Application of big data analysis technology in environmental monitoring

Yanyan Chen, Fanglei Ren and Shuhuan Wang, Linxuan Zhang*
Beijing Jiaotong University, Beijing 100044, China
*Corresponding author email: linxuanzhang@bjtu.edu.cn

Abstract. With the development of China’s social economy, industrialization and urbanization, the problem of environmental pollution is becoming more and more serious, and people pay more and more attention to the problem of environmental pollution. The traditional environmental monitoring has no way to meet the needs of the government and people, so the combination of big data and environmental monitoring is the need of the times and the inevitable trend of the development of environmental monitoring. This paper will focus on the application of big data analysis technology in atmospheric detection, analyze the significance of big data analysis technology in environmental monitoring, and promote the wide use of big data analysis technology in environmental monitoring.

Keywords: Big data, environmental monitoring, spatial classifier, time classifier.

1. Introduction
Environmental detection and protection work has been one of the key issues of government and people’s attention, so the requirements for environmental monitoring and the number of environmental monitoring are increasing day by day. Although the construction of automatic environmental monitoring station can solve some of the problems existing in the current environmental monitoring, due to the high cost of building automatic monitoring station and the high maintenance cost, many areas cannot afford the cost. For the above problems, use big data analysis technology to explore the monitoring data of local areas, so as to reflect the environmental situation and provide the government and people with the information needed.

2. Introduction to big data analysis technology

2.1. Contents of big data analysis technology
The content of big data analysis technology includes: data acquisition, parameter selection, parameter identification, system simulation, system intelligent self-study and neural network analysis. By analyzing a large number of data, the relationship logic between the data can be found, so as to provide useful information for solving environmental problems.

2.2. Process of big data analysis technology
(1) Determine the object and content of the research question
(2) Select data type and characteristic quantity
3. Application of big data analysis technology in atmospheric environment monitoring

3.1. Determine the object and content of the research question
In this study, the local area of a city was taken as the research object, and the average concentration of PM$_{2.5}$ was taken as the research content. First of all, the research object is divided into cell grids with the same area size (the size of the cell grid can be determined by the size of the research object), and then $g_1, g_2, g_3, ..., g_n$ are used to represent the above divided cells, then the research object can be expressed as $G(g_1, g_2, g_3, ..., g_n)$; the average concentration of PM$_{2.5}$ in corresponding cells can be expressed by $C_{g_1}, C_{g_2}, C_{g_3}, ..., C_{g_n}$, which means that the research content can be expressed as $C(g_1, g_2, g_3, ..., g_n)$. In the study, $J(C_i)$ is taken as the objective function of big data analysis. The purpose of this study is to use the regional data of known environmental conditions and use big data analysis technology to speculate the environmental conditions of areas without automatic monitoring stations. To sum up, $C_i$ should include $C_1$ and $C_2$, where $C_1$ refers to the data in the area with automatic monitoring station, and $C_2$ refers to the data to be calculated in the area without automatic monitoring station, i.e. the target value.

3.2. Selection type and characteristic quantity
When selecting data related to $C_2$, there are two principles for selection: possibility (only when data is available) and need (analysis of all kinds of data). The relationship between the data is simple and complex, with strong correlation and weak correlation. Most of the data in the field of environmental monitoring present a complex relationship of non-linear and diversity. In order to obtain the target value accurately, it is necessary to select a large number of data with certain correlation with $J(C_i)$. According to the relevant technical conditions and the basic principle of air pollution, the following seven types of data can be selected, including historical PM$_{2.5}$ average concentration data of detection device, traffic condition related data, human activity status data, road condition related data, meteorological condition related data, grid coordinates, pollution source related data, etc. The characteristic quantities corresponding to the seven types of data are: historical average concentration of PM$_{2.5}$ per hour, number of vehicles, vehicle speed, number of vehicles at different speeds, number of people moving in and out of the grid, length and width of the road, temperature, air pressure, wind speed, humidity, horizontal and vertical coordinates of the grid, number of pollution sources, pollution intensity, etc.

3.3. Classify the data and establish spatial and temporal data sets
Because the characteristic variables selected in this study have two attributes: they can change with the change of time and can change with the change of space. Therefore, the data can be divided into two data sets: temporal data set and spatial data set. Among them, road condition related data, grid coordinates, pollution source related data should be classified as spatial data set; historical PM$_{2.5}$ average concentration data with detection device, traffic condition related data, human activity data and meteorological condition related data should be classified as time data set.

3.4. Training spatial classifier and temporal classifier respectively
When analyzing the influence of spatial and temporal data sets on the objective function, the spatial classifier and the temporal classifier should be constructed respectively to calculate the objective function and get the expected results.
(1) Spatial classifier

The data contained in spatial classifier is static data, and the influence of the feature quantity on the objective function is multi-layer and multi-node, as well as its transfer path is linear, but the output at the node may be linear or nonlinear. Therefore, the spatial classifier conforms to the structure of static neural network, that is, the spatial classifier composed of artificial neural network part and input part can be used to predict the hourly average concentration of PM2.5. The function of input generation is to meet the requirements of artificial neural network part input self-value construction. In the grid with automatic monitoring station, two grid coordinates are randomly selected as \( a_1 \) and \( a_2 \), in which the characteristic quantity and hourly average concentration of PM2.5 are \( X_1, X_2, X_3, X_4, C_1 \) and \( C_2 \), respectively. The coordinates of the area to be predicted are \( L_i \), in which the characteristic quantity and hourly average PM2.5 concentration to be predicted are \( A_1, A_2 \) and \( C_i \), respectively. This paper analyzes the data generation process of the input part of the spatial classifier, and defines two algorithms. \( D \) is the algorithm of geometric distance, and \( P \) is the algorithm of Pearson distance. The specific calculation formulas are as follows:

\[
D_{si} = \arctan(a_x, a_y)
\]

\[
P_{fi} = \text{Pearson} - \text{Cor}(X_i, X_j)
\]

\[
P_{li} = \text{Pearson} - \text{Cor}(X_j, X_i)
\]

Respectively. After obtaining the partial data of artificial neural network, it will finally affect the target value through the neural network nodes and transmission.

