Forecast of Baoding Air Quality Index Based on Long-Short Term Neural Memory

Qingcheng Hu

International College, Hebei University, Baoding, China, 071000

a1040565026@163.com

Abstract. Accurate prediction of air quality in Baoding City will help residents take preventive measures against air pollution and relevant departments to understand the city's air quality situation and implement air pollution control policies. This article uses Python to crawl a total of 3457 sets of historical data on the concentration and air quality index of 6 pollutants including $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3 in Baoding City from 2014 to 2023, and establishes a long- and short-term neural network prediction model. Define the number of input layers, output layers and hidden nodes, and determine ReLU as the activation function. The tangible output results after applying the model are RMSE 23.6727, R2 0.8332, MAE 17.4298, and MBE 0.35743. The research results demonstrate that the model is training It shows excellent performance on both the test set and the test set. The evaluation results indicate that using the LSTM model to predict AQI in Baoding City has higher accuracy. At the same time, the model has strong robustness and can accurately forecast even in extreme situations.

Keywords: Air quality index, long-short-term neural memory network, recurrent neural network.

1. Introduction

In recent years, China's urbanization and industrialization processes have accelerated, and China's economy has developed rapidly, but the environmental problems it has caused have also become more and more serious [1]. The long-term presence of various air pollutants such as $PM_{2.5}$, NO_2, SO_2, CO, etc. It leads to the occurrence of acute and chronic diseases and also causes adverse effects and damage to the environment. As the city with the largest population and the widest territory in Hebei Province, Baoding City has significant location characteristics in the Beijing-Tianjin-Hebei region. By predicting the air quality in Baoding City, it can not only help residents take preventive measures against air pollution, but also provide the government with ideas for solving air environment problems, which is of profound significance for accurately controlling air pollution problems.

The Air Quality Index (AQI) is a quantitative indicator used to quantitatively describe the state of air quality, which enables a visual assessment of the pollution level of the atmosphere. In general, the higher the AQI value, the higher the concentration of pollutants in the environment, the more harmful it is to the body, and also causes a reduction in human comfort [2, 3].

Shen Hao Yang et al [4] used a BP artificial neural network to predict the concentration of the atmospheric pollutant SO2 in 2014, and this prediction successfully captured the pattern of change in the concentration of SO2.Kong et al. [5] used a coupled system of the 4D-LETKF model and the WRF-Chem model to forecast PM2.5 values in haze weather in Beijing, Tianjin, Hebei and its surrounding areas. The results show that the model has achieved significant improvements in prediction accuracy and operational efficiency, but has not yet taken into consideration other factors that may affect air quality. The BP neural network model established by Hao Yongchao et al. [6] based on 333 samples has an accuracy of 41.94% for predicting Xingtai City's air quality level in the subsequent month. Xie, Y. et al [7] used support vector machines (SVMs) to predict PM2.5 concentrations in cities in 2015, and the results showed that support vector machine regression (SVR) performs well toward for other machine models such as neural networks in short- and medium-term prediction, but once the feature dimensions are high, the model will be extremely complex. Si Zhijuan [8] established an improved gray neural network combination model to predict NO2 and other air pollutants in Tianjin, showing that the improved model was more accurate than the traditional gray

network model. Shi Xiaowen et al. [9] analyzed the dimensions of air quality levels in 2019 and concluded that the long-short-term memory (LSTM) air quality prediction model is better than the traditional back-propagation neural network, support vector machine and other classic algorithms. It is more efficient and has stronger generalization ability, which is especially suitable for urban air quality prediction. The above studies have the shortcoming of only using air quality index data for model training and prediction. However, air quality is the result of the interaction of multiple factors. Therefore, this study uses an improved long-short-term neural memory network to not only use the air quality index, but also use $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3 and additional data to improve the accuracy of AQI prediction results.

The LSTM prediction process proposed in this article is shown in Figure 1, which includes three steps: preprocessing, model establishment, and result analysis. The input is the acquired air quality data, and the output is the conclusion of the air quality prediction.

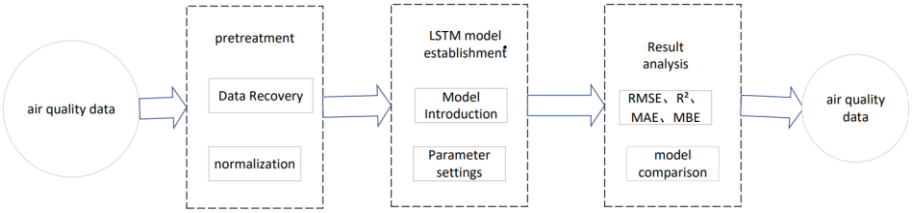

Figure 1. LSTM air quality prediction process

2. Data preprocessing

2.1. Data introduction

The data used in this study are a total of 3457 sets of historical air quality data in Baoding City from January 1, 2014 to December 14, 2023, crawled from the Internet. The daily average values of the air quality observations in Baoding City obtained include PM2. 5. PM10, CO, NO2, SO2, O3 and AQI, write the crawled data into a CSV file. Among the air quality pollutant factors, except for CO, the unit is mg/m3, the other pollutant factor units are ug./m3, see Table 1 for some samples.

