Research on water quality prediction method based on AE-LSTM

1st Huiqing Zhang ,
School of Automation, Department of Informatics
Beijing University of Technology
Beijing, China
zhq@bjut.edu.cn

2nd Kemei Jin
School of Automation, Department of Informatics
Beijing University of Technology
Beijing, China
jinkemei_999@163.com

***Abstract*—Aiming at the traditional prediction methods of related parameters that affect water quality, they usually only consider the temporal characteristics of the related parameters of water quality and ignore the problem that water quality changes are multivariate related, A prediction method of spatiotemporal correlation water quality parameters based on automatic encoder (AE) dimensionality reduction and long and short time memory (LSTM) neural network is proposed. Firstly, considering that water quality parameter changes have obvious time characteristics, a time series prediction model of water quality parameters is established based on LSTM. Secondly, considering that the water quality changes have multiple correlations, the upstream water quality will also affect the downstream water quality. If all the water quality parameters of the upstream station are added to the prediction model, redundant features will reduce the accuracy of parameter prediction. Therefore, the automatic encoder is used to reduce the dimensionality of the relevant parameters. Finally, the data set of Lang fang Water Quality Automatic Monitoring Station is applied to monitor the effectiveness of the method. By predicting the concentration of total phosphorus (TP) and total nitrogen (TN), the method is found to have better prediction accuracy and robustness.*

***Keywords*—long-short-term memory network, auto-encoder, water quality prediction, feature extraction, multivariate correlation**

I. INTRODUCTION

The prediction of important parameters in water quality is based on the historical monitoring data of the national control section, using modern technology to estimate and speculate the future change trend of water quality index concentration, and predicting the concentration of major pollutants in the water reflecting the degree of pollution is of great significance for making appropriate water environment management decisions [1]. Farid Khalil Arya [2] study the time characteristics of dissolved oxygen and temperature in water. M. Najafzadeh [3] used formulas based on evolutionary calculations to predict water quality parameters. Dai Zhijun [4] used grey theory to predict the main pollutants in the water environment.

In recent years, it has been found that long- and short-term memory networks (LSTM) [5] [6] [7] have achieved significant results in processing data with temporal characteristics and some nonlinear complex problems. Liu Jingjing [8] proposed the K-Similarity noise reduction LSTM neural network water quality prediction model. Li Zhenbo et al. [9] based on the SAE-LSTM mixed water quality prediction model, used a sparse encoder to reduce the dimension of the data, and then used LSTM to fit the data. These methods have achieved certain prediction effects in water quality prediction, but they ignore the spatiotemporal correlation of water quality changes, resulting in insufficient prediction accuracy. In order to better study the multivariate correlation [10] and time characteristics of water quality parameters, the automatic encoder (AE) [11] was used to select features of water quality-related features, and then the code of the automatic encoder was input into the LSTM neural network to establish a space-time-based Feature [12] prediction model. The prediction results are compared with the prediction results of LSTM's time series prediction model and LSTM's spatiotemporal correlation prediction model, and it is found that this method has higher prediction accuracy and better robustness.

In Section 2, we first introduce the related theories of auto-encoders and long-term and short-term memory networks used in this paper. The third part introduces the structure of the prediction model and model evaluation indicators. The fourth part is the experimental part based on the actual data set, introducing the experimental process and experimental results. The last section is summary.

II. METHODOLOGY

2.1 Automatic encoder

Self-encoding dimensionality reduction is an unsupervised learning neural network that encodes high-dimensional data into low-dimensional data [13] and reproduces the input signal as much as possible. The network can automatically learn features, and has been widely used in data compression and feature extraction projects [14]. When the prediction of water quality parameter concentration takes into account the influence between different water quality automatic stations, the input of the neural network will increase exponentially. Using automatic coding neural network, the pre-processed data is input into the encoder to automatically perform feature learning [15]. The automatic encoder is composed of encoder and decoder. The architecture is shown in Fig 1.

Corresponding author's telephone number: 18813030922
Water pollution control and treatment technology major special project (2018ZX07111005).

