International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 3, March-April 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Predicting Beijing Air Quality Data Based on LSTM Method

Zeng Guojing, Jin Renhao

School of Information, Beijing Wuzi University, Beijing, China

ABSTRACT

This paper studies the air quality data of Beijing from 2018 to 2020. On the basis of the correlation analysis of pollutant concentration, the circular neural network model based on LSTM algorithm is built to realize the prediction of AQI of Beijing. The results show that AQI index has a high correlation with PM2.5 and PM10, but only has a low negative correlation with O3. The prediction model of recurrent neural network shows high prediction accuracy. The research in this paper is helpful to promote the application of recurrent neural network model in air quality data and time series data.

KEYWORDS: *AQI; LSTM; Python; Keras; Pearson correlation*

How to cite this paper: Zeng Guojing | Jin Renhao "Predicting Beijing Air Quality Data Based on LSTM Method" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-3, April 2021, pp.774-777, URL: www.ijtsrd.com/papers/ijtsrd40000.pdf

IJTSRD40000

1. Research background

With the continuous development of economy and urban scale, Chinese development has entered a new era, and the people put forward higher requirements for urban air quality. As the sandstorm in March 2021, the air quality problem has once again become the focus of Beijing citizens. The monitoring and prediction of air quality is great practical significance in order to improve the air quality and the level of urban environmental construction.

In order to better monitor and predict air quality, the national environmental protection department began to use air quality index (AQI) to quantitatively describe air quality from 2012. AQI[1] is a kind of conceptual index which simplifies the concentration of several air pollutants in conventional monitoring into a single form, and represents the degree of air pollution and air quality status by classification. It is suitable for representing the short-term air quality status and change trend of cities. With the development of data mining, more and more machine learning models are applied to the prediction of air quality. Bai Heming[2] used BP neural network to forecast the AQI index for different seasons in Beijing. By comparing the forecast value and monitoring value of different seasons, they concluded that the forecast accuracy of autumn is the highest. Li Jinglu and Zeng Tian[3] used the principal component analysis method to study the air quality data of Beijing from 2000 to 2011, and concluded that the per capita GDP and the output value of the tertiary industry had the greatest correlation with air quality. Wang Mingjie and He Jiajia[4] used the method of mathematical statistics and typical circulation classification to study the AQI index. The results showed that the main pollutants causing weather pollution were NO_2、$PM_{2.5}$ and O_3.Li Ping and Ni Zhiwei[5] built a fractal popular learning support vector machine to predict AQI index. They adopt the method of fractal dimension first and then reduce the dimension, which improved the accuracy and stability of prediction. Xu Qi and Wu Qizhong[6] used the comprehensive scoring method to monitor and forecast the $PM_{2.5}$ concentration in the air. Based on the WRF-CMAQ model system, their evaluation results showed that the accuracy was better than the official forecast.

However, the air pollution index is a typical time series data. When using the traditional statistical model and the common neural network method to predict, the accuracy is not high enough and the calculation time is long. Recurrent neural network is a kind of neural network model with the input of time series data, which is more suitable for the modeling and prediction of time series data. LSTM solves the common problems of gradient disappearance and gradient explosion in traditional recurrent neural network. It is a common recurrent neural network algorithm and has many successful applications[7]-[10] in predicting time series data. But at present, the research on the application of recurrent neural network model based on LSTM algorithm in air quality prediction is still lacking, especially in Beijing data. Therefore, this paper uses Python deep learning library keras to build LSTM recurrent neural network model to realize the prediction of Beijing air quality data, and selects AQI as the main index of air quality as the prediction target variable.

2. Theoretical basis

2.1. Keras

Keras is a powerful high-level neural network API written for python. It can use tensor flow, theano and cntk as the interfaces of high-level applications. Keras is one of the commonly used machine learning tools, which has four advantages: user-friendly, modular operation, strong scalability, and high collaboration with Python. It contains a large number of functions and program optimizers and other components. The optimizer included in Keras can realize back propagation algorithm and adaptive gradient descent algorithm, which is convenient for the implementation of LSTM recurrent neural network algorithm.

2.2. Principle of LSTM neural network

Long term and short-term memory network (LSTM) is a variant algorithm of recurrent neural network (RNN). By using time back propagation training, it can solve the problems of gradient disappearance and gradient explosion in common neural network method. It is widely used in image video recognition, stock price trend prediction, disease prediction and other fields. LSTM algorithm uses memory cells to replace conventional neurons in RNN. Memory cells are more flexible components than neurons, and memory modules are introduced. Each storage unit is composed of forgetting gate, input gate and output gate, and its structure is shown in Figure 1.In Fig 1:t represents the specific time, x_t、$x_{t-1}$and x_{t+1} represent the input sequence at t time, $t-1$ time and $t+1$ time respectively;h_t、h_{t-1} and h_{t+1} represent the outputs of the memory cells at t time、$t-1$ time and $t+1$time respectively. The $tanh$ is the hyperbolic tangent function and σ is the sigmoid activation function. This function can transform to produce a smooth range value between 0 and 1, so as to observe the change of output value when the input value changes slightly.

