Research on Air Quality of Beijing-Tianjin-Hebei Region based on SVM and Regression Analysis

Li Tang
Department of Information Science and Technology
Tianjin University of Finance and Economics
Tianjin, China
tangli0831@tjufe.edu.cn

Caiyun Zhou
Department of Economics
Tianjin University of Finance and Economics
Tianjin, China
zhoucaiyun_0119@163.com

Li He
Department of Information Science and Technology
Tianjin University of Finance and Economics
Tianjin, China
renkeheli@163.com

Shuhua Zhang
Coordinated Innovation Center for Computable Modeling in Management Science
Tianjin University of Finance and Economics
Tianjin, China
shuhua55@126.com

***Abstract*—The air pollution of Beijing-Tianjin-Hebei region in China becomes increasingly serious. Focus on the air quality management of Beijing-Tianjin-Hebei region, this paper proposes a method based on SVM algorithm and multivariate linear regression to the prediction and evaluation of air quality. First, it builds the evaluation metrics based on the weather factors, the correlation of neighbor cities, and their combination. Furthermore, it applies the SVM algorithm to the classification and prediction of air quality of Beijing-Tianjin-Hebei region. By the application of multivariate linear regression method, the weather factor with no significant effect on air quality is removed to save the cost of calculation. Finally, the experimental results show that the method based on the combination of weather factors and correlation of neighbor cities is better than the other two methods. We draw the conclusion that the method is feasible and effective.**

Keywords—support vector machine (SVM); data mining; multivariate linear regression; air quality index (AQI)

I. INTRODUCTION

The environmental pollution of Beijing-Tianjin-Hebei region in China is becoming increasingly serious, especially the air quality. Environmental protection and management is related to the national economy and the livelihood of people. It is very important to strengthen the environmental management of cities and improve the air quality in the Beijing-Tianjin-Hebei region. It is also urgent to develop the research on the management of air quality of Beijing-Tianjin-Hebei region. The management of air quality of cities usually denotes the classification, prediction and evaluation of air quality.

With the development of digital city management and intelligent technology, there are a variety of big data related to ecological environment of the cities, such as meteorological data, environmental data, traffic flow and industrial data, etc. There is huge economic value and social benefit in the big data. The related technologies of big data have shown the broad prospects for development and application in the field of environmental management and protection [1].

Machine learning algorithm, as an effective analysis method for big data, can dig the intrinsic value with high efficiency. Therefore, under the environment of big data, we apply the machine learning algorithm to the research of city environmental management. There is an important research significance and great research value for the development of society. In this paper, we collect the meteorological data and air quality data of Beijing-Tianjin-Hebei region, and apply the machine learning algorithm and regression analysis method to the air quality classification, prediction and evaluation. Meanwhile, the multivariate linear regression method is adopted to find the weather indicator without significant effect on the air quality.

II. RELATED WORKS

There are some researches related on the urban environment in the recent years. Some related works focus on the air quality research of the cities.

Firstly, the researches analyze the influence of some factors on air quality, including the influence of emission reduction of APEC region on PM2.5 [2], the relationship analysis between economic development and the change of PM10 concentration [3], the relationship between the industrial agglomeration and the environment [4], the influence of meteorological elements on the air quality index [5], the quantitative analysis of automobile exhaust [6], and so on.

Secondly, the researchers introduced a variety of data mining algorithms to the analysis and prediction of air quality. Song Y. C. used the BP neural network algorithm and time series model to the air quality prediction [7]. Huang S. applied the multivariate linear regression method to the improvement of PM10 forecast [8]. Yu Z. introduced the linear regression

This research is supported by the Tianjin Social Science Foundation of China (TJYY15-017).

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

model and neural network spatial prediction model to predict PM2.5 of four cities in China [9]. According to the historical and real-time data (weather, traffic and mobility etc.), Yu Z. proposed a semi-supervised learning approach based on artificial neural network (ANN) and linear chain conditional random field(CRF) to deduce the air quality of Beijing and Shanghai [10].

Finally, there are some research methods on air quality evaluation, including the fuzzy-grey clustering method [11], fuzzy comprehensive evaluation method [12], principal component analysis [13], fuzzy analytic hierarchy process (AHP) method [14], etc.

Although some environmental research based on big data has carried out, it is necessary to develop the comprehensive data analysis and data mining for the air quality of Beijing-Tianjin-Hebei region.

