Pol. J. Environ. Stud. Vol. 32, No. 6 (2023), 5171-5183
DOI: *10.15244/pjoes/168940* ONLINE PUBLICATION DATE: 2023-09-25

Original Research

Prediction of Optimal Coagulant Dosage Based on FCM-ISSA-ANFIS Hybrid Model

Jingrun Liang[1, 2#], Lisang Liu[1, 2*#]

[1]School of Electronic, Electrical Engineering and Physics, Fujian University of Technology, Fuzhou 350118, China
[2]Fujian Province Industrial Integrated Automation Industry Technology Development Base, Fuzhou 350118, China

Received: 18 March 2023
Accepted: 28 June 2023

Abstract

Aiming at the shortcomings of traditional Adaptive Neural-Fuzzy Inference System (ANFIS) in water quality prediction, such as low learning efficiency and poor prediction accuracy, this paper proposed an optimal coagulant dosage prediction hybrid model based on fuzzy C-means clustering algorithm (FCM) and improved sparrow search algorithm (SSA). The hybrid prediction model is named as FCM-ISSA-ANFIS. Firstly, the water quality data of drinking water treatment plant (WTP) are statistically characterized and Pearson correlation analysis is used to determine the input variables and output variables of ANFIS. Then, the water quality data is divided into training set and test set, and the divided data sets are clustered and analyzed by FCM to determine the new fuzzy rule numbers of ANFIS. What's more, the improved SSA is used to train the antecedent parameters and consequent parameters of ANFIS to accelerate the convergence of the algorithm and improve the ability of jumping out the local optimum. Compared with the traditional ANFIS model based on subtractive clustering, the experimental results show that the root mean square error (RMSE), mean absolute error (MAE) and standard deviation (SD) of the proposed FCM-ISSA-ANFIS for predicting the annual coagulant dosage of drinking WTP are decreased by 45.24%, 66.34% and 54.21% respectively. The proposed algorithm can not only solve the shortcomings of traditional ANFIS, but also has fast convergence and high accuracy, which can meet the real-time production demand of drinking WTP.

Keywords: fuzzy C-Means clustering, Adaptive Neuro-Fuzzy Inference System, sparrow search algorithm, coagulant dosage, hybrid model

Introduction

As an important part of the water purification process, coagulant dosage will directly affect the water quality, the cost of water purification and profit of the drinking WTP. Coagulant dosage is a complex physical and chemical reaction process, which has the characteristics of time-varying, nonlinear and large time delay [1]. In the traditional drinking WTP, the human experiences and the jar tests are mainly used to determine the coagulant dosage [2]. The human experience method is that the operator adds the coagulant according to the turbidity of the raw water and his own experience. The jar test method is to simulate the coagulation dosing environment according to the raw water quality, so as to determine the coagulant dosage amount.

[#]These authors contributed equally to this work.
*e-mail: liulisang@fjut.edu.cn

Both methods share common disadvantages, such as large dosing errors, increasing dosing costs, and inefficient, which can't meet the requirements of drinking WTP [3]. Therefore, it is necessary to model the dosage of coagulant for drinking WTP. Srdjan et al. [4] simulated the actual coagulation pool environment, fitted the experimental data, generated polynomial equations and predicted the coagulant dosage. Although the equation can predict the coagulant dosage of drinking WTP and reduce the dependence of jar test, the accuracy and efficiency of prediction still need to be improved. In order to solve the measurement error problem of water quality measurement sensor, Liu et al. [5] proposed a measurement error detection model of software sensor, and the experimental results showed that the proposed method has high measurement error detection efficiency. Kim et al. [6] put forward the k-means-ANFIS hybrid model. The results showed that the prediction accuracy of k-means-ANFIS was better than the ANFIS model when cyclic input is adopted. Kote et al. [7] applied the Cascade Feed Forward Neural Network (CFFNN) to the coagulant dosage prediction of WTP. Zhang Yanyang et al. [8] proposed a HANN mixed model to predict the drinking water output of drinking WTP by combining artificial neural network (ANN) with genetic algorithm (GA). The experimental results proved that this model is better than the single ANN model. Zhang Jun et al. [9] put forward an improved multi-modal variable structure stochastic vector neural network algorithm (MM-P-VSRVNN) by combining rule base and principal component analysis method, and applied it to coagulant dosage in key production process of water purification process. Wang Hui et al. [10] established principal component regression (PCR), support vector regression (SVR) and Long ShortTerm Memory models to predict the influent quality and chemical dosage of drinking WTP. The experimental results proved that SVR and LSTM had higher prediction accuracy than PCR. Liu Yiqun et al. [11] proposed an improved LSTM prediction model based on automatic adjustment and time consistency. The proposed model can accurately predict the coagulant dosage of drinking WTP by combining the operator's experience and seasonal effect, but the model is too complicated and the prediction efficiency is not high.

In addition, related algorithms based on data-driven gradually attract the interest of scholars. Wang Kungjeng et al. [12] used data-driven method to combine the optimization based on GA algorithm and particle swarm optimization (PSO) algorithm with regression model analysis to optimize coagulant dosage in industrial wastewater treatment field. The results show that the proposed data-driven method improves the sewage quality and sludge level of printed circuit board manufacturing, and reduces the cost by 10%. Shi et al. [13] put forward a method to determine the dosage of coagulant in drinking WTP based on ultraviolet-visible spectrum and chemometrics. The research shows that the combination of online UV-Vis spectroscopy and chemometrics can simulate the operator's decision-making in determining coagulant dosage. Pang et al. [14] established a correlation model based on the relationship between disinfection by-products (DBPs) and physical and chemical parameters to predict and analyze the toxicity of DBPs. Volf et al. [15] established four time models to predict the water quality index (WQI) of drinking WTP. By using WEKA software to set the number of rules and related linear equations, the model has a good prediction ability for the peak value of WQI in raw water. Bressane et al. [16] proposed a fuzzy inference system (D2FIS) based on unmixed data to predict the coagulant dosage of WTP in real time, which can effectively reduce the operation and maintenance cost of WTP. The experimental results show that the proposed model is better than ANFIS, cascade correlation network and support vector machine. Bertone et al. [17] developed a data-driven Bayesian network model (BN) to predict the water quality of raw water. The experimental results prove that the proposed algorithm can effectively predict the water quality changes of dam discharge in different scenarios. However, those data-driven algorithms are too complicated to be used for real-time prediction.

Because of its excellent clustering ability, FCM algorithm is often used in data clustering, image recognition and image segmentation. Saberi et al. [18] proposed a FCM clustering model based on weighted density peak (DP-WFCM) to realize the anomaly detection in production process. Dhruv et al. [19] improved FCM by using PSO algorithm, and used the improved algorithm to segment the chest CT image of COVID-19 infected people. Wang Cong et al. [20] proposed an improved FCM algorithm (RFCM) based on residual drive for noise estimation of images. Compared with the deviation sparse FCM algorithm (DSFCM), this algorithm has higher accuracy and wider application range. Halder et al. [21] used FCM algorithm to detect the quality of groundwater basin suitable for agricultural irrigation. The algorithm determines different hydrochemical regions by comprehensively considering the basic parameters of groundwater, and the clustering results are better than those of hierarchical clustering, k-means clustering and condensed clustering.

