Research on water quality prediction based on PE-CNN-GRU hybrid model

Langlang Zhang1*, Jun Xie2, Xinxiu Liu2, Wenbo Zhang1, Pan Geng1

1College of Logistics Engineering, Shanghai Maritime University, Shanghai 201306, China
2Shanghai Urban Construction Design & Research Institute (Group) Co., Ltd., Shanghai

Abstract. Sewage treatment is a complex and nonlinear process. In this paper, a prediction method based on convolutional neural network (CNN) and gated recurrent unit (GRU) hybrid neural network is proposed for the prediction of dissolved oxygen concentration in sewage treatment. Firstly, akima’s method is used to complete the filling preprocessing of missing data, and then the integrated empirical mode decomposition (EEMD) algorithm is used to denoise the key factors of water quality data. Pearson correlation analysis is used to select better water quality parameters as the input of the model. Then, CNN is used to convolve the data sequence to extract the feature components of sewage data. Then, the CNN-GRU hybrid network is used to extract the feature components for sequence prediction, and then the predicted output value is obtained. The mean absolute error (MAE), root mean square error (RMSE) and mean square error (MSE) were used as evaluation criteria to analyze the prediction results of the model. By comparing with RNN model, LSTM model, GRU model and CNN-LSTM model, the results show that the PCA-EEMD-CNN-GRU (PE-CNN-GRU) hybrid model proposed in this paper has significantly improved the prediction accuracy of dissolved oxygen concentration.

1. Introduction
In recent years, with the rapid construction of cities, the problem of water environmental pollution has become increasingly serious[1,2]. Sewage treatment is the purification of sewage by degrading pollutants and separating them from sewage[3]. Sewage treatment process with some characteristics such as nonlinearity, randomness and strong coupling. The prediction of key water quality parameters is a prerequisite for ensuring the stable and efficient operation of sewage treatment plants[4,5]. In order to achieve this goal, it is necessary to establish a model that can accurately describe the sewage treatment process under different working conditions. Convolutional neural networks can automatically extract and learn features from one-dimensional sequence data very effectively[6,7]. Based on the comprehensive wavelet decomposition, autoregressive integrated moving average (ARIMA) and gated recurrent unit (GRU) model, it has better prediction accuracy, stability and robustness for conventional water quality indicators. [9] A deep hybrid model based on convolutional neural network-gated recurrent unit-support vector regression (CNN-GRU-SVR) was used to predict the water quality of the Ganges River using historical data[10]. Based on the input of PCA-BP neural network, the tendency of local optimal solution is effectively alleviated. [10] The parameters of water quality were predicted and analyzed by the nonlinear input and output fitting function of BPNN model. The variation of DO with time is revealed.

This paper proposes a DO prediction method based on PCA-EEMD-CNN-GRU. Firstly, the preliminary noise reduction preprocessing of the data is carried out. Then the processed data is input into the CNN model part, and the obtained one-dimensional result is input into the LSTM model. At each time step, the convolutional network structure is used to extract the input features. The CNN-GRU model has the advantages of both CNN and GRU. Compared with the results of a single prediction model, the PE-CNN-GRU model has higher prediction accuracy and can achieve accurate prediction of DO.

2. Data Processing
The sampling frequency of the sewage data sensor is different or the fault causes the data to be missing, and the working conditions change frequently. Based on the preprocessing of the total nitrogen, total phosphorus and ammonia nitrogen data of the real sensor, it is found that the data has problems such as repetition, loss, frequency change, and too large or too small.

*Corresponding author’s e-mail address: 1397731887@qq.com
Figure 1 shows the processed data distribution. For the problem of data repeatability, only the data that appears for the first time is retained. For the missing data, Python-based programming is used to implement linear smooth filling with akdima, and the frequency change and too large or too small data are discarded and filled. The box plot method is used to judge whether the data is too large or too small, and the missing, repeated, inconsistent frequency, and too large or too small data are discarded, and then nonlinear smooth filling is performed. Finally, the integrated empirical mode decomposition (EEMD) algorithm is used to denoise the key factors of water quality data.

