Prediction of Air Quality Index Based on LSTM

Yu Jiao[1]，Zhifeng Wang[1*], Yang Zhang[1]
1. School of Shanghai Polytechnic University
Shanghai, China
805720398@qq.com

***Abstract*—In view of the increasing attention paid by the state to environmental governance in recent years and the continuous deterioration of air quality, this paper proposes a prediction model of environmental quality based on Long Short Term Memory (LSTM). This paper uses data provided by the environmental protection department to predict Air Quality Index (AQI) through temperature, PM2.5, PM10, SO_2, wind direction, NO_2, CO and O_3. Firstly, this paper introduces the background, technical characteristics, development status and problems of air environment monitoring. Then, it will introduce the environmental prediction model. Finally, we make AQI prediction by using LSTM and analyze the error of the prediction results. The results show that LSTM can predict air quality index well.**

Keywords—LSTM; AQI prediction; error analysis

I. INTRODUCTION

The environment is the sum of all things that affect the existence of an organism or group of organisms, directly or indirectly, in space outside of a lifetime object or group. The air environment studied in this paper belongs to the environmental category of environmental impact assessment. The environment in the environmental impact assessment refers to the environment dominated by human, that is, the sum of all kinds of natural and social factors that revolve around the population space and can directly or indirectly affect human survival and development.

With the continuous improvement of China's development speed, people's awareness of environmental governance is becoming stronger and stronger. In 1979 the Environmental protection law of the People's Republic of China (trial) promulgated, It marks that China's environmental protection work has on the way, in 1982, the constitution provided the basis and guiding principles for environmental protection, The environmental impact assessment law of the People's Republic of China enacted in 2002 marks a new stage in China's environmental and resource legislation.

Atmospheric pollution sources can be divided into point source, plane source, line source and body source. Conventional sources of pollution mainly include sulfur dioxide (SO_2), particulate matter (PM), nitrogen dioxide (NO_2) and carbon monoxide (CO_2) [1]. In order to reduce the occurrence of air pollution incidents, China has closed down a large number of factories in recent years, and the air quality has improved step by step. By 2015, 1,436 air monitoring systems were built in cities above prefecture-level cities. So far, atmospheric environment monitoring stations have had some historical data. These data are mainly used for real-time monitoring, daily, weekly and monthly reports, etc. Real-time environmental information is also available on China's environmental monitoring website. However, with the continuous development of air pollution prevention and research, the tendency and rule of atmospheric pollutant concentration prediction has become the focus of people's attention. Therefore, this paper designed a prediction model of air quality index based on LSTM, we use the environmental monitoring station of Shanghai at the end of 2013 to September 2018 air quality data, by choosing the 90% data as the training set, the remaining 10% as test data sets. Finally, we use LSTM model predicted AQI and analyze the results.

II. CONSTRUCTION OF ENVIRONMENTAL PREDICTION SYSTEM BASED ON LSTM

A. Introduction of LSTM Cyclic Neural Network

LSTM is a kind of time Recursive Network Nerve (RNN), which is suitable for processing and predicting important events with relatively long intervals and delays in time series. LSTM is different from RNN mainly because it adds a processor to the algorithm to judge whether the information is useful or not. The structure under which processor is called cell.

A cell contains three gates, respectively called input gate, forget gate and output gate. When a piece of information enters the LSTM's network, the functions can be used to determine whether it is useful. Only the information that conforms to algorithm authentication can be left behind, inconsistent information is forgotten through the forget gate. It can solve the long - term problem of neural network under repeated operation [2,3]. The structure of LSTM is shown as Fig.1.

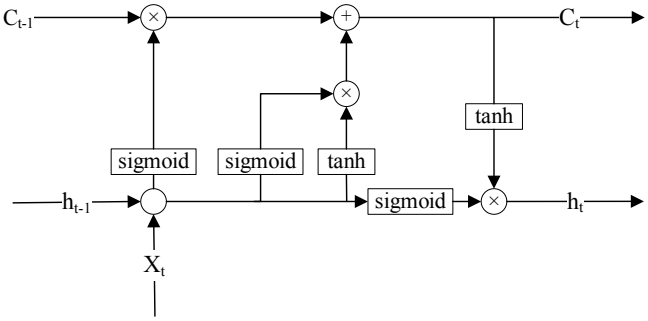

Fig.1. structure of LSTM

Forget gate. forget gate uses the previous output h_{t-1} and the current input x_t to output f_t by the function $sigmoid$. The output f_t range from 0 to 1. Then, the function sends f_t

to the current cell called C_{t-1}. The forget gate is shown as Fig.2.

