A Self-organizing LSTM-Based Approach to PM2.5 Forecast

Xiaodong Liu[1], Qi Liu[1], Yanyun Zou[2] and Guizhi Wang[3]

[1] School of Computing, Edinburgh Napier University, UK
q.liu@napier.ac.uk
[2] School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, China
[3]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, Jiangsu, China

Abstract. Nanjing has been listed as the one of the worst performers across China with respect to the high level of haze-fog, which impacts people's health greatly. For the severe condition of haze-fog, $PM_{2.5}$ is the main cause element of haze-fog pollution in China. So it's necessary to forecast $PM_{2.5}$ concentration accurately. In this paper, an artificial intelligence method is employed to forecast $PM_{2.5}$ in Nanjing. At the data pre-processing stage, the main factors among the air pollutants (O3, NO2, SO2, CO, etc.) as well as meteorological parameters (pressure, wind direction, temperature, etc.) that affect $PM_{2.5}$ are selected, and these factors of previous hours are as input data to predict $PM_{2.5}$ concentration of next hours. Considering the air pollutants and meteorological data are typical time series data, a special recurrent neural network, which is called long short term memory (LSTM) network, is applied in this paper. To determine the amount of nodes in the hidden layer, a self-organizing method is used to automatically adjust the hidden nodes during the training phase. Finally, the $PM_{2.5}$ concentrations of the next 1 hour, 4 hours, 8 hours, and 12 hours are predicted separately by using the self-organizing LSTM network based approach. The experimental result has been validated and compared to other algorithms, which reflects the proposed method performs best.

Keywords: Haze-fog, $PM_{2.5}$ Forecasting, Selecting Main Factors, Time Series Data, LSTM network, Self-organizing Algorithm.

1 Introduction

Haze-fog is not only related to meteorological conditions, but also has a non-negligible relationship with human activities. Once the emission exceeds the atmospheric circulation capacity and carrying capacity, the concentration of fine particles will be getting to high. As a result, it is easy to have a large range of haze-fog. The greatest impact of haze-fog is the human health, it’s easy to affect the respiratory tract of the body and causes various diseases. In Nanjing, there are many sources of pollution such as building construction, vehicle exhaust, coal power generation and so on. And because Nanjing is a downwind area, the pollution in the upwind area will be imported into Nanjing. As a result, Nanjing is one of the most air polluted cities. So this study aims to forecast $PM_{2.5}$ concentration of Nanjing.

The early $PM_{2.5}$ prediction methods were mainly based on the original statistical methods. Fuller et al. [1] use the average of the pollutants of API, and statistics the linear relationship between the factors and the $PM_{2.5}$ and PM_{10}, so as to realize the forecasting of $PM_{2.5}$. At the same time, this linear method is also used for the prediction of other air pollutants. Combined with the local climate, API air pollution can be predicted through empirical judgement and linear regression. Jian et al. [2] find the correlation between meteorological factors, that is, the humidity is positively related to haze-fog, and the wind speed is negatively related to haze-fog. It is proved that auto-regressive integrated moving average (ARIMA) model can effectively explore the relationship between haze and meteorological factors. Kibria et al. [3] use naive Bayes to integrate the distribution of $PM_{2.5}$ in space. Dong et al. [4] use the mathematical model based on the hidden Markov function to predict the $PM_{2.5}$ concentration. The prediction results show that the predicted value of $PM_{2.5}$ can fit the real value better.

To improve the accuracy of prediction, artificial neural network (ANN) has been widely used in this field. Zhu et al. [5] put forward an improved BP neural network algorithm, combining the auto-regressive and moving average (ARMA) model with BP neural network to predict $PM_{2.5}$ concentration. Venkadesha et al. [6] combine genetic algorithm and BP neural network to fuse multiple time domain meteorological factors, and determine the duration and resolution of prior input data, and improve the accuracy of prediction. Zheng Haiming et al. [7] use the radial basis function (RBF) neural network to predict the concentration of $PM_{2.5}$. The results show that the prediction accuracy is better than BP neural network. Mishra et al. [8] combine the Principle Component Analysis (PCA) and artificial neural network to get the correlation between meteorology and air pollutants variables, so as to predict the concentration of NO_2 in the air. Mishra et al. [9] use non-meteorological parameters (CO, O_3, NO_2, SO_2, $PM_{2.5}$) and meteorological parameters to make the fusion analysis combining artificial intelligence to forecast haze-fog, it is concluded that compared with the artificial neural network and multilayer perceptron model, NF model based on the artificial intelligence can better predict the urban haze-fog events in Delhi, India. Neto et al. [10] use artificial neural networks to recursively analyse residual residuals to find current patterns, so the accuracy of predicting the concentration of $PM_{2.5}$ and PM_{10} is improved. Shanshan Zhou et al. [11] use recurrent neural network (RNN) to predict $PM_{2.5}$ concentration. Compared with fuzzy neural network (FNN) and RBF feed-forward neural network, the experimental results show that RNN is outstanding. Bun Theang Ong et al. [12] put forward a new training method for automatic encoder, which is designed for time series prediction, to enhance the deep recurrent neural network (DRNN). The experiment shows that DRNN is better than the typical and most advanced automatic coder training method used in the time series prediction. Liu et al. [13] use the comprehensive prediction model to forecast the $PM_{2.5}$ concentration using the autoregressive moving average (ARIMA), artificial neural network (ANNs) model and exponential smoothing method (ESM).

