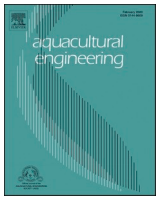
Aquacultural Engineering 91 (2020) 102122

Contents lists available at ScienceDirect

Aquacultural Engineering

journal homepage: www.elsevier.com/locate/aque



Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network

Xinkai Cao [a,b], Yiran Liu [a,c], Jianping Wang [d], Chunhong Liu [a], Qingling Duan [a,c,*]

[a] College of Information and Electrical Engineering in China Agricultural University, Beijing, 100083, China
[b] National Innovation Center for Digital Fishery, Beijing, 100083, China
[c] Beijing Engineering and Technology Research Center for Internet of Things in Agriculture, Beijing, 100083, China
[d] Ningbo Institute of Oceanography and Fisheries, Zhejiang, 315000, China

ARTICLE INFO

Keywords:
Dissolved oxygen prediction
K-means clustering
Gated recurrent unit neural network
Pond culture

ABSTRACT

Dissolved oxygen in water is an important ecological factor in ensuring the healthy growth of aquatic products, as hypoxic stress is known to restrict the growth of aquatic products. The accurate monitoring and prediction of dissolved oxygen is the key to precise regulation and control of pond aquaculture water quality. The current dissolved oxygen prediction model has some limitations, such as a short prediction period and inadequate prediction accuracy for actual production demands. Therefore, a prediction model of dissolved oxygen in pond culture was proposed based on K-means clustering and Gated Recurrent Unit (GRU) neural network. Firstly, the key factors affecting the changes in dissolved oxygen were selected by principal component analysis (PCA). The dissolved oxygen time series was then subjected to K-means clustering, and the dissolved oxygen prediction model was constructed using GRU. To improve the clustering effect, we enhanced the similarity calculation for the time series based on the variation of dissolved oxygen. This process combined the Euclidean distance with the dynamic time-warping distance. The proposed method can predict the dissolved oxygen content of aquaculture water over different time intervals according to the demands of real-world scenarios. The average absolute error of the 30-min interval model was 0.264, and the mean absolute percentage error was 3.5 %. Experimental results indicated that the proposed method achieves higher prediction accuracy and flexibility than the conventional approach.

1. Introduction

With the rapid growth of the world population and economic development, the human demand for animal protein continues to increase. As a result, more aquatic products have entered the market because of their unique nutritional value and rich protein content. In 2016, the total global output of aquatic products was about 202.1 million tons. The contribution of aquaculture to global fishing and aquaculture production increased gradually, from 25.7 % in 2000 to 46.8 % in 2016, and 20 % of the protein intake of the global population came from aquatic products (FAO, 2018).

Aquaculture water is an important habitat for aquatic products, and dissolved oxygen is a vital ecological factor in measuring water quality in the aquaculture environment. Dissolved oxygen directly affects the feed intake, bait conversion rate, and disease resistance of aquatic products, and any deficiencies in dissolved oxygen levels may affect production (Culp et al., 2017). Hypoxic stress adversely affects the normal growth of aquaculture products, and can even lead to large-scale deaths. This causes immeasurable losses to farmers, and is not conducive to the healthy development of the entire aquaculture industry. Therefore, monitoring the dissolved oxygen in water and regulating water quality are key factors in reducing aquaculture risks and ensuring the healthy growth of aquatic products.

Pond culture is the dominant mode of freshwater aquaculture. Because of the open environment of pond culture, the concentration of dissolved oxygen in aquaculture water is easily affected by a variety of ecological and environmental factors, and changes in dissolved oxygen levels are more complex. At present, methods of predicting the dissolved oxygen concentration in the aquaculture environment fall into two main categories (Duan et al., 2018). One is single-parameter prediction, whereby the dissolved oxygen is predicted using the change law of historical dissolved oxygen data. Single-parameter prediction only

* Corresponding author at: College of Information and Electrical Engineering, China Agricultural University, No. 17 Tsinghua East Road, Beijing, 100083, China.
E-mail address: dqling@cau.edu.cn (Q. Duan).

https://doi.org/10.1016/j.aquaeng.2020.102122
Received 11 October 2019; Received in revised form 28 March 2020; Accepted 3 September 2020
Available online 11 September 2020

considers changes in dissolved oxygen itself, neglecting the correlation with other parameters. The second approach is multi-parameter prediction, which uses water quality parameters and meteorological parameters as the input for prediction, fully considering the complexity of dissolved oxygen changes and the correlation with other environmental parameters. Hence, the multi-parameter prediction results are more accurate. Therefore, in studies of dissolved oxygen prediction, multi-parameter techniques have become very popular (Liu et al., 2013; Chen et al., 2018a,b).

The characteristics of dissolved oxygen in pond aquaculture water vary with meteorological and diurnal changes. Under better weather conditions, the dissolved oxygen content in water is higher during the day, but decreases significantly at night and in the early morning. To prevent water quality deterioration and aquatic product death caused by hypoxia, intelligent algorithms are used to predict the dissolved oxygen in water after a period of time. According to the predicted results, the aerator parameters can be set ahead of time. When the dissolved oxygen content falls below some threshold, the aerator starts automatically and the oxygen level will be increased.

In conclusion, dissolved oxygen in pond culture is affected by many water quality and meteorological factors, and diurnal variations have distinct characteristics. Under similar environmental conditions, variations in dissolved oxygen exhibit certain similarities. To improve the prediction accuracy and practicability, a flexible, accurate, and practical dissolved oxygen prediction model should be established. Such a model should not only consider the influences of multi-environmental factors on the change in dissolved oxygen levels, but also calculate appropriate model parameters of according to the data characteristics of different periods.

Based on the above considerations, a multi-parameter prediction model of dissolved oxygen in pond culture was established based on K-means clustering and gated recurrent unit neural network (GRU). In the model, principal component analysis (PCA) was used to select key parameters, and the dissolved oxygen was predicted based on K-means clustering and GRU.

Clustering is an unsupervised learning method (Zhu et al., 2019a, b). The purpose of clustering is to divide a set of physical or abstract objects into multiple classes consisting of similar objects. K-means clustering divides the sample set of historical time series into several classes with different data characteristics by calculating the similarity among the time series. GRU controls the extent to which the state information of the previous moment is brought into the current state by update gate, and controls the neglect degree of the state information of the previous moment by reset gate. It has the characteristics of simple structure, less parameters and strong convergence, and can effectively process time series data. These algorithms provide the theoretical basis for our research.

The main work of this paper includes: (1) According to the variation law and trend characteristics of dissolved oxygen in ponds, different variation characteristics of dissolved oxygen time series were mined by clustering. A GRU prediction model was constructed according to the variation law, ensuring that the proposed algorithm was adaptable and robust. (2) To improve the clustering effect, similarity calculation methods based on the Euclidean distance (ED) and dynamic time warping (DTW) were used to consider the similarity degree of values and time series trends. (3) In the construction of datasets, tag sets were constructed using dislocation interception, which freely selected input and output parameters of the model, as well as the prediction interval of parameters, ensuring that the model had high flexibility and versatility.

The remainder of this paper was organized as follows. Section 2 described some related studies, before Section 3 introduced the proposed method. Section 4 described specific experiments to evaluate the proposed method and analysed the results. Finally, Section 5 summarized the conclusions from this study.

