Prediction model of air pollutant concentration based on deep neural network: A case study of Fushun, Liaoning Province

Fei Lei¹, Deyu Gu²,³ and Xueli Wang²

¹Faculty of Information Technology, Beijing University of Technology, Beijing, China;
²College of Applied Sciences, Beijing University of Technology, Beijing, China
³Email:594647837@qq.com

Abstract. With the rapid development of the economy, the sources of air pollution are increasing, and the problem of air pollution is becoming more and more serious. Air quality prediction is a very effective means of predicting air pollution in the future, which helps the government regulatory authorities to provide early warning and protect people's physical and mental health. In this paper, a prediction model of air pollutant concentration based on deep neural network is proposed. With the concentration of PM₂.₅ as the prediction target, the neural network chooses bidirectional Long Short-Term Memory (LSTM) and fully connected neural network. First, historical meteorological data and the PM₂.₅ density from 2016 to 2017 were used as training data in this paper, which were obtained from website. Then, after pre-processing the input data, the data is transmitted to the network and trained multiple times to obtain network parameters that make the prediction effect better. Next, the network model is applied to the test set, and the test results are compared with the actual values to measure the prediction effect. Finally, by comparing with other prediction models, the results show that the proposed model performs better and has higher accuracy.

1. Introduction

In recent years, environmental pollution problems, especially air pollution, have received widespread attention [1-2]. Achieving a certain level of air pollution can cause varieties of health problems, especially for elderly, children, and people of all ages who are suffering from lung diseases such as asthma [3]. Airborne particulate matter (PM) is especially detrimental to health and has been estimated to cause between 3 and 7 million deaths every year, primarily causing cardiorespiratory disease [4].

For the prediction of the concentration of PM₂.₅, previously, statistical analysis models and chemical transport models were mostly adopted in air quality prediction. Statistical analysis and prediction models were mostly carried out by analyzing the laws of things and not knowing the internal change mechanism of things. Chemical transport models or dispersion models is based on the physical and chemical modeling of the atmosphere. These two methods are practical to some extent, and in recent years, some scholars have carried out further research [5-10]. However, due to the difficulty of modeling and the dependence on emission data, these two methods are difficult to achieve well prediction results. So it is very important to have such a good model for prediction the concentration of PM₂.₅, it is essential to make a model which describes and understand its complex relationships between PM₂.₅ concentrations as well as several variables affecting the concentration of PM₂.₅. For pollution forecasting, Machine Learning (ML) algorithms such as Support Vector Machine (SVM) and Decision Tree (DT) and Artificial Neural Network (ANN) were used [11-13]. However,
the existing methods have low accuracy and only use the meteorological data of one day and the pollutant concentration of the previous day as the input data of prediction model. Aiming at the problem of low accuracy of existing air quality prediction and considering that the meteorological conditions are constantly changing with the influenced relationship between before time and after time, a model named BD-LFC, which uses bidirection LSTM and fully connected network, is proposed in this paper.

The remainder of this paper includes the following contents, the second chapter introduces the experimental data preparation, the third chapter introduces the neural network selected in the text and the parameters used by the network, the fourth chapter compares the prediction results, and the fifth chapter is the conclusion and outlook.

2. Experimental data preparation

This chapter mainly explains the data source and data preprocessing method of this paper.

2.1. Historical meteorological data and air pollutant concentration data

The historical meteorological data of this article is mainly from the Weather Post Report website (http://www.tianqihoubao.com/), which provides historical meteorological and air quality information for 2,290 cities, counties and regions in 34 provinces and municipalities. The historical data of air quality comes from the national urban air quality trial release platform (http://beijingair.sinaapp.com/). The air quality data on the platform is updated daily, and the platform provides historical data of 1498 monitoring stations in 368 cities and regions across the country. This paper selects Liaoning Fushun as a typical resource-based city as the research object, and obtains historical meteorological data and air quality data from the above two channels. The platform provides historical data of six monitoring points. It also includes historical data of urban areas. In this paper, we uses urban historical data.