Although the above literatures can achieve good prediction results, air pollutants and meteorological data are typical time series data. It is difficult to reflect the correlation between data and time. In addition, how to decide the number of hidden nodes is still a major challenge for researchers.

2 Basic knowledge

2.1 Time Series Prediction

Meteorological data are typical time series data. Time series prediction is a prediction that extends to the future according to the historical data of the past. According to the process and regularity of the time series data, it establishes a mathematical model that is more accurate to reflect dynamic dependency relations. Then after the learning historical data period, this model can make a prediction of the trend of time series development.

The general steps of the time series prediction are as follows:

(1) After the collected historical data are reorganized, the time series data are formed after necessary pre-processing (noise reduction, removal of singularity, etc.). According to the composition and different influence factors of the time series, they're usually divided into four categories:

Long-term trend: the tendency to maintain steady growth or decline within a period of time is known as a long-term trend. For example, the recent growth in the price of bitcoin, or the declining price of electronic consumer goods.

Seasonal variation: the time series changes obviously according to the change of the four seasons. For example, the four seasons of sunspots, the company's sales volume changes within one year and the rainfall trend in this paper. And the time interval of the change is not fixed, it can be a month, a quarter, or even a day.

Cyclical change: cyclical change often appears inseparable with long-term trend. For example, the process of the replacement of the ancient Chinese dynasties usually includes the stages of recovery, prosperity, flourishing age, decline, and destruction.

Irregular change: irregular change refers to the part of random changes of the time series, and there is no law to follow between these data. It's often impossible to use mathematical models to fit their changes.

(2) Time series analysis. The formation of continuous values in time series is closely related to a number of factors, and it is usually not the result of only one factor. Therefore, when new time series data are obtained, it is usually necessary to analyse the interaction between their internal factors, and then get the implied features in the data.

(3) Characterizing the characteristics of extraction. According to the long-term trend, seasonal variation and cyclic variation of the time series, the approximate mathematical models are selected to characterize them.

(4) The training and prediction of the model. The correctness of the mathematical model is verified by historical time series, and the model is properly adjusted. After reaching the error requirement of the model, it can be used to predict the time series.

2.2 Long Short Term Memory Network (LSTM)

The input of recurrent neural network (RNN) hidden layer overlay the original data information as time goes by, which leads to the loss of contextual information. Therefore, in practical applications, the range of contextual information that general recurrent neural network structure can use is limited, resulting in gradient vanishing problem. To solve this problem, LSTM network is brought up. LSTM is very popular at the moment. It is not essentially different from the general RNN structure, but uses a different node, called "memory block" to replace a hidden layer node of general RNN.