Table 1. Some samples of pollutant concentration and air quality index

Data	AQI	$PM_{2.5}$	PM_{10}	CO	NO_2	SO_2	O_3
2024/1/1	117	77	185	1.8	57	64	48
20241/2	278	228	375	4.8	124	134	25
2024/1/3	251	201	347	4.4	97	111	30
2024/1/4	245	195	306	3	98	118	18
…	…	…	…	…	…	…	…
2023/12/11	46	32	40	0.6	24	3	38
2023/12/12	79	58	61	0.6	29	3	39
2023/12/13	94	70	75	0.7	37	3	22
2023/12/14	69	50	60	0.9	26	3	37

2.2. Data preprocessing

For the missing AQI and other attribute column data in the sample, the average value of Xn-1 and Xn+1 of the two days before and after is used as the value of the AQI and other attribute columns on that day. The calculation formula is:

$$x = \frac{x_{n-1} + x_{n+1}}{2} \tag{1}$$

For data that is missing for two consecutive days, calculate the average values of Xn-2, Xn-1, Xn+1, and Xn+2 for the four days before and after. There is an order of magnitude difference between

the attribute values of the sample data. If used directly to train the network model, it is easy to obtain weights with the same order of magnitude difference, which makes the model unstable and affects the accuracy of the prediction results. Therefore, the data is normalized before training the model. Some of the data are shown in Table 2.

Table 2. Some samples after normalization

AQI	$PM_{2.5}$	PM_{10}	CO	NO_2	SO_2	O_3
0.628821	0.41048	1	0	0.30131	0.33952	0.252183
0.737979	0.602917	1	0	0.321988	0.349001	0.054565
0.71979	0.573847	1	0	0.270286	0.31115	0.074723
0.79868	0.633663	1	0	0.313531	0.379538	0.049505
…	…	…	…	…	…	…
1	0.69163	0.867841	0	0.515419	0.052863	0.823789
1	0.732143	0.770408	0	0.362245	0.030612	0.489796
1	0.742765	0.796356	0	0.389068	0.024652	0.228296
1	0.720999	0.867841	0	0.368576	0.030837	0.530103

3. Model building

3.1. Model introduction

LSTM network is a variant of recurrent neural network (RNN), which effectively solves the problems of gradient explosion, gradient vanishing and degradation of long-term memory capacity of RNN by setting up three special unit structures, namely input gate, output gate and forgetting gate. This enables the recurrent neural network to better utilize temporal information over long time spans. The Air Quality Index (AQI) is involved in a variety of metrics and there are complex correlations between these metrics. The LSTM is able to capture the complex relationships between these features through its memory units and gating mechanisms, and continuously learns until it finds the appropriate weights to accurately predict the future air quality. Compared with other traditional machine learning methods, LSTM is more suitable for processing time-dependent sequential data, and is able to capture the nonlinear patterns and long-term dependencies in time-series data, thus increasing the accuracy of prediction.

The LSTM is a circular recursive neural network. All layers have the same weights. By introducing special neural units of memory cells to control the input of long-distance information, Enables neural networks to rely on long-span time-series information. This memory unit will have a connection to itself at the next step and copies the true value of its own state and accumulated external signals, Therefore it is more suitable for time series data forecasting. A single LSTM nerve cell mainly contains four layers of structure: forgetting gate, output gate, input gate and a unit state. Adding or deleting information by the controllers. Usually composed of a sigmoid neural network and a pair of multiplication operations, the sigmoid layer compresses the learned feature information into the range of 0 to 1, which determines how much information the current gate unit passes. The specific structure is shown in Figure 2 below.