III. Air Quality Analysis of Beijing-Tianjin-Hebei region based on SVM and multivariate linear Regression

A. Research Method

Big data mining and analysis of urban environment is an important issue for the management of society. The air quality of city is affected by a variety of external factors, including meteorological conditions, the air quality impact of neighbor cities, the impact of industry and economy, etc. In view of air quality of city, this paper applies the machine learning algorithm to the air quality classification and prediction of Beijing-Tianjin-Hebei region from three aspects, including the research based on weather factors; the research based on the air quality correlation of neighbor cities; the research based on the combination of weather factors and air quality correlation of neighbor cities. The main methods of this paper include the following two aspects.

Firstly, the data mining of air quality is based on Support Vector Machine (SVM) algorithm. We obtain the air quality index (AQI) data and meteorological data of Beijing-Tianjin-Hebei region, and preprocess the data. According to the air quality level, the class labels are established. If the air quality index (AQI) value of city is greater than 100, we defined the air quality level of city as 1. Otherwise, if the AQI value is small than 100, we defined it as 0. According to the various weather factors, we realize the feature selection and extraction, and construct the index system of air quality of Beijing-Tianjin-Hebei region. Then, we adopt the SVM algorithm for the data mining of air quality. The SVM algorithm is applied to the training sets to establish the prediction model. Finally, according to the prediction model and the given test sets, we run the SVM algorithm to realize the classification and prediction of air quality of Beijing-Tianjin-Hebei region. Then we calculate the prediction results of air quality level and performance parameters, including the accuracy (ACC), specificity (SPC) and sensitivity (SEN).

Secondly, the significant correlation analysis of air quality is based on the multivariate linear regression. At first, we analyze the influence of various weather factors on the air quality, and establish correlation model between the air quality and weather factors of the cities based on the multivariate linear regression method. After the modeling and calculation, we analyze the different impacts of weather indicators on air quality, and remove the weather factor which has not significant impact on air quality.

B. Evaluation Metrics

The evaluation metrics are built based on three kinds of characteristics, including the weather factors; the air quality correlation of neighbor cities; and their combinations.

Firstly, in the air quality research of Beijing-Tianjin-Hebei region based on weather factors, we propose six evaluation metrics, including average air pressure (0.1hPa), the average temperature (0.1℃), the average relative humidity (1%), average vapor pressure (0.1hPa), average velocity (0.1m/s) and the wind direction with the maximal wind speed. These six metrics are numerical data. The wind direction is coded as the value of 1-17. The weather data collected in this paper are from the data sets of the China International Exchange station provided by China Meteorological Data Network.

Secondly, in the air quality research based on the correlation of neighbor city, we collect the air quality index (AQI) data of four cities, including Tianjin, Beijing, Shijiazhuang and Chengde. The air quality data of four cities are from the on-line monitoring analysis Platform of China Air Quality. Here, we regard one of four cities as the forecasting object and let *AQI0* represent its air quality. Then *AQI1*, *AQI2* and *AQI3* denote the air quality index of the other three cities in the Beijing-Tianjin-Hebei region, and they are the evaluation metrics of air quality as the neighbor cities.

Thirdly, in the air quality research based on the combination of weather factors and the correlation of neighbor city, we combine the above indicators and define 9 evaluation metrics, namely, 6 weather indicators and 3 AQI data of neighbor cities. According to the 9 evaluation metrics, we can predict the AQI level of a city.

C. Algorithm Realization

The SVM algorithm proposed by Vapnik [15] is a supervised machine learning method which is widely used in statistical classification. The SVM algorithm can find the global optimal solution without the curse of dimensionality. In this paper, SVM algorithm is used to classify and predict the air quality of Beijing-Tianjin-Hebei region. The SVM algorithm used in this paper is derived from the LIBSVM software package developed by Lin [16]. In the SVM algorithm, there are two steps, including training process and prediction process. In the training step of SVM, we adopt the default kernel function. As for the other parameters, we also use the default values.

In addition, regression analysis is the most important statistical method in mathematical statistics. Through the regression model, we can analyze the relationship between two or more variables. The regression model includes linear regression model and nonlinear regression model. In this paper, the multivariate linear regression analysis method is used to analyze the relationship between the weather factors and air quality.