2. Related works

2.1. Similarity computations in cluster analysis

Clustering is usually an unsupervised learning method in which a large number of unlabeled data are divided into several categories according to the inherent similarity between them. Therefore, similarity calculation is the basis of cluster analysis (Zhang et al., 2015). A similarity distance is used to describe the real number of differences between different time series. Its value usually depends on the selected distance function. A smaller difference between two time series implies there is a smaller distance and a higher similarity between them. The similarity of sequences depends not only on the proximity of their values, but also on the similarity of the shapes of two sequences, that is, the similarity of their trends (Zhang and Pi, 2017). Currently, popular similarity calculation methods include ED (Junye et al., 2017; Kapil et al., 2016; Kapil and Chawla, 2017), feature subsequence distance (FSD) (Zolhavarieh et al., 2014), and DTW (Xi et al., 2017).

ED is the most commonly used similarity calculation method. It is simple and fast in dealing with equal-length time series, and satisfies the triangular inequality. However, it is not effective in dealing with unequal-length time series and local time warping. FSD uses local segments of time series, which largely avoids the influence of noise, but the huge time overhead hinders its wider application. DTW (Li, 2015; Ye et al., 2017) uses the idea of dynamic programming to find the shortest matching path. This can deal with time series of different lengths, is adaptable to noise, and supports amplitude translation and time axis expansion and bending. However, the computational complexity of the DTW distance is high, and so it is mainly applied to one-dimensional time series. Additionally, DTW does not satisfy the triangular inequality. To overcome these problems, many scholars have improved the DTW method, but it is not suitable for pond culture time series data.

Pond culture is the result of multiple production factors. The time series data produced in the process of pond culture are complex and closely related, having high-dimensional, dynamic, and uncertain characteristics. Dissolved oxygen in pond culture water is not only related to water quality factors, but also to meteorological factors. Traditional similarity calculation methods tend to neglect the iteration effect and internal relationship of production data, and are not suitable for the analysis of multi-dimensional pond culture time series data. Therefore, finding a similarity calculation method that is suitable for the characteristics of pond culture time series data is the basis of data analysis.

2.2. Prediction model of dissolved oxygen

The pond culture environment is a physical system with openness, variability, and complexity (Khani and Rajaee, 2016). Many scholars have developed prediction methods for solving the problem of pond aquaculture water quality (Csábrági et al., 2017). Some mechanism-based water quality prediction models have been constructed according to the physical and chemical properties of aquaculture. Other non-mechanism-based water quality prediction models have been constructed based on statistical methods (Qin et al., 2018).

In recent years, machine learning has been widely used in water quality prediction. These forecasting methods can be divided into two categories, single-parameter prediction models and multi-parameter prediction models. The main algorithms used include grey theory methods (Zhang et al., 2017; Li et al., 2012), neural networks (Tomić et al., 2018a,b; Ruben et al., 2018), support vector machines (Ji et al., 2017; Heddam and Kisi, 2018), and least-squares support vector machines (Rubio et al., 2015; Li et al., 2018; Yu et al., 2016).

2.2.1. Single-parameter dissolved oxygen prediction model

Single-parameter prediction methods have several limitations in the process of actual dissolved oxygen prediction, such as the single input

parameter and unstable prediction effects. Thus, there are few studies on single-parameter prediction at present (Zhu et al., 2017a,b; Huan et al., 2018a,b).

Yan et al. (Yan et al., 2014) used a water quality monitoring system based on a back-propagated (BP) neural network to detect and predict the trend in dissolved oxygen changes using historical real-time water quality monitoring data from crab pond culture. Huan et al. (Huan et al., 2018a,b) proposed a water quality prediction model based on empirical mode decomposition and a least-square support vector machine (LSSVM) in 2018. Empirical mode decomposition was used to obtain the variation characteristics of the water quality time series at different scales, and a Bayesian evidence framework was used to optimize the parameters of LSSVM to improve the prediction accuracy. Liu et al. (Liu et al., 2014) denoised the original dissolved oxygen time series and decomposed into several subsets by wavelet denoising, and an LSSVM was constructed for prediction. By decomposing the dissolved oxygen time series, each decomposed time series could be predicted, and then the final prediction results were obtained by combining multiple prediction values. However, the dissolved oxygen content in pond culture is not only related to water quality factors, but also to meteorological factors and aeration operations (Chen and Liu, 2014). The above-mentioned models do not take into account the close relationship between dissolved oxygen and other environment factors. As a result, the practicality of these models in the actual production and aquaculture process is not good.

2.2.2. *Multi-parameter dissolved oxygen prediction model*

Multi-parameter dissolved oxygen prediction models use key environmental parameters collected by the Internet of Things as input and predict the future dissolved oxygen content as output (Ahmed, 2017; Chen et al., 2018a,b).

Zhu et al. (Zhu et al., 2016) used a chaotic mutation process to optimize least-squares support vector regression (LSSVR). Liu et al. (Liu et al., 2012) optimized the model parameters of LSSVR by ant colony optimization, and a nonlinear prediction model was constructed for dissolved oxygen. The influencing factors were automatically obtained by the combination of the best parameters, thus improving the accuracy and efficiency of dissolved oxygen prediction. Li et al. (Li et al., 2017) took 16 predictive factors as inputs, including physical factors, nutrients, organic substances, and metal ions. Using 969 samples, the dissolved oxygen content was predicted by multiple linear regression (MLR), BP neural network (BPNN), and support vector machine (SVM) models. PSO was used to optimize the relevant parameters. The results showed that the PSO-SVM model outperformed the MLR and PSO-BPNN models.

Ta and Wei (Ta and Wei, 2018) adopted an improved convolutional neural network (CNN) model to predict dissolved oxygen from the conductivity, temperature, dissolved oxygen, and pH. The input vector was multiplied by its own transposition to form the input matrix, and the characteristics of the water quality parameter factors affecting dissolved oxygen were refined by two continuous convolutions of the input matrix. Yang et al. (Yang et al., 2014) selected air humidity, dissolved oxygen, water temperature, solar radiation, wind speed, and atmospheric pressure as inputs through PCA, and BPNN was employed to predict dissolved oxygen in crab culture ponds within an hour. These studies used modern intelligent algorithms to overcome, to a certain extent, the problem of low accuracy and unstable prediction results of single-parameter techniques. However, most of the prediction periods are fixed, which does not meet the actual needs of aquaculture production.

3. Proposed method

In this paper, a time-series-based dissolved oxygen prediction method was proposed. The method divided the time series using a sliding window to obtain the sample set. The sample set was classified by K-means clustering based on an improved similarity calculation method, and k sample sets with different data characteristics were obtained. The GRU model was constructed for different characteristics. The prediction process of the model is shown in Fig. 1.

Step 1: Data preprocessing. Data collected by Internet of Things sensors are usually vulnerable to environmental or human factors. This may lead to inaccurate data and inconsistent data formats. Therefore, data preprocessing was needed.

Step 2: Selection of key factors based on PCA. Using PCA, the key factors with the greatest influence on dissolved oxygen levels were selected as the input parameters of the model.

Step 3: K-means clustering based on ED-DTW. In this paper, a method based on ED and DTW was used. The optimal weight coefficients and number of clusters were found by comparing the clustering evaluation index and the parameters were adjusted accordingly.