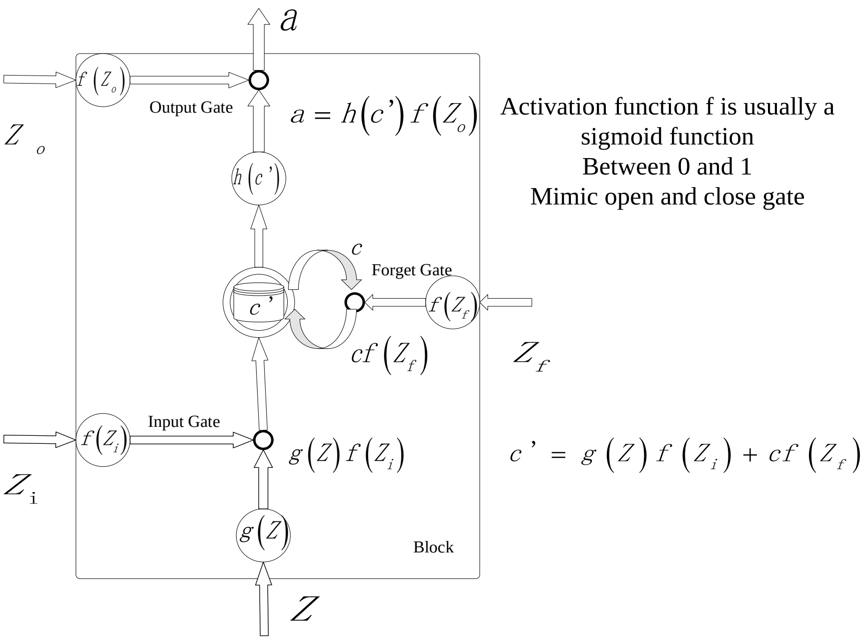


Fig. 1. Structure of memory block

Figure 1 shows the architecture of the memory block. There are three gates in the block. When the outside wants to write to memory cell, it must go through an input gate. Only when the input gate is opened, memory gate can be written. The input gate is opened or closed, which is learned by neural network itself. In the same way, there is an output gate, which is also learned by neural network. There is also a forget gate to decide when to forget the contents of memory cell, and the condition of opening or closing is also learned by neural network itself. Z, Z_i, Z_f, Z_o are scalers, which are derived from the inner product of input vectors and weight vectors adding to the bias. Weight vectors and bias are learned by gradient descent from training data. The function f often chooses sigmoid function, because the output value is between 0 and 1, which can indicate the opening degree of the gates (0 is closing; 1 is opening). When f(Z_i) = 1, g(Z) can be input. Instead, f (Z_i) =0 is equivalent to no input. Similarly, f(Z_o) controls the output of value. f(Z_f)=1 is equivalent to remember the previous value C in memory cell; when f (Z_f) =0, it is equivalent to forget value C. The values in the memory cell are updated by the formula in figure 2.

So the LSTM memory block can be seen as special neuron (4 inputs, 1 output), and 4 inputs refers to the signal that the outside wants to put into the memory cell as welas three control gates signal.

3 Self-organizing Algorithm

3.1 Sensitivity Analysis

In this paper, the amount of hidden nodes in LSTM network is adjusted by a self-organizing algorithm [14] during training phase. In this algorithm, a crucial method, sensitivity analysis (SA), is adopted to calculate how important the hidden nodes are. Sensitivity analysis is always used to judge the degree of dependence between output and input. So it's feasible to add or delete hidden nodes according to sensitivity analysis.

Before knowing the working mechanism of SA, it's necessary to understand the internal feedback dynamic characteristics of LSTM network. Figure 2 displays decomposition of network. It's divided into two parts. The first part is the self-circulatory part of hidden layer nodes. And the second part is a direct relationship between hidden layer nodes and output layer nodes.

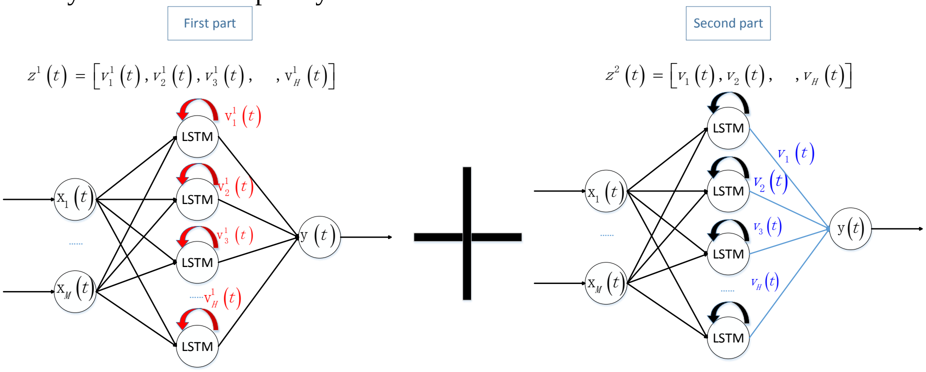


Fig. 2. Decomposition of LSTM network

The numerical calculation method adopted by this paper is as follows

$$S_h = \frac{Var[E(Y|Z_h)]}{Var(Y)} \quad , \quad \mathrm{h} = (1,2,\cdots,H) \tag{1}$$

where Z_h represents h_{th} input factor, Y is equal to the output of this model. $E(Y|Z_h)$ is the expected variance under the condition of output Y. And Var[E(Y|Z_h)] is the variance of all the feasible values of Z_h. $Var(Y)$ represents the variance of output Y. The result of this formula, S_h, is the general impact of that element on the reply.