2.2. Data content and pre-processing

As mentioned above, this paper uses the PM$_{2.5}$ data from the weather data and air quality data of Liaoning Fushun from 2016 to 2017. All data is pre-processed before entering the network. In the monitoring data, there will be accidental abnormal data, such as the occurrence of null data with all monitored values of zero. Before using the data, the abnormal data should be eliminated or interpolated so as not to affect the prediction results. Then, the data were normalized to allow data of different orders of magnitude to be valued within a range to avoid the prediction error caused by the difference of orders of magnitude. All sample data were divided into training set and test set in appropriate proportion.

The meteorological data indicators include daytime weather and night weather, the highest and the lowest temperature of the day, daytime and night wind direction, daytime and night wind power, a total of eight indicators. It is also important here to discuss about the various weather conditions included in this paper for the concentration of PM$_{2.5}$ forecast and modelling. Wind Direction sometimes has a substantial effect on the air quality of a particular city as well as a region. Air quality can either become better or worse based on wind direction. If the wind is coming from an area with extremely less pollution, then the air quality improves a significant amount. But if the wind is coming from a region that is highly polluted it is likely to become worse. Low wind speed for a highly polluted region with multiple sources of pollution is a problem because the pollution stays in the same region rather than blowing away in the direction of the wind [14]. Strong wind speeds generally promote the transport and travel of pollutants rapidly to distant places. High Temperature during the summer days contribute to photochemical reactions especially in the case of particulate matter and ozone. Whereas rain can clean the air but can cause problems of acid rain and soil pollution. Since some data in meteorological data are not numerical types, but category types, non-numerical data is converted into form of numerical values corresponding established numerical mapping. The feature weather is classified as partly cloudy, sunny, rainy, cloudy, snow, dust, haze and fog. These are represented by values one to eight. Wind direction has been classified as north, northeast, east, south-
east wind, southerly, north-west, west wind, south-west wind eight types. Table 1 and Table 2 shows the numerical representation of weather and wind direction. Wind Power has been summarized into five levels of wind power density based on the national standard of wind power in China as shown in Table 3.

3. Deep neural network model

3.1. LSTM neural network model
Recurrent Neural Network (RNN) model is a kind of neural network with cyclic structure proposed by Rumelhart in 1986 [15]. RNN is capable of memorizing data and calculating the relationship between multiple time step data. The output of the network hidden layer at every moment depends on the past information. However, since problems of gradient explosion and gradient disappearance in the training of the classical RNN model [16], the classical RNN model is difficult to deal with long-term dependence problems.

| Table 1. Factors of input feature: weather. | Table 2. Factors of input feature: wind direction. |
|--------------------------------------------|-----------------------------------------------|
| Weather                                   | Factors                                      |
| Sunny                                     | 1                                            |
| Partly Cloudy & Cloudy                    | 2                                            |
| Light Rain                                | 3                                            |
| Moderate & Heavy Rain                     | 4                                            |
| Snow                                      | 5                                            |
| Dust                                      | 6                                            |
| Haze                                      | 7                                            |
| Fog                                       | 8                                            |

| Table 3. Factors of input feature: wind power density. |
|------------------------------------------------------|
| Wind Power Levels | Wind Power Density(W/m²) | Factors |
|--------------------|--------------------------|---------|
| <3 & <=3           | <150                     | 1       |
| 3-4                | 150-250                  | 2       |
| 4-5                | 250-300                  | 3       |
| 5-6                | 300-400                  | 4       |
| 6-7                | 400-1000                 | 5       |

Figure 1. Schematic diagram of LSTM neurons. 

Figure 2. Bidirectional RNN structure diagram.
LSTM neural network model is an extension of RNN, which is more controllable in training and solves the problem of long-term dependence [17]. Different from the classical RNN network, the cyclic neurons of the LSTM network have different structures.