Step 4: Prediction model construction. For each training set, an GRU model was constructed, and the built model was tested by using test sets.

3.1. *Selection of key factors based on PCA*

The variation of dissolved oxygen in pond aquaculture water is affected by many water quality and meteorological factors. If all the collected factors were used as inputs, the model would be very complex, the prediction time would become very long, and the prediction effect may be poor. Therefore, PCA is used to reduce the dimension of the original acquisition parameters, and effective key factors are selected as input parameters to predict dissolved oxygen. PCA (Peng et al., 2017; Zhen et al., 2013) is a statistical analysis method that removes many

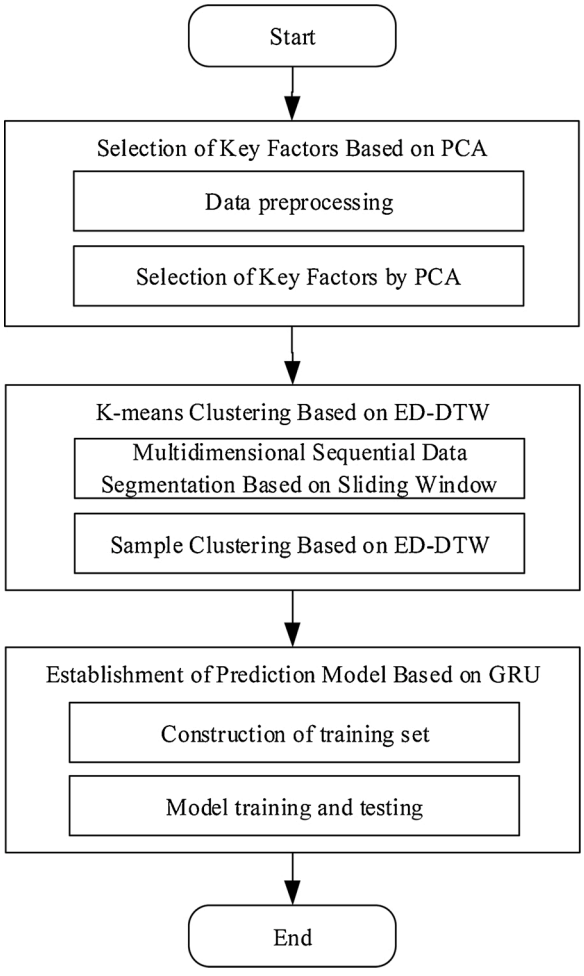


Fig. 1. Flowchart of prediction model.

variables and retains a few principal components (comprehensive variables) by dimension reduction technology. These principal components reflect most of the information of the original variables, and are usually expressed as some linear combination of the original variables. PCA (Peng et al., 2016; Zhu et al., 2017a,b) effectively reduces the number of predictive variables. The calculation steps of PCA are as follows.

(1) The original data are arranged in rows to form a matrix X, which contains m rows of data and n feature dimensions. In this paper, m denotes the number of samples collected and n denotes the types of environmental factors collected by the Internet of Things.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \quad (1)$$

(2) Normalize the entries of X to have a zero mean, and compute the normalized matrix Z.

$$Z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, i = 1, 2...., m; j = 1, 2, ..., n \quad (2)$$

where $\overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{s_j}$, $s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1}$.

(3) Find the covariance matrix C of X.

$$C = \frac{1}{m} Z^T Z \quad (3)$$

(4) Calculate the eigenvectors of the covariance matrix and rank the eigenvalues from large to small. Extract the first k columns of the eigenvector matrix to form the matrix P.
(5) Calculate $Y = ZP$ and obtain the reduced-dimension data matrix Y.
(6) Compute the contribution rate V_i of each feature root.

$$V_i = \frac{\lambda_i}{\sum \lambda_i} \quad (4)$$

where λ_i represents the eigenvalues.

In this study, the data collected by the Internet of Things were assumed to include water quality data such as water temperature, dissolved oxygen, pH, and meteorological data such as air pressure, wind speed, and solar radiation. We used PCA to reduce the dimensionality and selected key factors as input parameters for predicting the dissolved oxygen level.

3.2. K-means clustering based on ED-DTW

The diurnal variation trend of dissolved oxygen in pond culture has very obvious characteristics. If the time series data of dissolved oxygen are predicted directly, the prediction accuracy may be low because of the different data characteristics and variation trends, resulting in a poor prediction effect. For this reason, clustering was used to divide the dissolved oxygen time series data and cluster the data with similar trends. The prediction model was then constructed according to the data characteristics.

The similarity calculation is the basis for analyzing time series data. The similarity distance is a positive real number used to describe the difference between time series. Its value usually depends on the selected distance function. At present, a single distance index, such as ED or DTW, is used to calculate the similarity. However, because of the complex trend of dissolved oxygen under different environmental conditions, a single distance index is not suitable for calculating the similarity among dissolved oxygen time series. Therefore, a weighted combination of ED and DTW was used to calculate the similarity. This method not only considers the proximity of sequence values, but also the similarity of two sequence shapes, that is, the similarity of their trends. The calculation formula is as follows.

The ED between point (x_2, y_2) and point (x_1, y_1) is:

$$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (5)$$

where ρ represents the ED between two points.

For two time series $Q = \{q_1, q_2 \dots q_m\}$ and $C = \{c_1, c_2 \dots c_n\}$ with lengths m and n, the DTW distance γ can be calculated accumulatively using a dynamic programming method.

$$\gamma(i,j) = d(q_i, c_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (6)$$

$$d(q_i, c_j) = (q_i, c_j)^2 \quad (7)$$

where i and j represent any two points in Q and C, respectively. The formulas for calculating the similarity distance in the proposed model are as follows.

$$dist = \alpha \times d_1 + \beta \times d_2 \quad (8)$$

$$\alpha + \beta = 1$$

where, d_1 is the ED between two time series, d_2 is the DTW between two time series, α is the value similarity coefficient, and β is the shape similarity coefficient.

K-means clustering accepts an input value k, and then divides n data objects into k clusters. The basic idea of the algorithm is to assign the sample points to classes so that the similarity among samples in the same cluster is as high as possible and the similarity between samples in different clusters is as low as possible. Through an iterative process, the cluster partition $C = \{C_1, C_2 \dots C_k\}$ minimizes the squared difference E.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2 \quad (9)$$

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (10)$$

In (9), x is the eigenvector of cluster C_i, μ_i is the mean vector of cluster C_i, and $|C_i|$ is the number of eigenvectors of cluster C_i.

In this paper, we considered a sampling frequency of 30 min. Using the sliding window method, the sample set of time series was constructed with 10 consecutive samples as the sliding window length and 30-min intervals as the moving step length. For the constructed sample set, the best weight coefficients and the number of clustering centers k were determined by calculating the clustering evaluation index.

3.3. Prediction model based on GRU

GRU is a kind of Recurrent Neural Network (RNN). For tasks that involve sequential inputs, it is often better to use RNNs. RNNs process an input sequence one element at a time, maintaining in their hidden units a 'state vector' that implicitly contains information about the history of all the past elements of the sequence. RNNs are prone to extreme nonlinear behavior, including gradient vanishing and gradient explosion, when the error gradient propagates backward in several time steps. For RNNs, the above phenomena may occur as long as the sequence length is sufficient. Gating algorithm is a feasible method to cope with long-distance dependence of RNNs. The idea is that the gating unit gives the RNNs the ability to control its internal information accumulation, which can not only grasp long-distance dependence but also selectively forget information to prevent overload. Long Short-Term Memory

(LSTM) is the earliest gating algorithm of RNNs, which consists of three gating gates: input gate, forgetting gate and output gate. Because the three gates in LSTM have different contributions to improve the learning ability, omitting the gates with small contributions and their corresponding weights can simplify the structure of the neural network and improve its learning efficiency.

