Water Quality Prediction Method Based on LSTM Neural Network

Yuanyuan Wang, Jian Zhou, Kejia Chen, Yunyun Wang, Linfeng Liu

Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks
College of Computer, Nanjing University of Posts and Telecommunications
Nanjing, China
zhoujian@njupt.edu.cn

Abstract—**Water quality prediction has more practical significance not only for the management of water resources but also for the prevention of water pollution. It's a time series prediction problem which the traditional neural network isn't suitable. A new water quality prediction method based on long and short term memory neural network (LSTM NN) for water quality prediction is proposed in this paper. Firstly, a prediction model based on LSTM NN is established. Secondly, as the training data, the data set of water quality indicators in Taihu Lake which measured monthly from 2000 to 2006 years is used for training model. Thirdly, to improve the predictive accuracy of the model, a series of simulations and parameters selection are carried out. Finally, the proposed method is compared with two methods: one is based on back propagation neural network, the other is based on online sequential extreme learning machine. The results show that the method is more accurate and more generalized.**

Keywords-water quality prediction; LSTM NN; time series data; water quality indicators

I. Introduction

Water quality prediction provides a significant reference for dynamic regulation of water quality and sudden events. It's a basic work for the water resources management and the water pollution prevention. It adopts the prior-prevention instead of inferior-disposal in traditional management methods. The water indicators are influenced by various factors which are not fully structured. Besides, the non-deterministic and nonlinear feature determines the complexity of water quality prediction.

There are a variety of methods used for water quality prediction at home and abroad. These methods are mainly divided into the following four categories: mathematical statistics [1], gray theory [2], chaos theory [3] and neural network [4]. The method of mathematical statistics is effective on modeling, but the prediction is not ideal; The method of gray theory is just suitable for approximating exponential functions not for the complex nonlinear functions; The method of chaos theory just can be useful when the training data is very affluent; The traditional neural network whose advantages are non-linearity, abstraction, self-organization learning is very suitable for dealing with nonlinear, randomly data. Because of the structure of traditional neural network, it isn't suitable for processing the time series data.

Deep learning is a method of machine learning which can learn the characteristics of data automatically. In recent years, scholars have been trying to use deep-learning-based methods to solve time series prediction problem. The long and short memory neural network (LSTM NN) has 'memory' because of its own unique network structure. It has been applied in the field of time prediction successfully such as stock prediction [5] and traffic flow prediction [6].

Water quality indicators are the time series data and greatly affected by seasons with obvious seasonal diversity. Water quality prediction belongs to the time series prediction. Because of the traditional neural network isn't suitable for dealing with the time series data, this paper proposes a water quality prediction method based on LSTM NN. Firstly, a prediction model we introduced includes input layer, hidden layer and output layer. Secondly, the model is trained by historical water quality indicators. Thirdly, the predictive accuracy is improved by parameters selection and a series of simulations. Finally, the method based on LSTM NN for water quality prediction is compared with two methods: one is based on back propagation neural network (BP NN), the other is based on online sequential extreme learning machine (OS-ELM). The results verifies the effective of the method that we proposed.

The rest of the paper is structured as follows: We introduce the basic concepts of LSTM NN in Section 2. In Section 3, we describe a prediction model based on LSTM NN. Section 4 presents the results of training and performance of several selected models. Besides, the dataset is introduced in this section. Concluding and future envisions are described in Section 5.

II. Basic Concepts of LSTM NN

LSTM is based on recurrent neural network (RNN)[7]. It improves the structure of RNN. To deal with time series data, RNN adds a kind of special structure on hidden layer.

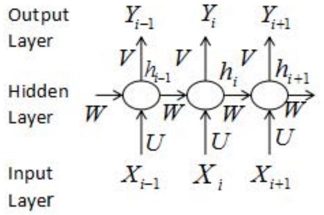

Fig 1. The structure of RNN

978-1-5386-1829-5/17/$31.00©2017 IEEE

As shown in Fig. 1 is the structure of RNN which includes input layer, hidden layer and output layer. The nodes in hidden layer are fully connected, the output of the hidden layer also becomes the input of the hidden layer at the next time. X_i is the input at *ith* time. Y_i is the output at *ith* time. h_i is the state of hidden layer at *ith* time. U is the weight between input layer and hidden layer, V is the weight between hidden layer and output layer, W is the weight between current hidden layer and hidden layer at next time. The computation of hidden layer at *ith* time is: $h_i = f(UX_i + Wh_{i-1})$.