The key part of LSTM network is the gates structure in its neural unit. In LSTM unit, information is added or deleted by gates structure which selectively allows information to pass through. LSTM cell has three gate structures which are used to maintain and update cell state included input gate, forgetting gate and output gate. The LSTM cell structure is shown in Figure 1.

3.2. Bidirectional LSTM neural network model
The bidirectional RNN model was proposed by Schuster in 1997 to solve the problem that unidirectional RNN cannot process the following information [18]. Unidirectional RNN can only process information in one direction. The basic idea of bi-directional cyclic neural network is that each training sequence is executed by two cyclic neural networks forward and backward respectively, and the output of the bi-directional RNN is the input of the next neural network layer. Figure 2 is the structure diagram of bidirectional RNN.

In 2005, Graves and Schmidhuber combined the BRNN with the LSTM cell and proposed the bidirectional LSTM. This model combines the advantages of the previous two models. In short term, the LSTM unit replaces the cyclic unit in the classical bidirectional RNN model. The bidirectional structure provides more efficient information through bidirectional reading and calculation of data, and has stronger capabilities of data processing [19].

3.3. BD-LFC model
In this paper, we propose a deep neural network model called BD-LFC using bidirectional LSTM and fully connected neural networks. The BD-LFC model structure is shown in Figure 3.

First, the bidirectional LSTM network first calculates the relation between the input data of two time steps to find out the internal relation between the sequence data, and then regressively calculates the predicted value of pollutant concentration through the subsequent full-connected neural network layer. We use the pre-processed meteorological data in the second part of this paper and the air PM2.5 concentration values from the previous day for each meteorological data as the input of model. In the fully connected neural network part, the ReLU activation function is used. Compared with the commonly used activation functions such as sigmoid and softplus, since the gradient of the non-negative interval of the ReLU function is constant, there is no gradient disappearance problem, so that the convergence speed of model is maintained at a stable state. Figure 4 shows the comparison.

3.4. Training and hyperparameter selection
In order to achieve a good prediction effect, it is very important to train the BD-LFC before implementing the prediction. We used stochastic gradient descent algorithm to train BD-LFC model with a small learning rate (0.0001).
In the process of model training, the problems of poor convergence and over-fitting are often encountered. Both of these problems seriously affect the prediction accuracy. Due to the insufficient number of neurons in deep neural network, the memory capacity of BD-LFC model is lacking, and the data of training set cannot be well fitted. This problem was solved when we increased the number of neurons per layer and reduced the learning rate. The over-fitting of deep neural network makes the generalization ability of model inadequate, and when the error on the training set is small, the error on the test set is large. We adopt two methods, early termination training and L2 parameter regularization, to solve the problem of over-fitting. When training large models that possess enough expressive ability and even become over-fitting, we often observe that training errors will gradually decrease over time, but test errors will rise again. Therefore, early termination is very effective for the selection of hyperparameters. Many regularization methods limit the learning ability of the model by adding a parameter norm penalty $\Omega(\theta)$ to the objective function $J$. The L2 parameter regularization strategy makes the weight closer to the origin by adding a regular term $\Omega(\theta) = \frac{1}{2} \| \omega \|^2_2$ to the objective function.

In the training, we take the sum of Mean Squared Error (MSE) and L2 regularization error as the final objective function to be optimized. The final objective function $f(x, x', \theta)$ is calculated as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (x_i - x'_i)^2$$

$$f(x, x', \theta) = MSE + L2 \cdot \Omega(\theta)$$

In formula (1), $n$ stands the total amount of data, $x_i$ is the true PM$_{2.5}$ concentration value, $x'_i$ stands the predicted output value of the model, and L2 stands L2 regularization coefficient.

In the training and application of deep neural network, it is difficult to select a set of appropriate hyperparameters, but it is certainly an important step, which directly affects the performance of neural network model. After several comparative experiments, we obtained a set of better model hyperparameters. The selection of hyperparameters is shown in Table 4.

