Research on Seasonal Prediction of PM2.5 Based on PCA - BP Neural Network

ZHANG Yi-wen, YUAN Hong-wu*, WU Hai-long, SUN Xin, DONG Yun-chun
College of Information Engineering, AnHui Xin Hua University, Hefei 230088
E-mail: yuanhongwu@axhu.edu.cn

Abstract: PM2.5 is a kind of data with strong time-series characteristics, so the current PM2.5 prediction methods mostly choose RNN, LSTM and other sequence models for prediction, but because RNN, LSTM and other models use the same weight to calculate the data input at different times, which does not conform to the brain-like design, resulting in a low accuracy of PM2.5 concentration prediction. In view of the above problems, a PM2.5 prediction method based on attention mechanism (at-rnn and at-lstm) is proposed. This method firstly establishes the encoder-decoder model based on deep learning. In the Encoder stage, attention mechanism is added, and attention weight is allocated to the input with time series characteristics, and then Decoder analysis and prediction are carried out. Through experiments, the effects of RNN, LSTM, at-rnn and at-lstm on the prediction of PM2.5 concentration in hefei are compared. The results show that the accuracy of the prediction method based on attention model is better than other methods, indicating the effectiveness of the prediction method based on attention model in the prediction of pollutants.

1. Introduction
With the increase of environmental pollution, the fog and haze weather continues to spread in most cities in China, and the impact of fine particles such as PM10 and PM2.5 on human health [1][2].

In recent years, a large number of researchers have studied the prediction of PM2.5 [3]. There are two main types of methods for PM2.5 predictive analysis. One type of forecasting method mostly uses linear computing. Such as Gene Expression Programming, Logistic regression model, LASSO regression model, etc. [4][5][6]. Using a linear model, the PM2.5 concentration value can be predicted from the associated data based on the correlation between the data. Another type of prediction method mainly uses nonlinear calculation methods. The representative method is to use neural network models for prediction. Because many atmospheric pollutants and meteorological factors and PM2.5 concentration are nonlinear characteristics [7][8], neural networks have better generalization ability and can better simulate the process of pollutants and meteorological factors, so the current selection. The neural network has made great progress in the simulation prediction [9][10].

With the wide application of deep learning in the fields of image processing, speech recognition, and natural language processing, a few researchers have used deep learning to predict weather or haze. For example, as the literature [11] selects the deep belief network, and considers PM2.5 and other meteorological and other complex feature relationships, using aerosol optical depth and meteorological parameters to predict the PM2.5 concentration value, compared with the neural network, improve the prediction accuracy rate. In [12] literature, the prediction of PM2.5 is sequential, the RNN prediction model based on time series is established, and RNN is optimized by LSTM. Compared with the recommended method of neural network, the proposed method improves the prediction accuracy. But
in general, the application of deep learning in PM2.5 forecasting is still in the early stage of research.

2. Related work

2.1 Time series based deep learning model

Traditional data mining methods have weaker learning ability for pollutants and meteorological factors, and have poorer perception of historical data such as PM2.5 concentration with time series characteristic pollutants. Since deep learning uses abstract representations of data hierarchies, data such as PM2.5 contaminants can be hierarchically partitioned in time blocks. Therefore, a time series-based deep learning method similar to Recurrent Neural Network (RNN) can be used to deeply mine complex pollutants and meteorological data, and to establish a time series based prediction method. The prediction model of RNN is shown in Figure 1:

\[
\begin{align*}
W & V \\
W & U \\
S & t-1 \\
O & t-1 \\
W & X_t-1 X_t \\
V & U \\
S & t+1 \\
O & t+1 \\
W & X_t+1 
\end{align*}
\]

Figure 1 RNN model

In Figure 1, the node data at each moment is composed of the input data at the current time and the data at the previous time, and each side of the input and output has the right weight: W, U, V.

At the same time, the Long Short-term Memory Networks (LSTM) method can also solve the problem of data dispersion in long time series. The structure of LSTM is shown in Figure 2:

Because PM2.5 has strong time series features, the common neural network methods have poor ability to perceive time series feature data. Therefore, using the Recurrent Neural Network (RNN) [13] model to predict the PM2.5 based on the time series deep learning method has achieved certain results. Although the RNN model can deeply learn the characteristics between various factors, due to the large amount of data and long time series, it is easy to cause gradient dispersion and gradient disappearance [14]. Using the model of Long Short-term Memory Networks (LSTM) [15], the problem of gradient dispersion and gradient disappearance can be solved in a certain procedure.
2.2 Attention Model
The current attention model is based primarily on the Encoder-Decoder model [16]. The attention model contains these parameters:

\[ Y_t = f(C_i, Y_1, ..., Y_{t-1}) \]  

\[ Y_t \] is the predicted output at time \( t \). The \( f \) function is the transformation method selected by the Decoder process, such as CNN, RNN, LSTM etc. The calculation of \( C_i \) is as shown in formula (2):

\[ C_i = \sum_{j=1}^{L_x} a_{ij} h_j \]  