In line with the above research, this paper focuses on the innovative, improvement and application of ANFIS. When the literature is examined, the research on improving ANFIS by FCM algorithm and improved SSA algorithm has not been found. Aiming at the shortcomings of traditional ANFIS based on subtractive clustering, this paper improves it by using advanced multi-strategies such as FCM and SSA, and proposes a novel prediction model of coagulant dosage, which is called FCM-ISSA-ANFIS hybrid model. Firstly, FCM is used to determine the number of fuzzy rules of ANFIS. Then, the improved SSA is used to train the antecedent parameters and consequent parameters of ANFIS. Finally, the annual raw water data is input into the

FCM-ISSA-ANFIS hybrid model to realize the real-time prediction of coagulant dosage for drinking WTP.

The present research aims to propose a hybrid prediction model of optimal coagulant dosage based on improved ANFIS, by coupling advanced multi-strategies such as FCM and SSA. The main contributions of this paper are outlined as follows:

1. The research statistically characterizes the water quality data of drinking WTP and analyzes the correlation through Pearson correlation method.
2. FCM algorithm is utilized to determine the new fuzzy rule number of ANFIS.
3. The SSA is enhanced by incorporating the Sine chaotic mapping method and adaptive weight method to train the antecedent and consequent parameters of ANFIS model, which contributing to faster convergence and greater capacity of escaping from local optimum.
4. The research proposes the FCM-ISSA-ANFIS hybrid model as a more promising solution to predict the optimal coagulant dosage of drinking WTP when compared to other algorithms.

The other sections of this paper are organized as follows: Section 2 discusses the FCM algorithm. Section 3 comprehensively explains the implementation of improved SSA algorithm to optimize ANFIS. Section 4 introduces the multi-strategy improved FCM-ISSA-ANFIS hybrid model. Section 5 compares experimental results of the proposed algorithm with the classical algorithm. Section 6 summarizes the research and highlights the future work.

Fuzzy C-Means Clustering Algorithm and Its Evaluation Index

Fuzzy C-Means Clustering Algorithm

Fuzzy C-means clustering algorithm usually employ a membership function to determine how much each data point belongs to a certain cluster. The core of the algorithm is to iteratively calculate and correct the cluster centers and classification matrices belonging to the membership function, so as to complete the cluster classification [22]. FCM divides n data vectors $X_i (i = 1, 2, ..., n)$ into C fuzzy groups, and calculates the clustering center of each fuzzy group under the premise that the objective function satisfying the dissimilarity index reaches the minimum. When FCM determines the degree of membership of each data point, the elements of its membership matrix U can take values on [0,1]. The data can be initialized so that the membership of a dataset sums to 1, which displays as follows:

$$\sum_{i=1}^{C} u_{ij} = 1, j = 1, 2, ..., n \tag{1}$$

The cost function of FCM is usually expressed as:

$$J(U, H_1, ..., H_c) = \sum_{i=1}^{C} J_i = \sum_{i=1}^{C} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} \tag{2}$$

In the formula, H_i is the cluster center point, U is the membership matrix; $d_{ij} = \|H_i + -X_i\|$ is the Euclidean distance from the center of the i class to the j class sampled data point. m is the weighted number of $[1, +\infty)$. In order to obtain the necessary conditions for the cost function of J to reach the minimum value, the Lagrangian maximal method [29] is used to solve and derivation of u_{ij} and X_j, and through continuous iterative solution operations, the required cluster centers and membership matrices are finally obtained.

Clustering Evaluation Index

The coagulant dosage in the drinking WTP is affected by many uncertain factors, the main factors are raw water flow, raw water temperature, raw water PH value and raw water turbidity. Due to the strong uncertainty and nonlinearity of these uncertain factors, it is generally difficult to calculate accurately. Therefore, FCM algorithm is used to cluster and analysis on the coagulant dosage in the drinking WTP, which reduces the number of fuzzy rules of ANFIS and improves its prediction efficiency. To assess the clustering effect of the FCM, an internal clustering evaluation index is introduced, such as Bezdek partition coefficient [23], Xie-Beni coefficient [24], reconstruction error rate [25] and clustering effectiveness indicator [26].

V_{PC} represents the Bezdek partition coefficient. Its main purpose is to recharacterize the membership degree of the divided data so that the sum of the squares of the membership degrees of all elements belonging to each category. V_{PC} is defined as follows:

$$V_{PC} = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} u^{2}_{ij} \tag{3}$$

Where, u is the element of the data, and n is the number of elements. V_{PC} quantifies the compactness and separability of clustering results, and the smaller the value, the better the clustering effect.

V_{XB} represents the Xie-Beni coefficient, which is defined as:

$$V_{XB} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u^{2}_{ij} \| x_j - v_i \|^{2}}{n(\min_{j \neq k} \| v_i - v_k \|^{2})} \tag{4}$$

Where, x is the element before division, and v is the element after division. V_{XB} includes the compactness and separability of clustering results and the influence of clustering number. The smaller the value, the better the clustering effect.

V_{RE} represents the reconstruction error rate of the data, which is defined as follows:

$$V_{\mathrm{RE}} = \frac{1}{n}\sum_{i=1}^{n} \| I(i) - I'(i) \| \tag{5}$$

Where, *I(i)* and *I'(i)* are the grayscale values of the *i*-th data before reconstruction and after reconstruction respectively. V_{RE} measures the difference between the clustering results and the original data, and the smaller the value, the better the clustering effect.

V_{PBMF} represents a clustering validity metric, which is defined as follows:

$$V_{\mathrm{PBMF}} = \left(\frac{1}{K} \times \frac{E_1}{E_{\mathrm{K}}} \times D_{\mathrm{K}}\right)^2 \tag{6}$$

Where, K is the given number of division classes, and E_1 is the given data set, which is a constant value. E_k is the sum of the fuzzy distances between the data to be classified and the cluster centers in the individual, and the value of V_{PBMF} will increase as E_k decreases. D_k is the maximum distance between all pairs of cluster centers in an individual. V_{PBMF} takes the accuracy and stability of clustering into account, and the larger the value, the better the clustering effect.

Improved SSA for Optimizing ANFIS

Adaptive Fuzzy Neural Inference System

Adaptive Fuzzy Neural Inference System is a fuzzy neural reasoning system proposed by Jang et al. [27], which will not fall into the local optimal limit, and the training effect of ANFIS is better than ANN when compared with ANN. ANFIS is widely used in control system recognition, pattern recognition and some nonlinear complex systems due to its decision-making ability of fuzzy system and self-learning ability of neural network.

In the traditional ANFIS, the If-Then fuzzy rule is usually used to express its output as a linear combination of fuzzy subsets of the input:

$$\begin{cases} If\ x\ is\ A_1\ and\ y\ is\ B_1, Then\ f_1 = p_1 x + q_1 y + r_1 \\ If\ x\ is\ A_2\ and\ y\ is\ B_2, Then\ f_2 = p_2 x + q_2 y + r_2 \end{cases} \tag{7}$$

The adaptive fuzzy neural inference system is usually represented by a five-layer feed-forward neural network. The network contains two inputs for (*x*, *y*) and one output for *f*. Its architecture can automatically generate If-Then fuzzy rules, and can achieve arbitrary precision in the process of approximating nonlinear functions. Its structure is shown in Fig. 1.