978-1-7281-9888-0/20/$31.00 ©2020 IEEE

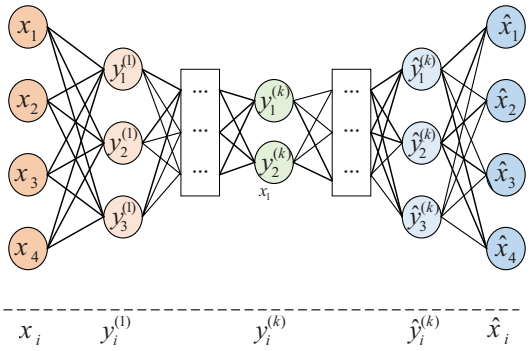

Fig 1. Structure diagram of auto encoder

In order to reduce the dimensions of highly nonlinear structures, multiple layers of encoding and decoding are required. The output of the first hidden layer is shown in (1), and the output of the kth hidden layer is shown in (2).

$$y_i^{(1)} = \sigma(W^{(1)} x_i + b^{(1)}) \tag{1}$$

$$y_i^{(k)} = \sigma(W^{(k)} y_i^{(k-1)} + b^{(k)}), k = 2, \ldots, K \tag{2}$$

Where σ is the activation function, commonly used sigmoid, $W^{(k)}$ is the weight matrix of the k-th layer, b^k is the offset vector of the k-th layer, K is the number of hidden layers, and $y_i^{(k)}$ is the characteristic representation of the network obtained after K hidden layers. In order to obtain the optimized model parameters, the reconstruction error of x_i and $\hat{x}_i$ needs to be minimized. In this paper, the mean square error is used to represent the loss function, as shown in (3).

$$L(x, z) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \tag{3}$$

2.2 long-short-term memory neural networks

LSTM has more complicated internal structure, and it can selectively remember or forget the information on the cell structure through the gated state. Remember information that needs long-term memory while forgetting unimportant information. A cell of LSTM contains three gating units. When dealing with timing prediction problems, LSTM networks converge faster and have higher accuracy than other traditional neural networks. The structure of LSTM neuron is shown in Fig 2.

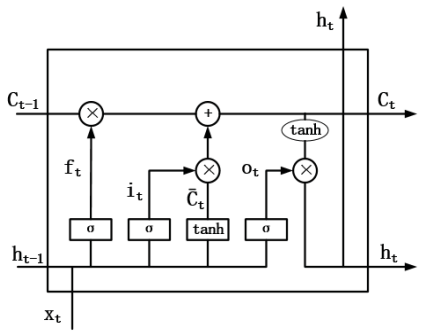

Fig 2 . LSTM neuron structure

In Fig 2, there are three inputs: the input value x_t of the training sample at time t, the long-term memory state of neuron C_{t-1} at time t-1, and the output value of neuron h_{t-1} at time t-1. There are two outputs: the long-term memory state C_t at time t, and the output h_t at time t. LSTM realizes long-term memory function by controlling three doors. First, LSTM discards some information that is not conducive to subsequent tasks through the forget gate. According to the output of the neuron at the previous moment, the input h_{t-1} at the current moment and the inputs at the current moment, through the activation function σ, the output f_t of the forget gate is obtained, as shown in (4).

$$f_t = \sigma(W_f [h_{t-1}, x_t] + b_f) \tag{4}$$

Where W_f is the input loop weight, b_f is the offset term. The input gate determines which part of the new information is stored . This process is divided into two steps: First, use the sigmoid function of the input gate to select the new information to be stored and record it as i_t , as shown in (5);Then according to the memory C_{t-1} and the input x_t , a new initial vector $\overline{C}_t$ is created using the function tanh, as shown in (6).

$$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\overline{C}_t = \tanh(W_c [h_{t-1}, x_t] + b_c) \tag{6}$$

Where i_t is the vector value of the input gate, $\overline{C}_t$ is the new information, W_c is the input weight, b_i and b_c respectively represents the offset term of their respective gates. Next, based on the forget gate and output gate coefficients, LSTM will update the current long-term memory state C_t . The calculation process of updating the cell state is shown in (7).

$$C_t = f_t \times C_{t-1} + i_t \times \overline{C}_t \tag{7}$$

Finally determine what information to output to participate in subsequent calculations. Use the Sigmoid function to calculate the neuron output, as shown in (8); then use the tanh function to limit the current long-term memory state value, which ranges from 0 to positive infinity. Finally, multiply the opening and closing degree o_t of the output gate to get the final output information h_t at the current moment, as shown in (9).