3. Construction of LSTM prediction model

3.1. data sources

This paper is based on the air quality data of Beijing from January 2018 to December 2020, and the data is from the website of China Weather Post (http://www.tianqihoubao.com/).A total of 1096 rows of observations were obtained. Data information includes daily AQI index and concentrations of six pollutants CO、NO_2、$PM_{2.5}$、SO_2、O_3、PM_{10}in Beijing.Due to the long sampling time and force majeure and other factors, some date data are missing. This paper uses the monthly mean of these seven kinds of data to borrow and supplement the missing values. The trend of AQI index and six kinds of pollutant values is shown in Figure 2.

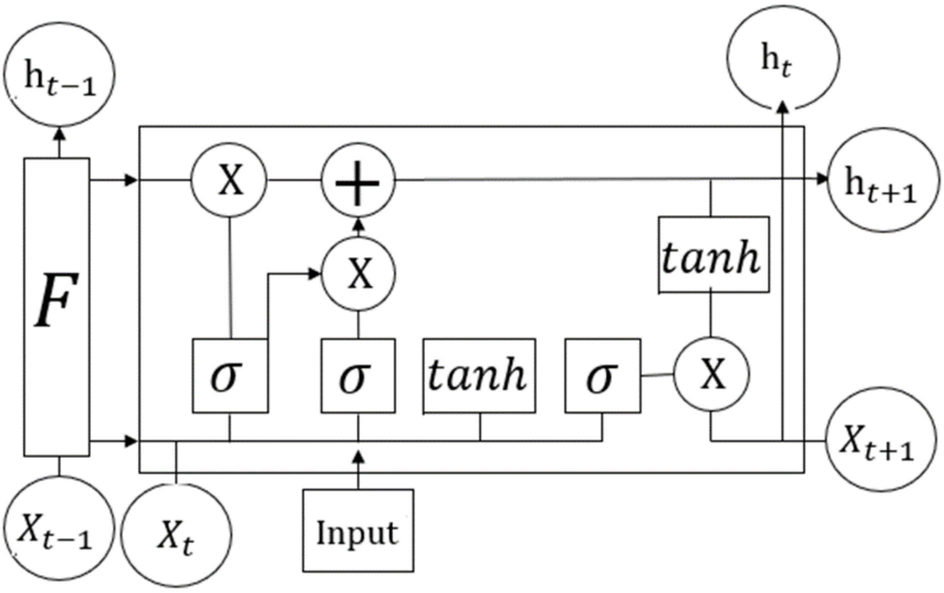

Fig 1 The structure of LSTM

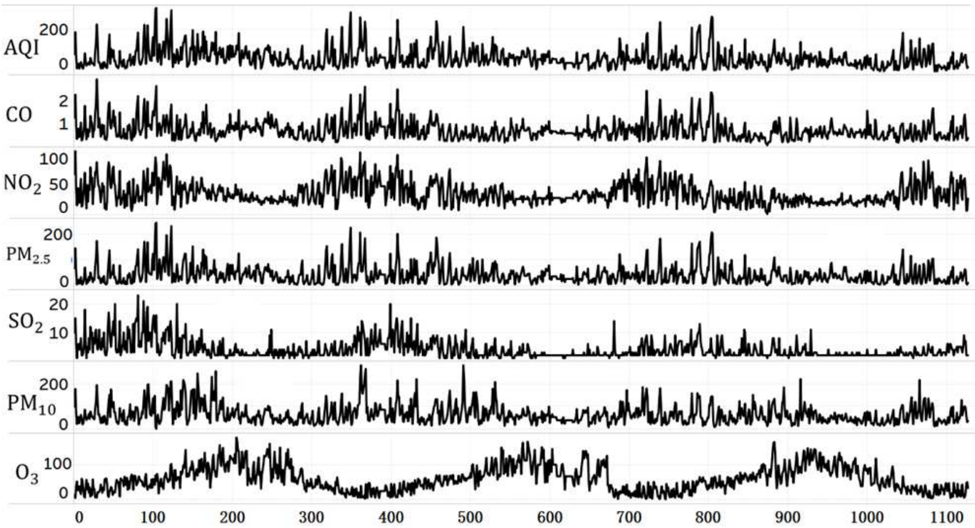

Fig 2 Variation trend of AQI index and six pollutants

3.2. Correlation analysis between AQI index and pollutants

It can be seen from Fig 2 that the change trend of AQI and the concentrations of CO、NO_2、$PM_{2.5}$、SO_2andPM_{10}in Beijing is roughly the same, When the AQI index becomes higher, the other five pollutants will also become higher. When the AQI index becomes lower, the other five pollutants will also become lower. Therefore, there is a positive correlation between AQI index and the concentrations of CO、NO_2、$PM_{2.5}$、SO_2andPM_{10}.However, when the AQI index becomes higher, the concentration of O_3 becomes lower, so there is a negative correlation between AQI index and O_3concentration.In order to further analyze the relationship between AQI and CO、NO_2、$PM_{2.5}$、SO_2、O_3、PM_{10}, the Pearson correlation coefficient of each index is shown in Table 1.There was a positive correlation between AQI index and the concentrations of CO、NO_2、$PM_{2.5}$、SO_2andPM_{10}, and a weak negative correlation between AQI index and O_3concentration, with the coefficient value of - 0.08.PM2.5 and PM10 had the highest positive correlation with AQI, and the correlation coefficients respective were 0.936 and 0.785.Therefore, in the study of air pollution control in Beijing, we can formulate relevant policies from the perspective of controlling the emission of these two pollutants, and take certain measures to reduce the concentration of these two pollutants.

