A Long Short-Term Memory Neural Network Model for Predicting Air Pollution Index Based on Popular Learning

Hong Fang[1(✉)], Yibo Feng[2,1], Lan Zhang[1], Ming Su[1], and Hairong Yang[3]

[1] College of Arts and Sciences, Shanghai Polytechnic University, Shanghai 201209, China
fanghong@sspu.edu.cn
[2] School of Mathematics and Statistics, Kashi University, Kashi 844000, China
[3] School of Mathematics and Statistics, Hefei Normal University, Hefei 230061, China

Abstract. With the acceleration of industrialization and modernization, the problem of air pollution has become more and more prominent, which causing serious impact on people's production and life. Therefore, it is of great practical significance and social value to realize the prediction of air quality index. This paper takes the Tianjin air quality data and meteorological data from 2017 to 2019 as an example. Firstly, random forest interpolation was used to fill in missing values in the data reasonably. Secondly, under the framework of deep learning in TensorFlow, Locally Linear Embedding (LLE) was used to choose multivariate data to reduce data dimensions and realize feature selection. Finally, a prediction model of the air quality index was established by using the Long Short-Term Memory (LSTM) neural network based on the data after dimension reduction. The experimental results show that the method has obvious effects in terms of dimensionality reduction and exponential prediction accuracy compared with Principal Component Analysis (PCA) and Back Propagation (BP).

Keywords: Air quality prediction · LLE · Deep learning · LSTM

1 Introduction

Due to climate change, industrial production and population migration, the air quality in the Beijing-Tianjin wing area is not optimistic. Air quality is closely related to people's health and production and life, haze often hits northern areas. The Air Quality Index (AQI) is an important indicator for people to understand the health of the air, especially PM2.5. The main pollutant which forms smog has become the primary goal of pollution control in Tianjin. Relevant research shows that in addition to pollutant emissions, air quality is also closely related to meteorological conditions such as wind speed, wind direction and temperature etc. Therefore, the analysis of the influence of meteorological changes on the concentration of atmospheric pollutants has important guiding significance for the control and management of pollution sources and the formulation of pollutant treatment plans.

Y. Nah et al. (Eds.): DASFAA 2020 Workshops, LNCS 12115, pp. 190–199, 2020.
https://doi.org/10.1007/978-3-030-59413-8_16

Li Xiaofei et al. [1] analyzed the characteristics of air quality changes in 42 cities in China from 2001 to 2010 and considered that there was a linear relationship between AQI and precipitation and wind speed. Wang liyuan et al. [2] took the air quality of Xichang city as an example, established a variable coefficient model and used the local linear estimation method to fit the model to quantitatively analyze the change of the influence degree of meteorological factors in Xichang city on local air quality with the seasonal change. The prediction based on statistical methods was based on statistics, and the future trend was predicted by analyzing historical air data. With the continuous development of machine learning, it is widely used in the prediction of linear and non-linear and time series models. Common methods mainly include clustering methods [3], support vector machines [4, 5] and neural networks [6] and so on. Among many deep learning models, the recurrent neural network (RNN) introduces the concept of time series into the design of the network structure, which makes it more adaptive in the analysis of time series data. Among many RNN variants, LSTM [7] model solves the problem of gradient disappearance and gradient explosion of RNN, so that the damaged neural network can effectively use long-term time series information.

According to literature review and related study, [8] mainly adopts six AQI indicators such as weather conditions, average temperature, maximum temperature, minimum temperature, wind speed and average wind speed as the initial data research. In view of the non-linear relationship among various indexes which affect the air pollution index, the LLE in the popular learning method is adopted in this paper to reduce the dimension of air data, and the reduced dimension is 8, which are respectively PM2.5, PM10, NO2, O3, weather conditions, minimum temperature, average temperature and average wind speed as input indexes. Based on the characteristics of air quality and meteorological data, an air pollution index prediction model based on LSTM was established in this paper. Compared with the BP neural network model, the LSTM neural network prediction has a lower error rate and it is more suitable for the prediction of air pollution time series data.