For LSTM network, the input data for SA is consist of 2 sections: the indirect and direct elements. The indirect input element is $z_1(t) = [v_1^1(t), v_2^1(t), v_3^1(t), \cdots, v_H^1(t)]$.

And the direct input element can be get by $z^2(t) = [v_1(t), v_2(t), \cdots, v_H(t)]$. Then the numerical definition of sensitivity for indirect elements is revised as:

$$S_h^1(t) = \frac{Var_h\left[E\left(y(t)|Z_h^1 = v_h^1(t)\right)\right]}{Var\left(y(t)\right)}, \tag{2}$$

Meanwhile, the numerical definition of sensitivity with direct elements is modified as:

$$S_h^2(t) = \frac{Var_h\left[E\left(y(t)|Z_h^2 = v_h(t)\right)\right]}{Var[y(t)]}, \tag{3}$$

where $v_h(t)$ is the output data of the h_{th} hidden layer node at time t, and the variances $Var[y(t)]$, $Var_h\left[E\left(y(t)|Z_h^2 = v_h(t)\right)\right]$ can be computed as in (3) and (4).

According (2) and (8), the overall sensitivity value ST_h of the h_{th} hidden node is:

$$ST_h(t) = \alpha S_h^1(t) + \beta S_h^2(t), \tag{4}$$

where α and β are the judgment constants, $\alpha \in$ [0, 0.2], $\beta \in$ [0.5,0.8].

3.2 Growing and Pruning Algorithm

The growing and pruning algorithm is to add or delete hidden layer nodes by using SA to achieve self-organizing ability of the LSTM network during training phase. By using this algorithm, the number of hidden layer nodes can be adjusted to the best. So the structure of the network is able to satisfy the high precision prediction condition.

The general idea of this method is: first, generate some hidden layer nodes randomly. Second, calculate sensitivity index of each node. Third, if the sensitivity index of a node is lower than a certain threshold, the node will be pruned; instead, if the result is not satisfied to get the ideal result, some new hidden nodes will be added to the network, according to those hidden nodes whose sensitivity values are large. At the same time, associated connection weights are updated.

Growing Step. At time t, if there are H nodes in the hidden layer, the RMSE for the neural network is $\mathrm{E(t)} > \delta(t)$ $(\delta(t) = t^{-0.65})$, so it suggests that the learning process is not able to get the ideal result and a new hidden node is needed for the current construction. The final sensitivity value is

$$ST_h(t) = max\{ST_j(t)\}\ , j = 1,2,\cdots,H \tag{5}$$

$$\mathrm{e}(t) = y_d(t) - y(t)$$

h is the node, which has the largest overall sensitivity value.

Pruning Step. Like the above step, at time t, if the overall sensitivity index is $ST_h < \tau$ (τ is the pruning threshold), then the hth node needs to be deleted. And the weights of the hth node are updated as well. Experiment

4 Experiment

4.1 Data Pre-processing

To predict PM2.5 concentration, this study uses hourly files from the ground automatic station in Nanjing, which include meteorological data as well as air pollutants. Raw data contains 30 factors, such as O_3, NO_2, CO, SO_2, pressure, relative humidity, temperature and so on. However, there are some missing values and outliers, this paper uses the average value of neighbour data to replace these values. Then, after the standardization, in order to select the main factors with respect to $PM_{2.5}$ concentration, a selection method called Mutual Information (MI) is adopted. This method can calculate that whether there is a relationship between the two variables X and Y, as well as the strength of the relationship. As a result, the MI values of these factors are over 2.5: O_3, NO_2, $PM_{2.5}$, pressure, wind direction of the instant maximum wind speed, temperature, wind direction of maximum wind speed, horizontal visibility, body temperature. Figure 3 shows the distribution of these factors.

