Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2021.27056

222 (2021) 168–181
May

Water plant optimization control system based on machine learning

Dongsheng Wang[a,*], Yan Wang[a], Rui Zhou[a], Yong Cao[b], Fuchun Jiang[b], Xue Zhang[b], Jinghua Li[a]

[a]*College of Automation and College of Artificial Intelligence, Jiangsu Engineering Lab for IOT Intelligent Robots, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, Tel. +8618913939776; email: wdsnjupt@163.com (D. Wang), Tel. +8618851172465; email: wangyan950906@163.com (Y. Wang), Tel. +8615722922821; email: zhourui1347@gmail.com (R. Zhou), Tel. +8602585866512; email: 3554813993@qq.com (J. Li)*
[b]*Suzhou Tap Co., Suzhou 215002, China, Tel. +86 13915570634; email: cyong.mail@126.com (Y. Cao), Tel. +8618115607225; email: chinajfc@163.com (F. Jiang), Tel. +8617761868288; email: shirden@eyou.com (X. Zhang)*

Received 26 August 2020; Accepted 18 January 2021

ABSTRACT

There are always some operational management problems in water plants using conventional manual control approaches, such as the stability of water treatment operation and the safety of treated water quality. With the increasing shortage of water resources and serious water pollution, how to ensure safe and energy-saving water supply is particularly concerned. This paper proposes a water plant optimization control system based on machine learning, which is able to design optimized schemes for pump group scheduling, backwashing and reagent dosing, etc. Based on machine learning algorithms, the system has the ability to model and configure parameters for water treatment operations, which makes the water treatment process work in an optimal mode. Our system not only ensures the clean and safe treated water quality but also benefits the rational use of water plant equipment. Through learning the historical operation big-data, the empirical models are established by the approaches of random forest and support vector machine. Thus, it is proposed of optimized configuration schemes with high stability, good water purification effect and low energy consumption. By the prediction results, the effectiveness of our system for water treatment optimization has been verified.

Keywords: Safe and energy-saving water supply; Optimization control system; Machine learning; Random forest; Support vector machine

1. Introduction

In traditional water plant water treatment, the operation of water plant is done through the experience of the workers. The backward water treatment technology always leads to unsafe and unstable operation of water production process, high water production cost and poor effluent quality. Therefore, it is necessary to design a system based on advanced technology to optimize the operation of water plant. The system must meet the following goals. First, it must keep water plant running. Second, satisfying the effluent requirements. Third, minimizing operation costs [1,2].

Due to the nonlinear and complex characteristics of water treatment process, it is difficult to establish an effective model for water treatment process with conventional mathematical model [3]. With the rapid development of machine learning algorithm, various algorithm models have been proposed and widely used in the research field of water treatment. For example, Szelag et al. [4] proposed to use the model established by data mining method to predict sewage water quality indicators: biochemical oxygen demand, chemical oxygen demand, total suspended solids. Yongeun et al. [5] presented a model of groundwater arsenic pollution prediction based on artificial neural network and

* Corresponding author.

support vector machine algorithm. The model is combined with water quality reference to predict groundwater arsenic pollution. Ding et al. [6] proposed to mine water pump data based on support vector machine in water plant. Sharafati et al. [7] used random forest to predict water quality parameters of water plant. But few people have proposed to use support vector machine and random forest to optimize the water production process of water plant, which is also the significance of this paper. By using support vector machines and random forests, the amount of data generated along drinking water treatment plants allows developing data-based models [8], which helps to predict operational parameters [9] and be incorporated into decision support system. Through learning the historical operation big-data, the system models and configures the parameters of water treatment operation. It achieves the optimization of the whole water supply system, including the pump unit scheduling system, the water production process and the secondary pump station. After optimization, the system sends instructions to complete the intelligent process from water intake, water purification to water delivery. It will avoid the problem of equipment aging [10] and inefficient operation caused by long-term unreasonable use of human factors, which ensures the safe and stable operation of water production process, improve work efficiency, reduce energy consumption and save cost.

At the same time, the system has greatly improved the water quality. Water plant purification methods include flocculation, sedimentation, sand filtration and chlorination [11]. Even though these purification methods may be effective, deterioration of source water quality may require modern optimization methods to ensure the effective purity of the water [12–14]. Coagulation, sedimentation, disinfection, etc. need to be completed by dosing. The dosage should be based on water quantity, water temperature, pH value, raw water turbidity, dissolved oxygen, oxygen consumption, and so on. But the general water plant will still be controlled by human observation and experience of the drug delivery, as the result of the dosage may not be reasonable enough. The system can detect the water condition in real time and predict the water quality more sensitively combined with the empirical data model, and predict the optimal dosage of chemicals quickly and accurately, so as to achieve the optimal water purification effect.

The major contribution of this work is the development of water plant optimization control system, which optimizes the water production process, water purification effect and save water production cost. The water treatment process, machine learning, random forest algorithm, and support vector machine algorithm are introduced in Section 2. The optimization results of each subsystem are given in Section 3. So far, the system has been successfully applied in Bai yang wan water plant.

2. Water treatment process and methods

2.1. Water treatment process

The production of tap water is inseparable from many water treatment processes, which make the raw water from turbidity, black, carrying a large number of bacteria and microorganisms to clear, safe tap water. There is also energy consumption in water treatment. The following will first describe the water plant water treatment process. The water treatment process includes water intake – coagulation – precipitation – filtration – disinfection – water supply. When raw water is pumped into the storage tank of water plant by water intake pumps, the next coagulation step will become the first step of water treatment. Coagulation process could make water treatment agent fully mixed with raw water, from which the water will be difficult to precipitate colloidal particles and micro suspended substances combined to form easy to precipitate flocs. Floc particles are easier to separate from water and precipitate, among which the mixing operation to generate floc particles needs to be carried out quickly after the drug is put into the water. By means of vigorous stirring by machinery, the water treatment agent can be evenly distributed into the raw water to make it fully reacted, leading to preliminary preparations for the next step of precipitation and filtration. Water treatment agent is generally composed of flocculant and disinfection drugs, of which the latter plays the role of algicides. Flocculants commonly used include alum, polyaluminum chloride, basic polyaluminum chloride, polyaluminum chloride, etc. Disinfection drugs commonly used include ozone, chlorine and chlorine dioxide. The operation of precipitation is carried out in the second step of water treatment: sedimentation tank, in which the flocs in the water will be separated from the water under the action of gravity, and the sludge will be formed at the bottom of the sedimentation tank. The sedimentation tank of water plant is shown in Fig. 1.