3. PM2.5 prediction method based on attention model
Due to PM2.5 has strong time series characteristics, RNN and LSTM are selected as the transformation methods of Encoder and Decoder stage respectively in the experiment. The attention-based PM2.5 prediction model is shown in figure 3:

![Figure 3. attention-based PM2.5 prediction model](image)

In the model shown in figure 3, the prediction method of encoder-decoder using RNN is denoted as ARNN, and the prediction method of encoder-decoder using LSTM is denoted as ALSTM. In the Encoder stage, the calculation method of the attention part is shown in formula (3):

\[ a_{ij} = 1 - |p\tilde{m}_i - pm_i| \sum_{l=1}^{L} |p\tilde{m}_i - pm_l| \]  

4. Experimental design and analysis of results

4.1 Data collection and preprocessing
Historical PM2.5 data of 10 observation points in Hefei were collected in the experiment. The data period was from January 1, 2016 to December 31, 2018, and pollutant data came from pm25. In [17].

Due to the time of data loss during data collection, if there is a data loss, the data should be filled according to formula (4):

\[ \text{Where is the missing hourly PM2.5 concentration data, the latest PM2.5 concentration data at the last moment of missing data, and the latest PM2.5 concentration data at the next moment of missing data.} \]
4.2 evaluation criteria
AAD, Average Absolute Deviation, RMSE, Root Mean Squared Error and R, Pearson correlation coefficient were used as evaluation criteria [18][19], as shown in equations (4), (5) and (6) respectively:

\[
AAD = \frac{1}{N} \sum_{i=1}^{N} |y_i - x_i|
\]  

ADD is the average relative deviation, represents the expected value, and represents the predicted value. The AAD index represents the size of the error, the AAD be smaller be better.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{\text{obs},i} - x_{\text{model},i})^2}
\]

It represents the ith predicted value, represents the ith true value, and n represents the prediction times. The RSME index represents, the prediction error, the RMSE be smaller be better.

4.3 experimental results and analysis
Data of the first 11 months of each year in 2016, 2017 and 2018 were selected as the training set, and data of December were selected as the test set for the experiment.

The above three groups of experiments were carried out in three years. The experimental training set of each group was the PM2.5 data of the first 11 months of each year, and the test set was the data of December each year. Due to the small amount of experimental data in each group, PM2.5 values from January 2016 to November 2018 were selected as the training set and data from December 2018 as the test set to further verify the model. The experimental results are shown in table 1:

| models         | evaluation standard | 2016    | 2017    | 2018     | 2016-2018 |
|----------------|---------------------|---------|---------|----------|-----------|
| RNN            | AAD                 | 0.535   | 0.486712| 0.700923 | 0.573513  |
|                | RMSE                | 38.595  | 46.3952 | 38.51398 | 33.01254  |
|                | R                   | 0.215   | 0.419958| 0.2714   | 0.3555    |
| AT-RNN         | AAD                 | 0.5     | 0.4012  | 0.54718  | 0.509786  |
|                | RMSE                | 37.69   | 40.165  | 33.68442 | 29.83526  |
|                | R                   | 0.3     | 0.5208  | 0.3123   | 0.429752  |
| LSTM           | AAD                 | 0.559   | 0.40904 | 0.62704  | 0.587758  |
|                | RMSE                | 39.34   | 45.6973 | 36.09344 | 33.20524  |
|                | R                   | 0.1507  | 0.433936| 0.323868 | 0.328122  |
| AT-LSTM        | AAD                 | 0.47    | 0.3912  | 0.50424  | 0.437676  |
|                | RMSE                | 35.3    | 35.626  | 31.28582 | 27.91342  |
|                | R                   | 0.267   | 0.6     | 0.34184  | 0.490116  |

As can be seen from table 1, all indexes of the model with an increased training set were improved on RNN, LSTM, at-rnn and at-lstm, indicating that the size of the training set directly affected the prediction accuracy of the model. In addition, except that the R value in 2016 is the optimal at-rnn, all other optimal indicators are at-lstm. The LSTM model can alleviate the problems of gradient disappearance and dispersion of data, while the attention-based model can further improve the LSTM structure, indicating that historical PM2.5 values at different times have different influences on the predicted PM2.5 values in the future.

5. Conclusions and future work
The accuracy of PM2.5 prediction was improved through the attention model, but it was found through data analysis that the small amount of data affected the prediction results. Meanwhile, the cause of
PM2.5 is complicated. In the model, only historical data of PM2.5 are used to predict the future PM2.5 value, without considering the influence of PM10, SO2, CO2, CO, O3 and other pollutants as well as meteorological factors on PM2.5. In the next step, data will be collected and meteorological data will be added to predict PM2.5 to further improve the prediction accuracy.