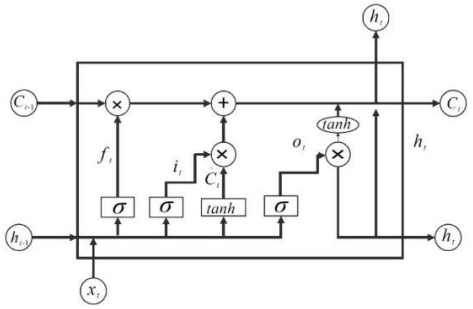

Figure 2. LSTM unit cell structure

The main formula for the door structure is:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right),\ i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{2}$$

$$\widetilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right),\ C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{3}$$

$$O_t = \sigma\left(W_O [h_{t-1}, x_t] + b_O\right),\ h_t = o_t * \tanh\left(C_t\right) \tag{4}$$

Among them, w_f, w_i, w_c and wo respectively represent the weight matrix of the forget gate, input gate, output gate and calculation unit state, b_f, b_i, b_c and b_o respectively represent the bias term of the above corresponding gate, ft represents the previous moment. Unit state pass rate, i_t represents the input gate update value, c_t represents the candidate value of the current state, and h_t denotes the output value of the current state. The forgetting gate determines how much information is retained from the cell state (c_{t-1}) in the previous moment based on the output value of the network at the previous moment (h_{t-1}) and the current input value (x_t). A value of 0 means all forgotten, and a value of 1 means all retained; the input gate passes sigmoid the layer function combines the network update value generated by the tanh layer function and the state candidate value c_t to determine which of the current network input values x_t are saved to the current unit state c_t. The forgetting gate and the input gate can be used to determine the current cell information is alternately updated and retained; the output gate determines the current output value h_t of the LSTM network by controlling the unit state h_{t-1} and the current input value x_t. The Sigmoid layer calculates the judgment conditions of the output gate, the tanh layer obtains the vector value of the current cell state, and the output of the current unit is obtained by multiplying the two layers.

3.2. Establishment of long- and short-term neural network models

The specific construction and implementation of the LSTM model is divided into three stages, as shown in the figure. 1) The data needs to be organized and the data set is divided; 2) The characteristic parameters and targets required for training and prediction need to be determined, and the LSTM The specific parameters of the network model and complete the construction of the model; 3) Use data to predict and test the model. The specific process is shown in Figure 3.

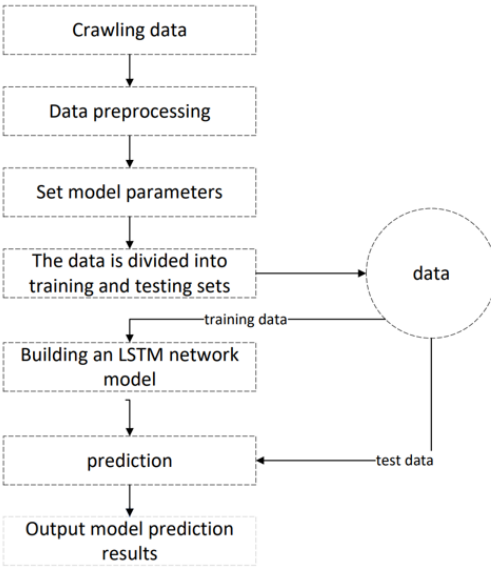

Figure 3. LSTM model prediction flow chart

The specific parameter settings of the model are shown in Table 3:

Table 3. Parameter settings of LSTM algorithm

Parameter	Parameter choice
Input node	6
Output node	5
Activation function	ReLU
Hidden layer function	10
Loss function	loss
Optimizer	Adam
Iterations	400
Time step	10
Initial learning rate	0.001
Learning decline factor	0.1

This article performs data preprocessing on the crawled data, cleans out outliers and supplements missing values, divides the training set and test set according to the ratio of 7:3, the input data is standardized and then sent to the LSTM neural network for processing. After multiple iterations, the LSTM neural network AQI prediction model is obtained.

4. Solving the model

4.1. Evaluation index

Taking RMSE, MAE, MBE and R2 as evaluation indicators, the smaller the RMSE, MAE and MBE, the closer the R2 value is to 1, indicating that the prediction accuracy of the model is higher. This paper chooses RMSE as the evaluation target. RMSE is the mean square root of the error between the prediction and the real. The formula for calculation is as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{m}} \tag{5}$$

MAE (mean absolute error, mean absolute error), the mean absolute error can better reflect the actual situation of the predicted value error. The calculation formula is as follows:

$$MAE = \sqrt{\frac{1}{m}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|} \tag{6}$$

MBE (mean bias error), the mean bias error is the average deviation is the mean of the predicted value and the real value. The formula for calculation is as follows:

$$MBE = \sqrt{\frac{1}{m}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)} \tag{7}$$

Correlation coefficient (R^2), which is the determinant, is the percentage of all the changes in a dependent variable that can be explained by a regression relation. The formula for calculation is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n} (\overline{y_i} - y_i)^2} \tag{8}$$

In the equation, n is the data quantity, the real value, the mean of the real value, and the prediction value.