The settled water in the sedimentation tank is collected from the collecting tank to the filter tank (shown in Fig. 2), and the water flow will pass through the granular filter material layer with gaps, which is the third step of water treatment: filtration. The filter layer removes fine suspended matter, bacteria and viruses due to its adhesion, at which point the water becomes clear. In the filtration process, excessive use of the filter material layer will cause blockage. The water plant always adopts the process of backwashing to make sure the flow of the filter material layer. Backwash can make the plug in the gap of the filter material remove reversely to clean it, and finally discharge the water. Backwash can use air and water, which can be used simultaneously or independently. So it is difficult to avoid the use of high power consumption of machine, such as fan, water pump. As the result, if the collocation of water pump and air blower frequency can be predicted in the current water quality situation back flush, energy consumption can be convenient to provide better reference configuration collocation to water management.

After the former several processes, the residual bacteria, viruses and microorganisms in the water will lose the turbidity that they can attach to, which is benefit to the final disinfection and sterilization process. Disinfection process can remove pathogenic pathogens to ensure drinking water bacteriological indicators. Generally, ozone, chlorine or sodium hypochlorite, etc. are used to oxidize and sterilize some microorganisms in water. Ozone is more effective than chlorine in oxidizing while the dosage is less, resulting in less chemical pollution, and the properties of water. Under normal circumstances, due to the long urban pipe network extension, ozone is easy to decompose in water.

Fig. 1. Sedimentation tank of water plant.

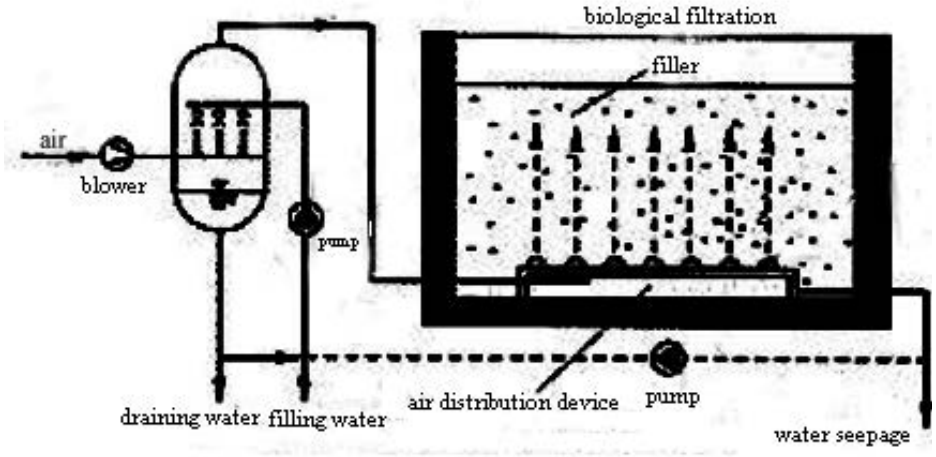

Fig. 2. Water plant filter tank.

In order to avoid the internal pollution of urban pipe network and ensure the residual disinfection level, a small amount of chlorine or chlorine ammonia will be added before the water into the pipe network mouth. Eventually the treated water will be pumped into the city's water grid and turned into tap water. The process flow of water plant is shown in Fig. 3.

Electric consumption runs through the water treatment process, among which the power consumption generated by the high-power mechanical processes such as the fetching pump (primary pump room), the lifting pump (secondary pump room), the backwashing system and the mud discharge system account for a high proportion. Of course, in addition to the electricity consumption, there is also reagent dosing in the water plant. For example, flocculant or even chlorine or ozone should be added during coagulation, and ozone or chlorine should be added during disinfection. Therefore, reasonable collocation can ensure the safe and stable operation of the water production process, improve the water quality and reduce the energy consumption of the water plant.

2.2. Optimized control system

2.2.1. System technical architecture design

It is based on spring boot framework in the realization of the optimal control system of water plant proposed

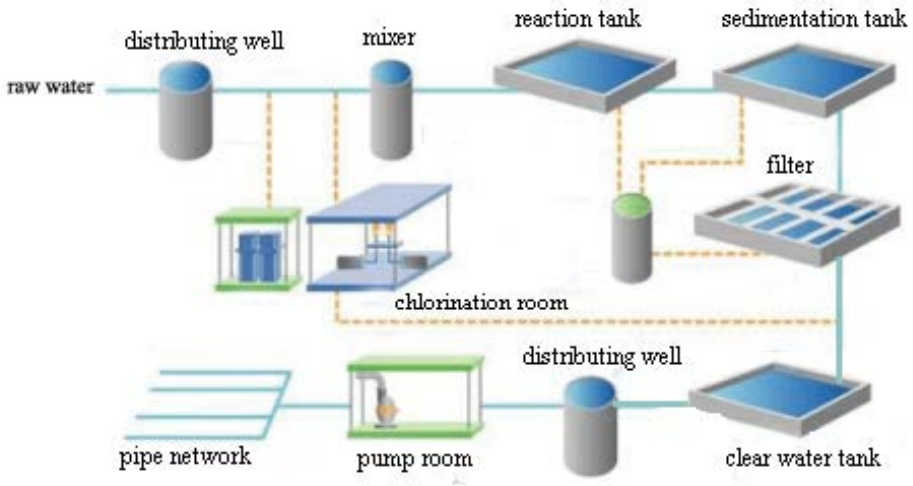

Fig. 3. Flow chart of water plant.

in this paper. The framework is newly provided by pivot team, who used specific ways to configure, so that defining the template configuration is unnecessary. According to the design structure of the water plant optimization system, the system operation is divided into visual layer, logic control layer, service layer and data layer. As shown in Fig. 4, the spring boot framework is shown, and the system operation principle is shown in Fig. 5.