Therefore, this paper proposes a method for the air quality prediction and assessment of Beijing-Tianjin-Hebei region based on SVM and multivariate linear regression analysis. In this paper, the AQI prediction and evaluation for Beijing-Tianjin-Hebei region is achieved by SVM algorithm. Furthermore, the multivariate regression analysis is performed to find the weather factor whose effect on the AQI is not significant. Then, that weather factor will be removed, and the new data sets are built. The AQI level will be predicted by SVM based on the new data. According to the results of prediction, we will evaluate the air quality. The detailed algorithm is as follows.

Step 1. According to the meteorological data set of Beijing-Tianjin-Hebei region, the LIBSVM algorithm is used to train the data sets and get the prediction model. Based on the model and the meteorological data, the AQI levels of Beijing, Tianjin and Hebei province are calculated, and the performance parameters ACC, SPC and SEN are obtained;

Step2. Air quality data based on the correlation of neighbor city in Beijing-Tianjin-Hebei region are built. The data sets are divided into the training data and test data. The LIBSVM algorithm runs the training data and obtains the prediction model of air quality. Then according to the model and the test data of neighbor cities, we predict the AQI of Beijing-Tianjin-Hebei region, and calculate the performance parameters of ACC, SPC and SEN;

Step3. The total data sets of Beijing-Tianjin-Hebei region combine the meteorological data sets and air quality data of the neighboring cities. Based on the data of their combination, we use LIBSVM algorithm to obtain the prediction model. Then, according to the model and the total data set, we predict the AQI level of Beijing-Tianjin-Hebei region and calculate the performance parameters of ACC, SPC and SEN;

Step4. Multiple regression analysis is used to remove the indicator without the significant impact on the AQI of Beijing-Tianjin-Hebei region. After removing the indicator, the new data sets are formed. According to the new data set, the SVM algorithm is applied to the prediction of AQI level of the Beijing-Tianjin-Hebei region.

Step 5. We compare and analyze the experimental results and draw the conclusions.

IV. Experimental Analysis

A. Data sets

The weather daily data and AQI daily data of four cities, namely, Tianjin, Beijing, Shijiazhuang and Chengde, are from January 1, 2014 to May 31, 2016. The experiment is conducted by open test. The experimental data sets of 2014 and 2015 are treated as the training data sets to build the prediction model. The data sets of 2016 are used as test data, and the AQI level of 2016 will be predicted.

B. Experimental results and analysis

We carry out the experiment from three aspects, which are based on the weather factors, based on the correlation of neighbor cities, and based on the combination of weather factors and correlation of neighbor cities. The experimental results are listed in Table 1.

TABLE I. Experimental Results by SVM

Performance parameters	**Air quality research of Beijing-Tianjin-Hebei region**		
	Weather factors	***Correlation of neighbor city***	***Combination of weather factors and correlation of neighbor city***
Accuracy	68.0921%	77.9605%	82.5658%
Specificity	84.8039%	77.6961%	91.4216%
Sensitivity	34%	78.5%	64.5%

From the above experimental results, we draw the conclusion that the method based on the combination of weather factors and correlation of neighbor cities is better than the other two methods. The results show that the combination method is feasible and effective.

By multivariate regression analysis, we find that the vapor pressure has not the significant relationship with the AQI, and remove it. Based on the new data sets, we apply the SVM algorithm to the prediction and evaluation of AQI of Beijing-Tianjin-Hebei region. The experimental results are listed in Table 2.

TABLE II. Experimental Results by Regression

Performance parameters	**Air quality research of Beijing-Tianjin-Hebei region**	
	Weather factors and correlation of neighbor city	***Weather factors and correlation of neighbor city without vapor pressure***
Accuracy	82.5658%	81.7434%
Specificity	91.4216%	90.9314%
Sensitivity	64.5%	63%

By the multivariate regression analysis, the vapor pressure has not the significant effect on AQI. After removing it, there is little difference between the new results and the original results. It verifies that the indicator of vapor pressure can be eliminated to save the calculation cost.

In conclusion, it is feasible and effective to predict and evaluate the air quality of Beijing-Tianjin-Hebei region based on SVM algorithm and multivariate linear regression analysis.

V. Conclusion

Focus on the air quality management and prediction of Beijing-Tianjin-Hebei region, this paper proposes a new method based on SVM algorithm and multivariate linear regression to predict the air quality. It adopts the SVM algorithm to the classification and prediction of air quality based on the three kinds of evaluation metrics, including the weather factors, the correlation of neighbor cities, and their combination. According to the multivariate regression method, the weather factor with no significant effect on AQI is removed. Finally, the results show that the method based on the combination of weather factors and correlation of neighbor cities is better than the other two methods. We draw the conclusion that this method is feasible and effective.