GRU is an algorithm based on the above idea. Its corresponding unit contains only two gates: update gate and reset gate. The extent to which the state information of the previous moment is brought into the current state is controlled by update gate, and the neglect degree of the state information of the previous moment is controlled by reset gate. Compared with LSTM model, GRU model is simpler, has fewer parameters, and fewer tensor operations, so it has faster convergence speed and shorter prediction time. With the same number of parameters, GRU has better comprehensive performance than LSTM model. The update mode is as follows.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{11}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{12}$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \tag{13}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{14}$$

where, h_t denotes the state of the system at t time, z denotes the update gate, r denotes the reset gate, σ denotes the Sigmoid function, and W denotes a set of weights.

In this paper, dissolved oxygen, water temperature, wind speed, atmospheric temperature, atmospheric pressure, solar radiation, relative humidity, and other parameters were used as input to the GRU prediction model. The dissolved oxygen content after different prediction time intervals was the output value. Based on this, an accurate dissolved oxygen prediction model was constructed.

4. Experiments and analysis

4.1. Data sources and preprocessing

The experimental data for this study were taken from Penaeus vannamei ponds in Ayue Aquaculture Farm, Fenghua, Ningbo City, Zhejiang Province, China. The total area of pond culture at this farm is 1.04×10^5 m^2. The experimental data were collected from September–October 2018. The measured water quality parameters were water temperature, dissolved oxygen, acidity, alkalinity, and conductivity. The meteorological parameters were atmospheric humidity, atmospheric temperature, atmospheric pressure, wind speed, wind direction, solar radiation, and rainfall. Water quality parameters were sampled every 10 min, giving a total of 10,936 sample data, and the meteorological parameters were sampled every 30 min, providing a further 6,384 sample data.

Data preprocessing includes the following four steps.

Step 1: Data cleaning. In this paper, the abnormal data was detected according to the Grubbs Criterion, and the error data were replaced using the global average of the parameter sequence.

Step 2: Data conversion. In the networking of aquatic products, the acquisition frequency of water quality parameters was 10 min, whereas that of meteorological parameters was 30 min. Most of the water quality parameters do not change much in 30 min, so the collected water quality parameters were converted to their average values over 30 min.

Step 3: Data fusion. The meteorological data and water quality parameters collected by the Internet of Things were not collected synchronously. Meteorological data collected at the closest time to each group of water quality data were matched, and then the two data points were fused.

Step 4: Data specification. The sequence of water quality parameters was normalized to obtain a normalized sequence of water quality parameters. The multidimensional water quality parameter sequence can be represented as an $m \times n$ matrix M, where m is the number of water quality parameters and n is the length of each water quality parameter sequence. The formula for normalizing M is as follows.

$$x = \frac{x' - x_{\min}}{x_{\max} - x_{\min}} \tag{15}$$

where x is the normalized value, x' is the original value, x_{min} is the minimum of the original values, and x_{max} is the maximum of the original values.

4.2. Evaluation index

4.2.1. Evaluation index of K-means clustering

The Davies–Bouldin Index (DBI) was used to evaluate the clustering performance. This is the sum of the average distances in any two classes is divided by the distance between the two clustering centers. Smaller values signify better performance. The calculation formula is as follows.

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{avg(C_i) + avg(C_j)}{d_{cen}(\mu_i, \mu_j)} \right) \tag{16}$$

$$avg(C) = \frac{2}{|C|(|C| - 1)} \sum_{1 \le i \le j \le |C|} dist(x_i, x_j) \tag{17}$$

$$d_{cen}(C_i, C_j) = dist(\mu_i, \mu_j) \tag{18}$$

where $avg(C)$ denotes the average distance within class C, $dist(\cdot, \cdot)$ denotes the distance between two samples, and μ denotes the center point of cluster C.

4.2.2. Evaluation index of water quality parameter prediction

The mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE) were used to evaluate the accuracy of the water quality prediction model. These indexes are calculated as follows.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \tilde{y}_i| \tag{19}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \tilde{y}_i}{y_i} \times 100\% \right| \tag{20}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2} \tag{21}$$

where y_i is the true value, $\tilde{y}_i$ is the predicted value, and N is the number of predicted values.

4.3. Experimental settings

To achieve the accurate prediction of dissolved oxygen, we designed the following experiments.

(1) Key factor selection using PCA.

The sensor data collected through the Internet of Things include dissolved oxygen, water temperature, pH, and meteorological data such as temperature, air pressure, and solar radiation. To reduce the computational complexity of the algorithm and increase the prediction accuracy of the model, the key influencing factors related to changes in dissolved oxygen were selected as the prediction factors through PCA.

(2) K-means clustering using ED-DTW.

The similarity calculation method based on ED-DTW was used to change and combine the weights of different distance coefficients. Through repeated experiments and validation, the distance evaluation indexes were compared, the optimal weight distribution of different distance coefficients was obtained, and the optimal number of clustering centers K was calculated.

(3) Prediction of dissolved oxygen using GRU

The K-means clustering was applied to obtain k datasets. In these datasets, 10 % of the data were selected as a test set, and the remainder were used as the training set.

The experimental coding environment was MyEclipse 2017 CI and Anaconda3 (64bit). The sample set clustering experiment based on K-means clustering was realized by MyEclipse 2017 CI, and the prediction of water quality parameters based on GRU was realized by Anaconda3 (64bit).

4.4. Experimental results and analysis

4.4.1. PCA experiments on original data sets

To reduce the complexity of the model and improve the accuracy of prediction, the PCA method was used to select the key factors, and the variance and principal component contribution rate are shown in Table 1.

From Table 1, it can be seen that the feature values of the first four components were all greater than 1. According to the principle of selecting the components with "feature value" greater than 1 as the principal components, the first four components were selected instead of the original variables.

The factor loadings of each feature to different principal components were obtained using the orthogonal rotation method with Kaiser standardization, and the obtained component matrix is shown in Table 2.

As shown in Table 2, water temperature, atmospheric humidity, and atmospheric temperature contributed the most to factor 1; atmospheric pressure contributed the most to factor 2, dissolved oxygen contributed the most to factor 3, and wind speed contributed the most to factor 4. In the selection of factors, the contribution rate of these factors was greater than 0.6, which can be considered to cover most of the variable information. Combined with relevant references (Chen et al., 2018a,b; Ta and Wei, 2018), the key influencing factors selected in this paper were dissolved oxygen, water temperature, wind speed, atmospheric temperature, atmospheric pressure, solar radiation, relative humidity, which were basically consistent with the key influencing factors selected by experts in aquaculture field based on their experience, reducing the data dimension and complexity.

Table 2
Rotating component matrix obtained by PCA.