Visual layer is a set of interfaces that interact with consumers. The main task of visual layer is human–computer or human–data interaction. In the visual layer, the system instructions triggered by users will be allocated by the control layer, which arrives at the specified business logic code and carry out business operations. The service layer usually refers to the interface and the implementation class of the interface. There is not much business logic in the service layer, whose majors is to realize the decoupling with the database, so as to make the system hierarchical and improve the security of the system to a certain extent. The main function of the interface in the service layer is to link with the data layer, and then add, delete, modify and query the database. The data layer can operate the tables in the database directly.

2.2.2. System structure design

Since the water treatment process is continuous, decoupled and each process can be carried out independently. Water plant optimization system proposed in this paper divides the water treatment process into independent,

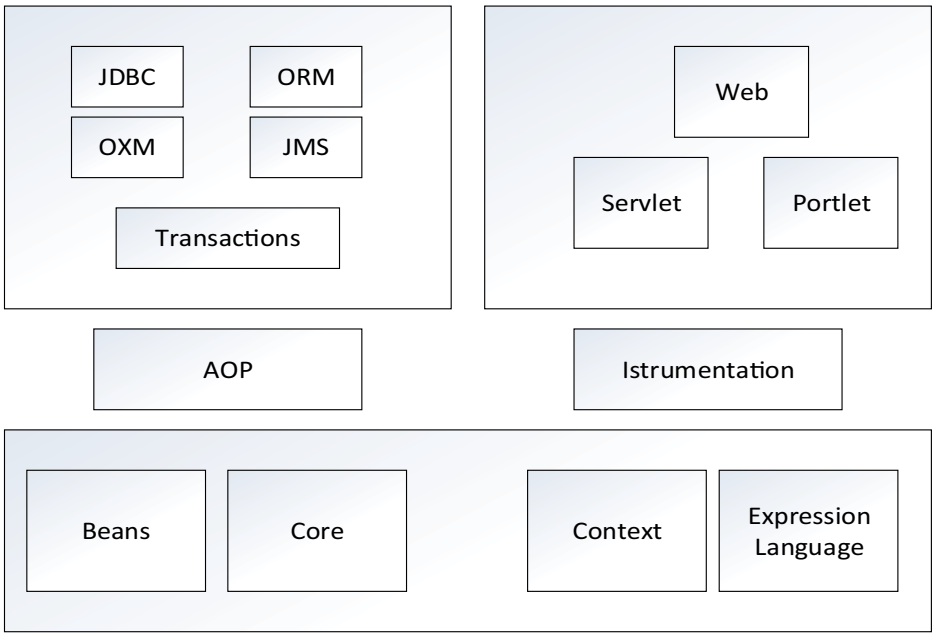

Fig. 4. Spring boot framework.

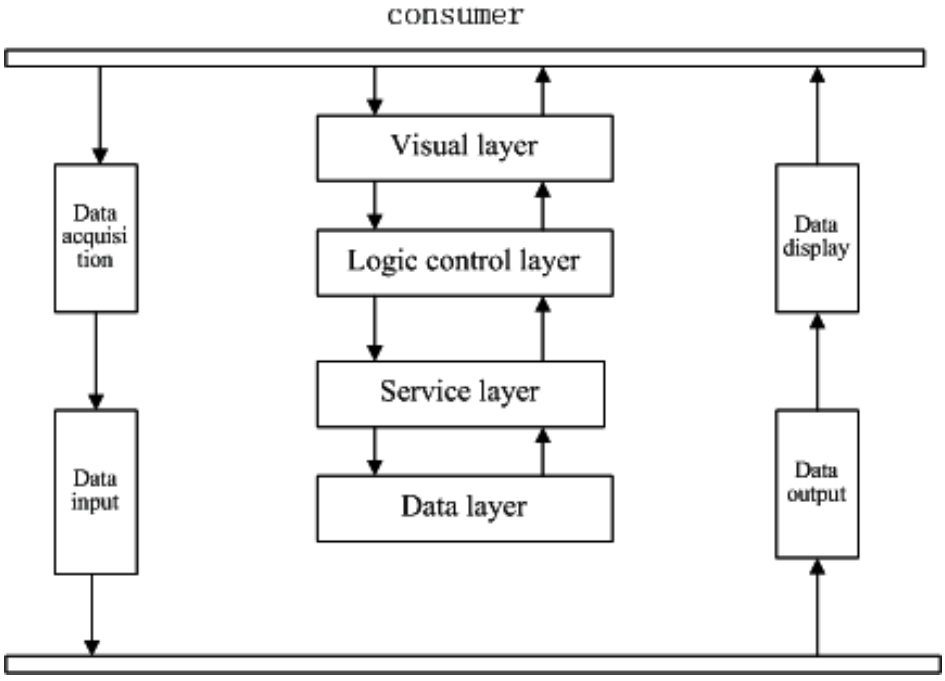

Fig. 5. System operation principle.

including the water intake process, reagent dosing, backwash and sludge discharge process and secondary water pump. Each subsystem can predict the corresponding energy consumption, and the prediction method is based on the historical data training model or empirical formula calculation. In addition, each subsystem will propose a reasonable parameter configuration scheme to support the water supply plant staff, which helps to provide more energy-saving optimal configuration scheme where the water treatment quality reaches the standard. The overall function design of water plant optimization system is shown in Fig. 6.