2 Algorithm Theory

2.1 LLE Algorithm

Dimension reduction in the field of machine learning refers to a mapping method to map data points in the original high-dimensional space to low-dimensional space. The essence of dimensionality reduction is to learn a mapping function $f : x- > y$, where x is the expression of the original data point, and the vector expression is currently used at most [9]. The common methods include PCA and linear discriminant analysis (LDA), but they are mainly used to deal with linear data sets, which have poor applicability for nonlinear data sets. In 2000, Saul et al. firstly proposed the LLE, which was an unsupervised learning algorithm and can be capable of calculating low-dimensional embedded coordinates of high-dimensional data, and traditional PCA, LDA and other sample-focused variance Compared with the linear dimensionality reduction method, LLE can efficiently maintain the linear characteristics of local samples during dimensionality reduction, and can be widely used in image recognition, high-dimensional data

visualization and so on. The main idea of LLE is to keep the linear reconstruction coefficients between the neighbors of high-order data and the linear reconstruction coefficients of the neighbors of low-order data to reduce the dimensionality. Based on this idea, LLE has the characteristics of global nonlinearity and local linearity.

The LLE algorithm is described as follows:

(1) Construct a neighborhood graph. The methods for determining the neighborhood usually include the KNN and ε-nearest neighbor methods. The K nearest neighbor points of each sample point are taken out, and each store is connected to its K nearest neighbor points to form a weighted neighborhood map in high-dimensional data. Solve the local linear matrix between the neighbors of the high-order data. This is achieved by solving the cost function $\Phi(w) = \sum_{i=1}^{N} \left| x_i - \sum_{j=1}^{k} w_{ji} x_{ji} \right|^2$ for getting the minimum w_{ij}. Considering s_i as a local covariance matrix, we can get

$$\begin{cases} S_i = (X_i - N_i)^T (X_i - N_i) \\ \Phi(w) = \sum_{i=1}^{N} w_i^T S_i w_i \end{cases} \tag{1}$$

Using the Lagrangian multiplier method, we can get

$$L(w_i) = w_i^T S_i w_i + \lambda\left(w_i^T 1_k - 1\right) \tag{2}$$

The value of local reconstruction coefficient can be obtained by differentiating and solving:

$$\begin{cases} \frac{\partial L(w_i)}{\partial w_i} = 2S_i w_i + \lambda 1_k = 0 \\ w_i = \frac{S_i^{-1} 1_k}{1_k^T S_i^{-1} 1_k} \end{cases} \tag{3}$$

(2) Map to a lower dimensional space. In order to optimally maintain the low-dimensional embedded coordinate Y of the weight matrix W, the cost function is:

$$\arg\min_{Y} \psi(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{k} w_{ji} y_{ji} \right\|^2 \tag{4}$$

The output is the matrix Y of low dimensional vectors: $Y = \left[y_1, y_2, \ldots, y_N\right]$, $d \times N$, W is represented by an N by N coefficient matrix w, for the near point i of j: $W_{ji} = w_{ji}$, so you can get: $\sum_{j=1}^{N} w_{ji} y_{ji} = \sum_{j=1}^{k} w_{ji} y_{ji} = YW_i$. Based on matrix knowledge, $A = [a_1, a_2, \ldots, a_N]$, we can get: $\sum_i (a_i)^2 = \sum_i a_i^T a_i = tr\left(AA^T\right)$, use the Lagrangian multiplier method again and derive: $MY^T = \lambda' Y^T$. We can see that Y is actually

a matrix formed by the eigenvectors of M. In order to reduce the data to the d dimension, we only need to take the eigenvectors corresponding to the minimum d non-zero eigenvalues of M, and the first smallest eigenvalue is generally close to 0. We discard them and finally keep the feature vector corresponding to the first $[2, d+1]$ feature values from small to large.

The classic Swiss roll data set in popular learning is used as the experimental data set [10], and the dimensionality reduction using LLE is shown in Fig. 1 to maintain the original topology.

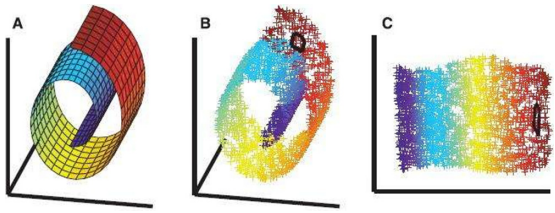

Fig. 1. Schematic diagram of LLE dimension reduction

2.2 Ten-fold cross check

K-fold cross-validation[11] divides the training set into K sub-samples. A single sub-sample is retained as the data of the validation model, while the other $K-1$ samples are used for training. The cross-validation was repeated K times, once for each subsample, with an average of K results, or using other combinations, resulting in a single estimate. The advantage of this method is that randomly generated subsamples can be repeatedly used for training and verification at the same time, and the results are verified once each time. Figure 2 is the schematic diagram of the ten-fold cross test.