By analyzing a neural network node, the following relationship can be obtained

\[
\varphi(W_{li} \cdot X_i + W_{2i} \cdot X_2) + b_i
\]

Where: \( \varphi \) — Transformation function of nonlinear output (sigmoid function is the most commonly used one)

\( b_i \) — The output input offset of the node

\( W_{li}, W_{2i} \) — The input weight

When the node output is linear output, \( \varphi = 1 \). From the above formula, it can be concluded that in order to make the spatial classifier constructed by neural network meet the requirements of air pollutant concentration prediction, in the whole neural network, the weight of each sensing layer and node, the output and input offset of node and the transformation of function need to be solved. In order to solve the above problems, the inversion method can be used. The specific idea is that, contrary to the actual neural network experience, the residual error is allocated to each weight according to the size of the input value from left to right, and the effect of \( \varphi \) is to add a proportional coefficient to the weight.

When applied to atmospheric monitoring, the data of grid with automatic monitoring device can be selected. By repeatedly training the spatial classifier, the residual value will continue to decrease, that is, the spatial classifier is constantly improved.

(2) Time classifier

The characteristic quantity of time data set can be represented by \( X_j \), where \( i \) is grid point and \( j \) is characteristic quantity, then \( X = \{x_1, x_2, ..., x_n\}, x_i = \{x_1, x_2, ..., x_j, t\} \), \( t \) represents a certain time point. The concentration of air pollutants can be expressed by \( y_i \), then \( Y = \{y_1, y_2, ..., y_n\} \). When \( X \) is easy to determine, \( y_i \) is only related to \( y_{i+1} \) and has no relationship with other local atmospheric pollutant concentrations. On this basis, if \( X \) is clear, the probability of \( Y \) occurrence is a normal distribution function, which is determined by the state characteristic function \( \exp(u \cdot s(y_i, x_i)) \) and the transition characteristic function \( \exp(\lambda \cdot t(y_{i+1}, y_i, x_i)) \), where \( u \) and \( \lambda \) are two undetermined parameters ( \( \lambda \) is weight parameters). The process of time classifier construction is actually the analysis process of conditional random field. In practical application, conditional probability function can be transformed.
into linear function. Like spatial classifier, time classifiers will be improved in the process of continuous training.

3.5. Comprehensive calculation results and output
Through the analysis and training of the above-mentioned large data sets, the optimal time classifier and spatial classifier are obtained. The integrated time classifier and spatial classifier can calculate the concentration of air pollutants in the grid without automatic monitoring device.

4. Conclusion
At present, Chinese environmental problems are complex, pluralistic and nonlinear. For example, there are a large number of interactive factors among the atmosphere, water and soil, and there are also a lot of interaction between different regions. Therefore, in the case of interaction of various data, big data analysis is a very useful tool, which can help to study some climate and environmental issues.

4.1. Improve the ability of environmental early warning
With big data technology, all kinds of environmental monitoring data can be integrated and analyzed, which is conducive to improving the accuracy of data. It can also predict the future environmental changes scientifically, which is conducive to improving the scientificity and effectiveness of environmental monitoring. What is more, environmental monitoring combined with big data technology can also achieve real-time monitoring, information and data sharing and transmission.

4.2. Provide scientific basis for decision-making
In the process of big data technology application, the monitoring results can be visually presented with the help of big data model, so as to improve the effectiveness of data. And combine the environmental governance scheme and the model to analyze, so as to select the best governance scheme, which can effectively avoid the mistakes of decision-making and reduce the cost of governance. Data sharing can encourage people to participate in the work of environmental protection and contribute their own strength to environmental protection.

References
[1] Wang Qiong, Huang Yanmin, Yi Min, Yao Shenjun, Wu Jianping, Chen Jie. Performance evaluation of atmospheric environmental monitoring data sharing website from the perspective of users [J]. Environment and sustainable development, 2019, 44 (06): 140 - 144.
[2] Wang Yujiao. Discussion on environmental monitoring and control measures of air pollution [J]. Environment and development, 2019, 31 (10): 132 + 134
[3] Zhao Yinlong. Development of PM$_{2.5}$ real time online concentration and exposure system and its application in atmospheric environmental health research [C] 2019 (06): 109 – 112.
[4] Wang yuesi, Gong Zhengyu, Liu Zirui. Construction and application of comprehensive stereo observation network for air pollution in Beijing Tianjin Hebei region and its surrounding areas [J]. Environmental science research, 2019, 32 (10): 1651 - 1663.
[5] Yuan Xue, Jiang Jie. Application of big data analysis technology in atmospheric environmental monitoring [J]. Low carbon world, 2017, (28).
[6] Liu Wenqing, Chen Zhenyi, Liu Jianguo, et al. Stereo monitoring technology and application of atmospheric environment in China [J]. Science Bulletin, 2016, 61 (30).
[7] Li Wei, Hu Hao, Xu Fuchun, et al. Application of big data analysis technology in atmospheric environment monitoring [J]. China environmental monitoring, 2015, 31 (3).
[8] Lu Ju, Li Ling, Zhou Ruiming, et al. Research progress of atmospheric PM$_{2.5}$ [J]. China Construction, 2015, (3).