Prediction of Air Quality Based on KNN-LSTM

Zepeng Qin, Chen Cen, Xu Guo
Beijing Engineering Research Center for IoT Software and Systems, Beijing University of Technology, Beijing 100124, China
zepengqin@emails.bjut.edu.cn

Abstract. Since most of the existing air quality index (AQI) predicting models focused on prediction of the time series data of a single target monitoring station, they failed to consider the correlation and mutual influence among the air quality monitoring station sites and the spatio-temporal characteristics of air quality. And this will lead to a certain one-sidedness during air quality prediction of a particular site. Aimed at this problem, a short-term air quality prediction model based on K-nearest neighbor (KNN) and Long Short-Term Memory (LSTM) was proposed. The model firstly used KNN algorithm to select the time and space-related monitoring stations, then the air quality index sequences of these stations were constructed into data sets, followed by training and testing processes in the LSTM model, and finally the model was verified with real data. It is suggested that the prediction accuracy of the hybrid prediction model constructed in this paper is acceptable in terms of the space-time correlation, and it could be an alternative for further application of air quality prediction.

1. Introduction
In recent years, there have been more serious air pollution problems in many areas of China especially in winter, in which haze frequently invaded in many cities. Thus, monitoring and forecasting of air pollution indicators are of great importance. Intelligent Environmental Protection required researchers to rely on modern sensing technology, network technology and other means to effectively analyze and speculate environmental issues such as air quality for the purpose of providing a basis for scientific and rational decision-making. Since 2013, the air quality index (AQI) has been adopted in China to evaluate the quality of air quality and classify the city’s air quality. A larger AQI value means a more serious air pollution condition. On the foundation of environmental monitoring, the evaluation and prediction of air quality have become a topic of increasing concern to scholars at home and abroad. In this study, an integral short-term air quality prediction model based on K-nearest neighbor and Long Short-Term Memory was proposed. The AQI index was forecasted using the monitoring data of the provincial control station in Beijing as a sample, to accurately understand the trend of air quality changes.

2. Long Short-term Memory Neural Network
As a special type of RNN, the Long Short-Term Memory (LSTM) model was come up with solve the problem of the disappearing gradient of RNN in dealing with long-term dependence. LSTM added memory cells to each neural unit in the hidden layer, thereby making the memory information in the time series controllable. Each memory unit contained one or more memory cells and three gate controllers. Once the information is transmitted between the units of the hidden layer, it could control the memory and forgetting degree of the previous information and the current information through
several controllable doors, which contribute to the long-term memory function of RNN has and its important role in the practical application.

A schematic diagram of the basic structure of the LSTM model was depicted in Fig.1.

![Basic structure of the LSTM model](image)

Fig.1 Basic structure of the LSTM model

In this model, $x_t$ denoted the data input at time $t$, $h_{t-1}$ is the output data at time $t-1$, $s_{t-1}$ represented the state of the cell memorized at time $t-1$, and $h_t$ represented the output data at time $t$. The blue circle represented the point-to-point operation, and the yellow box was the learning neural network layer.

The LSTM memory unit concluded four elements: an input gate, an output gate, a Forgetting gate, and a loop-connected memory cell.

3. KNN Algorithm

K-nearest neighbor (KNN) is one of the machine learning algorithms that can be used for classification. Its main idea was to predict the correlation between samples through calculating the distance between different types of samples. The commonly used method for measuring distance included Euclidean distance, Markov distance, Manhattan distance, etc. Since its simplicity and easiness in implementation and fast in training, our research chose Euclidean distance to calculate and measure the correlation of AQI between ness degree between the target prediction station and nearby stations and further to select the target station. Several monitoring stations with higher degrees were combined with their air quality data for prediction. The Euclidean distance is calculated with equation (1) as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

I represented the target monitoring station, $j$ represented the $j$th monitoring station except the target monitoring station, $d_{ij}$ represented the Euclidean distance between the two monitoring stations, $k$ represented the specific time, $x_{ik}$ and $x_{jk}$ represented the AQI of $j$ monitoring station and $i$ monitoring at $k$ time.