2.3. Machine learning

Alpaydin put forward his definition of Machine learning in 2004: Machine learning is programming computers to optimize a performance criterion using example data or past experience. Machine learning algorithm is generally called training data. According to the different training data, machine learning algorithm can be divided into supervised learning and unsupervised learning, semi supervised learning, and enhanced learning [15,16]. The common methods of machine learning includes regression algorithm, neural network, support vector machine and random forest. Although the principle of regression algorithm is simple, it cannot adapt to nonlinear prediction. Neural network has good generalization ability as a design pattern classifier, but it has many disadvantages such as over learning and local optimization. Meanwhile, its classification performance is far inferior to that of support vector machine and random forest. Support vector machine (SVM) has many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition, and has good classification performance. The generalization ability of random forest is better than that of SVM. However, random forest is inferior to support vector machine in imbalanced classification. In terms of classification performance, SVM and random forest have their own advantages and disadvantages. Therefore, support vector machine and random forest are selected as the mathematical models of pump unit scheduling optimization system, secondary pump subsystem and reagent dosing system, respectively.

2.3.1. Random forests

Ho [17] proposed the concept of random forests in 1995. Random forest is a tree based on machine learning, which uses the power of multiple decision trees to make decisions. Decision tree (DT) is a supervised machine learning algorithm used in solving classification and regression problems. Decision tree can be simply understood as a series of decisions to achieve a certain result. In machine learning, decision tree (DT) [18] classification algorithm is a prediction model. It describes a mapping relationship between object attributes and object values, and is composed of decision nodes, branches and leaf nodes.

The decision tree is shown in Figs. 7 and 8.

The value of entropy describes the degree of complexity and confusion of information contained in a set of data sets. The higher the value of entropy is, the more chaotic the information in the data set will be. The calculation formula of entropy is as follows:

$$H(S) = -\sum_{n \in N} p(n) \log_2 p(n) \tag{1}$$

In the formula, S plays the role of calculating entropy of the data set, N is a collection of classes, $p(n)$ is the

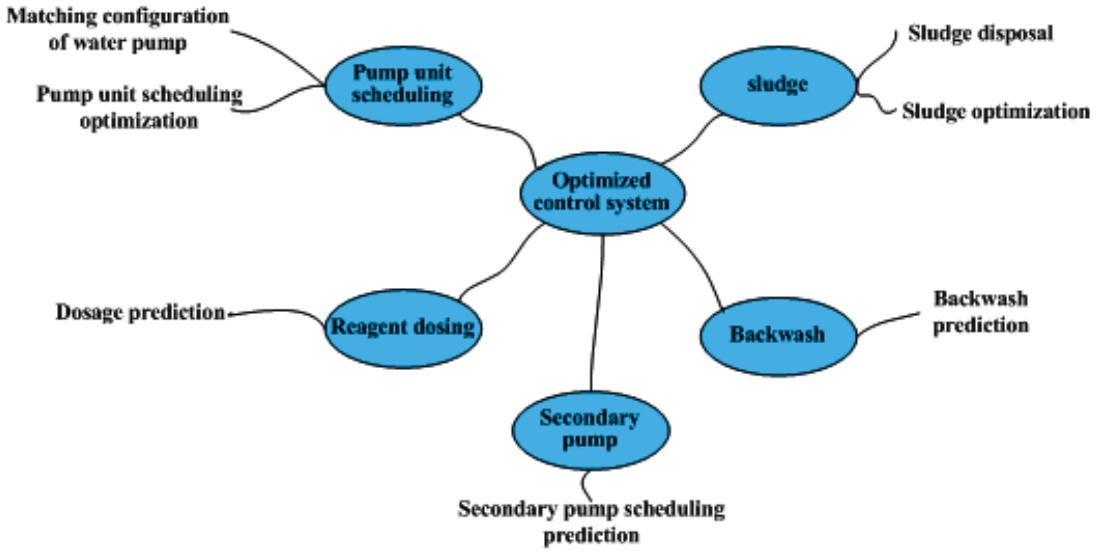

Fig. 6. Overall function design of water plant optimization system.

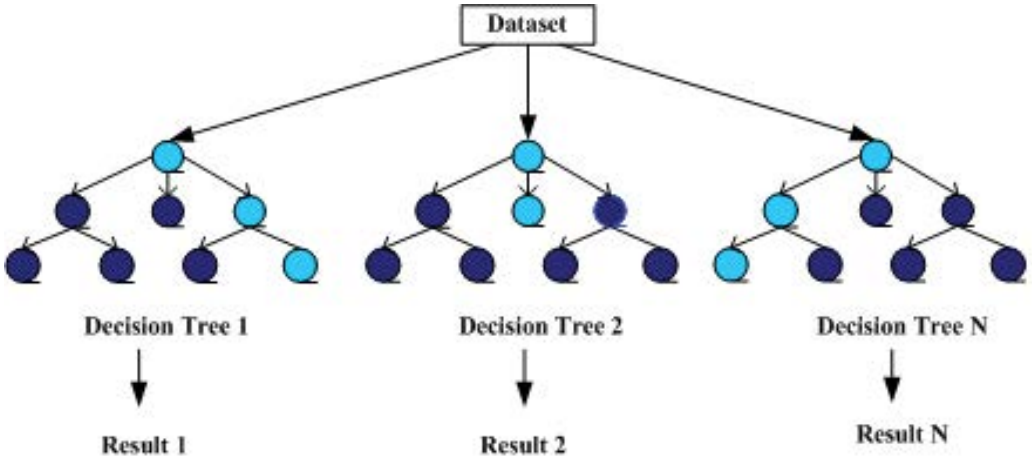

Fig. 7. Decision tree.

probability of the data set as part of the class N. When all the elements in the collection belong to the same class, the collection of information will be minimized. While the entropy of a node is equal to 0, it won't split in the ID3 algorithm. Selection of the root node selected while the attribute information gain maximum but not randomly, the information gain formula is shown below.