Table1. Correlation coefficient matrix of AQI index and six pollutants in Beijing

	AQI	$PM_{2.5}$	SO_2	NO_2	PM_{10}	O_3	CO
AQI	1	0.936	0.438	0.580	0.785	-0.080	0.757
$PM_{2.5}$	0.936	1	0.492	0.659	0.624	-0.043	0.857
SO_2	0.438	0.492	1	0.619	0.413	-0.258	0.624
NO_2	0.580	0.659	0.619	1	0.503	-0.453	0.718
PM_{10}	0.785	0.624	0.413	0.503	1	-0.003	0.474
O_3	-0.080	-0.043	-0.258	-0.453	-0.003	1	0.474
CO	0.757	0.857	0.624	0.718	0.464	-0.172	1

4. Research on AQI prediction

According to the correlation analysis of AQI index and six kinds of common pollutants, the air quality of the next day can be predicted by the historical data of these pollutant concentration indexes. This paper establishes a model for AQI, which is the main index to measure air quality. The next day's AQI index value is used as the prediction target variable, and the AQI index and the historical index value of six pollutants are used as the model input variables. The LSTM neural network algorithm program is supported by using Keras module in Python. Due to the difference of data scale between each index value, this paper uses the method of maximum and minimum to realize the normalization of each index data. In the LSTM model, there are 100 neurons in the hidden layer and only one neuron in the output layer; the first 70% of the sample data is used as training data, and the last 30% as test data. Finally, when comparing the difference between the predicted results of the model and the real values, the predicted results are de normalized. The fitting curve between the predicted value and the real value on the training set and the test set is shown in Figure 3. It can be seen from the figure that the prediction error of LSTM model on the training set and the test set is small, indicating that the model has high prediction accuracy. The average absolute error of the model in the training set and the test set are 3.31 and 5.17 respectively, and the average absolute error rate in the training set and the test set are 4.13% and 4.91% respectively, which further shows that the model has high prediction accuracy. In Figure 3, green represents the training set and red represents the test set