**Table 4. Hyperparameter selection table.**

| Parameter      | Description                                           | Value |
|----------------|-------------------------------------------------------|-------|
| learning_rate  | Learning rate of stochastic gradient descent algorithm | 0.0001|
| time_steps     | Time step of input data                               | 2     |
| batch_size     | The number of data used for each gradient descent calculation | 8     |
| train_steps    | Total amount of training steps                         | 40000 |
| L2             | L2 regularization coefficient                         | 0.0003|
| LSTM_size      | LSTM unit size                                        | 60    |
| hidden1        | The number of nodes in the first layer of fully connected neural networks | 25    |
| hidden2        | The number of nodes in the second layer of fully connected neural networks | 15    |
| hidden3        | The number of nodes in the third layer of fully connected neural networks | 8     |
| output_unit    | The number of nodes in the output layer               | 1     |

4. Experiments and results analysis

The experimental environment parameters of this paper are as follows: Intel(R) Core(TM) i9-9900X CPU @3.5GHz; NVIDIA GeForce GTX 1080Ti 11GB GPU; Windows 10 operating system; 64GB of memory. All models, including the comparison algorithm model, are built under the TensorFlow (version 1.9) deep learning development framework. We use MSE to compare the accuracy between different algorithm models. The following algorithm model is used for comparison:
1. Back-Propagation (BP) Neural Network: BP neural network is a multilayer feedforward neural network trained according to the error back propagation algorithm. BP neural network is a multi-layer feedback propagation neural network algorithm. This kind of network can classify arbitrary complex models and perform well in the multi-dimensional function mapping problem.

2. Support Vector Machines (SVM): SVM is a machine learning algorithm that constructs hyperplanes for separating different classes and is generally used for analyzing data that has a categorical output variable. Whereas in case of a continuous numeric output variable we use regression analysis in place of classification called Support vector regression (SVR). An SVR model is used to obtain an approximate function \( g(x) \) from a given complex sample data \( G = \{ (x_i, y_i) \}_{i=1}^N \). The main idea is to first map the non-linearly separable data into a higher dimensional linearly separable feature space and then using this feature space for computation using linear programming [20].

3. BD-LFC model: the deep neural network model proposed in this paper

We compared the accuracy of different models in the test set. As shown in Table 5, BD-LFC model has higher accuracy compared with other algorithm models.

Table 5. Prediction accuracy of different models in the test set.

| Method        | Model      | MSE     |
|---------------|------------|---------|
| BP Neural Network |            | 0.012049|
| SVM           | Linear SVR | 0.017532|
|               | SVR        | 0.015612|
| BD-LFC        |            | **0.001115** |

To better illustrate the idea of our hybrid model, we tested the final prediction results of predicted value of the concentration of PM\(_{2.5}\) by different models. Figure 5 shows the comparison of prediction errors of SVM, BP neural network and BD-LFC model. It can be seen that the BD-LFC model has minor error.

Figure 6 shows the final prediction results of our model. We randomly selected more than 4 months from the average concentration of PM\(_{2.5}\) of each day of Fushun, Liaoning Province in 2018 and plotted the real value curve, which was shown by blue. The x-coordinate represents each day. In addition, according to the time steps and meteorological parameters mentioned above, the daily mean concentration value of PM\(_{2.5}\) in the first two days of a certain day and the corresponding meteorological parameters constitute the input data input model for prediction, and the predicted value is drawn into the predicted value curve, which is expressed by orange. It can be seen that the prediction results from BD-LFC model can track the trend of the concentration of PM\(_{2.5}\) and obtain an effective prediction of future air quality.