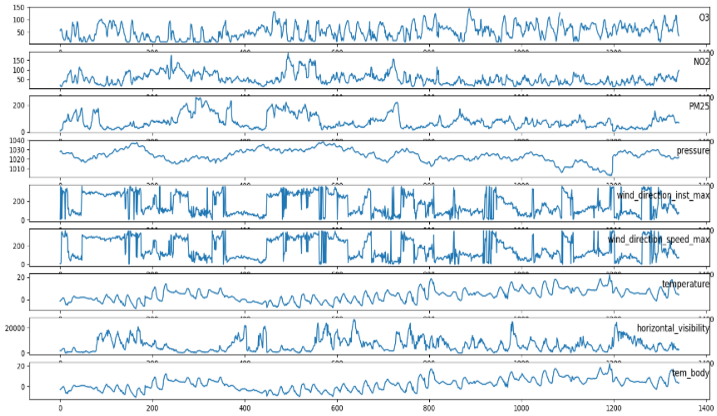
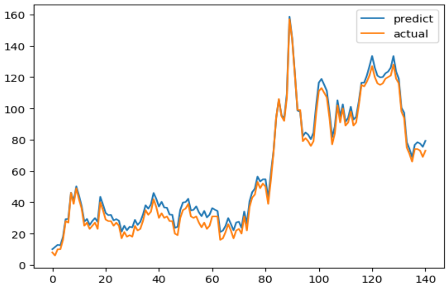


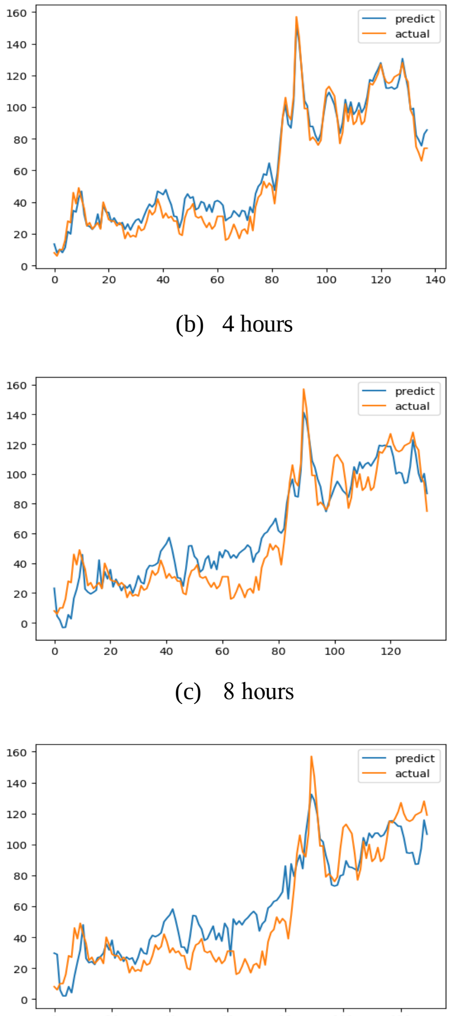
Fig. 3. Distribution of main factors

4.2 Experimental Results

According to the self-organizing algorithm, the number of hidden layer nodes is 30. And training dataset and test dataset are set to be 1200 hours and 141 hours separately. The epochs are equal to 50 times.



(a) 1 hour



(b) 4 hours

(c) 8 hours

(d) 12 hours

Fig. 4. Actual and predict $PM_{2.5}$ concentration for the evaluation dataset for one-, four-, eight-, and twelve-hour

This experiment uses self-organizing LSTM network to predict the PM2.5 concentration of next 1 hour, 4 hours, 8 hours and 12 hours. Figure 4 displays the curves of actual value and predict value. The X-axis is the quantity of the test samples, and the Y-axis is the concentration value of $PM_{2.5}$. From these four pictures, we can see that the fitting degree of one-hour prediction is the highest, then that of four hours prediction is lower. And, the fitting precision of eight hours and twelve hours decreases gradually. We can draw conclusion that with the increase of time domain, the prediction precision is reduced. So this method has advantages in nowcasting.

And table 1 shows the comparison between SO-LSTM (self-organizing LSTM) and the algorithm in [6]. The evaluation parameter is coefficient of determination (R^2).

Table 1. The comparison between SO-LSTM and GA-ANN

Hours / Algorithms	1	4	8	12
GA-ANN	0.992	0.965	0.935	0.915
SO-LSTM	**0.999**	**0.991**	**0.937**	**0.918**

From this table, it’s obvious that the SO-LSTM shows a higher coefficient of determination from 1 hour to 12 hours than GA-ANN.

Table 2 shows the performance of MLR, ANN, NF [9] and SO-LSTM in the case of predicting the PM2.5 concentration after 39 hours.

