result of RBF neural network classification was same as the ones of grey clustering modified (and merely a different in a few evaluation sites in winter). The RBF neural network was taking the classification of pollutants in evaluation grades as starting point, while the improved grey clustering was taking the membership of pollutants belong to evaluation grades as starting point, these above two methods were inconsistent on the evaluation results. By using the powerful functions of machine learning, the neural networks could speed up the calculation greatly. The evaluation methods researched above could also be generalized to the applications of other methods of environmental quality evaluation and would be of great practical values. On the other hand, the relationship between the evaluation results and human health were needed more research and exploration.

Acknowledgements

This research is supported by the Science Fund of Education Department of Yunnan Province (2018JS034) and the Scientific Research Staring Foundation of Kunming University of Science and Technology (KKSY201767034).

References

Chen H., Li Q., Yang Y.P., 2012, Research on the method for city air quality assessment based on fractal model, China Environmental Science, 32, 954-960, DOI: 10.3969/j.issn.1000-6923.2012.05.028

Chen S., Liu P., Li Z., 2018, Regional analysis of air quality control in china's power sector, Chemical Engineering Transactions, 70, 619-624, DOI:10.3303/CET1870104

Deng J., 1982, Introduction to grey system theory, System and Control Letter, 1, 2882-2941, DOI: 10.1007/978-3-642-16158-2_1

Han B., Wang B. D., 2006, Comments on two methods for comprehensive assessment of ecological environment quality in estuarine and coastal waters, Advances in Marine Science, 24, 254-258, DOI: 10.1016/S1872-2040(06)60041-8

Liu J., Yang P., Lv W. S., 2015, Environmental air quality evaluation method based on the six pollutants in the urban areas of Beijing, Journal of Safety and Environment, 15, 310-315.

Luo D. G., Wang X. J., Guo Q., 2004, The application of ANN realized by MATLAB to underground water quality assessment, Acta Scientiarum Naturalium Universitatis Pekinensis, 40, 296-302, DOI: 10.1111/j.1744-7909.2005.00184.x

Rava E. M. E., Chirwa Evans M. N., 2017, Prediction of performance of the moving-bed biofilm pilot reactor using back-propagation artificial neural network (BP-ANN), Chemical Engineering Transactions, 61, 1189–1194, DOI: 10.3303/CET1761196

Reddy M. K., Rama Rao K. G., Rammohan Rao I., 2004, Air quality status of Visakhapatnam (India)-indices basis, Environmental Monitoring and Assessment, 95, 1-12.

Thomas P., Derigent W., Suhner M.C., 2016, A classifier ensemble for classification of dynamic data. Application to an indoor air quality problem, Journal Europeen des Systemes Automatises, 49(3), 375-391, DOI: 10.3166/JESA.49.375-391

Xie Y., Wang W. J., Li B. C., Zhao Z. W., He L., Wang Y. X., 2017, Analysis of SO_2 pollution in Baoding based on MATLAB grey model, Chemical Engineering Transactions, 59, 901-906, DOI: 10.3303/CET1759151

Xu W. G., Zhang Q. Y., Guo H., 2006, Application of grey clustering modified model in system total amount control, China Environmental Science, 26, 546-549, DOI: 10.1016/S0379-4172(06)60102-9