	Component			
	1	2	3	4
Dissolved oxygen	.195	.486	**.673**	.103
Water temperature	**.751**	−.365	.040	−.101
pH	.598	−.317	.502	.182
Conductivity	−.076	−.150	.540	−.379
Atmospheric humidity	**−.827**	−.249	−.090	−.048
Atmospheric temperature	**.878**	−.216	−.136	.002
Atmospheric pressure	−.006	**.824**	.110	−.203
Wind direction	−.027	−.011	.589	.328
Wind speed	.142	−.208	.181	**.675**
Solar radiation	.406	.254	.078	.464
Rainfall	−.234	−.143	−.070	.457

4.4.2. K-means clustering based on ED-DTW

In order to find the best cluster center number and the best coefficient, we set ED coefficient, DTW coefficient and cluster center number k to calculate the DBI values under different combinations of variables, and published the results of different combinations as shown in Table 3.

At the same time, we drew the change of DBI corresponding to different K values under different coefficient combinations, as shown in Fig. 2.

It can be seen from the figure that when $k = 2$, the coefficient combination was $ed = 0.3$, $DTW = 0.7$, the corresponding DBI was the lowest, indicating that the clustering effect was the best under sub-combination.

4.4.3. Prediction of dissolved oxygen based on GRU

To prove that the method proposed in this paper achieves high prediction accuracy and has practical value, we not only compared it with the classical method, but also performed simulation experiments with more advanced methods in recent years over different prediction times and using different evaluation indexes.

Table 3
Different weight coefficients and their corresponding DBI.

ED coefficient	DTW coefficient	2 DBI	3	4	5	6
0	1	2.112	2.175	2.398	2.294	2.163
0.1	0.9	2.083	2.146	2.262	2.275	2.286
0.2	0.8	2.170	2.241	2.213	2.362	2.396
0.3	**0.7**	**1.986**	2.072	2.206	2.261	2.144
0.4	0.6	2.115	2.066	2.270	2.170	2.207
0.5	0.5	2.117	2.353	2.341	2.274	2.359
0.6	0.4	2.212	2.267	2.257	2.388	2.665
0.7	0.3	2.213	2.262	2.345	2.425	2.553
0.8	0.2	2.305	2.399	2.416	2.588	2.534
0.9	0.1	2.754	2.597	2.744	2.710	2.575
1	0	2.602	2.639	2.796	2.731	2.823

Table 1
Variance and principal component contribution rate.

Component	Initial eigenvalue			Extract the sum of load squares		
	Total	Variance percentage	Accumulate %	Total	Variance percentage	Accumulate %
1	3.403	30.935	30.935	3.403	30.935	30.935
2	2.329	21.175	52.109	2.329	21.175	52.109
3	1.258	11.440	63.549	1.258	11.440	63.549
4	1.004	9.131	72.680	1.004	9.131	72.680
5	.898	8.164	80.844			
6	.627	5.697	86.540			
7	.574	5.214	91.755			
8	.486	4.422	96.177			
9	.231	2.097	98.274			
10	.123	1.116	99.390			
11	.067	.610	100.000			

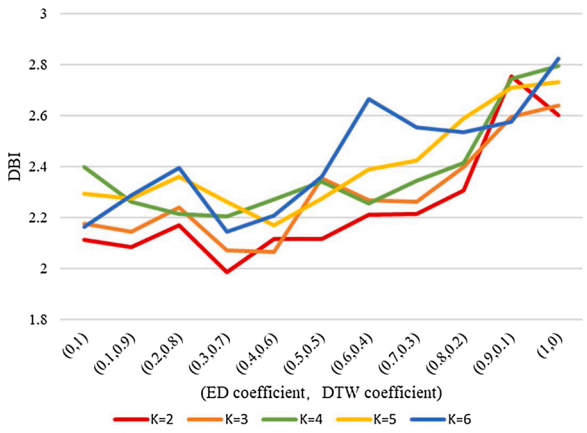


Fig. 2. Changes of DBI under different combinations.

The results of different models using a 30-min prediction interval are presented in Table 4.

From the above table, when the interval of dissolved oxygen prediction was 30 min, the proposed method outperformed the non-clustered GRU model by 6.4 % in terms of MAE, 0.9 % in terms of MAPE, and 7.6 % in terms of RMSE. Compared with the PCA-LSTM model, the improvements were 28.3 %, 27.6 %, and 22.1 %, respectively. Compared with the PCA-CNN model, the improvements were 35.3 %, 37.2 %, and 33.6 %, respectively. Compared with the PCA-ELM model, the improvements were 27.9 %, 23.4 %, and 28.4 %, respectively. Compared with the PCA-KNN model, the improvements were 35.5 %, 26.4 %, and 32.5 %, respectively.

The results of different models using a 60-min prediction interval are presented in Table 5.

From the above table, when the interval of dissolved oxygen prediction was 60 min, the proposed method outperformed the non-clustered GRU model by 2.5 % in terms of MAE, 5.4 % in terms of MAPE, and 22.6 % in terms of RMSE. Compared with the PCA-LSTM model, the improvements were 40.9 %, 35.7 %, and 40.4 %, respectively. Compared with the PCA-CNN model, the improvements were 39.8 %, 30.4 %, and 39.1 %, respectively. Compared with the PCA-ELM model, the improvements were 27.8 %, 21.8 %, and 30.7 %, respectively. Compared with the PCA-KNN model, the improvements were 51.1 %, 41.5 %, and 51.1 %, respectively.

The results of different models using a 90-min prediction interval are presented in Table 6.

From the above table, when the interval of dissolved oxygen prediction was 90 min, the proposed method outperformed the non-clustered GRU model by 8.2 % in terms of MAE, 6.3 % in terms of MAPE, and 15 % in terms of RMSE. Compared with the PCA-LSTM model, the improvements were 19.7 %, 18.2 %, and 21.5 %, respectively. Compared with the PCA-CNN model, the improvements were 17.0 %, 16.2 %, and 19.2 %, respectively. Compared with the PCA-ELM model, the improvements were 53.9 %, 54.7 %, and 49.3 %, respectively. Compared with the PCA-KNN model, the improvements were 62.1 %, 57.6 %, and 60.0 %, respectively.

The results of different models using a 120-min prediction interval are presented in Table 7.

From the above table, when the interval of dissolved oxygen prediction was 120 min, the proposed method outperformed the non-clustered GRU model by 14.5 % in terms of MAE, 12.1 % in terms of MAPE, and 4.0 % in terms of RMSE. Compared with the PCA-LSTM model, the improvements were 29.8 %, 26.0 %, and 19.8 %, respectively. Compared with the PCA-CNN model, the improvements were 31.0 %, 31.8 %, and 23.6 %, respectively. Compared with the PCA-ELM model, the improvements were 42.8 %, 44.8 %, and 34.6 %, respectively. Compared with the PCA-KNN model, the improvements were 60.9 %, 60.9 %, and 54.7 %, respectively.

The results of different models using a 300-min prediction interval are presented in Table 8.

From the above table, when the interval of dissolved oxygen prediction was 300 min, the proposed method outperformed the non-clustered GRU model by 10.6 % in terms of MAE, 17.1 % in terms of MAPE, and 16.4 % in terms of RMSE. Compared with the PCA-LSTM model, the improvements were 33.7 %, 38.6 %, and 33.0 %, respectively. Compared with the PCA-CNN model, the improvements were 41.9 %, 50.1 %, and 42.2 %, respectively. Compared with the PCA-ELM model, the improvements were 42.2 %, 45.5 %, and 36.9 %, respectively. Compared with the PCA-KNN model, the improvements were 65.5 %, 65.7 %, and 63.0 %, respectively.