Table 2. The comparison between SO-LSTM and other algorithms

	MLR	ANN	NF	SO-LSTM
R^2	0.51	0.53	0.72	**0.79**

From table 2, we can see that SO-LSTM performs the best.

5 Conclusion and Future Work

In this paper, a self-organizing LSTM network is employed to predict $PM_{2.5}$ concentration.

First, by using the mutual information algorithm, nine main factors (O_3, NO_2, $PM_{2.5}$, pressure, wind direction of the instant maximum wind speed, temperature, wind direction of maximum wind speed, horizontal visibility, and body temperature) that affect $PM_{2.5}$ in Nanjing are as input data. Because these data are typical time series data, recurrent LSTM network is suitable to apply in this case for its memory capability.

Then, to determine the number of hidden nodes of this network, a self-organizing algorithm called grow-prune is adopted. This method can add or delete nodes during training phase. And in this paper, the number of hidden nodes is determined to be 30.

Finally, after training and test steps, the model is built to predict $PM_{2.5}$. According to the experimental results, the SO-LSTM model performs better than MLR, ANN, NF, GA-ANN. The accuracy is improved.

With time interval growth, the accuracy of prediction is getting lower. For future work, we intend to further improve on the accuracy of long-term forecasting with more excellent pre-training methods.

Acknowledgements

This work was supported by Major Program of the National Social Science Fund of China (Grant No. 17ZDA092). It has also received funding from the European

Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 701697

References

1. Fuller, G.W., Carslaw, D.C., Lodge, H.W.: An empirical approach for the prediction of daily mean PM_{10} concentrations. Atmospheric Environment 36(9), 1431-1441 (2002).
2. Jian, L., Zhao, Y., Zhu, Y.P., Zhan, M.B., Bertolatti, D.: An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. Science of the Total Environment 426(2), 336-345 (2012).
3. Kibria, B.M.G, Sun, L., Zike, J. V.: Bayesian Spatial Prediction of Random Space-Time Field With Application to Mapping $PM_{2.5}$ Exposure. Journal of the American Statistical Association 97(457), 112-124 (2002).
4. Dong, M., Yang, D., Kuang, Y.: $PM_{2.5}$ concentration Prediction using hidden semi-Markov model-based time series data mining. Expert Systems with Application 36(5), 9046-9055 (2009).
5. Zhu, H., Lu, X..: The Prediction of $PM_{2.5}$ Value Based on ARMA and Improved BP Neural Network Model. 2016 International Conference on Intelligent Networking and Collaborative Systems (ICINCS), 515-517 (2016).
6. Venkadesh, S., Hoogenboom, G., Potter, W.: A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks. Applied Soft Computing 13(5), 2253-2260 (2013).
7. Zheng, H., Shang, X.: Study on prediction of atmospheric $PM_{2.5}$ based on RBF neural network. 2013 Fourth International Conference on Digital Manufacturing & Automation (ICDMA), 1287-1289 (2013).
8. Mishra, D., Goyal, P.: Development of artificial intelligence based NO_2, forecasting models at Taj Mahal, Agra. Atmospheric Pollution Research 6(1), 99-106 (2015).
9. Mishra, D., Goyal, P., Upadhyay, A.: Artificial intelligence based approach to forecast PM $_{2.5}$, during haze episodes: A case study of Delhi, India. Atmospheric Environment, 102, 239-248 (2015).
10. Neto, P., Cavalcanti, G., Madeiro, F.: An Approach to Improve the Performance of PM Forecasters. Plos One 10(9), 1-23 (2015).
11. Zhou, S., Li, W., Qiao, J.: Prediction of $PM_{2.5}$ Concentration Based on Recurrent Fuzzy Neural Network. Proceedings of the 36th Chinese Control Conference (PCCC), 3920-3924 (2017).
12. Bun, T., Komei, S., Koji, Z.: Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $PM_{2.5}$. Neural Computing & Applications 27(6), 1553–1566 (2016).
13. Liu, D.J., Li, L.: Application Study of Comprehensive Forecasting Model Based on Entropy Weighting Method on Trend of $PM_{2.5}$ Concentration in Guangzhou, China. International Journal of Environmental Research & Public Health 12(6), 7085-7099 (2015).
14. Han, H., Li, Y., Guo, Y., Qiao, J.: A soft computing method to predict sludge volume index based on a recurrent self-organizing neural network. Applied Soft Computing 38 (C), 477-486 (2016).