4.2. Experimental results

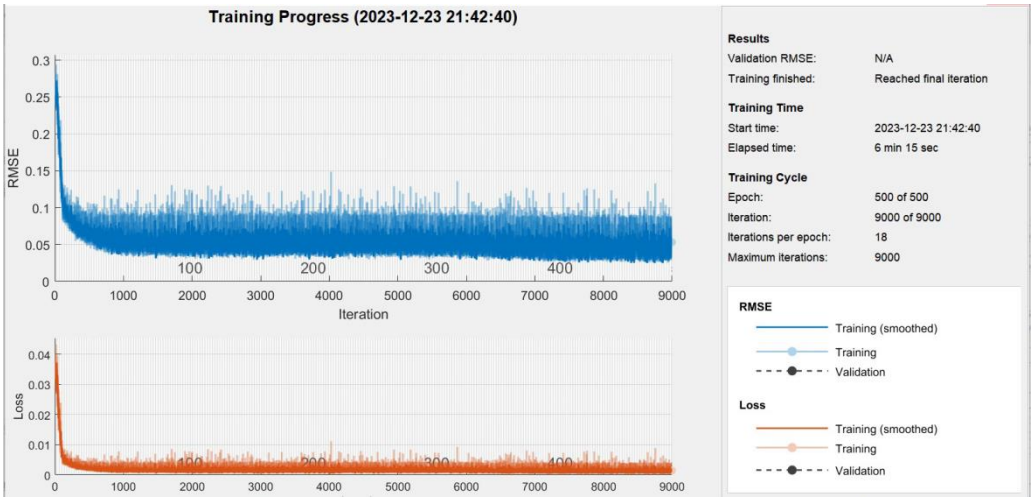

Figure 4. RMSE and loss function trend chart

As can be seen from Figure 4 above, the RMSE and loss function continue to decrease with the number of iterations and gradually reach convergence.

Prediction of air quality is carried out based on the divided data sets and determined network model parameters. Figures 5 and 6 below respectively show the comparison between the predicted values of the training set and the true value prediction results, and the comparison of the predicted values of the test set and the true value prediction results.

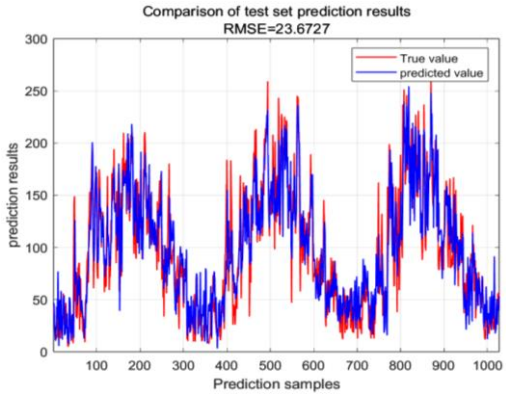

Figure 5. Comparison of test set prediction results

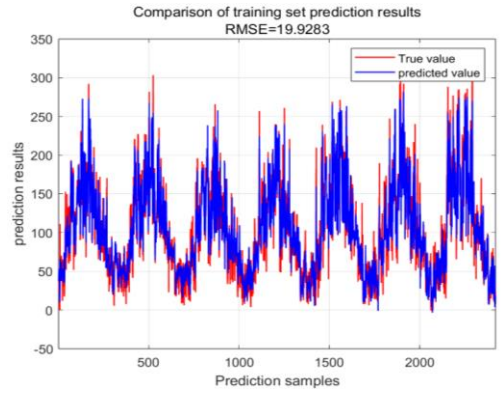

Figure 6. Comparison of training set prediction results

From the diagram, it can be seen that the RMSE of the forecast result of the training set is 19.9283, and the predicted RMSE of the test set is 23.6727, which shows that the forecast result is quite accurate.

Additional assessment parameters for the test suite and the training package are shown in Table 4:

Table 4. Other evaluation indicators

Evaluation index	Test set	Training set
R^2	0.8332	0.89075
MAE	17.4298	13.7233
MBE	-0.35743	-0.3855

Based on the model prediction, the four precision indexes RMSE, R2, MAE, and MBE of the LSTM model were 23.6727, 0.8332, 17.4298, -0.35743, and the four precision indexes were 19.9283, 0.8332, 17.4298, and - respectively. 0.3574. In general, this model has a higher precision and can predict more accurately.