Acknowledgement
This work is supported by the Key Program in the Youth Elite Support Plan in Universities of Anhui Province (gxyqZD2018087), the Project of Anhui Xinhua College Research Institute (yjs201707), Anhui university provincial natural science research project (KJ2019A0877).

References
[1] ZHOU Shuxue, WANG Xing, GONG Zhongqiang, SHI Chune. Transport patterns of PM2.5 in the western Yangtze River Delta district [J]. Acta Meteorologica Sinica, 2017, 75(06):996-1010.
[2] Makkonen U, Hellén H, Anttila P, et al. Size distribution and chemical composition of airborne particles in south-eastern Finland during different seasons and wildfire episodes in 2006 [J]. Science of the Total Environment, 2010, 408(3):644-51.
[3] LI Shenxin, ZOU Bin, LIUXingquan, et al. Pollution status and spatial-temporal variations of PM2.5 in China during 2013-2015 [J]. Research of Environmental Sciences, 2017(5):678-687.
[4] Spatiotemporal distribution characteristics of PM2.5 concentration and its main control factors in China based on multivariate data analysis [J]. Climatic and Environmental Research (in Chinese), 2018, 23(5).
[5] GU Yangyang, SUGuijin, CHAI Tao, et al. Influencing factors and prediction model of PM2.5 concentration in Beijing area [J]. Environmental Chemistry, 2018(3):397-409.
[6] LUO Hongyuan, WANG Deyun, LIUYanling, et al. PM2.5 concentration forecasting based on two-layer decomposition technique and improved extreme learning machine [J]. Systems Engineering—Theory & Practice, 2018, 38(5):1321-1330.
[7] QIAO Junfei, CAI Jie, HAN Honggui. Study on Prediction of PM2.5 Based on T-S Fuzzy Neural Network [J]. Control Engineering of China, 2018, 25(3).
[8] WU Chunlin, LI Qi, HOU Junxiong, et al. PM2.5 concentration prediction using convolutional neural networks [J]. Science of Surveying and Mapping, 2018, 242(8):72-79.
[9] LI Xiaoli, MEI Jianxian, ZHANG Shan. PM2.5 concentration prediction using BP-Adaboost neural network based on modified particle swarm optimization [J]. Journal of Dalian University of Technology, 2018, 58(3):99-106.
[10] ZHOU Binbin, LI Wenjing, QIAO Junfei. Prediction of PM2.5 concentration based on self-organizing recurrent fuzzy neural network [J]. CAAI Transactions on Intelligent Systems, 2018, 72(4):21-28.
[11] CUI Xianghui, XIE Jianfeng, ZHANG Feng, DING Lin, LI Zengshun, HAO Zhenhuan, LIU Yong, ZHAO Qichao. Establishment of PM2.5 Forecasting Model Based on Deep Learning [J]. Beijing Surveying and Mapping, 2017(06):22-27.
[12] Fan Junxiong, LI Qi, ZHU Yajie, HOU Junxiong, FENG Xiao. Aspatial-temporal prediction framework for air pollution based on deep RNN [J]. Science of Surveying and Mapping, 2017, 42(07):76-83+120.
[13] LIE Seng, SHI Zhengwei, SHI Tianyang, et al. Prediction of storm surge based on recurrent neural network [J]. CAAI Transaction on Intelligent Systems, 2017, 12(5):640-644.
[14] WANG Xing, WU Ji, LIU Chao, et al. Exploring LSTM based recurrent neural network for failure time series prediction [J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(4):772-784.
[16] PENG Min, YANG Zhaoxiong, ZHU Jiahui. Semantic Enhanced Topic Modeling by Bi-directional LSTM [J]. Journal of Chinese information processing, 2018, 32(4): 40-49.

[17] YANG Bo, SU Xiaohong, WANG Yadong. A Hybrid Algorithm Based on Attention Model [J]. Journal of Software, 2005, 16(6): 1073-1080.

[18] Atmospheric data collection [EB/OL] . http://www.weather.com.cn/

[19] WU Yunyun, ZHU Jianjun, ZUO Yanying. Investigation on the Square Root of the Linear Combination of the Square of RMSE and Smoothness for the Vondrank Filter’s Evaluation [J]. Geomatics and Information Science of Wuhan University, 2012, 37(10): 1212-1214.

[20] YANG Jianguo, ZHAO Hong, CEN Kefa. Optimized BP network model for predicting igniting temperature of pulverized-coal based on genetic algorithm [J]. Journal of China coal society, 2006, 31(2): 211-214.