Fig. 2. Schematic diagram of 10-fold cross-validation

2.3 LSTM Algorithm Principle

RNN is a powerful deep neural network, which has been widely used in natural language processing in recent years because of its remarkable effect on time-series data processing with long-term dependence. Training RNN when using back propagation algorithm [12] over time, in order to solve the problems in dealing with long-term dependence on the disappearance of the gradient, Hochreiter & Schmidhuber proposed LSTM model [13], compared with the traditional RNN, it has a more elaborate information transmission mechanism, can effectively solve the problem of long time dependence. At the same time, as the basic subunit in the encoder-decoder framework, LSTM can also realize the encoding and decoding of time series data, and replace the hidden layer neurons in RNN with memory units to realize the memory of past information. Each memory unit contains one or more memory cells and three door controllers. The structure of LSTM is shown in Fig. 3.

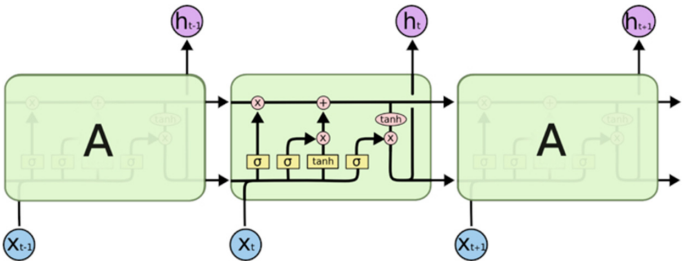

Fig. 3. LSTM structure diagram

The basic structure of LSTM and RNN is similar, with chain structure. However, LSTM has a very useful mechanism, namely forgetting gate. In LSTM, not only a single network layer is used, but four modules are interacted in a special way [14], as shown in Fig. 4.

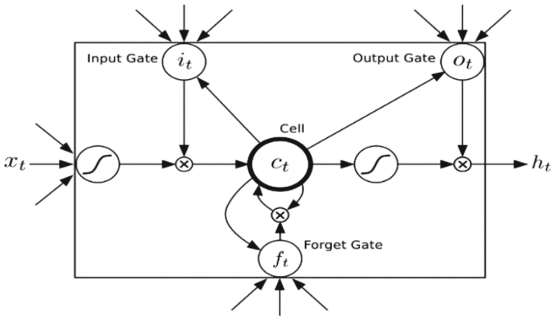

Fig. 4. Schematic diagram of LSTM neural network structure

In the LSTM model, X_t, h_t are respective the input and output of the LSTM network at time t, $t = 1, 2, 3, \cdots$ and the output h_t is the LSTM memory unit, iteratively calculated

by the following formula:

$$i_t = \sigma\left(\sum W_{xi}x_t + \sum W_{hi}x_{t-1} + \sum W_{ci}x_{t-1} + b_i\right) \tag{5}$$

$$f_t = \sigma\left(\sum W_{xf}x_t + \sum W_{hf}x_{t-1} + \sum W_{cf}x_{t-1} + b_f\right) \tag{6}$$

$$o_t = \sigma\left(\sum W_{xo}x_t + \sum W_{ho}x_{t-1} + \sum W_{co}x_{t-1} + b_o\right) \tag{7}$$

$$\begin{aligned} c_t &= f_t c_{t-1} + i_t \tanh\left(\sum W_{xc}x_t + \sum W_{hc}x_{t-1} + b_o\right) \\ h_t &= o_t \tanh(c_t) \end{aligned} \tag{8}$$

Among them, i_t, f_t, c_t, o_t represent the vector values of the input gate, forget gate, output gate, and memory unit, w is the weight of various input loops, b is the offset vertex, σ is the sigmoid function, which is used to control the flow of units Weight, ranging from 0 to 1, t is the intercellular activation vector. *Tanh* is a hyperbolic tangent function, and $\odot$ represents an element-wise multiplication operation.

3 Prediction Method of Long-Term and Short-Term Memory Neural Network Based on Popular Learning

3.1 Random Forest Interpolation

Because of the inconsistency between the air quality data and the meteorological data recording time, there are missing values in the acquired data. In order to avoid the impact of missing values on the prediction of the data, choose to use the random forest interpolation method [15] after comparing with other interpolation methods. This method can effectively process high-dimensional data and effectively extract auxiliary variable information, which is suitable for processing issuing data in the context of big data.