$$\mathrm{IG}(A,S) = H(S) - \sum_{t \in T} p(t)H(t) \tag{2}$$

Including entropy $H(S)$ is A collection of S, T is obtained by A attribute space S collection of subsets, the number of elements in $p(t)$ is A subset of T and S the ratio of the number of objects in the set $H(t)$ is A subset of entropy as A root node chooses. The next decision is to select the remaining property of the attribute information, which gain maximum, and then each decision point selection attribute of information gain in accordance with the residual properties of the rules of the maximum. Drop speed of every decision between information entropy is the fastest by the above rules. However, ID3 algorithm has some defects, for example, when a property under many different categories, each attribute category contains less elements and his information entropy is very low. If the property has no strong correlation between the results and even no correlation, it will lead to the failure. In order to solve the problem, C4.5 algorithm is put forward, which is according to the information gain rate to choose decision node and selecting information gain rate by selecting decision points every time. Calculating the information gain rate formula is as follows:

$$\mathrm{IG}(A) = \frac{\mathrm{IG}(A,S)}{H(A)} \tag{3}$$

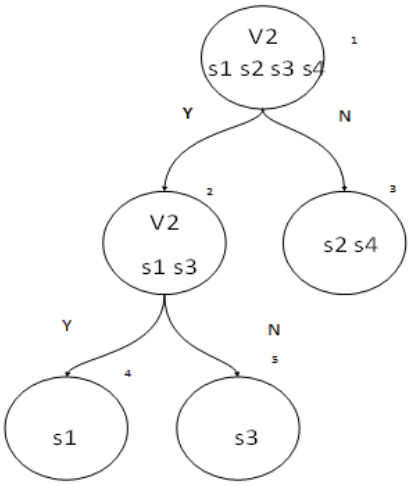

Fig. 8. Tree structure.

where IG(A) represents the information gain of attribute A, $H(A)$ represents the information entropy of attribute A. The information gain ratio of attribute A to its own entropy is the information gain rate.

2.3.2. Multi-output least squares support vector machine

SVM is a data-based machine learning model based on structural risk minimization (SRM) [19,20]. The SRM minimizes the empirical error and model complexity simultaneously which contribute to the improvement of generalization ability of the classification or regression problems [21]. It uses the maximum edge hyperplane as the decision plane for binary classification of data. The loss function and regularization method are used to calculate the empirical risk and optimize the risk structure. SVM can use kernel function to classify low dimensional data, in order to achieve the purpose of high-dimensional nonlinear classification. The solution objective and constraint conditions of support vector machine are shown in Eq. (4).

$$\begin{cases} \max\limits_{w,b} \dfrac{1}{\|w\|} \min\limits_{\substack{x_i \\ i=1,..,n}} y_i\left(w^T x_i + b\right) \\ \text{s.t } y_i\left(w^T x_i + b\right) > 0, i = 1,...,n \end{cases} \tag{4}$$

where y_i is the output value, $w^T x + b$ is hyperplane.

However, the standard support vector machine is long in training time and the computational complexity is high. In order to solve this problem, we think of its conventional improved least squares support vector machine first. The difference between the least squares support vector machine and the standard support vector machine is that the loss function adopts the two norm of error and changes the inequality constraint into equality constraint, which greatly reduces the computational complexity and training time. But the least squares support vector machine is only for the case of single output. When dealing with the multi output pump system, multiple single output least squares support vector machines are often used to estimate the model. It will also increase the complexity of the model. Therefore, an improved SVM multi output least squares regression algorithm is proposed. It not only has the characteristics of transforming the original non-equality constraints of support vector machines into the current equality constraints but also has better generalization ability and lower computational complexity when dealing with multi output systems. It can produce multiple arguments, to adapt to the complex multiple input multiple output system, its optimization objectives and constraints are as shown below:

$$\begin{cases} \min \dfrac{1}{2}\sum\limits_{j=1}^{k}\|w_j\|^2 + \dfrac{1}{2}C_0\sum\limits_{i=1}^{n}\lambda_i^2 + \dfrac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k}C_j\xi_{i,j}^2 \\ \sum\limits_{j=1}^{k}\left(y_{i,j} - w_j^T\varphi\left(x_i\right) - b_j\right) = \lambda_i \\ y_{i,j} - w_j^T\varphi\left(x_i\right) - b_j = \xi_{i,j} \\ j = 1,2,...,k;\ i = 1,2,...,n; \end{cases} \tag{5}$$

$$f_j\left(x\right) = \sum_{i=1}^{l}\left(\beta_i + \alpha_{i,j}\right)K_m\left(x_i, x_z\right) + b_j \tag{6}$$

where C_j is a single output of fitting error of punish coefficient, C_0 is fitting for all the output error of punish coefficient, λ_i is the fitting error of i dimension input to all output dimensions, $\eta_{i,j}$ is the fitting error between the i dimension input and the j dimension output, $K_m(x_i,x_z)$ is a kernel function. On one hand, Eq. (6) takes the loss function of input variables into account the fitting errors of each component in the model, which makes each individual error be a certain role to the objective function and then optimize the overall. On the other hand, it can reduce the influence of noise data and improve the anti-noise performance.

During the modeling of the pump subsystem in this paper, radial basis function (RBF) was adopted, as shown below:

$$K\left(x,z\right) = \exp\left(-\gamma\|x - z\|^2\right) \tag{7}$$

RBF is widely used because of the advantage of strong adaptability and just require few parameters, which are the default kernel function of libsvm. In the process of building the model, the combination of C and γ parameters has an important influence on the performance and accuracy of the model. In this paper, genetic algorithm [22] is adopted to solve the optimal parameter combination of the model (C, γ).

3. Prediction results and discussion

3.1. Prediction results of pump unit scheduling optimization subsystem

Between water pump room and water distribution throughout the valve in the pipe, the system designs the selection of the valve switch interface. When entering the parameters of the valve switch, system will return the raw water pressure, flow into the factory, and then the pump is

tie-in. The last step includes the open frequency of variable frequency pump, water pump, water pump efficiency and prediction of power consumption, which benefits for providing a valve switch in input combination cases of several pump collocation efficiency, power consumption, and so on. In this part, the realization of the function is mainly to provide history data from water plant and professional management reference configuration data water supply factory. After running the program, the system will collect the operation data in real time, and find all the information of the same valve configuration in the database. Then it display these data in order from low to high according to power consumption. Fig. 9 shows the parameter input prediction interface of the slurry pump subsystem. The operation of the interface is easy to understand and operate, only need to input the required water flow and the approximate range of pressure.