At the same time, we drew the prediction error curves of some test sets on different prediction models, as shown in Fig. 3.

In the above figures, Y-axis represented the error value of real value

Table 4
Results of different models with 30-min prediction interval.

Model	Evaluating indicator		
	MAE	MAPE	RMSE
PCA-K-means-GRU	**0.264**	**3.509**	**0.353**
PCA-GRU	0.282	3.540	0.382
GRU	0.575	8.414	0.713
PCA-LSTM	0.368	4.849	0.453
PCA-CNN	0.408	5.584	0.532
PCA-ELM	0.366	4.579	0.493
PCA-KNN	0.409	4.767	0.523

Table 5
Results of different models with 60-min prediction interval.

Model	Evaluating indicator		
	MAE	MAPE	RMSE
PCA-K-means-GRU	**0.311**	**4.351**	**0.436**
PCA-GRU	0.319	4.597	0.563
GRU	0.871	11.865	1.098
PCA-LSTM	0.526	6.767	0.732
PCA-CNN	0.517	6.255	0.716
PCA-ELM	0.431	5.562	0.629
PCA-KNN	0.636	7.434	0.892

Table 6
Results of different models with 90-min prediction interval.

Model	Evaluating indicator		
	MAE	MAPE	RMSE
PCA-K-means-GRU	**0.347**	**4.538**	**0.476**
PCA-GRU	0.378	4.844	0.560
GRU	1.141	17.406	1.493
PCA-LSTM	0.432	5.546	0.606
PCA-CNN	0.418	5.417	0.589
PCA-ELM	0.753	10.02	0.938
PCA-KNN	0.916	10.701	1.187

Table 7
Results of different models with 120-min prediction interval.

Model	Evaluating indicator		
	MAE	MAPE	RMSE
PCA-K-means-GRU	**0.459**	**5.817**	**0.691**
PCA-GRU	0.537	6.616	0.720
GRU	1.622	25.148	1.895
PCA-LSTM	0.654	7.861	0.862
PCA-CNN	0.665	8.528	0.905
PCA-ELM	0.803	10.540	1.057
PCA-KNN	1.173	14.872	1.527

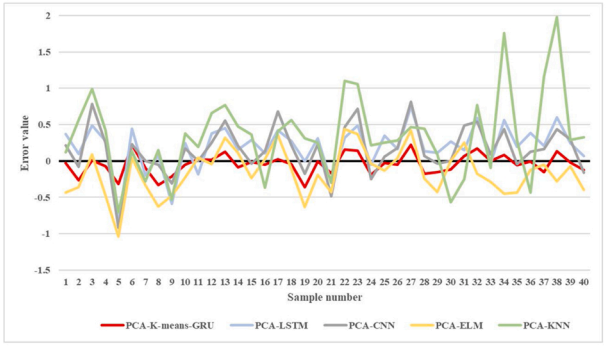
Table 8
Results of different models with 300-min prediction interval.

Model	Evaluating indicator		
	MAE	MAPE	RMSE
PCA-K-means-GRU	**0.941**	**10.956**	**1.256**
PCA-GRU	1.053	13.217	1.503
GRU	4.213	68.777	4.898
PCA-LSTM	1.419	17.852	1.874
PCA-CNN	1.619	21.938	2.172
PCA-ELM	1.629	20.120	1.989
PCA-KNN	2.729	31.908	3.398

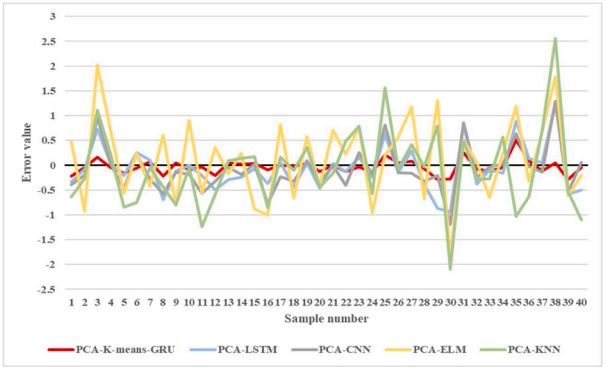
and predicted value, and X-axis represented some test set samples. It can be seen from the above figures that in the short prediction period, the error value of PCA-Kmeans-GRU model basically fluctuated slightly near the 0 axis, indicating that the error between the predicted value and the real value was about 0, while the fluctuation range of other models was relatively large, indicating that the error between the predicted value and the real value of the model was unstable and the difference was large;

When the prediction time was long, although the prediction error of each model increased, the PCA-K-means-gru model had a more stable error curve and a smaller fluctuation range, which indicated that the prediction accuracy of the model was higher and the prediction result was stable, while the prediction error of other models was larger and the error curve fluctuates obviously. The experiment results indicated that compared with other models, the PCA-K-means-GRU model had a higher prediction accuracy and stability.

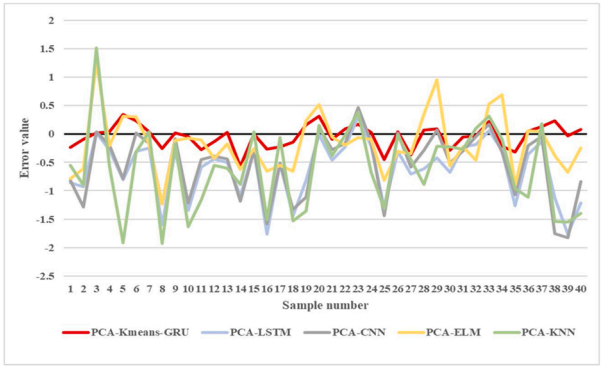
From the analysis above, we can conclude that when the prediction time of dissolved oxygen was short, the prediction accuracy of various algorithms was relatively high. As the prediction time interval became larger, the proposed method can still maintain high prediction accuracy, indicating that the proposed model offered high stability. For different prediction time intervals, the K-means-GRU model represented an



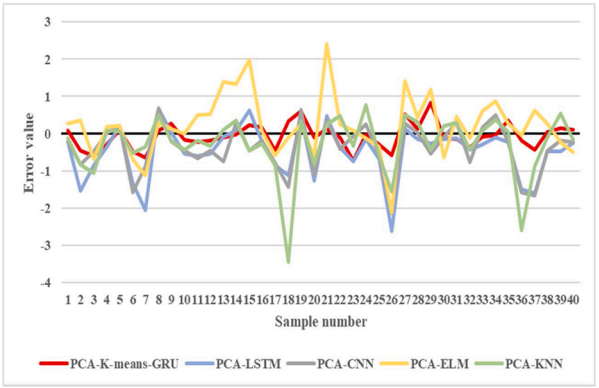
(a) Prediction error curve of each model in 30min



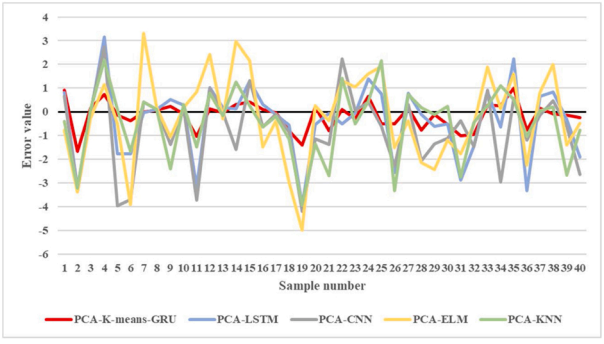
(c) Prediction error curve of each model in 90min



(b) Prediction error curve of each model in 60min



(d) Prediction error curve of each model in 120min



(e) Prediction error curve of each model in 300min

Fig. 3. Prediction error curve of each model under different prediction time length.

obvious improvement over GRU and the other models in each evaluation index, which indicated that the clustering method allowed the model to achieve higher prediction accuracy. Additionally, the model can flexibly set the prediction time according to demand, which greatly improved the generalization ability of the model.