3. Model Design

3.1 Modeling environment and strategy

This model is based on the Windows operating system, using Python3.7 version, encoding, using Pycharm programming tools, the establishment of RNN, LSTM, CNN-LSTM, CNN-GRU model using Tensorflow2.1 framework to achieve. The water quality data of the experiment were derived from the key water quality monitoring data of Shanghai Zhuyuan Sewage Treatment Plant from January 2020 to December 2020. The data were divided into training set and test set according to the ratio of 7:3. In order to obtain various factors related to dissolved oxygen content in water quality data, combined with the sample data of existing water quality factors, DO is used as the target to be predicted, and COD, NH4-N, PH, TN, TP, flow, T, SS, DO are used as influencing factors. In this paper, the cumulative variance contribution rate of PCA is used to determine the selection of characteristic variables. SPSS software is used to analyze the Pearson correlation coefficient of the sensor data. The larger the value obtained by the analysis, the higher the correlation degree. The results are shown in table 1. The Pearson correlation coefficient is used to measure the degree of correlation between different variables. The operating principle is shown in Equation (1):

\[ r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} \]  

\[ r \]: Pearson correlation coefficient; \(X_i, Y_i\) The first sample i a numerical said corresponding characteristics; \(\bar{X}\) and \(\bar{Y}\) respectively indicate the mean value.

According to the correlation degree of Table 1, the correlation degree of water inflow is 0.64, and the correlation degree of total nitrogen is 0.07, which is the lowest. Based on various factors, COD, NH4-N, PH, TN, water inflow, temperature and SS were selected as the input parameters of the prediction model.

3.2 Model evaluation index

Use mean absolute error (MAE), root mean square error (RMSE), mean square error (MSE) and mean absolute percentage error (MAPE) as evaluation criteria to judge the accuracy of model prediction.

MAE calculation formula:

\[ MAE = \frac{1}{m} \sum_{i=1}^{m} |(y_i - \tilde{y}_i)| \]  

RMSE calculation formula:

\[ RMSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \tilde{y}_i)^2 \]  

MSE calculation formula:
\[ MRSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \]  

(4)

Where \( y_i \) represents the true value and \( \hat{y}_i \) represents the predicted value.

### 3.3 Model construction and analysis

The convolutional neural network model can extract the local features of water quality parameters well. GRU has higher operating efficiency than LSTM. This paper combines the CNN model with the GRU model (CNN-GRU model), which combines the extraction of water quality data feature information by CNN and the sensitivity of GRU to time series data. The structure of the constructed neural network model is shown in Figure 2.

**Fig.2** The basic structure of CNN-GRU neural network model

It can be seen from Fig.2 that the CNN-GRU neural network model mainly includes input layer, CNN layer, fully connected layer, GRU layer and output layer. The input layer is to input the correlation amount that affects the TN data at a certain time and integrate the TN data into a vector, and use the sliding window to form a new time series. The new time series data enters the CNN layer for feature extraction of local data. The CNN part is composed of one-dimensional convolution (1D-CNN) and maximum pooling layer, and Relu function is used as activation function. The fully connected layer converts the multi-dimensional feature matrix into a one-dimensional vector and passes it to the GRU network. The GRU layer fully learns and memorizes the incoming time series information data, predicts the TN data, and outputs the results after denormalizing the predicted values through the Dense layer.

**Fig.3** Prediction process based on PE-CNN-GRU model

The process, as shown in Figure 3, is mainly divided into the following four steps: 1) The input water quality data set is divided into training set and test set, and normalized; 2) Network parameters were set, MSE was used as loss function, and Adam optimization algorithm was used to update the weight; 3) Train the model through the training set. If it meets the fitting requirements, enter the next step, otherwise return to step 2), adjust the network parameters and optimize the algorithm. 4) Using the trained model to test the test set, and the predicted output is normalized to obtain the predicted value of the water quality data.

### 4. Experimental results and analysis

After many experiments, the optimal parameters of each part are determined as follows table 2. The network structure uses one-dimensional convolution (1D-CNN), in which the convolution layer and the pooling layer are used to extract and screen the main features of the data. The data is processed in time series. The convolution layer traverses the input water quality data information, and the convolution operation is performed with the weight of the convolution kernel and the local sequence
of the water quality data information to obtain a preliminary feature matrix. The pooling layer takes the obtained feature matrix as input, slides on the matrix sequence with the pooling window, and moves once to obtain a maximum value for outputting a more expressive feature matrix. In this paper, Dropout technology is used to prevent the occurrence of overfitting. The Dropout parameter is set to 0.2. The purpose of using Dropout technology is to prevent the input data from losing important features when passing through the pooling layer, so as to facilitate the extraction of high-dimensional features of the data, and then reduce the dimension of the data through the Flatten layer.