Based on massive historical data and professional configuration data, the system selects the pump plan according to the current pump configuration, and finally outputs it to the system page for employees' reference.

Fig. 10 shows the optimal pump matching scheme. Due to the experience, the water staff could obtain best solution to reduce unnecessary power consumption, which helps to improve the efficiency of the pump set and ensure safe and stable water production.

For solving the problem of pump matching efficiency, the system uses the multi output least squares SVM model through historical data. In the sewage pump subsystem, the input is the state of each valve in the pipe network, and the matching mode of six pumps is shown as the output parameter. At last, the selection of water supply pump from high flow rate to low energy consumption is provided. As shown in Fig. 10, the optimal pump unit combination given by the system is consistent with the optimal value of historical data (group 20 data) in Table 1. So, the matching results given by the system are identical and the system prediction is accurate.

3.2. Prediction results of sludge discharge optimization subsystem

The sludge discharge optimization subsystem could predict the period and power consumption of sludge discharge. In this subsystem, input is raw water flow, external drainage volume and raw water turbidity. Meanwhile, the output consists of the sludge discharging period, total

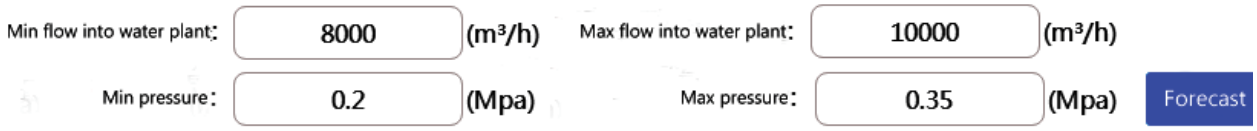

Fig. 9. Input parameter interface of slurry pump subsystem.

Table 1
Experience configuration and power consumption of pump unit scheduling[TS: Please check the entry "4□6(7Hz)".]

	Muddy water pressure (Mpa)	Water inflow (m^3/h)	Water pump matching scheme	Efficiency (%)	Power consumption (kwh/km^3)
1	0.29	9,438.89	3(45.9 Hz), 4	74.64	101.68
2	0.29	9,415.74	3(46.9 Hz), 5	76.81	101.69
3	0.29	9,438.89	3(45.9 Hz), 4	76.93	101.45
4	0.29	9,412.85	3(45.9 Hz), 4	77.01	101.43
5	0.29	9,427.31	3(45.9 Hz), 4	77.17	101.32
6	0.27	9,025.12	3(7 Hz), 4, 6(7 Hz)	77.43	95.4
7	0.27	9,004.86	3(48 Hz), 4	77.52	95.27
8	0.27	9,019.33	3(40.69 Hz), 4	77.85	95.16
9	0.27	9,019.33	3(48.69 Hz), 4	77.98	94.91
10	0.27	9,013.54	3(40.69 Hz), 4	78.03	94.68
11	0.27	9,007.75	3(48.69 Hz), 4	78.05	95.03
12	0.27	9,033.8	3(48.69 Hz), 4	78.11	94.73
13	0.27	9,016.43	3(48.69 Hz), 4	78.17	94.7
14	0.27	9,042.48	3(40.69 Hz), 4	78.22	94.94
15	0.24	8,177.31	3(30.69 Hz), 4	79.01	84.23
16	0.24	8,258.33	3(40.69 Hz), 4	79.6	83.3
17	0.24	8,240.97	3(30.69 Hz), 4	79.61	83.5
18	0.24	8,232.29	3(30.69 Hz), 4	79.67	83.5
20	0.24	8,500	3(49.9 Hz), 6(45.9 Hz)	79.67	83.2

Pump pressure (Mpa)	Water flow (m^3/h)	Collocation pattern						Efficiency (%)	Power consumption (kwh/km^3)
		1	2	3	4	5	6		
0.24	8500			49.9Hz			45.9Hz	79.6	83.2

Fig. 10. Prediction results of pump unit scheduling.

displacement, water yield ratio, and electricity consumption of sludge removal.

Mud drainage mainly includes two processes: siphon mud drainage and perforation mud drainage. In this part of the function, the prediction of mud discharge is mainly based on empirical judgment, empirical data and corresponding calculation formulae.

When the original water quantity is 300,000 m^2, the following empirical data are obtained for the discharge period, the single discharge water quantity and the average unit power consumption of the two processes.

- mud discharge period

The interval of raw water turbidity has the following relation with the mud discharge period of the two sludge discharge processes.

The turbidity of raw water is inversely proportional to the period, which means the more turbidity the water is, the more frequently the mud discharge word should be carried out. According to this table, we can get the siphon drainage period T_1^* and perforated drainage period T_2^* under the inlet water amount of 300,000.

- a single discharge of mud water, average unit power consumption

According to experience, the water volume and average unit power consumption of a single discharge of mud in two sludge discharge processes are shown in the following table:

When we know the parameters such as the period, the single discharge volume and the average unit electricity consumption of the two types of sludge discharge. We are able to calculate the period, water production ratio, water quality, electricity consumption and sludge discharge unit consumption required for sludge discharge under a certain amount of water. The calculation formula of mud discharge period is shown in formula (8):

$$\begin{cases} T_1 = 300{,}000 / L \times T_1^* \\ T_2 = 300{,}000 / L \times T_2^* \end{cases} \tag{8}$$

where T_1 is the siphon mud drainage period, T_2 is the perforated mud drainage period, L is the actual original water input, and the unit is m^2.

The calculation formula of single mud discharge water is shown in formula (9):

$$\begin{cases} \mathrm{Lm1} = 900 \times T_1 / 24 \\ \mathrm{Lm2} = 150 \times T_2 / 24 \end{cases} \tag{9}$$

where Lm1 is the siphon water quantity of a single mud discharge, Lm2 is the perforation water quantity of a single mud discharge. Fig. 11 shows the prediction of sludge discharge period and the amount of single sludge.

Finally calculate the total displacement, water production ratio, electricity consumption, mud discharge per consumption.

The calculation formula of total displacement is shown in Eq. (10):

$$\mathrm{Lm} = \left(\mathrm{Lm}_1 + \mathrm{Lm}_2\right) \times S \tag{10}$$

where Lm is the total displacement and S is the total number of pools.

The calculation formula of water yield ratio is shown in Eq. (11):

$$\mathrm{Lr} = \frac{\left(L - \mathrm{Lm} - \mathrm{Lh}\right)}{L} \tag{11}$$

where Lr is the ratio of water production, Lh is the amount of water input and external drainage for reuse.

The calculation formula of power consumption is shown in Eq. (12):

$$\mathrm{Ec} = \left(\mathrm{Lm} / 1{,}000\right) \times 34.5 \tag{12}$$

where Ec is the power consumption.

The calculation formula of mud discharge per consumption is shown in Eq. (13):

$$\mathrm{ec} = \mathrm{Ec} / \left(L \times \mathrm{Lr} / 1000\right) \tag{13}$$

where ec is the mud discharge per unit.

Fig. 12 shows input parameters of sludge subsystem. When the raw water is input every day, the water is drained back to the raw water and becomes turbid. It can be used to predict the period and power consumption of different sludge discharge methods.

As shown in Fig. 13, the results of sludge discharge optimization scheme can clearly guide the water plant staff to regularly carry out sludge discharge work.

The period of sludge discharge, the amount of sludge discharged in a single time, energy consumption and efficiency of each sludge discharge process are the results obtained by optimizing the above formula through empirical data (The data are obtained based on the condition that the inlet water amount of the studied water plant is 300,000 m^2, as shown in Tables 2 and 3). If the sludge discharge period and energy consumption of other water

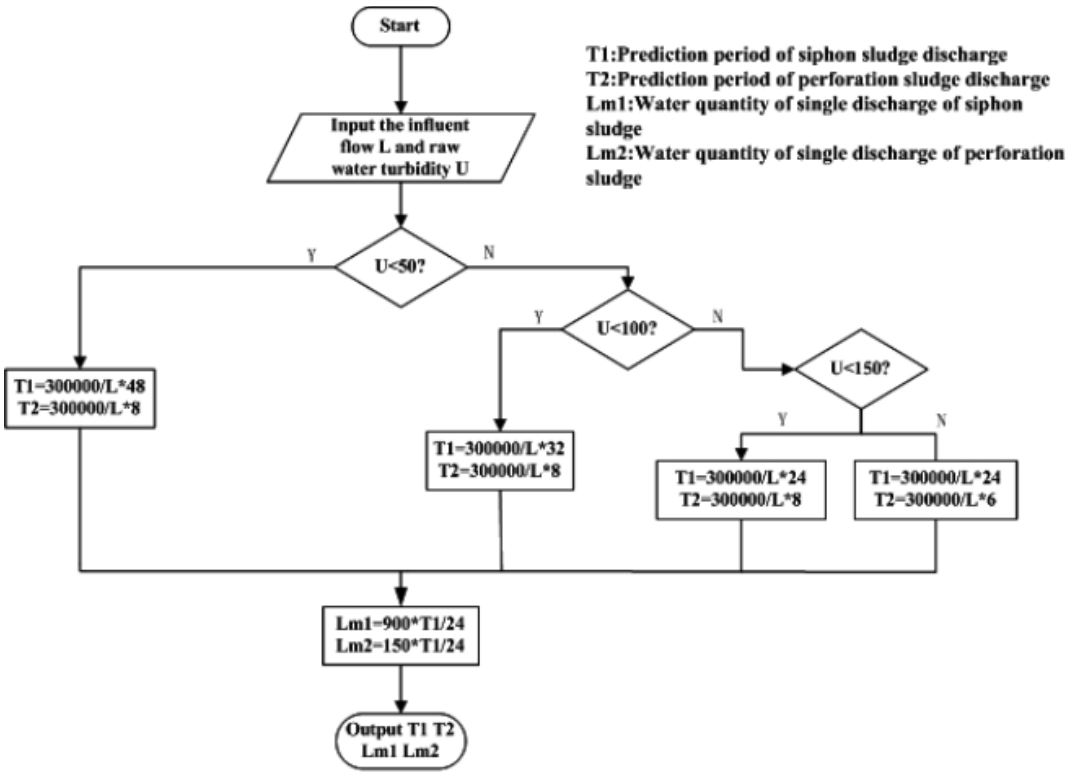

Fig. 11. Flow chart of sludge discharge prediction.

Fig. 12. Input parameters of sludge subsystem.

Sludge discharge optimization scheme

Raw water quantity (m3/d)	Discharged water quantity of recycling (m3/d)	Perforated method period (h)	Siphon method period (h)	Total sludge dischange water quantity (m3/h)	Power consumption (kwh)	Water production rate (%)	Average power consumption (kwh/km3)
120000	40	20	120	18500.00	638.250	84.55	6.29

Fig. 13. Optimization scheme of sludge subsystem.

plant are predicted, it should be adjusted and optimized of the calculation formula of energy consumption and period according to the actual working conditions. At present, the scheme has been applied to the water plant with remarkable results.

3.3. Prediction results of backwash subsystem

In the backwash subsystem, the backwash period can be predicted by inputting the raw water flow, the surface

Table 2
Mud discharge period division

Raw water turbidity interval (NTU)	Siphon mud	Perforated mud
<50	48 h	
50–100	32 h	8 h
100–150	24 h	
>150	24 h	6 h

Table 3
Single discharge of mud water and average unit power consumption

	Every time the sludge discharge quantity (m^3)	Average unit power consumption (kwh/km^3)
Siphon mud	150	34.5
Perforated mud	900	34.5

density of algae and the turbidity of precipitation effluent. The power consumption can be predicted by inputting pump and blower frequency of air, water mixed flush, single air flush and water flush.