RNN has a vanishing gradient problem when adopts the back-propagation algorithm. To solve this problem, Sepp Hochreite and Jurgen Schmidhuber proposed LSTM NN [9], which is a new deep-learning neural network based on RNN. The structure of LSTM NN's neurons called memory block that maintains a constant error flow. So it is better to deal with time series data.

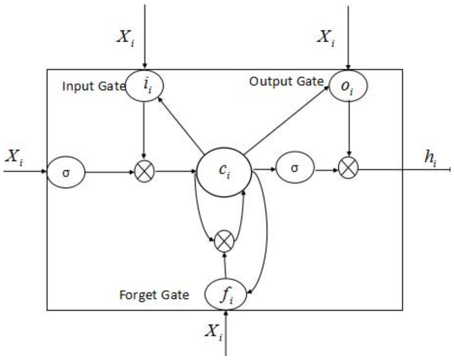

Fig 2. The structure of memory block

Fig. 2 shows the internal structure of the memory block. The function of Memory blocks is same to the neurons in traditional neural networks. In the rest of paper, we named it neurons. In the Fig. 2, the meaning of X_i, Y_i, h_i is same in Fig.1. c_i is the state of the memory block at *ith* time, which determines the update and output of the network. Besides, it plays a critical role in processing the time series data.

To protect and control c_i, there are three gates in the memory block.

1) Input gate: it decides what kind of new information should be stored in the cell, i_i is the value of input gate at *ith* time, the formula as follows:

$$i_i = \sigma(W_{ix}X_i) + W_{hh}h_{i-1} + b_i \qquad (1)$$

W_{ix} is the weight between the input gate and the input layer, W_{hh} is the weight between the hidden layer at the last time and hidden layer at current time. b_i is the bias vector. σ is the activation function.

2) Forget gate: it decides what kind of information should be dropped from the cell, f_i is the value of forget gate at *ith* time, the formula as follow:

$$f_i = \sigma(W_{fx}X_i) + W_{fh}h_{i-1} + b_f \qquad (2)$$

W_{fx} is the weight between forget gate and input layer, W_{fh} is the weight between forget gate and hidden layer. b_f is the bias vector.

3) Output gate: it decides what kind of information should be output from the cell, o_i is the value of output gate at *ith* time, the formula as follow:

$$o_i = \sigma(W_{ox}X_i + W_{hh}h_{i-1} + W_{oc}c_{i-1} + b_o) \qquad (3)$$

W_{ox} is the weight between the output gate and the input layer, W_{oc} is the weight between output gate and the state of memory block at *ith* time. b_o is the bias vector.

The computation of c_i as follow:

$$c_i = f_i * c_{i-1} + i_i * \sigma(W_{cx}X_i + W_{hh}h_{t-1} + W_{cc}c_{i-1} + b_c) \qquad (4)$$

W_{cx} is the weight between input layer and the state of memory block at *ith* time, W_{cc} is the weight between the state of memory block at *ith* time and at next time, b_c is the bias vector.

III. WATER QUALITY PREDICTION MODEL BASED ON LSTM NN

In order to make an accurate prediction of water quality indicators, we establish a model based on LSTM NN for water quality prediction as shown in Fig. 3. Compared with the traditional neural network, the nodes in hidden layer are fully connected and adopt the structure of memory block.

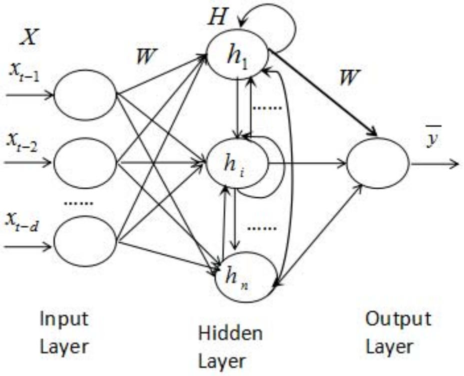

Fig 3. The water quality prediction model based on LSTM NN

As shown in Fig.3, the input of the model is $X = (x_{t-1}, x_{t-2}, x_{t-d})$, x_{t-d} is the value of water quality indicators in *d* months before *tth* month. The number of neurons in input layer is determined by the time step (*d*). $H = (h_1, h_2, ...h_i, ...h_n)$ is the value of hidden layer, h_i is the value of hidden layer at *ith* time. The number of neurons in hidden layer is selected by simulation. $\overline{y} = \overline{x_t}$ is the prediction for *tth* month which is the output of the model. The prediction of this model is single step, so the number of neurons in output layer is 1. The activation function of each neuron is sigmoid function.The model does the computation as follows:

$$h_i = H(W_{hx}X + W_{hh}h_{t-1} + b_n) \qquad (5)$$

$$\overline{y} = W_{hx}X + b_y \tag{6}$$

W_{hx} is the weight between hidden layer and input layer, W_{hh} is the weight between hidden layer at the last time and current hidden layer. b_y is the bias vector.

Since the root mean square error (*RMSE*) can reflect the distance between predictive value and true value. It is used to evaluate prediction performance. The smaller *RMSE* is, the higher predictive accuracy is. The formula as follow:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e^2} \tag{7}$$

e is the error between predictive value $\overline{y}$ and true value, $y = x_t, t = 1,2,3....n$ is true value of the *tth* month. The formula as follow:

$$e = y - \overline{y} \tag{8}$$

In order to improve the predictive accuracy of the model, the model is trained by historical data of water quality indicators. In training, the model selects the best performance with different time step (*d*), the number of iterations (*epoch*) and the number of neurons in hidden layer *(Hiddnum*).

IV. SIMULATION AND ANLYSIS

The data set used in this paper is the monthly observed data of Taihu Lake from 2000 to 2006. A prediction model based on LSTM NN is established by Keras which is a deep-learning framework. The model is used for predicting the value of dissolved oxygen (DO) and total phosphorus (TP) in Taihu Lake which includes 657 recordings. The proportion between training set and test set is 9:1. Finally, the method based on LSTM NN is compared with the method based on BP NN and OS-ELM according to *RMSE*.

A. Data set

The dataset used in this paper is provided by the Nanjing Institute of Geography, Chinese Academy of Sciences. It contains the observed data from eight sites in Taihu Lake. The time series of DO and TP from Jan. 2000 to Dec. 2006 are shown in Fig. 4 and Fig. 5. Fig. 4 and Fig. 5 reveal the following: DO and TP steadily increase in spring period; In the summer, DO and TP reach their maximum value and usually there are no significant changes; DO and TP are the lowest in the late winter and mid-autumn seasons. So the water quality indicators are the time series data and influenced by season.

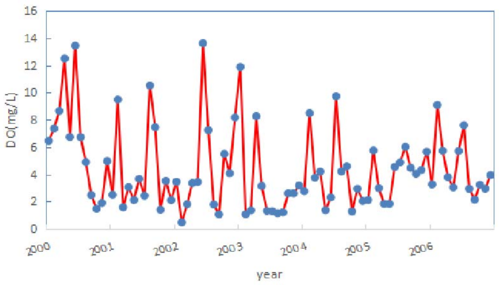

Fig 4. DO time series (from Jan.2000.to Dec.2006.)

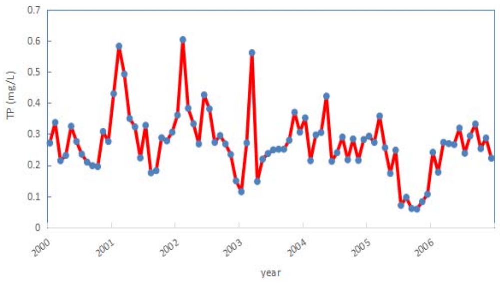

Fig 5. TP time series (from Jan.2000.to Dec.2006.)

B. Parameters selection

In order to improve the predictive accuracy of the model, the model makes a parameters selection among time step (*d*), the number of iterations (*epoch*) and the number of neurons in hidden layer (*Hiddnum*). In this paper, several simulations are carried out. The results are shown in TABLE I.