3.2 Dimension Reduction

The data used in this article is the meteorological observation data of 30 national meteorological observatories in Tianjin from January 1, 2017 to December 31, 2019. All of the data comes from Tianjin Meteorological Bureau, including weather conditions, wind, wind direction, average wind speed, minimum temperature, maximum temperature and average temperature. The air quality data of Tianjin used in this paper are the daily air quality index of 35 environmental monitoring stations in Tianjin and the concentration data of 6 pollutants (PM2.5, PM10, SO2, CO, NO2, O3). The data comes from Tianjin Environmental Protection Bureau and Tianjin Environmental Protection Monitoring Center.

In this paper, the average classification accuracy before or after dimension reduction is compared by the 10-fold cross-validation technique to verify the effectiveness of the dimension reduction method. The data dimension after dimension reduction is 8 and

they are PM2.5, PM10, NO2, O3, weather conditions, minimum temperature, average temperature and average wind speed.

As can be seen from Fig. 5, the accuracy of PCA is higher than that of LLE only when the dimension is reduced to 2. When the dimension is reduced to other number, the accuracy of LLE is significantly better than that of PCA. And the accuracy of the model reached the highest when the dimension is reduced to 8 with an average classification accuracy of 85.07 %. Therefore, LLE is used in the dimensionality reduction of data. This method is effective and the dimension of the data is reduced to 8 with reducing the data volume by about 33 %.

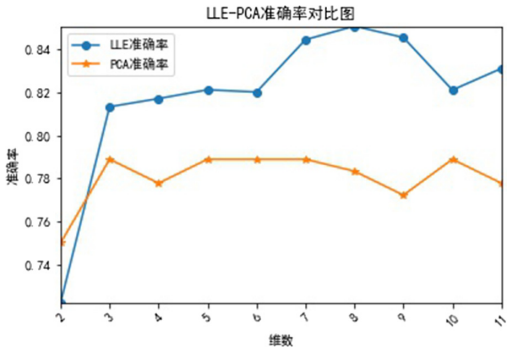

Fig. 5. Accuracy results of LLE and PCA in different dimensions

3.3 LSTM neural network prediction

We choose to use the LSTM neural network and take the first 600 data as the training set, the last 100 data as the test set and then normalize the data. Set the time step, batch size and learning rate for each batch of training samples to 2, 32 and 0.0006 respectively. Then define the neural network variables, use the hyperbolic tangent function (*tanh*) as the activation function in the hidden layer and also use the average absolute error (MAE) loss function and the efficient Adam gradient version of Adam to compile the model. When training this model, the number of iterations (epochs) is 1000 and the comparison between the predicted value (green line) and the real value (red line) is shown in Fig. 6.

3.4 BP neural network prediction

To predict the value of the air quality index AQI, through the previous use of LLE for dimension reduction, 8 factors such as PM2.5, PM10, NO2, weather and wind speed will be used as the impact factors of the API prediction model of the day. Normalizing the data, because of different types of data having different magnitudes and dimensions. Finally, each principal component is used as the input data of the BP neural network model.

The BP neural network prediction model in this paper uses three layers to build, one input layer, one hidden layer, and one output layer. The first layer of activation functions

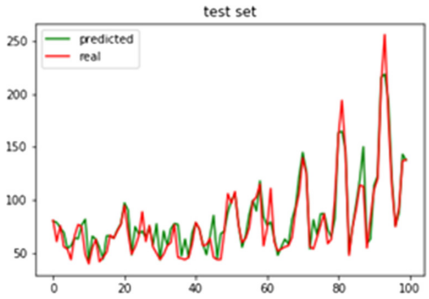

Fig. 6. Prediction Results by LSTM (Color figure online)

uses exponential, and the input variables are two-time steps with 8 characteristics. The second layer uses the softmax activation function for output and the last layer uses the relu activation function for output. Finally, the average absolute error (MAE) loss function and the efficient Adam version of stochastic gradient descent were used to compile the model. The training of the model uses the first 600 data sets as the training set and the last 100 data sets as the prediction set. The image comparison between the final prediction result and the real result is shown in Fig. 7.