The first layer is the fuzzification layer, the nodes of this layer fuzzify the exact input into several fuzzy subsets, and use the membership function to describe the degree of its membership to a subset. The formula is as follows:

$$O_i^{(1)} = \begin{cases} \mu_{Ai}(x_1) \\ \mu_{Bi}(x_2) \end{cases}, i = 1, 2 \tag{8}$$

Where, $x_j (j = 1, 2)$ is the node as the exact input for *j*. A_i or B_i as its corresponding fuzzy subset, μA_i or μB_i as its membership function which shape is determined by the antecedent parameters.

The second layer is the rule reasoning layer, which is responsible for calculating the incentive strength of fuzzy rules. The formula is as follows:

$$O_i^{(2)} = \omega_i = \prod_{i=2}^{2} O_i^{(1)} = \mu_{Ai}(x_1)\mu_{Bi}(x_2), i = 1, 2. \tag{9}$$

The third layer is the normalization layer, which is responsible for normalizing the excitation intensity. The formula is as follows:

$$O_i^{(3)} = \overline{\omega_i} = \omega_i / \sum_{i=1}^{2} \omega_i \tag{10}$$

The fourth layer is the output layer of fuzzy rules, which can adaptively generate *If-Then* fuzzy rules. The formula is as follows:

$$O_i^{(4)} = \overline{\omega_i} f_i = \overline{\omega_i}(p_i x_1 + q_i x_2 + r_i) \tag{11}$$

Where, $\{p_i, q_i, r_i\}$ is the consequent parameter.

The fifth layer is the output layer, which is responsible for converting the fuzzy output into an accurate output. The formula is as follows:

$$O_i^{(5)} = \sum_{i=1}^{2} \overline{\omega_i} f_i \tag{12}$$

The improved SSA algorithm can be used to train and learn for the ANFIS model. Firstly, the antecedent parameters are fixed. Then, the system output of the ANFIS model can be expressed as a linear combination of the consequent parameters, which can be expressed by the following formula:

$$\begin{aligned} f &= \sum_{i=1}^{2} \overline{\omega_i} f_i = \sum_{i=1}^{2} \overline{\omega_i}(p_i x_1 + q_i x_2 + r_i) \\ &= (\overline{\omega_1} x_1) p1 + (\overline{\omega_1} x_2) q_1 + (\overline{\omega_1}) r_1 + (\overline{\omega_2} x_1) p_2 + (\overline{\omega_2} x_2) q_2 + (\overline{\omega_2}) r_2 \\ &= \times\theta \end{aligned} \tag{13}$$

Where, the vector θ constitutes the consequent parameter set of {*p*1, *q*1, *r*1, *p*2, *q*2, *r*2,}, which can be estimated and adjusted by the improved SSA. Similarly,

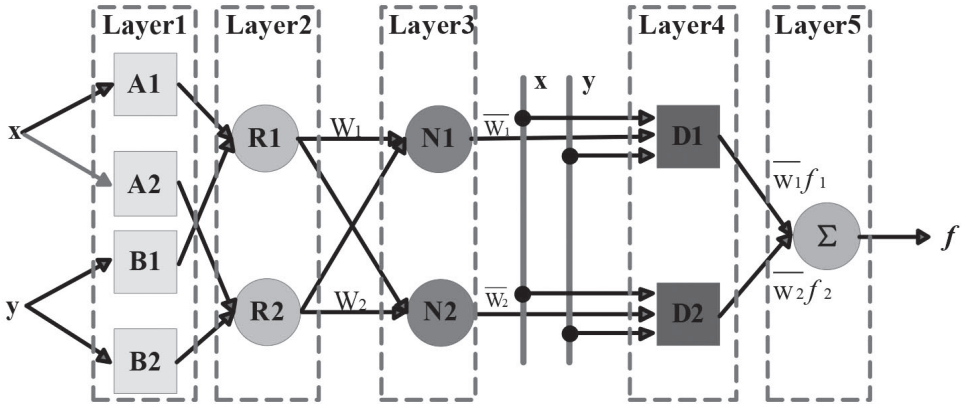

Fig. 1. Basic ANFIS structure.

the antecedent parameters in the fuzzy rules and the update of the connection weights can be completed by the improved SSA, and the training speed and parameter learning efficiency of ANFIS model are improved.

Sparrow Search Algorithm and Its Improvements

Inspired by the predatory and anti-predatory behaviors of sparrows, the sparrow search algorithm was put forward by Xue et al. [28] in 2020, which has the advantages such as simple structure, few adjustment parameters and easy implementation. However, it also has disadvantages of uneven population distribution, poor initial solution quality and easy to fall into local optimum [29]. Therefore, the standard SSA algorithm is improved by using Sine chaotic mapping and adaptive weight method.

Assume that the sparrow population is represented by a matrix as follows:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,j} \end{bmatrix} \tag{14}$$

Where, i = 1, 2, …, p. j = 1, 2, …, d. p represents the number of sparrow populations, and d represents the dimension of the problem. The fitness value of sparrow population can be expressed as follows:

$$F(X) = \begin{bmatrix} f(x_{1,1}) & f(x_{1,2}) & \cdots & f(x_{1,j}) \\ f(x_{2,1}) & f(x_{2,2}) & \cdots & f(x_{2,j}) \\ \vdots & \vdots & \vdots & \vdots \\ f(x_{i,1}) & f(x_{i,2}) & \cdots & f(x_{i,j}) \end{bmatrix} \tag{15}$$

The sparrow that finds the food source first will become the producers, and its location will be updated as follows:

$$X^{t+1}_{i,j} = \begin{cases} X^{t}_{i,j} \cdot \exp(\dfrac{-i}{\partial \cdot t_{\max}}), R_2 < ST \\ X^{t}_{i,j} + \Phi \cdot B, R_2 \geq ST \end{cases} \tag{16}$$

Where, t represents the iterations. $X^{t}_{i,j}$ represents the position of the i-th sparrow in the j-th dimension. $t_{\max}$ denotes the maximum iterations. $\partial \in (0,1)$ indicates random numbers. $ST \in [0.5, 1]$ indicates the safety value. $R_2 \in [0,1]$ indicates the warning value. $\Phi \in (0,1)$ indicates the normal distribution. B represents the $1 \times d$ matrix. If $R2 < ST$, it indicates that the producers are safe for foraging. Otherwise, if $R2 \geq ST$, it denotes that the natural enemies were discovered and the whole population should move to other regions.

In the standard SSA, the initial population of sparrows is generated by random function, which has some limitations, such as uneven population distribution and poor population diversity. Sine chaotic mapping, as a kind of chaotic mapping, has the characteristics of nonlinearity, randomness and ergodicity, which can be used to initialize the population, enrich the population size and improve the diversity of the population. Compared with Tent chaotic mapping and logistic chaotic mapping, Sine chaotic mapping is only generated by deformed *sin* function, and its structure is simpler and more operable. Therefore, this paper uses Sine chaotic mapping to initialize sparrow population and enrich the diversity of sparrow population. The Sine chaotic mapping formula [30] is as follows:

$$x_{i+1} = S(x_i) = \eta \cdot \sin(\pi \cdot x_i) \tag{17}$$

Where, $\eta \in (0,1)$ represents the chaotic factor.