Tab. 2 Parameters setting of CNN-GRU.

<table>
<thead>
<tr>
<th>Model parameter</th>
<th>Parameter value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-D convolution layer filter</td>
<td>80</td>
</tr>
<tr>
<td>1-D Kernel size of the convolution layer</td>
<td>1</td>
</tr>
<tr>
<td>The 1-D convolution layer is filled</td>
<td>Same</td>
</tr>
<tr>
<td>Pool layer size</td>
<td>2</td>
</tr>
<tr>
<td>Pool layer filling</td>
<td>Same</td>
</tr>
<tr>
<td>GRU layer activation function</td>
<td>Relu</td>
</tr>
<tr>
<td>Batch size</td>
<td>1024</td>
</tr>
<tr>
<td>Time_step</td>
<td>20</td>
</tr>
<tr>
<td>Epoch</td>
<td>150</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Loss</td>
<td>Mean squared error</td>
</tr>
</tbody>
</table>

The prediction effects of RNN, LSTM, CNN-LSTM and CNN-GRU are compared. The results show that the accuracy of PE-CNN-GRU model is obviously better than that of other single models. According to the comprehensive table 3 information, the mean square error of the PE-CNN-GRU neural network model proposed in this paper is reduced by 26.02 %, 25.60 %, 29.45 % and 17.32 % respectively compared with RNN, LSTM, CNN-LSTM and CNN-GRU neural networks. The average absolute error decreased by 51.71 %, 50.24 %, 52.95 % and 49.5 % respectively. The root mean square error ratios were 21.72 %, 15.5 %, 21.06 % and 17.24 % lower, respectively.

Tab. 3 Comparison of prediction accuracy of each model.

<table>
<thead>
<tr>
<th>model</th>
<th>MSE</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>0.079806</td>
<td>0.227864</td>
<td>0.282499</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.079356</td>
<td>0.224512</td>
<td>0.261702</td>
</tr>
<tr>
<td>CNN-LSTM</td>
<td>0.076423</td>
<td>0.233839</td>
<td>0.280143</td>
</tr>
<tr>
<td>CNN-GRU</td>
<td>0.071410</td>
<td>0.217883</td>
<td>0.267227</td>
</tr>
<tr>
<td>PE-CNN-GRU</td>
<td>0.059035</td>
<td>0.110014</td>
<td>0.221132</td>
</tr>
</tbody>
</table>

The data from January 1, 2020 to December 31, 2020 is divided into training set and test set in a ratio of 7 : 3. The comparison of the prediction curves of each model is shown in Fig.4. It can be seen from Fig.4 that the PE-CNN-GRU hybrid neural network model proposed in this paper has a high degree of fitting with the actual load curve and has good prediction accuracy.

5. Conclusion

Aiming at the prediction of water quality parameter DO in sewage treatment process, a water quality prediction model based on PE-CNN-GRU hybrid neural network is constructed. Multiple sets of data are used as input to predict the time series of DO, and CNN is used to extract the time series characteristics of the input data. The iterative efficiency of the model is improved, and then the DO is predicted in a simpler and more efficient GRU network. The experimental results show that the PE-CNN-GRU hybrid model has higher prediction accuracy than the traditional single model RNN and LSTM, can well approximate the real water quality data. The MAE, MSE and RMSE are reduced to 0.110,0.059 and 0.221, and the prediction error fluctuates less. The number of training iterations required in the training process is less, and the prediction efficiency is improved. It is more suitable for the time series prediction of total nitrogen in sewage water quality parameters.

Acknowledgments

This work was supported in part by the Chinese Government under Shanghai 2020 Science and Technology Innovation Action Plan for Social Development Science and Technology Tackling Project 20dz1204600 Research and Demonstration of Key Technology for Intelligent Treatment of Combined Flow System Large Sewage Plan.
Reference:


