Research of Air Pollutant Concentration Forecasting Based on Deep Learning Algorithms

YongMing Pan, YaJie Wang* and MingZhao Lai

Tianjin University of Technology, Tianjin, China

*Corresponding author e-mail: 807063173@qq.com, 79411339@qq.com, 13820923179@163.com

Abstract. In order to accurately predict the concentration of air pollutants in Shanghai, a prediction model of the concentration of air pollutants in Shanghai based on Wavelet Transform and Long Short-Term Memory (LSTM) was established to predict the concentration of six air pollutants in Shanghai. Firstly, the historical time series of daily air pollutant concentration is decomposed into different frequencies by wavelet decomposition transform and recombined into a set of high-dimensional training data. Secondly, LSTM prediction model is trained with high-dimensional data sets, and parameters are adjusted repeatedly to obtain the optimal prediction model. The results show that the combined model is more accurate than the traditional LSTM model in predicting pollutant concentration.

1. Introduction

With the rapid development of urbanization and industrialization, tremendous economic achievements have been achieved, but at the same time, pressure has been exerted on resources, energy and environment [1]. Air pollution has become an important problem that restricts economic and social development and affects human health. Monitoring and predicting pollutant concentration can effectively avoid the health threat caused by excessive pollutant concentration. At the same time, the prediction of major air pollution concentration can be used as a policy tool for environmental protection departments to regulate social and economic activities such as transportation, industry and urban construction under extreme air pollution conditions. Therefore, in order to support decision-making of environmental protection management and avoid serious accidents caused by air pollution, it is urgent to establish an accurate and stable pollutant concentration prediction model to guide the release of air pollutant control measures and public health protection.

At present, the research on prediction of atmospheric pollutant concentration mainly focuses on the application of deterministic model and computational model [2]. Computational models usually require a large number of historical measurements under various meteorological conditions, and establish historical pollutant data and by means of regression and neural networks. The relationship between predictive variables. M.A. Elangasinghe [3] extracts key information from daily available meteorological parameters and year-round source emission patterns, and establishes a physical-based ANN air pollutant prediction tool, which can fully capture the temporal variation of air pollutant concentration in specific scenarios. H.P. Corporation combined with Elman Neural Network and Autoregressive Integrated Moving Average Method to establish the model can still obtain excellent
prediction results when the concentration of particulate matter is high [4]. The traditional neural network model can not meet the need of high precision air quality prediction gradually, so researchers try to improve the prediction accuracy by improving the structure of input variables [5]. Qingping Zhou [6] constructed a hybrid EMMD-GRNN model based on data preprocessing and analysis to improve the data dimension of input variables, so as to quickly and accurately predict the PM$_{2.5}$ concentration in the next day. Ping Wang [7] effectively improves the prediction accuracy of artificial neural network and support vector machine by correcting the error terms of traditional methods. Baolei Lv [8] established empirical regression models for PM$_{2.5}$ and O$_3$ air pollutant concentration prediction in three large cities of Shanghai, Nanjing and Guangzhou. The prediction model was empirical non-linear regression model, and was designed for automatic data retrieval and prediction platform.

In this paper, wavelet decomposition transform is used to analyze the time series data of six atmospheric pollutant concentrations in Shanghai to get high-dimensional information [9]. Based on the long-term and short-term neural memory network, a prediction model is constructed to predict the atmospheric pollutants in Shanghai. Compared with the dynamic characteristics of air pollution index, the long-term and short-term neural memory network can effectively solve the adverse effects of the spatial and temporal evolution of air pollution index [10]. Therefore, based on the ability of LSTM to analyze and forecast time series data, this paper mainly completes the following three aspects: (1) using wavelet decomposition to elevate the dimension of data, optimize input variables, and improve the prediction accuracy of LSTM model; (2) developing Wavelet-LSTM air pollutant concentration prediction model. (3) Comparing the traditional LSTM prediction results with the actual data, the Wavelet-LSTM prediction results verify that the prediction accuracy and stability of the LSTM model can be improved by the wavelet decomposition transformation of the data.

2. Method

2.1. LSTM neural network

Long Short-Term memory [11], as a popular recursive neural network algorithm, was first proposed by Hochreite and Schmidhuber (1997), which promoted the memory ability of long (static) and short (cyclic) dynamic features of time series. Different from the traditional cyclic neural network model, LSTM has a special neuron structure called "memory unit". The hidden layer of LSTM network constructed by this structure can store information of any length of time and obtain more accurate time series model [12]. Assuming that at the time of $t$, the input, output and state of a memory cell module are $x_t$, $h_t$ and $C_t$, the input gate, forgetting gate, output gate, input conversion, updating of cell state and output formula of hidden layer of the memory cell module are shown in Formula (1)–(6), respectively.

\begin{align*}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) \\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f) \\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o) \\
c'_t &= \tanh(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot C'_t \\
y_t &= w_m m_t + b_y
\end{align*}
2.2. Wavelet Transport

Wavelet decomposition transform uses window adjustment method to achieve the purpose of localization analysis. The input signal can be decomposed into low-frequency signal reflecting the real trend of change of signal data and high-frequency signal reflecting random fluctuation. The time series data of atmospheric pollutant concentration is a set of non-linear and non-stationary time series data. The information characteristics of the time series data of atmospheric pollutant concentration in different time and frequency can be extracted by wavelet decomposition transformation. The pollutant concentration data are decomposed into sequence groups composed of different dimension data signals by wavelet decomposition transform. Compared with the original data, these sequence groups have more stable variance and fewer singular value points, which can express the original signal information more effectively and accurately, and then achieve the purpose of improving the prediction accuracy.

\[
\psi_{j,k}(t) = 2^{j/2} \varphi(2^j t - k) \quad (7)
\]

\[
\varphi_{j,k}(t) = 2^{j/2} \varphi(2^j t - k) \quad (8)
\]

In the formula, J and K are scale parameters and translation parameters respectively. The signal can be expressed by formulas (7) and (8) as follows:

\[
y(t) = \sum_{k} c_{j_0}(k)2^{j_0/2} \varphi(2^{j_0} t - k) + \sum_{j} \sum_{k} d_{j,k}(t)2^{j/2} \varphi(2^j t - k) \quad (9)
\]

In the formula, the approximate coefficients and the detail coefficients are respectively used. The data of pollutant concentration can be decomposed into:

\[
y(t) = A_{mt} + D_{ht} + \Lambda + D_{mt} \quad (10)
\]

In formula 10, the approximate information \( A_{mt} \) set represents the information characteristics of the original data, and the high-frequency information \( D_{ht} \) \( \Lambda \) \( D_{mt} \) set represents the noise part of the original information.

3. Research Data

The object of this study is Shanghai, located in eastern China. Daily historical data of air pollutant concentration in Shanghai from January 1, 2013 to December 31, 2017 were collected, totaling 1,460. Daily surface measurements of pollutants include PM\(_{10}\), PM\(_{2.5}\), NO\(_2\), SO\(_2\), O\(_3\) and CO. Data were collected from China National Environmental Monitoring Center and Ministry of Ecology and Environment of the People's Republic of China. In order to verify the predictive performance of the prediction model for atmospheric pollutant concentration, 1,095 data sets from January 1, 2014 to December 31, 2016 were used as training data sets of the prediction model, and 365 data sets from January 1, 2017 to December 31, 2017 were used as test data sets.

4. Result and Discussion

In order to verify that the Wavelet-LSTM model proposed in this paper has higher prediction accuracy than the traditional LSTM model, the time series data of Shanghai air pollutant concentration from 2014 to 2016 are trained with the traditional LSTM model and the combined model respectively, and the air pollutant concentration data of 2017 are used to test the predictability of the LSTM prediction model after adding the wavelet decomposition transform. As shown in Table 1, when PM\(_{10}\) is used as dependent variable, MAPE can reach 7.54%, while when PM\(_{2.5}\) is used as dependent variable, MAPE
can reach 17.25%. Although LSTM can perform well in time series prediction. However, the MAPE standard deviation of the prediction results is large, which indicates that the prediction stability of the traditional LSTM model can not be guaranteed when the independent variables are different.

Because LSTM model can efficiently express the high-dimensional non-linear relationship between input variables and predicted targets through the kernel function, using appropriate high-dimensional information as input variables can more effectively and accurately describe the information characteristics, indicating that the prediction ability of the model depends largely on the selection of input variables in model design. In this paper, the input variable data is upgraded to a high-dimensional data set through wavelet decomposition transformation, which can more comprehensively and effectively represent the trend of data change, thus improving the prediction accuracy. Six atmospheric pollutants (PM$_{10}$, PM$_{2.5}$, NO$_2$, SO$_2$, O$_3$, CO) were used as input variables of LSTM model by using the approximate low-frequency information and high-frequency information obtained by wavelet decomposition transformation in the MATLAB toolbox, and the high-frequency information group and the low-frequency information group were obtained. LSTM model uses wavelet decomposition transform to get data as a new training data. The experimental results show that the average MAPE is 10.71%. Compared with the traditional LSTM model, the MAPE is reduced by 1.91%. The variance of the model for different prediction targets is significantly reduced, and the stability of the model is significantly improved. The low-frequency information subset as the trend of data change can be obtained by wavelet decomposition transform. The LSTM prediction model has a deeper learning level for pollutant concentration data. After using wavelet decomposition transform to process input variables, the prediction accuracy is high and the stability is good under different target pollutants. According to the prediction results in 2017, the model has strong applicability for predicting the abnormal points of inhalable particulate matter (PM$_{10}$), but poor ability for predicting the abnormal values of SO$_2$. It shows that the generalization ability of pollutant concentration prediction model will be affected by the difference of data values and units.