Since the producers' fitness value will gradually decline in the later iterations, the standard SSA algorithm will have shortcomings such as insufficient search ability and easy to fall into local optimum, which breaks the balance between global exploration and local exploitation. Therefore, this paper uses adaptive weight method to update the location of the producers, which makes the algorithm have better global search ability

in the initial iteration and stronger local exploitation ability in the later iteration. The adaptive weight method is updated by the following formula:

$$\alpha = 0.4 + \frac{\sin(\pi \cdot t)}{0.4 + e^{t/t_{\max}}} \tag{18}$$

Then, the producers' position is updated by the adaptive weight method, which can be expresssed by the following formula:

$$X^{t+1}_{i,j} = \begin{cases} \alpha \cdot X^{t}_{i,j} \cdot \exp(\frac{-i}{\partial \cdot iter_{\max}}), R_2 < ST \\ (1-\alpha) \cdot X^{t}_{i,j} + \Phi \cdot B, R_2 \geq ST \end{cases} \tag{19}$$

The variables have been described in formula (16), and will not be repeated here again.

A FCM-ISSA-ANFIS Hybrid Model

FCM-ISSA-ANFIS Modeling

Firstly, the raw water quality data of drinking WTP are preprocessed, and the missing values in the raw water data are replaced by K-Nearest Neighbors (KNN). Then, the preprocessed data set is divided into training set and test set, and the divided data set is clustered by FCM algorithm to determine the number of ANFIS's new fuzzy rules. What's more, the improved SSA algorithm is used to train the ANFIS model to be an optimal coagulant dosage prediction model, which is called the FCM-ISSA-ANFIS hybrid model. Finally, the improved ANFIS model is used to predict the annual coagulant dosage of drinking WTP. The proposed FCM-ISSA-ANFIS hybrid model mainly includes the following structures: FCM clustering, data division, fuzzy system generator, fuzzy reasoning system and adaptive fuzzy neural network. The flow chart of FCM-ISSA-ANFIS hybrid model is shown in Fig. 2.

The modeling steps of the proposed FCM-ISSA-ANFIS hybrid model are as follows:

Step 1: Pearson correlation method is used to analyze the correlation of raw water quality data of drinking WTP, and the input variables of the ANFIS are determined to be raw water turbidity, raw water temperature, PH value and influent flow, and the output variable of the ANFIS is coagulant dosage.

Step 2: The raw water quality data of drinking WTP is divided into training set and test set according to the ratio of 7:3. The training set is used to train the ANFIS prediction model, and the test set is used to verify the performance of ANFIS.

Step 3: FCM is used to cluster the divided data set and get the cluster center. By setting the initial position of fuzzy sets, the membership matrix of FCM is used to calculate the similarity of each fuzzy set. According to the similarity, the fuzzy rules are merged to be a new fuzzy rule of ANFIS. The initial cluster center number of FCM is set to 28. The maximum iterations of FCM is set to 200. The partition matrix index is set to 3, and the target error is set to 10E-6.

Step 4: Initialize the parameters of SSA, including the number of sparrow populations and the maximum iterations. The ANFIS is set to the objective function. The training data sets are input of ANFIS, and the parameters of ANFIS are trained and optimized by improved SSA algorithm, so that the appropriate membership function, rule weight and consequent parameters are gradually adjusted and found. If the training error meets the requirements or reaches the maximum iterations, the training is terminated.

Step 5: After training and optimizing the ANFIS model, the optimal prediction model is named as the FCM-ISSA-ANFIS hybrid model. Then, the test set is used to verify the performance of the proposed FCM-ISSA-ANFIS hybrid model, and its structural parameters are restricted to develop its performance.

Step 6: The FCM-ISSA-ANFIS hybrid model is used to predict the optimal coagulant dosage of drinking WTP in the next year, and RMSE, MAE and SD are calculated to further evaluate its performance.

Performance Metrics

To further evaluate the prediction accuracy of the proposed FCM-ISSA-ANFIS hybrid model, RMSE, MAE and SD are introduced as its performance metrics. RMSE describes the degree of deviation between the predicted value and the real measured data. The calculation formula of RMSE is as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{n} (y_k - \hat{y}_k)^2}{N}} \tag{20}$$

MAE describes the absolute average value of error between the predicted value and the real measured data, which is expressed by the following formula:

$$\text{MAE} = \frac{\sum_{k=1}^{n} abs(y_k - \hat{y}_k)^2}{N} \tag{21}$$

SD describes the dispersion degree of the prediction error, which is expressed as follows:

$$\text{SD} = \sqrt{\frac{\sum_{k=1}^{n} (y_k - \mu)^2}{N}} \tag{22}$$

Where, N is the number of samples. μ is the arithmetic average of prediction value. $\hat{y}_k$ represents the average of prediction value. y_k represents the prediction value.

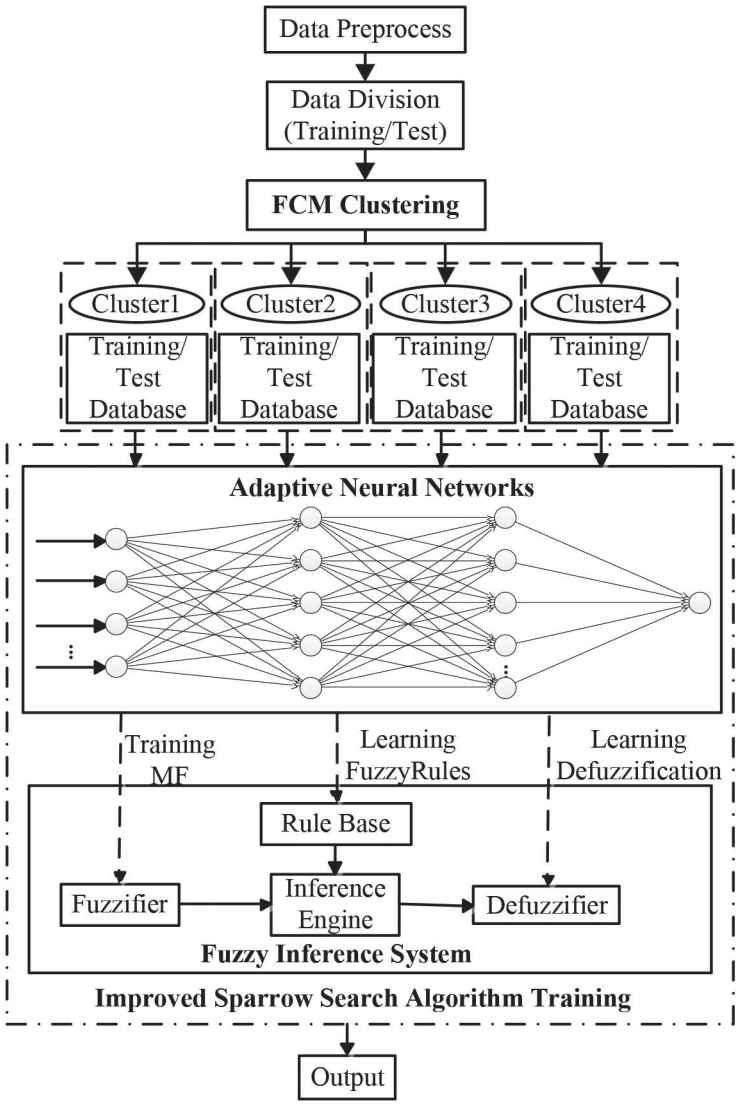

Fig. 2. Flow chart of FCM-ISSA-ANFIS hybrid model.