5. Conclusion

This paper has described a model for predicting the dissolved oxygen in pond culture. Firstly, the key water quality and meteorological factors with the greatest influence on dissolved oxygen were selected by PCA. The time series of dissolved oxygen were then clustered, and prediction models based on GRU were constructed. To make the clustering results more reasonable and accurate, the ED and DTW similarity measures were synthetically combined, and the similarity degree of both the values and trends in the dissolved oxygen time series were considered. Finally, the dissolved oxygen content of aquaculture water in time intervals of 30 min – 300 min was predicted by the model. Compared with PCA-LSTM, PCA-ELM and PCA-CNN models, the proposed model produced more accurate prediction results over different time intervals. The key prediction factors related to the target parameters can be flexibly selected when the model was being built, and the prediction time can be set according to actual demand, ensuring high universality and high practical value.

In future work, different optimization algorithms will be used to optimize the parameter selection of the GRU model in an attempt to improve the experimental accuracy.

Funding

This research was supported by the Jiangsu Agricultural Science and Technology Innovation Fund [grant number CX(19)1003]; and the project of Ningbo Public Welfare Science and Technology [grant number 202002N3034].

CRediT authorship contribution statement

Xinkai Cao: Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing - original draft. **Yiran Liu:** Formal analysis, Methodology, Software. **Jianping Wang:** Data curation, Resources, Investigation. **Chunhong Liu:** Validation, Writing - review & editing. **Qingling Duan:** Funding acquisition, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aquaeng.2020.102122.

References

Ahmed, A.A.M., 2017. Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). J. King Saud Univ. Eng. Sci. 29 (2), 151–158.

Chen, W.B., Liu, W.C., 2014. Artificial neural network modeling of dissolved oxygen in reservoir. Environ. Monit. Assess. 186 (2), 1203–1217.

Chen, Y., Cheng, Q., Fang, X.M., Yu, H., Li, D., 2018a. Prediction of dissolved oxygen in aquaculture water by principal component analysis and long-term memory neural network. Agric. Eng. 34 (17), 183–191.

Chen, Y., Yu, H., Cheng, Y., Cheng, Q., Li, D., 2018b. A hybrid intelligent method for three-dimensional short-term prediction of dissolved oxygen content in aquaculture. PLoS One 13, e01924562. https://doi.org/10.1371/journal.pone.0192456.

Csábrági, A., Molnár, S., Tanos, P., Kovács, J., 2017. Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. Ecol. Eng. 100, 63–72. https://doi.org/10.1016/j.ecoleng.2016.12.027.

Culp, J.M., Luiker, E., Glozier, N.E., Meding, M., Halliwell, D., Wrona, F.J., 2017. Dissolved oxygen relationships of Under-Ice water column and pore water habitat: implications for environmental guidelines: dissolved oxygen levels under river-ice. River Res. Appl. 33 (3), 461–468.

Duan, Q.L., Liu, Y.R., Zhang, L., Li, D.L., 2018. Research progress and development trend analysis of big data technology in aquaculture. Agric. Mach. 49 (6), 1–16.

FAO, 2018. The State of world fisheries and aquaculture 2018. In: Meeting the Sustainable Development Goals. Rome. Licence: CC BY-NC-SA 3.0 IGO. http://www.fao.org/3/i9540en/I9540EN.pdf.

Heddam, S., Kisi, O., 2018. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J. Hydrol. 559, 499–509. https://doi.org/10.1016/j.jhydrol.2018.02.061.

Huan, J., Cao, W., Qin, Y., 2018a. Prediction of dissolved oxygen in aquaculture based on EEMD and LSSVM optimized by the Bayesian evidence framework. Comput. Electron. Agric. 150, 257–265.

Huan, J., Cao, W., Qin, Y., Wu, F., 2018b. Dissolved oxygen prediction of aquaculture water based on run-length detection reconstructing collective empirical modes. Agric. Eng. 34 (8), 220–226.

Ji, X., Shang, X., Dahlgren, R.A., Zhang, M., 2017. Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of Wen-Rui Tang River, China. Environ. Sci. Pollut. Res. 24 (19), 1–15.

Junye, G., Liu, M., Zhang, D., 2017. Application of effective distance in clustering algorithms. J. Front. Comput. Sci. Technol. 11, 406–413.

Kapil, S., Chawla, M., 2017. Performance evaluation of K-means clustering algorithm with various distance metrics. Int. J. Comput. Appl. 110, 12–16.

Kapil, S., Chawla, M., Ansari, M.D., 2016. On K-means data clustering algorithm with genetic algorithm. In: Fourth International Conference on Parallel, Distributed and Grid Computing. Waknaghat, India, 22-24 December 2016, pp. 202–206.

Khani, S., Rajaee, T., 2016. Modeling of dissolved oxygen concentration and its hysteresis behavior in rivers using wavelet Transform-based hybrid models. Clean Soil Air Water 45, 15003952. https://doi.org/10.1002/clen201500395.

Li, H., 2015. On-line and dynamic time warping for time series data mining. Int. J. Mach. Learn. Cybern. 6 (1), 145–153.

Li, Z., Jiang, Y., Yue, J., Zhang, L., Li, D., 2012. An improved gray model for aquaculture water quality prediction. Intell. Autom. Soft Comput. 18, 557–567. https://doi.org/10.1080/10798587.2012.10643265.

Li, X., Sha, J., Wang, Z., 2017. A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. Hydrol. Res. 48 (5), 1214–1225. https://doi.org/10.2166/nh.2016.149.

Li, X., Ai, J., Lin, C., Guan, H., 2018. Prediction model of dissolved oxygen in ponds based on ELM neural network. In: 2nd International Conference on Energy Engineering and Environmental Protection 2018. Sanya, China, 20-22 November 2018.

Liu, S., Xu, L.Q., Li, D., Z, L, 2012. Prediction model of dissolved oxygen in crab culture based on ant colony optimization least squares support vector regression. Agric. Eng. 28 (23), 167–175.

Liu, S., Xu, L., Li, D., Li, Q., Yu, J., Tai, H., Zeng, L., 2013. Prediction of dissolved oxygen content in river crab culture based on least squares support vector regression optimized by improved particle swarm optimization. Comput. Electron. Agric. 95 (4), 82–91.

Liu, S., Xu, L., Jiang, Y., Li, D., Chen, Y., Li, Z., 2014. A hybrid WA-CPSO-LSSVR model for dissolved oxygen content prediction in crab culture. Eng. Appl. Artif. Intell. 29, 114–124. https://doi.org/10.1016/j.engappai.2013.09.019.