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t \times \tanh(C_t) \tag{9}$$

Where W is the input weight, b_o is the offset term.

III. PREDICTION MODEL

Historical water quality information, meteorology, and upstream diffused water flow all have an impact on regional water quality. When analyzing and predicting local water quality parameters, it is necessary to fully consider the multiple correlations existing in the temporal and spatial dimensions of the water quality parameters. In this paper, the water quality parameter data of the upstream monitoring stations are also considered into the neural network. The structure diagram of the water quality parameter prediction model is shown in Fig 3.

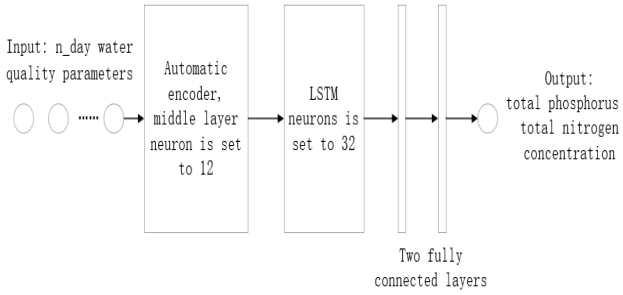

Fig 3 . Recurrent neural network structure diagram

In the water quality prediction model constructed in this paper, the middle layer neurons of the autoencoder are set to 12, the LSTM layer neurons are set to 32, and two fully connected layers are set. The model predicts the concentration of TP and TN on the 16th day of the Qin ying yang shui station based on the water quality parameter data of the three water quality automatic stations in the previous 15 days. The model input contains the characteristics of three water quality automatic stations, each station has 9 characteristic parameters, and the outputs of the model are total phosphorus and total nitrogen concentration. Neural network parameters are updated by Adam optimizer, relu is used for LSTM activation function, and MSE is used for loss function.

This paper studies regression problems. When training the above model, this paper uses two evaluation indicators, root mean square error (RMSE) and coefficient of determination (Coefficient of determination, R^2 score), to evaluate the prediction effect. Their calculation formulas are shown in (10) and (11):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{11}$$

Where $\hat{y}_i$ is the predicted value, y_i is the actual value, and n is the amount of data. The smaller the value of RMSE, the smaller the prediction error of the water quality parameter model. The closer the value of R^2 is to 1, the better the fit between the predicted water quality parameter and the actual value.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, the data of water quality automatic stations in Hebei area are collected as experimental evaluation data to verify the effectiveness of the spatio-temporal correlation prediction method based on AE-LSTM. The experimental data was collected from Lang fang Water Quality Automatic Station from July 25, 2018 to December 7, 2018. In the experiment, a 16-day sliding window is used to divide the data, and the data of the previous 15 days is used as the input of the model to predict the data of the 16th day. After the sliding window divides the data, each water quality automatic station has 150 sets of data, randomly selects 30 sets as the test set, and the remaining 120 sets of data as the training set.

The data set contains a total of 13 variables, including 9 water quality indicators, physical water quality indicators: TEMP(water temperature), conductivity, turbidity; chemical water quality indicators: PH(potential of hydrogen),TN(total nitrogen),TP(total phosphorus), DO (dissolved oxygen), CODMn (Permanganate Index) , NH3-N (Ammonia nitrogen). Other variables: station name, monitoring time, longitude, latitude.

According to the longitude and latitude positioning, analyze the positional relationship between the automatic water quality monitoring stations, and determine the Lao-Xia-An-Road Station, Qin-Ying-Yang-Shui station and Wang-Jia-Bai stations that constitute the water quality prediction spatial data. The elevation of Wang-Jia-Bai station is 17.08 meters, the elevation of Qin-Ying-Yang-Shui station is 11.01 meters, the separation of Wang-Jia-Bai station and Qin-Ying-Yang-Shui station is 9.01 km, and both stations are located in the North Canal, so it can be judged that Wang-Jia-Bai station is located upstream of Qin-Ying-Yang-Shui station. Lao-Xia-An-Road Station has an elevation of 19.5 meters and is located in Feng-gang-jian-river, which is a tributary of the North Canal. Lao-Xia-An-Road station and Qin-Ying-Yang-Shui station are separated by 4.76 kilometers. It can be judged that Lao-Xia-An-Road station is also located upstream of Qin-Ying-Yang-Shui station. It can be judged that both Wang-Jia-Bai station and Lao-Xia-An-Road station are located upstream of Qin-Ying-Yang-Shui station. Wang-Jia-Bai, Lao-Xia-An-Road are located at the upstream of Qin-Ying-Yang-Shui station and are close to each other. Therefore, the water quality data of the three stations are selected for experimental research.