Wavelet-LSTM forecasting model can accurately predict the target pollutant concentration of the next day according to the five air pollutant concentration values of the previous day, and also can accurately predict the numerical abrupt change due to seasonal change by memorizing and learning the air pollutant concentration data of Shanghai from 2014 to 2016. Therefore, it can be proved that the combined model can not only provide more accurate pollutant concentration prediction values, but also be used to explore the development trend of pollutant concentration.

| Table 1. Forecasting mape of LSTM and Wavelet-LSTM (%) |
|-----------------|-------|-------|-------|-------|-------|-------|
|                 | PM$_{10}$ | PM$_{2.5}$ | NO$_2$ | SO$_2$ | O$_3$ | CO | Average |
| LSTM            | 7.54     | 17.25   | 12.63  | 15.73  | 9.05  | 13.53 | 12.62   |
| Wavelet-LSTM    | 7.78     | 12.59   | 10.73  | 11.02  | 11.77 | 10.37 | 10.71   |

5. Conclusion

(1) The combined forecasting model uses wavelet decomposition to decompose the time series data of atmospheric pollutant concentration. The decomposed low-frequency data and high-frequency data are simultaneously taken as input variables, and the data dimension is increased. Through the information representation of pollutant concentration time series data at different frequencies, the data information characteristics can be better described.

(2) Using LSTM neural network, which is a high-dimensional non-linear learning algorithm, to build a prediction model. It can be applied to the prediction of Shanghai atmospheric pollutant concentration time series data. However, due to the low dimension of pollutant concentration time series data, the information representation is incomplete, and the prediction ability of six pollutants is different, so the generalization ability of the prediction model is affected.

(3) From the experimental results, it can be seen that the combined forecasting model has significantly improved the accuracy and stability of pollutant concentration prediction. Especially in the prediction of burst data points, the combined forecasting model of original data processed by
wavelet decomposition transform can accurately predict the concentration of air pollutants. Therefore, it can be considered that the high-dimensional input variables consisting of low-frequency information and high-frequency information obtained from the time series data of pollutant concentration after wavelet transform can express the data information more accurately.

(4) By comparing the actual data, the predicted average MAPE value is reduced to 10.71%, and the predicted model has the best predictive ability for PM$_{10}$. Compared with the complex mechanism model with high computational cost, it is more suitable for the prediction environment with strong complex uncertainties. Therefore, the mixed prediction model has strong applicability and high application value in predicting the concentration of air pollutants.

References

[1] Künzli N, Kaiser R, Medina S, et al. Public-health impact of outdoor and traffic-related air pollution: a European assessment [J]. Lancet, 2000, 356(9232):795-801.

[2] Zhang Y, Bocquet M, Mallet V, et al. Real-time air quality forecasting, part I: History, techniques, and current status [J]. Atmospheric Environment, 2012, 60(32):632-655.

[3] Elangasinghe M A, Salmond J A, Dirks K N, et al. Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis [J]. Atmospheric Pollution Research, 2014, 5(4):696-708.

[4] H.P. Corporation. Forecasting SO2 pollution incidents by means of Elman artificial neural networks and ARIMA models. Abstr. Appl. Analysis, 2013 (4):1728-1749

[5] Wu Q, Xu W, Shi A, et al. Air quality forecast of PM10 in Beijing with Community Multi-scale Air Quality Modeling (CMAQ) system: emission and improvement [J]. Geoscientific Model Development, 2014, 7(5):2243-2259.

[6] Zhou Q, Jiang H, Wang J, et al. A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network [J]. Science of the Total Environment, 2014, 496(2):264-274.

[7] Wang P, Liu Y, Qin Z, et al. A novel hybrid forecasting model for PM 10, and SO2, daily concentrations[J]. Science of the Total Environment, 2015, 505(505C):1202-1212.

[8] Lv B, Cobourn W G, Bai Y. Development of nonlinear empirical models to forecast daily PM2.5 and ozone levels in three large Chinese cities [J]. Atmospheric Environment, 2016, 147:209-223.

[9] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8):1735-1780.

[10] Luo H, Huang M, Zhou Z. Integration of Multi-Gaussian fitting and LSTM neural networks for health monitoring of an automotive suspension component [J]. Journal of Sound & Vibration, 2018, 428:87-103.

[11] Alizadeh M J, Kavianpour M R. Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean [J]. Marine Pollution Bulletin, 2015, 98(1-2):171-178.

[12] Nourani V, Baghanam A H, Adamowski J, et al. Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review [J]. Journal of Hydrology, 2014, 514:358-377.