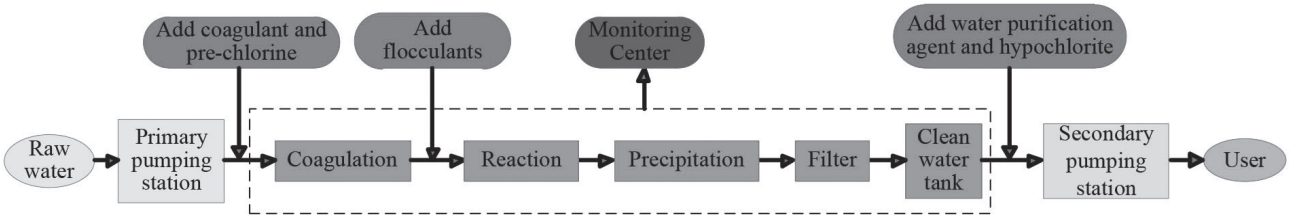

Fig. 3. Flowchart of water treatment process in drinking WTP.

Results Analysis and Discussion

Data Description

The drinking WTP treats natural water such as reservoirs by a series of physical and chemical methods to remove impurities and some harmful substances in the water, so that the water quality meets the requirements of domestic water. The flow chart of water treatment process in drinking WTP is shown in Fig. 3.

In the flow chart of water treatment, the reaction tank is an indispensable part of the water treatment process, especially for the treatment of high turbidity water sources. It uses specific chemical reagents to remove impurities and pollutants in raw water, which makes the chemical reaction more complete and improves the efficiency of sewage purification.

In addition, as can be seen from the Fig. 3, in order to achieve more thorough purification of raw water, the water treatment system in this study used

polyacrylamide (PAM) as the flocculants to flocculate raw water, which removed suspended matter and increased the cleanliness of the raw water.

The dataset utilized in this study was obtained from the online operation data of coagulation dosing system of Fuzhou drinking WTP in China in 2021. Water samples were collected from a large reservoir in Fuzhou, Fujian, China. Water samples were collected by pumping method at different stages of water treatment process, including the quality of raw water, after adding coagulant, before and after filtration. The sampling frequency was once every 30 minutes, and the average concentration of water samples collected within 24 hours was determined for further analysis of raw water quality. In order to ensure the consistency and integrity of samples, all sampling equipment is thoroughly cleaned before sampling to avoid any cross-contamination, thus ensuring the reliability and accuracy of analysis results. The statistical characteristics of the data set are shown in Table 1.

As can be inferred from the Table 1, there is low turbidity in winter and high turbidity in summer. The influent flow is greatly influenced by the seasons, with a smaller inflow in winter and a larger inflow in summer. The PH value of raw water does not change greatly due to the influence of seasons.

To further analyze the linear correlation between raw water turbidity, raw water temperature, PH value, influent flow, coagulant dosage and effluent settled water turbidity, Pearson correlation method was used to analysis the annual data set of drinking WTP, which improved the prediction accuracy and efficiency of the FCM-ISSA-ANFIS hybrid model.The results of Pearson correlation analysis for water quality data is shown in Table 2.

From the analysis of the results in the Table 2, it can be known that the coagulant dosage has a strong linear correlation with raw water turbidity, raw water temperature, PH value and influent flow. There is a strong positive correlation between raw water turbidity and raw water temperature, but a negative correlation between raw water turbidity and effluent settled water turbidity.

Cluster Results Analysis

Pearson correlation analysis results show that the coagulant dosage in drinking WTP is influenced by raw water flow, raw water temperature, PH value and raw water turbidity. Since the factors fluctuate widely and have strong uncertainty and nonlinearity, it is generally difficult to calculate them accurately. Therefore, the FCM is used to cluster the divided data sets in drinking WTP, which reduces the number of fuzzy rules n FCM-ISSA-ANFIS hybrid model and improves its prediction efficiency. The cluster analysis results of FCM are shown in Table 3.

Table 1. Statistical characteristics of the annual operation data of the drinking WTP in 2021.

Classification	Variable	Mean	Standard deviation	Coefficient of variation	Min	Max
Raw water	Turbidity(NTU)	14.9406	0.8943	0.0599	10.73	16.57
	Temperature (ºC)	23.2006	3.9378	0.1697	11	35
	PH	7.8328	0.2545	0.0325	6.9	8.7
	Influent flow(m^3/d)	87584.454	1460.5527	0.0167	85004	89992
Effluent settled water	Turbidity(NTU)	0.4247	0.3622	0.8528	0.1	2.50
Operational parameters	Dosage(mg/L)	35.8569	2.1465	0.0599	25.76	39.76

Table 2. Pearson Correlation analysis.

Pearson Correlation	Raw Water Turbidity	Raw Water PH	Raw Water Temperature	Influent Flow	Coagulant Dosage	Effluent Settled Water Turbidity
Raw Water Turbidity	1	.017	.686**	.035	.026	-.032
Raw Water PH	.017	1	.040	-.014	-.035	-.004
Raw Water Temperature	.686**	.040	1	.044	.027	-.017
Influent Flow	.035	-.014	.044	1	-.006	-.013
Coagulant Dosage	.026	-.035	.027	-.006	1	-.039
Effluent Settled Water Turbidity	-.032	-.004	-.017	-.013	-.039	1

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3. Results of FCM for four clustering evaluation indexes.

Index \ Clusters	2	3	4	5	6
V_{PC}	0.8560	0.7737	0.7504	0.7505	0.7532
V_{XB}	0.0821	0.1486	0.1455	0.2083	0.2667
V_{RE}	1.4518E-29	1.6139E-29	2.3795E-29	2.3174E-29	2.1105E-29
V_{PBMF}	0.1649	0.2254	1.5444	0.0521	0.0379

As can be seen from the Table 3, when the number of designated clusters is gradually increased from 2 to 4, the V_{PC}, V_{XB} and V_{RE} are gradually increased, while the V_{PBMF} is gradually decreased, which indicating that the clustering effect of FCM becomes better. However, the value of 4 is a threshold of clustering. When the number of clusters continues to increase from 4 to 6, the values of V_{PC}, V_{XB} and V_{RE} begin to decrease, and the V_{PBMF} begins to increase, which indicates that the clustering effect of FCM is getting worse. Therefore, it can be inferred that the clustering effect of FCM is the best when the value of clustering number is 4.

Prediction Results Analysis

It is well-known that choosing the appropriate research parameters is critical to solve the problem of predicting the coagulant dosage. In this study, the raw water turbidity was choosed as the main research parameter and other parameters listed in Table 1 as secondary parameters. There are several reasons for this. First of all, the turbidity of raw water has been widely accepted by multiple regulatory agencies as a crucial indicator affecting the quality of raw water in the drinking WTP. Secondly, turbidity is a vital metric for assessing the number of suspended particles in water. Finally, continuous monitoring of turbidity in treated water is relatively simple and inexpensive, making it a convenient parameter for routine monitoring and operation control.