The input of the water quality parameters prediction model based on AE-LSTM is the 15-day water quality data of the Lao-Xia-An-Road station, Wang-Jia-Bai station and Qin-Ying-Yang-Shui station after dimensionality reduction by AE. Enter data of 1 day at each time step, and loop 15 times to complete an iteration of the neural network. The input of the spatiotemporal correlation prediction model based on LSTM is 15-day data from three automatic water quality stations. The input of single time series prediction model based on LSTM is 15-day water quality parameter time series data of Qin-Ying-Yang-Shui station. Three models are used to predict the concentration of TP and TN, and the fitting results are shown in Fig 4, Fig 5.

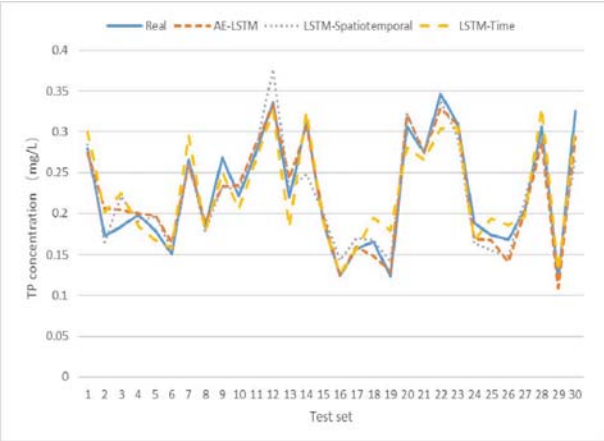

Fig 4. Comparison chart of prediction results of total phosphorus concentration

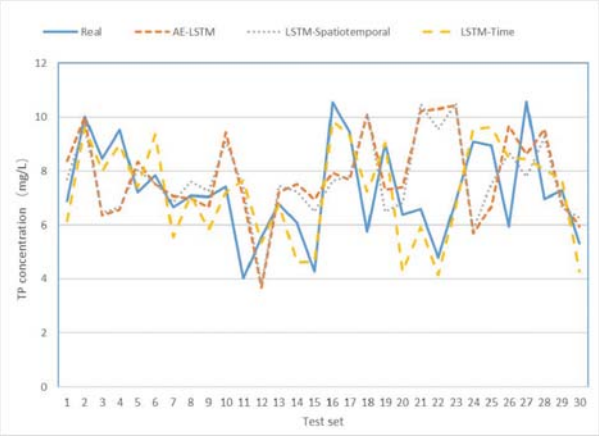

Fig 5. Comparison chart of prediction results of total nitrogen concentration

Compare the results of the three models. Tables 1 and 2 list the model evaluation indicators of the three models.

Table 1. Comparison of evaluation indexes of three models for predicting TP concentration

Prediction model of TP	R^2	RMSE
AE-LSTM	0.924	0.0002
LSTM- Spatiotemporal	0.887	0.0005
LSTM-Time	0.891	0.0005

Table 2. Comparison of evaluation indexes of three models for predicting TN concentration

Prediction model of TN	R^2	RMSE
AE-LSTM	0.909	0.024
LSTM- Spatiotemporal	0.828	0.496
LSTM-Time	0.772	0.512

It can be concluded from Tables 1 and 2 that the RMSE of the water quality spatiotemporal correlation prediction model based on AE-LSTM is less than the single time series prediction model based on LSTM and spatiotemporal correlation prediction model based on LSTM, and R^2 is greater than the LSTM time prediction model and the LSTM time-space water quality prediction model. And the prediction accuracy of TP and TN are both greater than 0.9, indicating that the water quality spatiotemporal correlation prediction model AE-LSTM is robust and the prediction accuracy is higher, and the error between the predicted TP and TN concentration values and the actual TP and TN concentration values is more small. Using AE for feature dimensionality reduction, combined with the superiority of LSTM algorithm for processing time series, the spatiotemporal correlation prediction of water quality based on AE-LSTM is performed.