4.3. Conclusion of the experimental part

(1) By repairing and normalizing the crawled data, ensuring the quality of the data and eliminating the potential impact of the dimensionality of the original variables on the prediction results, the accuracy of the model prediction is indirectly improved

(2) The prediction model proposed by this research institute exhibits excellent performance on both the training and prediction sets, with determination coefficients reaching high levels exceeding 0.8, respectively. This indicates that the model can explain about 80% or more of the variance in the target variable, thus successfully capturing key patterns and trends present in the data. This high degree of fit strongly supports the reliability and predictive ability of the model, and the model's performance on the prediction set is also satisfactory, which confirms its generalization ability to new data and enhances its credibility in practical applications.

(3) This model has strong generalization ability, and the fusion of LSTM and memory network innovation can not only improve the memory ability of the original neural network model, but also enrich the application types and fields of memory networks. In language processing tasks, the proposed model is well suited to processing high time sequences, such as machine translation, dialogue generation, coding and decoding. LSTM is an effective method to capture long term dependence of stock price and trading volume, which can improve forecast performance. In addition, LSTM can also be used to handle factors such as nonlinear relationships, seasonal changes, and unexpected events that affect the stock market, thereby improving the robustness and generalization of the model. In terms of traffic flow prediction, traffic flow data usually has obvious time series characteristics, and LSTM can effectively capture the dependency relationships between different time points and model the highly dynamic nature of traffic flow. Introducing LSTM into image processing tasks can achieve modeling of temporal image sequences, such as video analysis and action recognition.

5. Conclusion

The LSTM neural network model can deeply dig into and remember the relationship between the changes in the input sample parameters themselves and the AQI index, making learning more complete and prediction accuracy more accurate. The results obtained from the prediction model can not only help people comprehensively understand the emissions of air pollution sources in their area, but also help people timely grasp the factors affecting urban air quality, providing scientific basis for air quality monitoring, early warning and regulation.

References

[1] Cheng Hao. China's urban and rural human settlements: achievements and challenges [J]. Human Habitat, 2016 (03): 22-23.

[2] Gao Qingxian, Liu Junrong, Li Wentao, et al. Comparative study and enlightenment of air quality index (AQI) between China and the United States [J]. Environmental Science, 2015, 36 (04): 1141-1147.

[3] Ding Yuxian, Sun Weina, Niu Jiangang. Research on AQI characterization of Hohhot-Baotou urban agglomeration from 2014 to 2019 [J]. Inner Mongolia Science and Technology and Economy, 2020 (14): 7-8.

[4] Shen Haoyang, Wei Anlei, Wang Xiaowen, etc. Application of BP artificial neural network in prediction of SO2 mass concentration in ambient air [J]. Environmental Engineering, 2014, 32 (06): 117-121.

[5] Kong Y W, Sheng L F, Li Y P, et al. Improving PM2.5 forecast during haze episodes over China based on a coupled 4D-LETKF and WRF-Chem system [J]. Atmospheric Research. Volume 249, Issue. 2021. PP 105366.

[6] Hao Yongchao, Hou Xinrui, Yu Ling, etc. Research on the construction of air quality prediction model in Xingtai City based on neural network [J]. Computers and Information Technology, 2018, 26 (04): 69-74.

[7] Xie Yonghua, Zhang Mingmin, Yang Le, et al. Prediction of urban PM2.5 concentration based on support vector machine regression [J]. Computer Engineering and Design, 2015, 36 (11): 3106-3111.

[8] Si Zhijuan, Sun Baosheng, Li Xiaofang. Air quality prediction based on improved gray neural network combination model [J]. Journal of Environmental Engineering, 2013, 7 (09): 3543-3547.

[9] Shi Xiaowen, Jiang Hongxun. Research on LSTM model for air quality prediction with high accuracy and strong robustness [J]. Statistics and Decision Making, 2019, 35 (16): 49-53.

[10] Shi Xueliang, Li Liang, Zhao Qinghua. Air quality index prediction based on improved LSTM network [J]. Statistics and Decision Making, 2021, 37 (16): 57-60.

[11] Sun Chaoyun, Du Yaohui, Pei Lili et al. Air quality index prediction method based on inverse variance multiple model fusion [J]. Environmental Engineering, 2023, 41 (02): 197-204.

[12] Gao Song, He Zhuojun, Liu Ziyue, et al. Application of machine learning based differential fusion analysis in air quality prediction [J]. Electronic Measurement Technology, 2021, 44 (18): 85-92.

[13] Zhu Juxiang, Ren Mingyu, Gu Wei, et al. Air quality prediction modeling based on CEEMDAN-IGWO-CNN-LSTM [J/OL]. Computer simulation: 1-11 [2024-02-19] http://kns.cnki.net/kcms/detail/11.3724.tp.20230922.1501.004.html.