4. KNN-LSTM Mixed Air Quality Index Prediction Model

4.1. Experimental data source and pretreatment

This paper mainly tests the hourly air pollutant concentration data of 35 provincial control stations in Beijing (including 12 national control stations). Fig.2 demonstrated a schematic diagram showing the spatial distribution of latitude and longitude of each site. The data used in this article is stored in a csv file and processed after being read by the python tool. The site name and number in the data were
marked and a timestamp was added to the data to simplify the manipulation of the data in subsequent experiments.

![Fig.2 Geographic distribution map of provincial air quality monitoring stations in Beijing](image)

The data set of this paper is the total air quality data of 35 provincial control stations in Beijing during 2017-2018. The data includes site information such as site name, site latitude and longitude, as well as the hourly concentration values of atmospheric pollutant indicators: PM2.5, PM10, SO2, NO2, O3, CO so on. The AQI value of each pollutant index was calculated using the relevant formula in the Technical Regulations for Ambient Air Index (Trial), and then select the max index as the air quality index of the site at that time. Select the site numbered S020, namely “front door”, as the target site to predict the air quality index at the time t of the station. The first 12 hours of the 35 city control stations in Beijing were selected to form a data matrix X, as shown below:

\[
X = \begin{bmatrix}
  x_1^{t-12} & x_2^{t-12} & \ldots & x_{M-1}^{t-12} & x_M^{t-12} \\
  x_1^{t-11} & x_2^{t-11} & \ldots & x_{M-1}^{t-11} & x_M^{t-11} \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  x_1^{t-6} & x_2^{t-6} & \ldots & x_{M-1}^{t-6} & x_M^{t-6} \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  x_1^{t-2} & x_2^{t-2} & \ldots & x_{M-1}^{t-2} & x_M^{t-2} \\
  x_1^{t-1} & x_2^{t-1} & \ldots & x_{M-1}^{t-1} & x_M^{t-1}
\end{bmatrix}
\]

\(x_i\) represented the data of the i-th workstation, \(\{x_i^{t-12}, x_i^{t-11}, \ldots, x_i^{t-2}, x_i^{t-1}\}\) represented the air quality indices sequence of i-th monitoring station in the first 12 hours of time t.

4.2. Experimental process
The core experimental process of the KNN-LSTM algorithm model used in this paper were divided into three parts. The flow chart of the whole experiment was illustrated in Fig.3:
In the first part, the air quality index of each monitoring point in the original data matrix $X$ was taken according to the equation (2) within the first 12 predicted hours, and the average air quality index of each monitoring station in the time period was obtained. Then, using the KNN algorithm according to equation (1), the Euclidean distance between the average air quality index of each monitoring station and the target monitoring station were calculated, and then these Euclidean distances were ordered according to the increasing distance.

$$\bar{x}_j = \frac{1}{35} \sum_{t=1}^{12} x^t_j$$  \hspace{1cm} (2)

In the second part, the air quality index of the target station was predicted by the LSTM model by inputting the data of the first 12 hours of the top $M$ stations which were closely related to the target station. The specific algorithm implementation process was listed as follows: $M$ takes 2 to 35 in turn, and each time the data of $M$ stations closely related to the target monitoring station were selected to form the following data matrix $I$:

$$I = \begin{bmatrix}
    x_1^{t-12} & x_2^{t-12} & \cdots & x_M^{t-12} \\
    x_1^{t-11} & x_2^{t-11} & \cdots & x_M^{t-11} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_1^{t-6} & x_2^{t-6} & \cdots & x_M^{t-6} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_1^{t-2} & x_2^{t-2} & \cdots & x_M^{t-2} \\
    x_1^{t-1} & x_2^{t-1} & \cdots & x_M^{t-1}
\end{bmatrix}$$

$x^t$ represents the air quality index set of $M$ monitoring points at time $t$, and $x^t_i$ represented the air quality index of the $i$-th station at time $t$ in the $M$ monitoring stations after screening by the KNN algorithm.

The data of the $M$ stations in October 2018 were selected as the training set and input into the LSTM model (indicated by $\Phi$) for prediction using the equation (3), and the error analysis is performed with the predicted values.
\[ \hat{X}_i^t = \Phi([x_1 \ x_2 \ \ldots \ x_{M-1} \ x_M]^T) \] (3)

In the final part, the M value with the corresponding smallest error value was output as the predicted value and the final prediction results were obtained.