V. Conclusion

Aiming at the spatiotemporal correlation of water quality, this paper proposes a prediction method of spatiotemporal correlation based on AE-LSTM. According to longitude, latitude, elevation, distance, the water quality automatic stations with upstream and downstream relations were selected to establish the spatiotemporal characteristics. Experimental results show that the prediction model whose input was spatiotemporal features and based on AE-LSTM has better prediction effect and higher prediction accuracy than spatiotemporal prediction model based on LSTM and prediction model based on LSTM that only considered time characteristics, and can effectively predict water quality parameters.

Acknowledgment

This work is supported by Water pollution control and treatment technology major special project (2018ZX07111005).

References

[1] Lv Jiake,Wang Xuan,Zou Wei. A Hybrid Approach of Support Vector Machine with Differential Evolution Optimization for Water Quality Prediction[J]. Journal of Convergence Information Technology,2013,8(2).

[2] Farid Khalil Arya,Lan Zhang. Time series analysis of water quality parameters at Stillaguamish River using order series method[J]. Stochastic Environmental Research and Risk Assessment,2015,29(1).

[3] M. Najafzadeh,A. Ghaemi,S. Emamgholizadeh. Prediction of water quality parameters using evolutionary computing-based formulations[J]. International Journal of Environmental Science and Technology,2019,16(10).

[4] Dai Zhijun, Peng Xiaochun, Huang Hu. Application of gray model theory in prediction of river water pollution [J]. Environmental Protection, 2002 (01): 28-29.

[5] Unjin Pak,Chungsong Kim,Unsok Ryu,Kyongjin Sok,Sungnam Pak. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction[J]. Air Quality, Atmosphere & Health,2018,11(8).

[6] C. Okan Sakar,S. Olcay Polat,Mete Katircioglu,Yomi Kastro. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks[J]. Neural Computing and Applications,2019,31(10).

[7] Liu Jun,Zhang Tong,Han Guangjie,Gou Yu. TD-LSTM: Temporal Dependence-Based LSTM Networks for Marine Temperature Prediction.[J]. Sensors (Basel, Switzerland),2018,18(11).

[8] Liu Jingjing, Zhuang Hong, Tie Zhixin, Cheng Xiaoning, Ding Chengfu.K-Similarity noise reduction LSTM neural network multi-

factor prediction model for water quality [J] .Application of Computer Systems, 2019,28 (02): 226-232.

[9] Zhenbo Li,Fang Peng,Bingshan Niu,Guangyao Li,Jing Wu,Zheng Miao. Water Quality Prediction Model Combining Sparse Auto-encoder and LSTM Network[J]. IFAC PapersOnLine,2018,51(17).

[10] Hamid R. Safavi,Kian Malek Ahmadi. Prediction and assessment of drought effects on surface water quality using artificial neural networks: case study of Zayandehrud River, Iran[J]. Journal of Environmental Health Science and Engineering,2015,13(1).

[11] Lianfa Li. A Robust Deep Learning Approach for Spatiotemporal Estimation of Satellite AOD and PM2.5. 2020, 12(2)

[12] Petr Hurtik,Vojtech Molek,Irina Perfilieva. Novel dimensionality reduction approach for unsupervised learning on small datasets[J]. Pattern Recognition,2020,103.

[13] Yuan-Yuan Liu,Lei Li,Ye-Sen Liu,Pak Wai Chan,Wen-Hai Zhang. Dynamic spatial-temporal precipitation distribution models for short-duration rainstorms in Shenzhen, China based on machine learning[J]. Atmospheric Research,2020,237.

[14] Monsalve Jonathan,Rueda-Chacon Hoover,Arguello Henry. Sensing Matrix Design for Compressive Spectral Imaging via Binary Principal Component Analysis.[J]. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society,2019.

[15] Zeinab Tirandaz,Gholamreza Akbarizadeh,Hooman Kaabi. PolSAR image segmentation based on feature extraction and data compression using Weighted Neighborhood Filter Bank and Hidden Markov random field-expectation maximization[J]. Measurement,2020,153.