ACKNOWLEDGMENT

The research is supported by the Tianjin Social Science Foundation of China (TJYY15-017).

REFERENCES

[1] S.Q. Lu, G. Xie, Z. Chen, et al. “The Management of Application of Big Data in Internet of Thing in Environmental Protection in China,” IEEE First International Conference on Big Data Computing Service and Applications. IEEE Computer Society, pp. 218-222, 2015.

[2] J.L. Li, L.X. Wu, C. B. Ren, et al. “Impact of APEC Regional Emission Reduction on Spatio-Temporal Variation of PM2.5 Concentration in Beijing,” Geography and Geo-Information Science, vol. 32, no. 3, pp. 110-115, 2016. (In Chinese)

[3] X.F. Kang, G. Wang, M.H. Zhang, et al. "The Grey Relational Analysis about Economic Development and Change of PM10's Concentration in Beijing-Tianjin-Hebei Region," Environmental Monitoring in China, vol. 5, pp. 1-6, 2015. (In Chinese)

[4] X.Y. Qian and Y.Wang,"The Coupling relationship between industrial agglomeration and ecological environment in Beijing-Tianjin-Hebei region," Statistics and decision, no. 3, pp. 103-106, 2016. (In Chinese)

[5] H.M. Bai, H.D. Shi, Q.X. Gao, et al. “Re-Ordination of Air Pollution Indices of Some Typical Cities in Beijing-Tianjin-Hebei Region Based on Meteorological Adjustment,” Journal of Ecology and Rural Environment, vol. 31, no. 1, pp. 44-49, 2015. (In Chinese)

[6] X.H. Huang, G.C. Wan and C. Chen, “A Study of Air Quality Classification and Evolution of Automobile Exhaust Diffusion in Beijing, Tianjin and Hebei,” Journal of Jiamusi University (Natural Science Edition), vol. 34, no. 4, pp. 617-620, 2016. (In Chinese)

[7] Y.C. Song and S. Zhen, “The application of BP neural network and time series model to air quality forecast for Baotou,” Journal of Arid Land Resources and Environment, vol. 27, no. 7, pp. 65-70, 2013. (In Chinese)

[8] S. Huang, X. Tang, W.S. Xu, et al. “Application of ensemble forecast and linear regression method in improving PM10 forecast over Beijing areas,” Acta Scientiae Circumstantiae, vol. 35, no. 1, pp. 56-64, 2015. (In Chinese)

[9] Y. Zheng, X. Yi, M. Li, et al. “Forecasting Fine-Grained Air Quality Based on Big Data,” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 2267-2276, 2015.

[10] Y. Zheng, F. Liu and H.P. Hsieh. “U-Air:when urban air quality inference meets big data,” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1436-1444, 2013.

[11] H. Ding, Y. H. Liu and S. X. Cao. “Research on assessment of urban air quality based on fuzzy-grey clustering method,” Environmental Science & Technology, vol. 36, no. 12, pp. 374-379, 2013. (In Chinese)

[12] L. Y. Lv and H. Y. Li, “Air Quality Evaluation of Beijing-Tianjin-Hebei Region of China Based on the Fuzzy Comprehensive Evaluation Method,” Acta Scientiarum Naturalium Universitatis Nankaien, no.1, pp. 62-68, 2016. (In Chinese)

[13] D. M. Huang, X. Q. Chen and T. Xiao, “The Comprehensive Evaluation of the Air Quality for Baoding City Based on Principal Component Analysis and Fuzzy Comprehensive Evaluation,” Journal of Baoding University, no. 2, pp. 119-126, 2015. (In Chinese)

[14] S. J. Huang, K. Q. Li and G. Y. Zhu, “Evaluation and analysis on the indoor air quality of colleges based on fuzzy analytic hierarchy process,” Environmental Engineering, vol. 32, no. 5, pp. 90-94, 2014. (In Chinese)

[15] Vapnik. The nature of statistical learning theory. Berlin:Springer, 2000.

[16] C.C. Chang, C. J. Lin, “LIBSVM: a library for support vector machines,” http://www.csie.ntu.edu.tw/~cjlin / libsvm/.