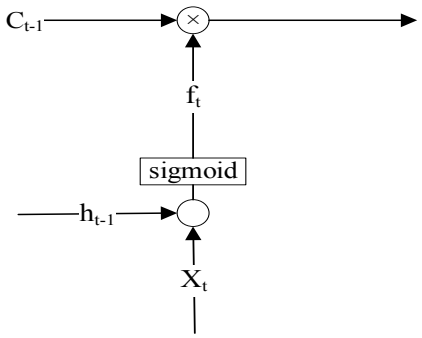

Fig.2. The forget gate

Input gate. Input date and the tanh function coordinate control what new information is added. This function creates a new candidate vector $\tilde{C}_t$, the input gate will give each $\tilde{C}_t$ a value in [0, 1], it used to control how much new information is added. By using the output of the forget gate f_t to control the previous unit on the extent of the forgotten, in combination with the output i_t which from the input date to control how much new information is added. In this way we can update the current memory unit. The input gate is shown as Fig.3.

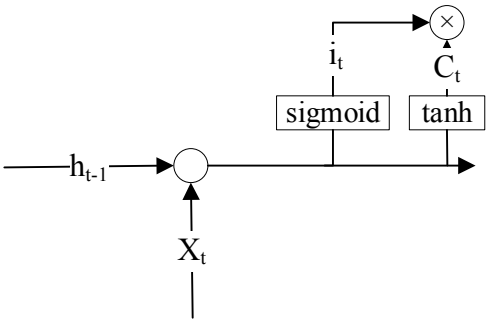

Fig.3. Input gate

Then the old cells will update the status, the old state C_{t-1} times f_t, it can discard information that needs to be discarded. By add the information which given by $i_t \times \tilde{C}_t$, it creates a new candidate value. The update function is shown as Fig.4.

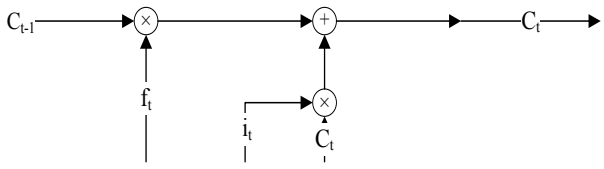

Fig.4. Update function

Output gate. The output gate outputs the value that we need to output, and the output will be based on the state of the cell, but it's also the filtered version. First of all, we run a sigmoid function to determine which part of the cell state will be output. Then, we manipulate the cell state through tanh (to get a value between -1 and 1) and multiply it by the output of the sigmoid gate, in the end, we just output the part that we determine the output is what we want [4]. The output gate is shown as Fig.5.

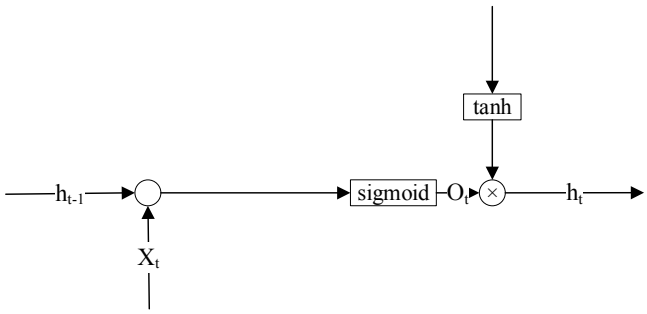

Fig.5. Output gate

The three gates functions are shown below:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (6)$$

III. DATA SELECTION AND PREPROCESSING

China's current Air Quality Index (AQI) is used to characterize air quality, which contains pollutants has six kinds: SO_2, NO_2, O_3, CO, PM10 and PM2.5. According to the six parameters concentration, we can calculate the Individual Air Quality Index (IAQI). Chooses the maximum IAQI for Air Quality Index as AQI. Because of the wide sources and complex composition of six pollutants, the forecast of air quality is full of uncertainty.

A. Distribution of AQI

We collect the air information from the end of 2013 to September 2018. AQI distribution is shown as Fig.6. we can see the AQI value which greater than 150 mainly concentrated in November to February every year. It mainly due to heating and set off firecrackers during Chinese New Year. At the same time, we collected the data of maximum temperature and minimum temperature and wind data, we think the AQI changes associated with these factors, so we add the parameters of the temperature and wind direction in this prediction system, the data is shown as Fig.7.