![Figure 5. PM\(_{2.5}\) prediction errors (MSE) of different models.](image1)

![Figure 6. Prediction results of BD-LFC model.](image2)
5. Conclusions
In this paper, after data preprocessing of meteorological data and PM$_{2.5}$ concentration data in the past two years, the two kinds of data were combined as input data to train the constructed deep neural network. After optimizing objective function by using the stochastic gradient descent algorithm, a model named BD-LFC of PM$_{2.5}$ concentration prediction with higher precision is obtained.

Considering that the changes of meteorological conditions and the concentration of PM$_{2.5}$ are continuous processes and have influenced relationship before and after, we adopt bidirectional LSTM and fully connected neural network to build model. We mainly predict the concentration of PM$_{2.5}$ in a day, and use the data of two time steps for modeling. After comparing several different algorithms, it is proved that the proposed model has higher precision.

We mainly predict the concentration of PM$_{2.5}$ in a day, and use the data of two time steps for modeling. In the following experiments, we will pay attention to the influence of changing the prediction time step on the prediction accuracy, and whether the model we proposed can still have a better prediction effect when more historical training data are obtained, or PM$_{2.5}$ concentration value per hour in a historical period of time is taken as input. These are all questions that need further consideration.

Reference
[1] Kelly F J, Fussell J C 2015 Air pollution and public health: Emerging hazards and improved understanding of risk Environmental Geochemistry and Health 37(4) 631-649
[2] Mclaren J, Williams I D 2015 The impact of communicating information about air pollution events on public health Science of the Total Environment 538 478-491
[3] Osama A. Ghoneim , Doreswamy , B.R. Manjunatha 2017 Forecasting of Ozone Concentration in Smart City using Deep Learning International Conference on Advances in Computing, Communications and Informatics (ICACCI)
[4] Rohde RA, Muller RA 2015 Air Pollution in China: Mapping of Concentrations and Sources PLoS ONE 10 8
[5] Xu Bin, Luo Liangqing, Lin Boqiang 2016 Ecological Indicators 63 346-358
[6] Boznan M, Lesjak M, Mlakar P 1993 Atmospheric Environment Part B Urban Atmosphere 27(2) 221-230
[7] N Pepe, G Pirovano, A Balzarini, A Toppetti, G M Riva, F Amato, G Lonati 2019 Atmospheric Environment: X 2 100020
[8] Belis C A, Karagulian F, Larsen B R, Hopke P K, 2013 Atmospheric Environment 69 94-108
[9] Hopke P K, 2016 J. Air Waste Manag. Assoc. 66 237-259.
[10] Ciarelli G, Aksoyoglu S, El Haddad I, et al. 2017 Atmos. Chem. Phys. 17 7653-7669.
[11] Uwe Schlink, Olf Herbarth, Matthias Richter, Stephen Dorling, Giuseppe Nunnari, Gavin Cawley, and Emil Pelikan 2006 Environmental Modelling & Software 21(4) 547-558
[12] D Birant. 2011 Journal of Environmental Informatics 17 1
[13] AS Luna, MLL Paredes, GCG de Oliveira, and SM Correa 2014 Atmos. Enviro. 98 98-104
[14] Bing-Chun Liu, Arihant Binaykia, Pei-Chann Chang, et al. 2017 PLoS ONE 12 7
[15] Rumelhart D E, Hinton G E, Williams R J 1986 Learning representations by back-propagating errors Nature 323(6088) 533-536
[16] Hochreiter S 1998 The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions International Journal of Uncertainty, Fuzziness and Knowledge-Based, Systems 6(2) 107-116
[17] Hochreiter S,Schmidhuber J 1997 Long short-term memory Neural Computation 9(8) 1735-1780
[18] Schuster M,Palwal K K 1997 IEEE Transactions on Signal Processing 45(11) 2673-2681
[19] Graves A & Schmidhuber J 2005 Neural Networks 18(5) 602-610
[20] Scho¨lkopf B, Smola AJ, Williamson RC, Bartlett PL 2000 New support vector algorithms Neural Computation 12 1207-1245