The national water effluent standards mandates that total coliforms should be non-detectable, suspended matter should not exceed 0.5 NTU, and pH values should be between 6.5 and 8.5. Though additional parameters shown in Table 1, like pH values, flow rate, and temperature, impact water quality, their influence primarily remains indirect due to the following reasons. Firstly, influent flow rate has a minimal impact on coagulant dosage, which is mainly determined on the basis of water turbidity levels. Secondly, changes in temperature and pH may influence the solubility or reaction rate of some chemicals in water, which affects

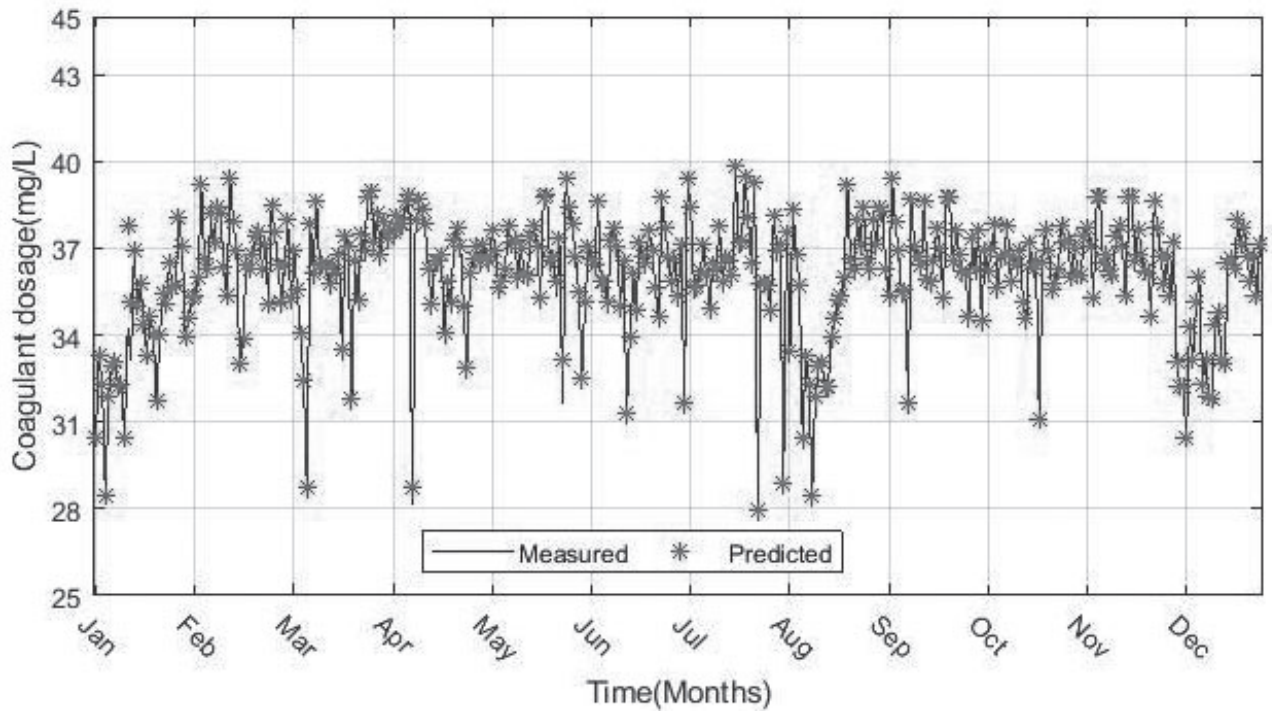

Fig. 4. Comparison results of FCM-ISSA-ANFIS predicted value and actual value.

coagulant dosage. However, the influence is generally complex and challenging to quantify, requiring special experiments for an accurate assessment.

In summary, although influent flow rate, temperature, and pH of raw water impact coagulant dosage, they cannot serve as the primary research parameters. Turbidity, in contrast, accurately and directly reflects the water quality and is thus commonly utilized as main operational parameter for predicting coagulant dosage.

Additionally, the type of coagulant to be employed in water treatment design is also a crucial parameter to consider. The coagulant used in this study is Poly Aluminum Chloride (PAC), which has the following advantages, such as good coagulation effect on fine particles, high purity, easy dissolution, stable quality and no negative impact on downstream treatment facilities.

To intuitively observe the prediction effect of FCM-ISSA-ANFIS hybrid model, the predicted value and real measured data of coagulant dosage are shown in Fig. 4. The blue curve represents the real measured data of coagulant dosage in drinking WTP, while the red point represents the predicted value of coagulant dosage in FCM-ISSA-ANFIS hybrid model.

As can be seen from the scatter chart, the red dot can basically cover most of the blue curve, which shows that the prediction error between the FCM-ISSA-ANFIS hybrid model and the real coagulant dosage is very small, and the proposed FCM-ISSA-ANFIS hybrid model has a high prediction accuracy for the coagulant dosage of drinking WTP. In addition, the predicted value of the proposed FCM-ISSA-ANFIS hybrid model and the actual measured data are analyzed by R^2 regression, and the regression result is shown in Fig. 5.

As can be seen from the Fig. 5, the predicted value of the proposed FCM-ISSA-ANFIS hybrid model has a linear relationship with the actual measured value,

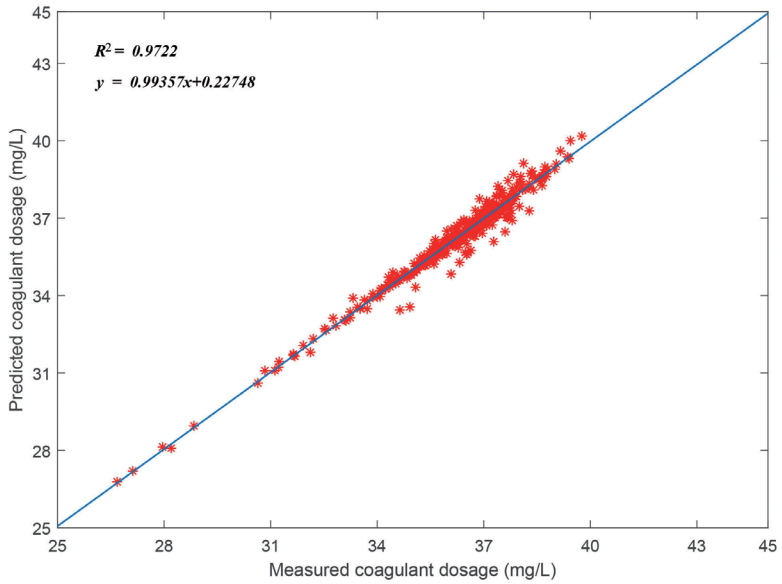

Fig. 5. Scatter plot of prediction error of FCM-ISSA-ANFIS.

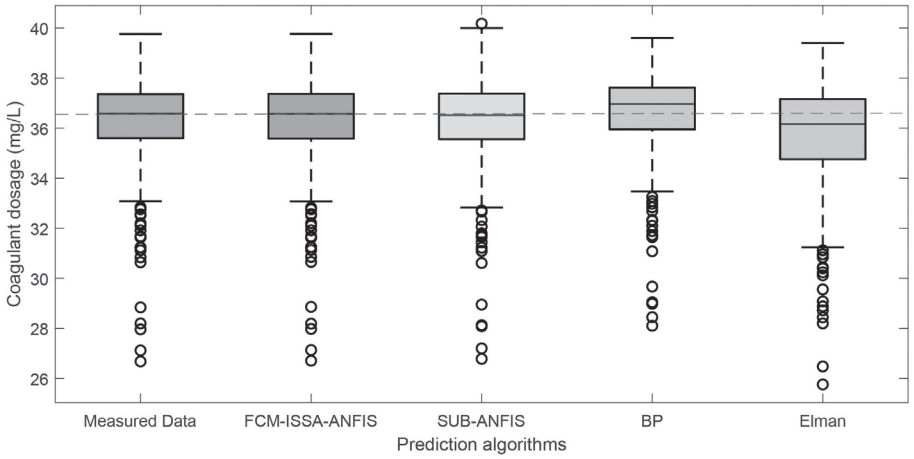

Fig. 6. Box chart of coagulant dosage prediction under different algorithms.

Table 4. Comparison performance results of different models.

Performance / Algorithm	RMSE	MAE	SD
BP	2.2949	1.5223	1.9236
Elman	2.9932	2.2372	2.2144
SUB-ANFIS	2.2232	1.5406	1.8781
FCM-ISSA-ANFIS	1.2174	0.5185	0.8599

and the equation $y = 0.99357x + 0.22748$ fitted by linear regression is very close to the linear equation $y = x$, which proves that the predicted value of the proposed FCM-ISSA-ANFIS hybrid model is very close to the actual measured data of drinking WTP. And $R^2 = 0.9722$ further proves that the proposed FCM-ISSA-ANFIS hybrid model has high prediction accuracy.

To further verify the outstanding performance of the proposed FCM-ISSA-ANFIS hybrid model, BP neural network algorithm, Elman neural network algorithm and ANFIS model based on subtractive clustering in Ref. [31] (SUB-ANFIS) are selected to compare with it. The above-mentioned model is used to predict the annual coagulant dosage for drinking WTP, and compared with the actual coagulant dosage for drinking WTP, and a box chart is drawn as shown in Fig. 6.

As can be seen from the Fig. 6, the first box chart is the actual coagulant dosage for drinking WTP, the second box chart is the predicted value of the proposed FCM-ISSA-ANFIS hybrid model. The third, fourth and fifth box chart are SUB-ANFIS model, BP algorithm and Elman algorithm, respectively. As for the proposed FCM-ISSA-ANFIS hybrid model, its predicted value is the closest to the actual coagulant dosage of drinking WTP, and the median, quartile and abnormal value of the box chart basically coincide with the first box chart. As for the SUB-ANFIS model, there is an abnormal value in its upper limit, which shows that the prediction ability of the SUB-ANFIS model is limited and it is impossible to accurately predict the actual coagulant dosage of drinking WTP. In addition, the median and quartile of the box chart of BP algorithm and Elman algorithm are quite different from the first box chart, and the number of outliers is also large. It is not difficult to infer from the comparison results of box charts that the prediction accuracy of the proposed FCM-ISSA-ANFIS hybrid model is outperformed the BP algorithm, Elman algorithm and SUB-ANFIS model.

To observe the advantages of the proposed FCM-ISSA-ANFIS hybrid model obviously, the RMSE, MAE and SD of the prediction errors of the proposed FCM-ISSA-ANFIS hybrid model, SUB-ANFIS model, BP algorithm and Elman algorithm are calculated, respectively, and the results are shown in Table 4.

As can be seen from the Table 4, the order of RMSE from small to large is FCM-ISSA-ANFIS, SUB-ANFIS, BP and Elman. The order of MAE from small to large is FCM-ISSA-ANFIS, BP, SUB-ANFIS and Elman. The order of SD from small to large is FCM-ISSA-ANFIS, SUB-ANFIS, BP and Elman. Therefore, when RMSE, MAE and SD are considered as the measurement standards of the algorithm, the prediction performance of the proposed FCM-ISSA-ANFIS hybrid model is outperformed the BP algorithm, Elman algorithm and the SUB-ANFIS model in Ref. [31]. Specifically, compared with the SUB-ANFIS model, the RMSE, MAE and SD of the proposed FCM-ISSA-ANFIS hybrid model in predicting coagulant dosage decreased by 45.24%, 66.34% and 54.21%, respectively. Therefore, the proposed FCM-ISSA-ANFIS hybrid model has better prediction accuracy and reliability than the SUB-ANFIS model in Ref. [31]. It is suggested that the proposed FCM-ISSA-ANFIS hybrid model be used to predict the coagulant dosage of WTP drinking water in real time, so as to reduce the dosage cost and labor cost and improve the efficiency of drinking WTP.

Conclusions

In this paper, a hybrid prediction model based on FCM-ISSA-ANFIS is proposed to predict the optimal coagulant dosage for drinking WTP. First of all, the water quality data of drinking WTP is statistically characterized and analyzed. Then, the Pearson correlation method is used to calculate and analysis the variables correlation of water quality data. What's more, the raw water quality data set of drinking WTP is divided into training set and test set, and the FCM is used to cluster the divided raw water quality data set to determine the new ANFIS fuzzy rule number. Finally, the improved SSA is used to train the antecedent parameters and consequent parameters of ANFIS, and its performance is verified by test set. When RMSE, MAE and SD are considered as the measurement criteria of the algorithm, the proposed FCM-ISSA-ANFIS hybrid model outperforms BP, Elman and SUB-ANFIS in predicting the annual coagulant dosage for drinking WTP. The results show that the proposed FCM-ISSA-ANFIS hybrid model not only solves the shortcomings of SUB-ANFIS in Ref. [31], but also accelerates the convergence and increases the ability of jumping out the local optimum. The proposed FCM-ISSA-ANFIS hybrid model has only been verified in the simulation experiment of drinking WTP in local area. To apply it to practical engineering problems and verify its effectiveness, future research should focus on extending it to the real coagulation dosage prediction of drinking WTP in other areas.

Acknowledgments

This work was financially supported by Natural Science Foundation of Fujian Province on the project (Grant no. 2022H6005).

Conflict of Interest

The authors declare no conflict of interest.