TABLE I. THE SIMULATION RESULTS OF THE PROPOSED MODEL

			DO		TP	
d	***Hiddnum***	***epoch***	***RMSE***	***MaxError***	***RMSE***	***MaxError***
3	10	10	0.066	0.160	0.043	0.116
	10	20	0.061	0.156	0.048	0.155
	13	25	0.058	0.151	0.042	0.118
	15	**20**	0.051	0.138	**0.041**	**0.103**
	15	25	0.102	0.215	0.054	0.188
	17	30	0.060	0.225	0.043	0.094
4	10	20	0.053	0.152	0.044	0.118
	13	25	0.054	0.162	0.047	0.135
	15	20	0.049	0.127	0.042	0.130
	15	25	0.050	0.181	0.046	0.108
	17	30	0.054	0.153	0.046	0.097
5	10	20	0.057	0.198	0.043	0.112
	13	25	0.061	0.196	0.041	0.121
	15	20	0.053	0.143	0.042	0.103
	15	**25**	**0.046**	**0.116**	0.047	0.125
	17	30	0.048	0.127	0.043	0.124

In TABLE I, *MaxError* is the maximum error between the value of prediction and true data. As can be seen from TABLE I, *RMSE* gradually decreases as *Hiddnum* increases, when *Hiddnum* is greater than a certain value, *RMSE* begins to increase. In the same value of *Hiddnum*, as the *epoch* increases, *RMSE* gradually decreases, when *epoch* exceeds a certain value, *RMSE* begins to increase.

As for DO time series, *RMSE* is the smallest when the time step is 5, *Hiddnum* is 15 and *epoch* is 25; As for TP time series, *RMSE* is the smallest when the time step is 3, *Hiddnum* is 15, *epoch* is 20. The above results indicated by black bold font in TABLE I.

C. Comparative Results

To compare results obtained from LSTM NN, BP NN and OS-ELM, the same test set is used. The results are shown in Fig. 6 and Fig. 7. Fig. 6 shows the prediction results of DO. Fig. 7 shows the prediction results of TP.

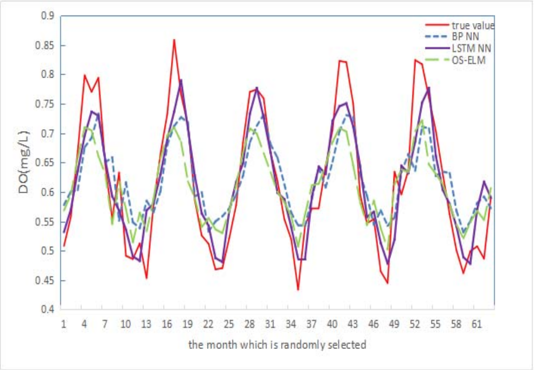

Fig 6. Prediction results of DO

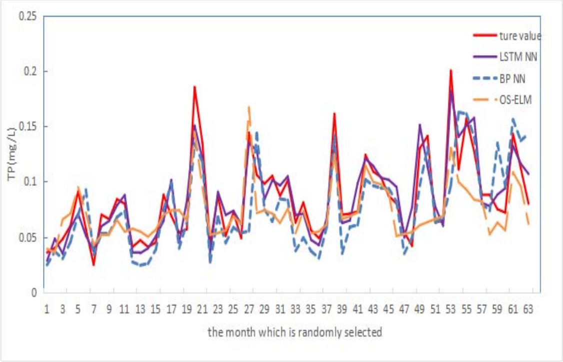

Fig 7. Prediction results of TP

As shown in Fig. 6 and Fig. 7, when the fluctuation of data is small, the predictive accuracy of BP NN is higher than OS-ELM. When the fluctuation of data is large, the predictive accuracy of OS-ELM is higher than BP NN. The predictive accuracy of LSTM NN is always higher than the first two in both cases. It indicates that the structure of BP NN and OS-ELM is not suitable for dealing with time series data, LSTM NN can be a good choice for dealing with sequential prediction problem.

The comparison of *RMSE* from LSTM NN, BP NN and OS-ELM with different time steps is shown in Fig. 8 and Fig. 9. Fig. 8 shows *RMSE* of DO with different time steps. Fig. 9 shows *RMSE* of TP with different time steps.