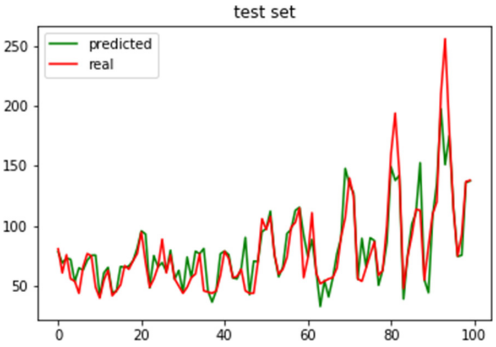

Fig. 7. Prediction Results by BP

To compare LSTM and BP more clearly, we use the indexes of RMSE and Pre-Rate. Test RMSE (standardized) is the root mean square error of the unit data, it represents the model error value of the normalized data with the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2} \tag{9}$$

where y_i is the normalized value and $\hat{y}_i$ is the normalized mean. Test RMSE is the root mean square error of the actual data, it represents the model error value of the real data. Pre_rate is defined by the following formula:

$$\text{Pre\_rate} = 1 - \sum_{i=1}^{n}\left(y_{pre_i} - y_{real_i}\right)^2 \Big/ \sum_{i=1}^{n}\left(y_{pre_i} - \overline{y}_{real}\right)^2 \tag{10}$$

where y_{pre_i} represents the predicted AQI and y_{real_i} represents the real AQI. Pre_rate represents the similarity between the predicted data and real one. We calculate the predicted value of the root mean square error of normalized data set, root mean square error of real data set and the degree of match between the predicted value and the true value in the prediction results of BP and LSTM, as shown in Table 1, from which we can see that the LSTM model has higher accuracy and robustness.

Table 1. Comparison of BP and LSTM results

Method	Predicted data set test RMSE (Standardized)	Real data set test RMSE	Pre_Rate
BP	0.057	18.284	0.7654394507408142
LSTM	0.047	13.308	0.8757250383496284

4 Conclusion

Aiming at predicting the air quality data and meteorological data of Tianjin, this paper uses LLE method to reduce the dimension, and proposes an air pollution index prediction model based on LSTM. The method is based on a deep learning network, which uses environmental big data to train the model, which fully exploits the semantic features in air quality data, and realizes air pollution prediction based on environmental big data. By comparing the root mean square error and prediction accuracy of LSTM and BP, the validity of LSTM model in the prediction of air pollution in this paper is verified.

Acknowledgments. This paper is funded by the program of the key discipline "Applied Mathematics" of Shanghai Polytechnic University (XXKPY1604).

References

1. Li, X., Zhang, M., Wang, S., Zhao, A., Ma, Q.: Analysis on the change characteristics and influential factors of China's air pollution index. Environ. Sci. **33**(06), 1936–1943 (2012)
2. Wang, L., Li, Y.: Analysis of meteorological factors affecting air quality in Xichang city-a study based on variable coefficient model. J. Yangtze Normal Univ. **33**(05), 98–102+112 (2014)
3. Çevik, H.H., Çunkaş, M.: Short-term load forecasting using fuzzy logic and ANFIS. Neural Comput. Appl. **26**(6), 1355–1367 (2014). https://doi.org/10.1007/s00521-014-1809-4
4. Li, P., Tan, Z., Lili, Y., et al.: Time series prediction of mining subsidence based on a SVM. Int. J. Min. Sci. Technol. **21**(4), 557–562 (2011)
5. Pai, P.F., Hong, W.C.: Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. Electric Power Syst. Res. **74**(3), 417–425 (2005)
6. Yu, W., Chen, J.: Application of BP artificial neural network model in forecasting urban air pollution. Pollut. Control Technol. **26**(3), 55–57 (2013)

7. Graves, A.: Long short-term memory. In: Graves, A. (ed.) Supervised Sequence Labelling with Recurrent Neural Networks, pp. 37–45. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-24797-2_4
8. Zheng, Y., Liu, F., Hsieh, H.P.: U-Air: when urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2013)
9. Ma, R., Wang, J., Song, Y.: Multi-manifold learning based on nonlinear dimension reduction based on local linear embedding (LLE). J. Tsinghua Univ. (Sci. Technol.) **48**(04), 582–585 (2008)
10. Lan, W., Wang, D., Zhang, S.: Application of a new dimensionality reduction algorithm PCA_LLE in image recognition. J. South Central Univ. Nationalities (Nat. Sci. Ed.) **39**(01), 85–90 (2020)
11. Tang, Y., Xu, Q., Ke, B., Zhao, M., Chai, X.: SVM model optimization of blasting block size based on cross-validation. Blasting **35**(03), 74–79 (2018)
12. Zeng, H.: Research on prediction model of environmental pollution time series based on LSTM. Huazhong University of Science and Technology (2019)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
14. Duan, D., Zhao, Z., Liang, S., Yang, W., Han, Z.: Prediction model of PM2.5 concentration based on LSTM. Comput. Meas. Control **27**(3), 215–219 (2019)
15. Meng, J., Li, C.: Interpolation of missing values of classification data based on random forest model. Stat. Inf. Forum **29**(09), 86–90 (2014)