References

1. WEI, LI, JUNQI, WANG, HUIYU, XIN, TING, LI, JINMING, DUAN, DENNIS, MULCAHY. Determination of cost-effective optimum coagulant dosage for removal of disinfection by-product precursors in water treatment based on the theory of elasticity. Journal of Water Process Engineering. **47**, 102782, **2022**.
2. SUBIN, LIN, JIWOONG, KIM, CHUANBO, HUA, MI-HYUN, PARK, SEOKTAE, KANG. Coagulant dosage determination using deep learning-based graph attention multivariate time series forecasting model. Water Research. **232**, 119665, **2023**.
3. LI, LEI, RONG, SHUMING, WANG, RUI, YU, SHUILI. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. Chemical Engineering Journal. **405**, **2021**.
4. M., SRDJAN, B., ZIJAH, T., LJILJANA, V., DEJAN. The Equation for the Optimum Dosage of Coagulant for Water Treatment Plant. Tehnicki vjesnik - Technical Gazette. **26** (2), **2019**.
5. LIU W., RATNAWEERA H., KVAAL K. Model-based measurement error detection of a coagulant dosage control system. International Journal of Environmental Science and Technology. **16** (7), 3135, **2018**.
6. KIM, CHAN MOON, PARNICHKUN, MANUKID. Prediction of settled water turbidity and optimal coagulant dosage in drinking water treatment plant using a hybrid model of k-means clustering and adaptive neuro-fuzzy inference system. Applied Water Science. **7** (7), 3885, **2017**.
7. KOTE*, DR A.S., WADKAR D.V. Application of Feed Forward Neural Network for Prediction of Optimum Coagulant Dose in Water Treatment Plant. International Journal of Innovative Technology and Exploring Engineering. **8** (12), 1853, **2019**.
8. ZHANG, YANYANG, GAO, XIANG, SMITH, KATE, INIAL, GOULVEN, LIU, SHUMING, CONIL, LENNY B., PAN, BINGCAI. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. Water Research. **164**, **2019**.
9. ZHANG, JUN, LUO, DA-YONG. Multimodal Control by Variable-Structure Neural Network Modeling for Coagulant Dosing in Water Purification Process. Complexity. **2020** 1, **2020**.
10. WANG, HUI, ASEFA, TIRUSEW, THORNBURGH, JACK. Integrating water quality and streamflow into prediction of chemical dosage in a drinking water treatment plant using machine learning algorithms. Water Supply. **22** (3), 2803, **2022**.
11. LIU, YIQUN, HE, YIWEI, LI, SHUMAO, DONG, ZHENGHUI, ZHANG, JUNPING, KRUGER, UWE. An Auto-Adjustable and Time-Consistent Model for Determining Coagulant Dosage Based on Operators' Experience. IEEE Transactions on Systems, Man, and Cybernetics: Systems. **51** (9), 5614, **2021**.
12. WANG, KUNG-JENG, WANG, PEI-SHAN, NGUYEN, HONG-PHUC. A data-driven optimization model for coagulant dosage decision in industrial wastewater treatment. Computers & Chemical Engineering. **152**, **2021**.
13. SHI, ZHINING, CHOW, CHRISTOPHER W. K., FABRIS, ROLANDO, LIU, JIXUE, SAWADE, EMMA, JIN, BO. Determination of coagulant dosages for process control using online UV-vis spectra of raw water. Journal of Water Process Engineering. **45**, **2022**.
14. PANG, ZHEN, ZHANG, PEIFENG, CHEN, XINYI, DONG, FEILONG, DENG, JING, LI, CONG, LIU, JUNPING, MA, XIAOYAN, DIETRICH, ANDREA M. Occurrence and modeling of disinfection byproducts in distributed water of a megacity in China: Implications for human health. Science of The Total Environment. **848**, **2022**.
15. VOLF, GORAN, SUŠANJ ČULE, IVANA, ŽIC, ELVIS, ZORKO, SONJA. Water Quality Index Prediction for Improvement of Treatment Processes on Drinking Water Treatment Plant. Sustainability. **14** (18), **2022**.
16. BRESSANE, ADRIANO, GOULART, ANA PAULA GARCIA, MELO, CARRIE PERES, GOMES, ISADORA GURJON, LOUREIRO, ANNA ISABEL SILVA, NEGRI, ROGéRIO GALANTE, MORUZZI, RODRIGO, REIS, ADRIANO GONçALVES DOS, FORMIGA, JORGE KENNETY SILVA, DA SILVA, GUSTAVO HENRIQUE RIBEIRO, THOMé, RICARDO FERNANDES. A Non-Hybrid Data-Driven Fuzzy Inference System for Coagulant Dosage in Drinking Water Treatment Plant: Machine-Learning for Accurate Real-Time Prediction. Water. **15** (6), **2023**.
17. BERTONE, EDOARDO, ROUSSO, BENNY ZUSE, KUFEJI, DAPO. A probabilistic decision support tool for prediction and management of rainfall-related poor water quality events for a drinking water treatment plant. Journal of Environmental Management. **332**, **2023**.
18. SABERI, HOSSEIN, SHARBATI, REZA, FARZANEGAN, BEHZAD. A gradient ascent algorithm based on possibilistic fuzzy C-Means for clustering noisy data. Expert Systems with Applications. **191**, **2022**.
19. DHRUV, BHAWNA, MITTAL, NEETU, MODI, MEGHA. Hybrid Particle Swarm Optimized and Fuzzy C Means Clustering based segmentation technique for investigation of COVID-19 infected chest CT. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. **11** (2), 197, **2022**.
20. WANG, CONG, ZHOU, MENGCHU, PEDRYCZ, WITOLD, LI, ZHIWU. Comparative Study on Noise-Estimation-Based Fuzzy C-Means Clustering for Image Segmentation. IEEE Transactions on Cybernetics. 1, **2022**.
21. HALDER, SUDIPA, BHATTACHARYA, SHUVOSHRI, ROY, MALABIKA BISWAS, ROY, PANKAJ KUMAR. Application of fuzzy C-means clustering and fuzzy EDAS to assess groundwater irrigation suitability and prioritization for agricultural development in a complex hydrogeological basin. Environmental Science and Pollution Research. **2023**.
22. SUJIL, A., KUMAR, RAJESH, BANSAL, RAMESH C. FCM Clustering-ANFIS-based PV and wind generation forecasting agent for energy management in a smart microgrid. The Journal of Engineering. **2019** (18), 4852, **2019**.
23. BEZDEK J.C. Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology. **1** (1), 57, **1974**.
24. XIE X., L., BENI G. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. **13** (8), 841, **1991**.

25. GRAVES, DANIEL, PEDRYCZ, WITOLD. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. Fuzzy Sets and Systems. **161** (4), 522, **2010**.
26. PAKHIRA, MALAY K., BANDYOPADHYAY, SANGHAMITRA, MAULIK, UJJWAL. A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. Fuzzy Sets & Systems. **155** (2), 191, **2005**.
27. JANG J.S.R. ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics. **23** (3), 665, **1993**.
28. XUE, JIANKAI, SHEN, BO. A novel swarm intelligence optimization approach: sparrow search algorithm. Systems Science & Control Engineering. **8** (1), 22, **2020**.
29. LIU, LISANG, LIANG, JINGRUN, GUO, KAIQI, KE, CHENGYANG, HE, DONGWEI, CHEN, JIAN. Dynamic Path Planning of Mobile Robot Based on Improved Sparrow Search Algorithm. Biomimetics. **8** (2), **2023**.
30. ZHENG, YUANZHOU, LI, LEI, QIAN, LONG, CHENG, BOSHENG, HOU, WENBO, ZHUANG, YUAN. Sine-SSA-BP Ship Trajectory Prediction Based on Chaotic Mapping Improved Sparrow Search Algorithm. Sensors. **23** (2), **2023**.
31. NARGES, SHAKERI, GHORBAN, ASGARI, HASSAN, KHOTANLOU, MOHAMMAD, KHAZAEI. Prediction of the optimal dosage of coagulants in water treatment plants through developing models based on artificial neural network fuzzy inference system (ANFIS). Journal of Environmental Health Science and Engineering. **19** (2), 1543, **2021**.