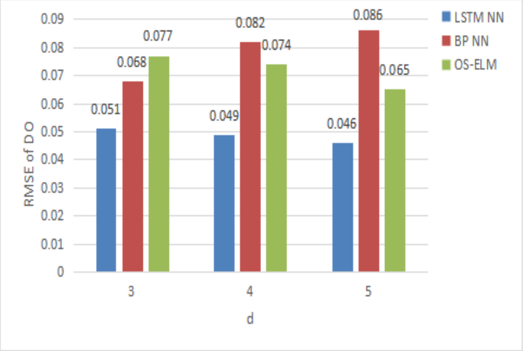

Fig 8. RMSE of DO with different time steps

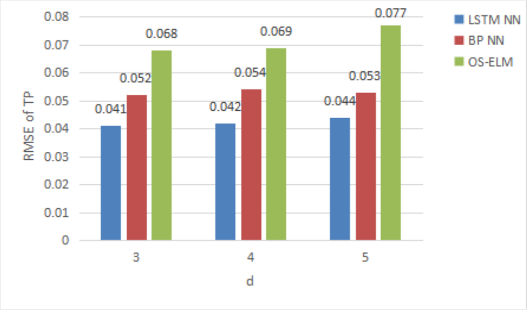

Fig 9. RMSE of P with different time steps

It can be seen from Fig. 8, *RMSE* of OS-ELM is higher than BP NN and LSTM NN on the prediction of DO. When the time step is 4 or 5, *RMSE* of OS-ELM is higher than LSTM NN, but lower than BP NN. *RMSE* of LSTM NN always be the smallest with each time step.

It can be seen from Fig. 9, *RMSE* of OS-ELM always be the largest on the prediction of TP. As the increase of time step, *RMSE* of BP increases steadily, LSTM NN has been relatively stable which always be the smallest with each time step. All experimental results demonstrate that LSTM NN is generalized and effective for the prediction of water quality indicators.

V. CONCLUSION

Water quality prediction has more practical significance not only for the management of water resources but also for the prevention of water pollution. Based on the sequential characteristics of water quality indicators, this paper proposes a new method based on LSTM NN for water quality prediction and sets up a prediction model. The model is trained by the historical data of water quality indicators in Taihu Lake. To improve the predictive accuracy of the model, several simulations and parameter selection are carried out. Compared with BP NN and OS-ELM, the predictive accuracy of LSTM NN is higher. Besides, LSTM NN is more generalization. Considering the disadvantage of long training cycle, a more effective memory block will be designed in the future work.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 71301081, 61373139, 61572261), Natural Science Foundation of Jiangsu Province (No. BK20130877, BK20150868), Postdoctoral Science Foundation of China (No. 2014M551637), Postdoctoral Science Foundation of Jiangsu Province (No.1401046C), Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (No.17KJB520027).

REFERENCES

[1] Liu Fang and Hu Caihong, “Establishment and Application of Forecast Model for Precipitation Based on Mathematical Statistics,” Meteorological and Environmental Sciences, vol. 37, 2014, pp. 89-93, doi: 10.3969/j. ssn. 1673-7148.2014.02.015.

[2] Sun Yuanhuan and Hu Yuzhuo, Application of Improved Gray Neural Network Model for Water Quality Prediction, Chongqing: Chongqing University, 2010.

[3] Sun Xihao and Yan Lei, “Prediction Model of Taihu Using Improved Fuzzy Time Series,” Science and Technology&Innovation, vol. 21, 2016, pp. 15-16, doi: 10.15913/j. Cnki. kjycx.2016.21.015.

[4] Yuan Honglin and Gong Ling, “Using BP Neural Network for The Prediction of Soap River Water Quality,” Journal of Security and Environment, vol. 13, Apr. 2013, pp. 106-110, doi: 10. 3969/j. issn. 1009-6094. 2013. 02.023.

[5] Sun Ruiqi, A Study on Price Forecasting of US Stock Based on LSTM Neural Network, Beijing: Capital University of Economics and Business, 2015.

[6] X. Ma and Z. Tao, “Long Short Term Memory Neural Network for Traffic Speed Prediction Using Remote Microwave Sensor Data,” Transportation Research Part C Emerging Technologies. Vol. 54, May 2015, pp. 187-197.

[7] http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

[8] A. Graves, Supervised Sequence Labeling with Recurrent Neural Network, Poland: Studies in Computational Intelligence, July 2 012, pp. 37-44.

[9] Hochreiter, S and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, 1997, pp. 1735-1780.

[10] Yoshua Bengio and Patrice Simard.“ Learing Long-Term Dependencies with Gradient Descent is Difficult,” Transactions on Neural Networks , 1994, pp. 157-166.

[11] Bunchingiv Bazartseren and Gerald Hildebrabt, “Short Term Water Level Perdiction Using Neural Networks and Neuro-fuzzy Approach,” Neurocomputing, vol. 55, 2003, pp. 439-450.

[12] G.Peter, “Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model,” NeuroComputing. Vol. 50, 2003, pp.15 9-175.