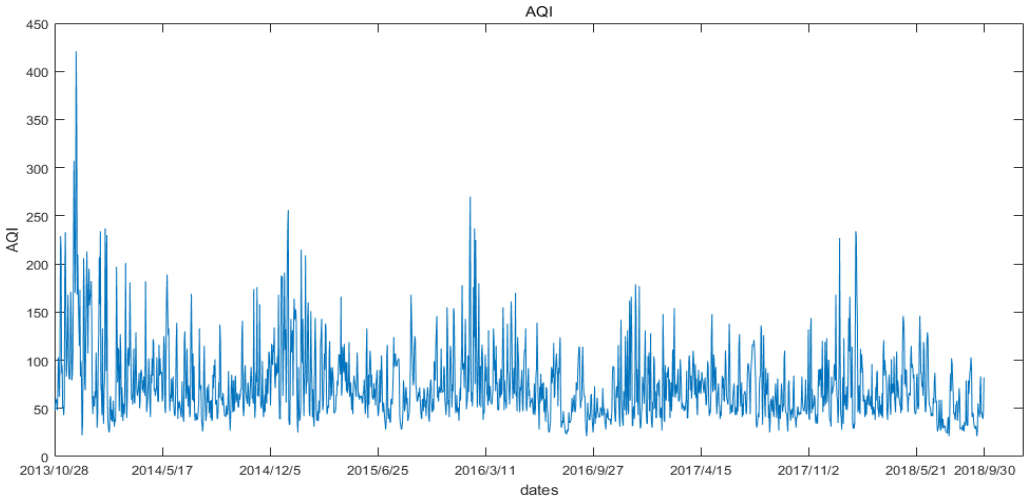

Fig.6. AQI distribution

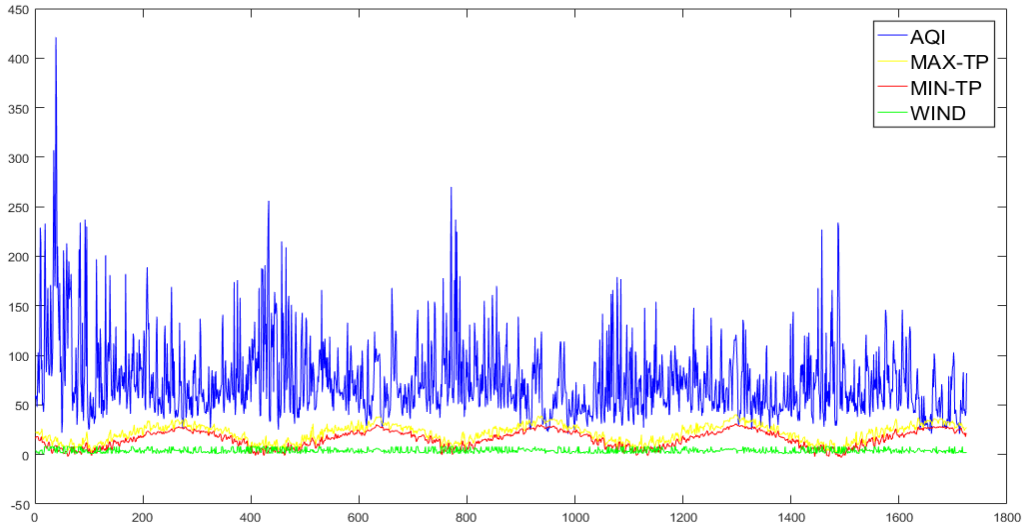

Fig.7. temperature, wind direction and AQI distribution

B. Correlation Analysis

Pollutants are the direct cause of air quality decline. Meanwhile, pollutants are also dispersed in the atmosphere with air circulation. In this paper, air quality is predicted by adding the highest temperature, the lowest temperature and the wind. The correlation analysis of these 9 parameters and AQI is shown as Fig.8. From Fig.8. We analyze PM2.5 and PM10, SO2, NO2, CO, O3, the maximum temperature, minimum temperature and wind direction and the AQI. Those correlation coefficients are 0.95, 0.89, 0.7, 0.64, 0.81, 0.023, 0.19, 0.26 and 0.39. Although there is a very small negative correlation coefficient, these parameters have a direct or indirect relationship with the AQI, so the paper uses these 9 parameters in the model of air quality prediction for training.