In this study, the evaluation of the predicted results was processed using root mean square error (RMSE) and mean absolute percentage error (MAPE), which is defined as:

\[ \text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (x_t - \hat{x}_t)^2} \] (4)

\[ \text{MAPE} = \frac{1}{N} \sum_{t=1}^{N} \left| \frac{x_t - \hat{x}_t}{x_t} \right| \times 100\% \] (5)

### 4.3. Experimental results and analysis

In this paper, the data of October 2018 in the data set were selected, one data for each station per hour, and a total of 26040 data were selected as the training set training model. And the data from November 1, 2018 to November 20, 2018 were selected as the test model of the test set.

The comparison between the predicted value and the actual value within 48 hours in a certain two days were extracted in the experiment, as demonstrated in Fig. 4. It was depicted that the predicted value and the actual value of the mixed model show a good fitting degree, and the predicted value time series were basically in consistent with changing trend of the actual value, which denote that the predicted results are relatively accurate.

![Comparison diagram of experimental results](image)

**Fig. 4** Comparison between the predicted value and the real value of AQI in Beijing

In the experiment, different M values were selected to predict different results. In order to test the influence of different M values on the predicted results, M was taken from 2 to 35 in turn. Fig. 5 showed the root-mean-square error when M takes different values.

![RMSE at different values of M](image)

**Fig. 5** Root-mean-square error of prediction results when M took different values
The error analysis of 16800 iterations of the AQI in the KNN-LSTM hybrid algorithm model were shown in table 1.

Table 1. Error analysis table of mixed prediction model prediction AQI

| Test model | Root mean square error | Mean absolute percentage error | Time consuming |
|------------|------------------------|-------------------------------|----------------|
| KNN-LSTM   | 2.372                  | 6.592                         | 90.56          |

By comparing Fig.5 and table 1, it is suggested that the prediction model mixed with KNN-LSTM algorithm for atmospheric pollution prediction have a better prediction effect, and the prediction means were effective. As depicted in Fig.6, different values that M taken had different influences on the prediction results. The experiment showed that when M was 12, the prediction results have the smallest error and the highest prediction efficiency.

5. Conclusion
As atmospheric pollutants concentration was not only related to the pollutant source, but also associated with many factors such as atmosphere, terrain, it is difficult to accurately predict the atmospheric pollution. Therefore, this study came up with a kind of space-time prediction of air quality situation through the KNN-LSTM hybrid model of atmospheric pollution index –AQI. The main results were concluded as follows:

(1) In this study, the air quality index was predicted with the data of provincial air pollution monitoring stations in Beijing, and an air quality index prediction method based on the KNN-LSTM hybrid algorithm model was proposed, whose effectiveness was further verified.

(2) On the basis of time dimension, this study further took the influence of spatial distribution on the concentration of atmospheric pollutants into consideration. Through comparative experiments, it was verified that the prediction results obtained in this study were closer to the actual AQI value with a higher prediction accuracy.

(3) However, considering that the concentration of atmospheric pollutants was restricted by many factors, the prediction accuracy may be inaccurate when other factors showed a greater impact. Therefore, we suggested that the influence of meteorological, geographical and other factors should be studied in subsequent research. Using the combination of relevant data, suitable prediction models should be explored to improve the prediction accuracy.

References
[1] Wen Y, Yang X, Xu Y. Cloud-computing-based framework for multi-camera topology inference in smart city sensing system[C]. Proceedings of the 2010 ACM multimedia workshop on Mobile cloud media computing. ACM, 2010: 65-70.
[2] Ni X Y, Huang H, Du W P. Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data[J]. Atmospheric environment, 2017, 150: 146-161.
[3] Chen X, Liu X, Xu P. IOT-based air pollution monitoring and forecasting system[C]. International Conference on Computer and Computational Sciences. IEEE, 2015:257-260.
[4] Akhter Mohiuddin Rather,Arun Agarwal,V.N. Sastry. Recurrent neural network and a hybrid model for prediction of stock returns[J]. Expert Systems With Applications,2015,42(6).
[5] Jia H. Investigation into the effectiveness of long short term memory networks for stock price prediction[J]. arXiv preprint arXiv:1603.07893, 2016.
[6] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7): 2038-2048.