Peng, X., Liang, C., Yu, Y., Wang, D., 2016. PCA-GRNN-GA based PH value prediction model applied in penaeus orientalis culture. In: 2016 6th International Conference on Digital Home. Guangzhou, China, 2-4 December 2016, pp. 227–232.

Peng, X., Xie, S., Yu, Y., Wu, Z., 2017. Fuzzy neural network based prediction model applied in primary component analysis. Clust. Comput. 20 (1), 1–10.

Qin, R., Long, Z., Wei, Y., Li, D., 2018. A method for predicting dissolved oxygen in aquaculture water in an aquaponics system. Comput. Electron. Agric. 151, 384–391.

Ruben, G.B., Ke, Z., Bao, H., Ma, X., 2018. Application and sensitivity analysis of artificial neural network for prediction of chemical oxygen demand. Water Resour. Manag. 32 (1), 1–11.

Rubio, F.C., Fernández, F.G.A., Pérez, J.A.S., Camacho, F.G., Grima, E.M., 2015. Prediction of dissolved oxygen and carbon dioxide concentration profiles in tubular photobioreactors for microalgal culture. Biotechnol. Bioeng. 62 (1), 71–86.

Ta, X., Wei, Y., 2018. Research on a dissolved oxygen prediction method for recirculating aquaculture systems based on a convolution neural network. Comput. Electron. Agric. 145, 302–310. https://doi.org/10.1016/j.compag.2017.12.037.

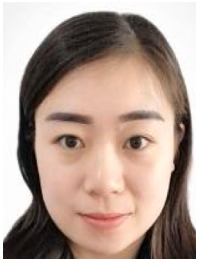
Tomić, A.Š, Antanasijeviä, D., Ristiä, M., Periä Grujiä, A., Pocajt, V., 2018a. A linear and non-linear polynomial neural network modeling of dissolved oxygen content in surface water: inter- and extrapolation performance with inputs' significance analysis. Sci. Total Environ. 610, 1038–1046. https://doi.org/10.1016/j.scitotenv.2017.08.192.

Tomić, A.Š, Antanasijević, D., Ristić, M., Perić-Grujić, A., Pocajt, V., 2018b. Application of experimental design for the optimization of artificial neural network-based water quality model: a case study of dissolved oxygen prediction. Environ. Sci. Pollut. Res. Int. 25 (10), 9360–9370.

Xi, W., Jiang, M., Sheng, C., Chao, Y., Wei, J., Guo, Z., 2017. A hybrid time series matching algorithm based on feature-points and DTW. In: 9th International

Symposium on Computational Intelligence & Design. Hangzhou, China, 10-11 December 2016, pp. 171–175.
Yan, J.X., Yu, L.J., Mao, W.W., Cao, S.Q., 2014. Study on prediction model of dissolved oxygen about water quality monitoring system based on BP neural network. Adv. Mater. Res. 912-914, 1407–1411.
Yang, Y., Tai, H., Li, D., 2014. Real-time optimized prediction model for dissolved oxygen in crab aquaculture ponds using back propagation neural network. Sens. Lett. 12 (3), 723–729, 7.
Ye, Y., Niu, C., Jiang, J., Ge, B., Yang, K., 2017. A shape based similarity measure for time series classification with weighted dynamic time warping algorithm. In: 4th International Conference on Information Science & Control Engineering. Changsha, China, 21-23 July 2017, pp. 104–109.
Yu, H., Chen, Y., Hassan, S.G., Li, D., 2016. Dissolved oxygen content prediction in crab culture using a hybrid intelligent method. Sci. Rep. 6, 27292 https://doi.org/10.1038/srep27292.
Zhang, M., Pi, D., 2017. A novel method for fast and accurate similarity measure in time series field. In: IEEE International Conference on Data Mining Workshops. New Orleans, the United States of America, 18-21 November 2017, pp. 569–576.
Zhang, H., Li, Z., Sun, Y., Zhang, H., 2015. Hierarchical segmentation and similarity measure of time series. Comput. Eng. Appl. 51 (10), 147–151. https://doi.org/10.3778/j.issn.1002-8331.1305-0437.
Zhang, F., Xue, H., Ma, X., Wang, H., 2017. Grey prediction model for the chemical oxygen demand emissions in industrial waste water: an empirical analysis of China. In: 13th Global Congress on Manufacturing and Management. Zhengzhou, China, 28-30 November 2016, pp. 827–834.
Zhen, M., Song, X., Rong, W., Lei, G., 2013. A modified water quality index for intensive shrimp ponds of Litopenaeus vannamei. Ecol. Indic. 24, 287–293.
Zhu, C., Liu, X., L, H, H, J, Y, N, 2016. Optimization of dissolved oxygen prediction model for industrial aquaculture. Agric. Mach. 47 (1), 273–278.
Zhu, C., Liu, X., Ding, W., 2017a. Prediction model of dissolved oxygen based on FOA-LSSVR. In: 36th Chinese Control Conference. Dalian, China, 26-28 July 2017, pp. 9819–9823.
Zhu, X., Li, X., Zhang, S., Xu, Z., Yu, L., Wang, C., 2017b. Graph PCA hashing for similarity search. IEEE Trans. Multimed. 19 (9), 2033–2044. https://doi.org/10.1109/TMM.2017.2703636.
Zhu, X., Zhang, S., He, W., Hu, R., Lei, C., Zhu, P., 2019a. One-step multi-view spectral clustering. IEEE Trans. Knowl. Data Eng. 99, 1-1.
Zhu, X., Zhang, S., Li, Y., Zhang, J., Yang, L., Fang, Y., 2019b. Low-rank sparse subspace for spectral clustering. IEEE Trans. Knowl. Data Eng. 99, 1-1.
Zolhavarieh, S., Aghabozorgi, S., Teh, Y.W., 2014. A review of subsequence time series clustering. Sci. World J. 2014, 1–19. https://doi.org/10.1155/2014/312521.

Xinkai Cao received the B.S. degree in computer science and technology from China Agricultural University, Beijing, China, in 2018, where he is currently pursuing the M.S. degree in computer science and engineering from the College of Information and Electrical Engineering. His research interests include machine learning and water quality prediction.



Yiran Liu received M.S. degree in computer science and technology from China Agricultural University, Beijing, China. She is currently a teacher of College of information science and engineering, Shanxi Agricultural University. Her research interests include machine learning and time series modeling.

Jianping Wang received the B.S. degree in Biology Education from Hangzhou Normal University, Zhejiang, China, in 1989. He is currently a researcher of Ningbo Marine and Fisheries Research Institute, and he is employed by the Ministry of Agriculture and Zhejiang Province as experts in the prevention and control of aquatic diseases. His research interests include study on aquatic diseases and quality analysis of aquatic products.

Chunhong Liu received her Ph.D. degree from the School of Information and Communication Engineering of Harbin Engineering University in 2005, and she completed her post-doctoral work in Beijing University of Aeronautics and Astronautics in 2008. She is currently an associate professor and master's tutor at China Agricultural University. Her main research interest is signal and information processing.

Qingling Duan received the Ph.D. degree in agricultural electrification and automation from China Agricultural University, Beijing, China, in 2011. She is currently a Professor of computer science and technology and the Deputy Director of the Department of Computer Engineering, College of Information and Electrical Engineering, China Agricultural University. She was involved in the research domains of information processing and artificial intelligence.