IV. Experiment Design And Result Analysis

According to the data provided by the environmental monitoring department, we repeated the iterative experiments. We use the multiple features of time series model based on LSTM, fitting LSTM by using multivariate input features. Firstly, we set the model of the hidden layer as 10, input layer as 9 and input layer as 1. By choosing the 90% data as the training data set, the remaining 10% as a test set. Using the standardization of data processing to improve the accuracy of the training. We set the training sample set as 60, the time step as 15, learning rate as 0.0006, then we define the neural network variables. At the same time, set we the number of iterations for 5000. Finally, we get the square deviation and mean square error are respectively 7.56 and 10.95.

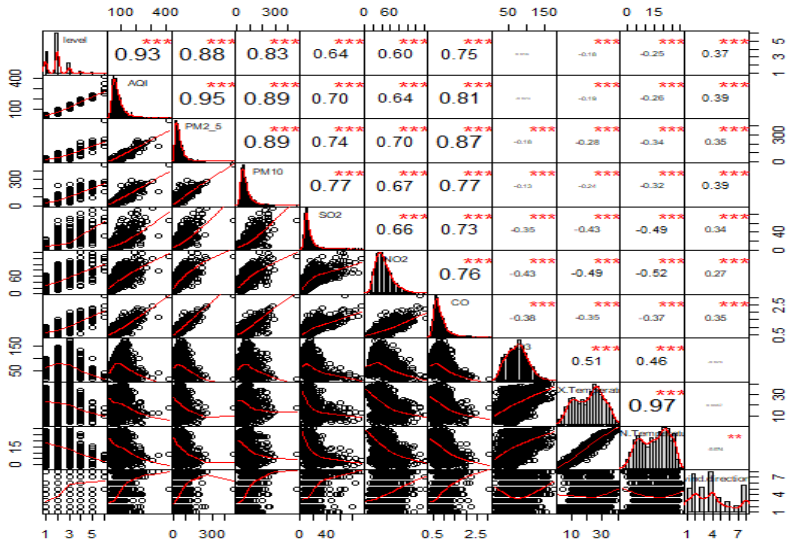

Fig.8. correlation analysis

Then, we get the predict data. The comparison chart of predicted results and test samples is shown as Fig.9. (in which blue is the original data and yellow is the prediction data obtained by training.)

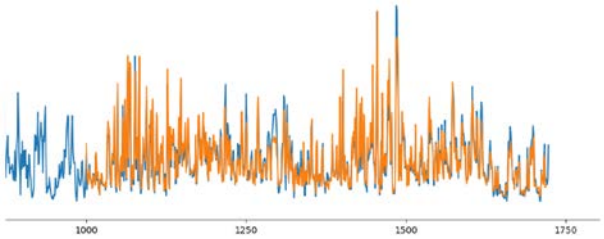

Fig.9. predict result

V. CONCLUSIONS

This article by studying the Shanghai air quality data and analyzing coefficient of correlation. By using LSTM cycle neural network to predict AQI based on 9 parameters. Based on the characteristics of time series, we can solve the problem of multiple input time variable very well. This experiment based on LSTM time series model to predict the Shanghai AQI has high precision, long range prediction and strong adaptive ability. It can approximate nonlinear mapping well. This mode can also be used in other multivariable input time series prediction problem, also has been widely used in life.

There is much to improve in the future, we can see, through the experiment LSTM, it can predict AQI very well. But through data analysis, we can see AQI data which bigger than 150 is far less than the AQI below 150, the samples of imbalance cause prediction model over-fitting, we will balance those imbalance data by Synthetic Minority Oversampling Technique (SOMTE) [5].

ACKNOWLEDGMENT

This work was supported by Shanghai Polytechnic University Graduate Project Fund No. EGD18YJ0053.

REFERENCES

[1] Environmental Engineering Assessment Center of the Ministry of Environmental Protection, Environmental impact Assessment of Construction Projects, China Environmental Science Press, 2012

[2] Ian Goodfellow, DEEP LEARNING, Posts & Telecom Press, 2017

[3] Peter Harrington, Machine Learning in Action, Posts & Telecom Press ,2013

[4] Chunlu Zhang, Application of tensorflow-based LSTM model in forecasting taiyuan Air Quality Index. Journal of Chongqing University of Technology, vol. 32, pp.137-141, 2018

[5] Bing Cheng, Yidan Su, Classification of unbalanced data based on km-smote and random forest. Computer Technology and Development. Vol. 25, pp.17-18, 2015