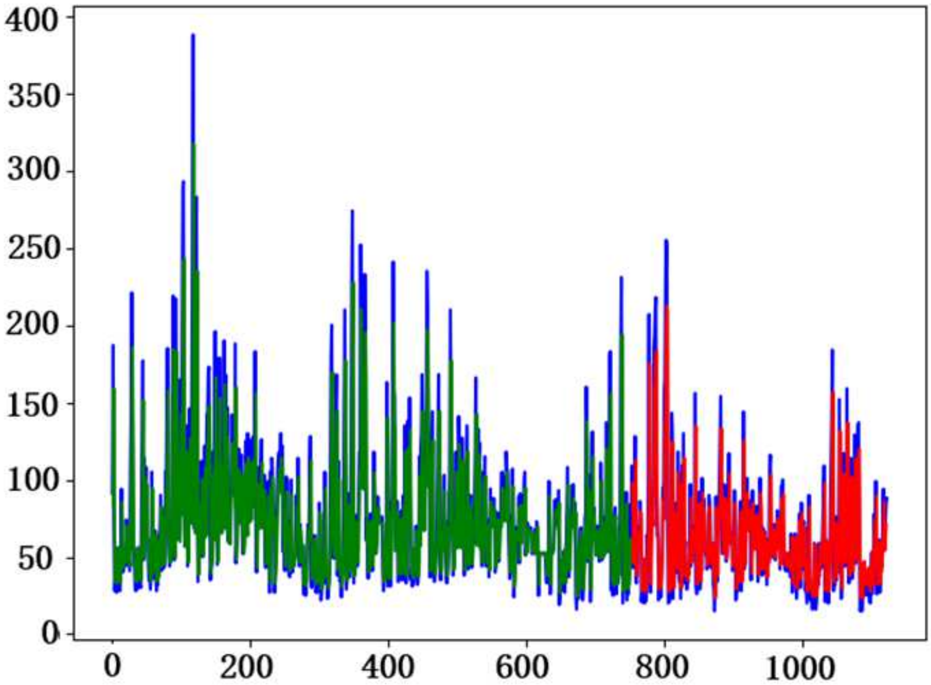

Fig 3 Prediction effect of LSTM model on training set and test set

5. Conclusion

Based on the analysis of the concentration of air pollutants in Beijing from January 2018 to December 2020, this paper analyzes the air pollution index, the concentration change trend of six pollutants and the correlation of each pollutant index. The results show that there is a positive correlation between AQI and the concentrations of CO、NO_2、$PM_{2.5}$、SO_2、O_3、PM_{10}, and a negative correlation between AQI and O_3.Due to the nonlinear relationship between the AQI index and the concentration of these pollutants, the traditional statistical prediction method cannot achieve the ideal prediction accuracy. In this paper, the recurrent neural network model is used to establish the prediction model, and the long-term and short-term memory network (LSTM) is used for model operation. The results show that the model has high prediction accuracy. The results show that the recurrent neural network can be widely used in the area of air quality data prediction, and can also be extended to more time series data.

Acknowledgment

This paper was supported in part by General Project of Science and Technology Plan of Beijing Municipal Education Commission (KM201910037002), and Beijing Excellent Talents Training Funding (2017000020124G051).

Reference

[1] Liu-Jie, Yang-Peng, Lu Wen-sheng, et al. Environmental air quality evaluation method based on the six pollutants in the urban areas of Beijing [J]. Journal of Safety and Environment, 2015, 15(1): 310-315

[2] BAI Heming, SHEN Runping, SHI Huading, et al. Forecasting model of air pollution index based on BP neural network[J]. Environmental Science & Technology, 2013, 36(3): 186-189

[3] LI Jinglu, ZENG Tian. Analysis on the Principal Component of Factors Affecting Air Quality in Beijing: From 2000-2011 Years of Experience Data[J]. Ecological Economy, 2017, 33(1): 169-173

[4] WANG Mingjie, HE Jiajia, WANG Shuxin, ZHANG Lei. 2018. Atmospheric pollution characteristics and typical circulation pattern in Shenzhen based on AQI [J]. Ecology and Environmental Sciences, 27(2): 268-275.

[5] LI Ping, NI Zhiwei, ZHU Xuhui, WU Zhangjun. Air pollution index prediction model of SVM based on fractal manifold learning[J]. Journal of Systems Science and Mathematical Sciences, 2018, 38(11): 1296-1306.

[6] XU Qi, WU Qizhong, LI Dongqing, et al. 2020. Assessment of the Air Quality Numerical Forecast in the Main District of Beijing (2018) [J]. Climatic and Environmental Research (in Chinese), 25 (6): 616–624.

[7] ZHANG Zhen, ZHU Quanjie, LI Qingsong, et al. Prediction of mine gas concentration in heading face based on keras long short time memory network[J]. Safety and Environmental Engineering, 2021, 20(1): 61-67

[8] Yang Taichun, Tao Jianfeng, Yu Honggan, Liu Chengliang. Real-time prediction of torque of cutter head of shield machine based on LSTM[J]. 2020, 16（6）: 1801-1808.

[9] Zhiling Tang, Qianqian Liu, Minjie Wu, Wenjing Chen, Jingwen Huang. WiFi CSI Gesture Recognition Based on Parallel LSTM-FCN Deep Space-Time Neural Network[J]. China Communications, 2021, 18(03): 205-215.

[10] ZHANGLin, HUANG Yanwen, XUAN Jie, FU Xiong, LIN Qiaomin, WANG Ruchuan. Trust Evaluation Model Based on PSO and LSTM for Huge Information Environments[J]. Chinese Journal of Electronics, 2021, 30(01): 92-101.