Table 4 shows the prediction results of various parameters in the filter on the backwash period.

The first thing that the water plant staff should do is to configure the frequency and power of the water pump and blower, and then input the frequency and time in the prediction interface to calculate the power consumption based on the configuration data. The system designs a configuration interface for backwashing, in which we can configure the frequency F and efficiency P of water, air and mixed flushing pumps and blowers, at the same time the input frequency f and time t (min) can be calculated according to Eq. (14):

$$\mathrm{Ecb} = \left(f/F\right) \times P \times \left(t/60\right) \tag{14}$$

where Ecb is backwash power consumption.

Table 5 shows the prediction of power consumption of backwash equipment under different collocation conditions. Through the above calculation formula (14), the power consumption of different flushing modes can be calculated. It can make the water plant monitoring personnel know the flushing method with the lowest energy consumption.

3.4. Prediction results of reagent dosing subsystem

Reagent dosing is mainly in the process of coagulation and disinfection. Based on the historical data and medication habits of a water plant, the system could predict the amount of alum, pre-ozone, main ozone and sodium hypochlorite in the water plant. In this subsystem, the dosage can be predicted after the input of raw water flow, turbidity, water temperature, pH value, dissolved oxygen and oxygen consumption and other parameters. For predicting ozone, we should input raw water smell, water ozone standard and ozone concentration before exhaust destruction to the system. Then the system will return the optimal quantity of primary ozone and pre-ozone. Sodium hypochlorite is mainly used for maintaining the residual chlorine level of the tap water in the urban pipe network, so that the contamination of the tap water caused by microorganisms in the pipeline network can be avoided. It is usually put into the municipal pipe network before the tap water enters through the lift pump. After predicting the amount of sodium hypochlorite, input raw water oxygen consumption, raw water ammonia nitrogen, raw water temperature, factory water residual chlorine control standard and other parameters of the system, the optimal dosage of sodium hypochlorite could be output.

Table 4
Backwash parameters and prediction of period

Raw water quantity	250,000 m^3
Algae density of raw water	44 ten thousand/L
Effluent turbidity of sedimentation tank	20 NTU
Prediction period	14 h

The interactive design diagram of water treatment agent dosage prediction is shown in Fig. 14. The prediction of reagent dosing mainly refers to the establishment of prediction model based on random forest according to the historical data of dosage. Through a large number of previous empirical data and the external environmental parameters that affect the dosage, the trained model system will compare the input parameters with the historical data to obtain the optimal dosage. The application of the model system can regulate the dosing of chemicals and avoid the safety of water quality caused by human error operation.

3.5. Prediction results of secondary pump subsystem

The secondary water pump system is responsible for delivering the sterilized water to every household. It is necessary to control the water volume and increase the water pressure, for keeping the water used normally at a certain height, which contribute to avoiding the inconvenient use of tap water in the upper floors due to low water pressure. The best prediction of lifting level and power consumption of the system is to output the level curve by inputting required parameters such as lifting water level and lowest and highest level before lifting.

As shown in Fig. 15, when inputting parameters such as lifting water amount and lifting level range, the system will search the relevant data within the range in the database, and display the relationship curve of water level, power consumption and efficiency, so as to provide advice to the staff. At the same time, the optimal liquid level is found by random forest algorithm.

Table 6 shows the historical data of secondary pump operation. It includes raising water level, liquid level, working efficiency and power consumption. The prediction of the subsystem is to query the data in the database according to the input lifting water volume, such as the input in the above figure: the lifting water volume is 4,900, the min liquid level is 3, and the max liquid level is 4. When the lifting water volume is 4,900 and the liquid level is between 3 and 4, the lowest power consumption is regarded as the best lifting level. Fig. 15 is the graph fitting according to the found data. At present, it has been put into

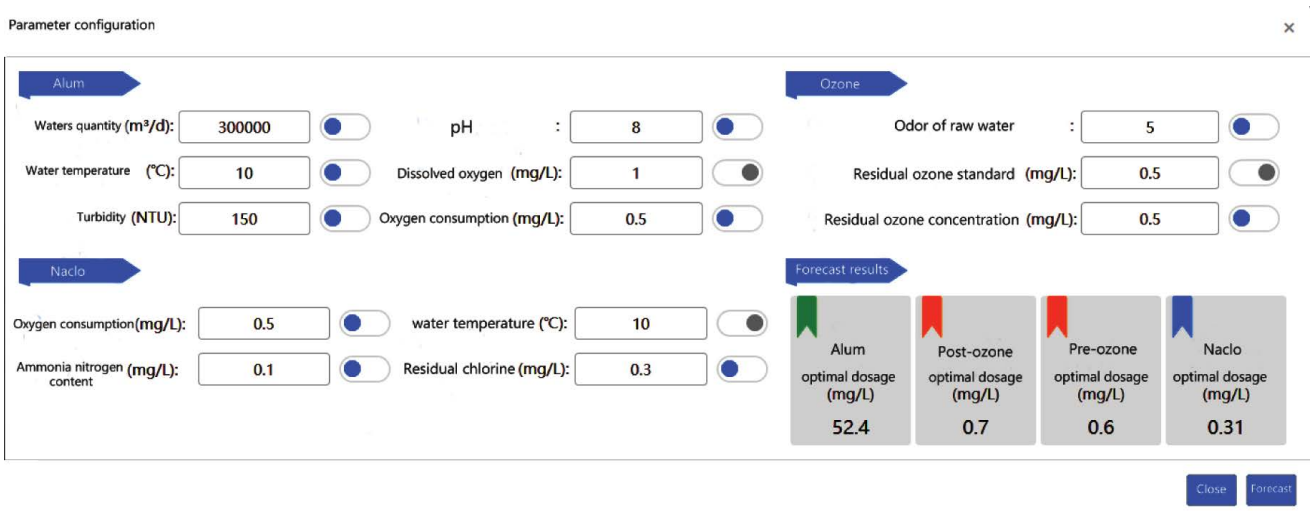

Fig. 14. Prediction of dosage of water treatment reagent.

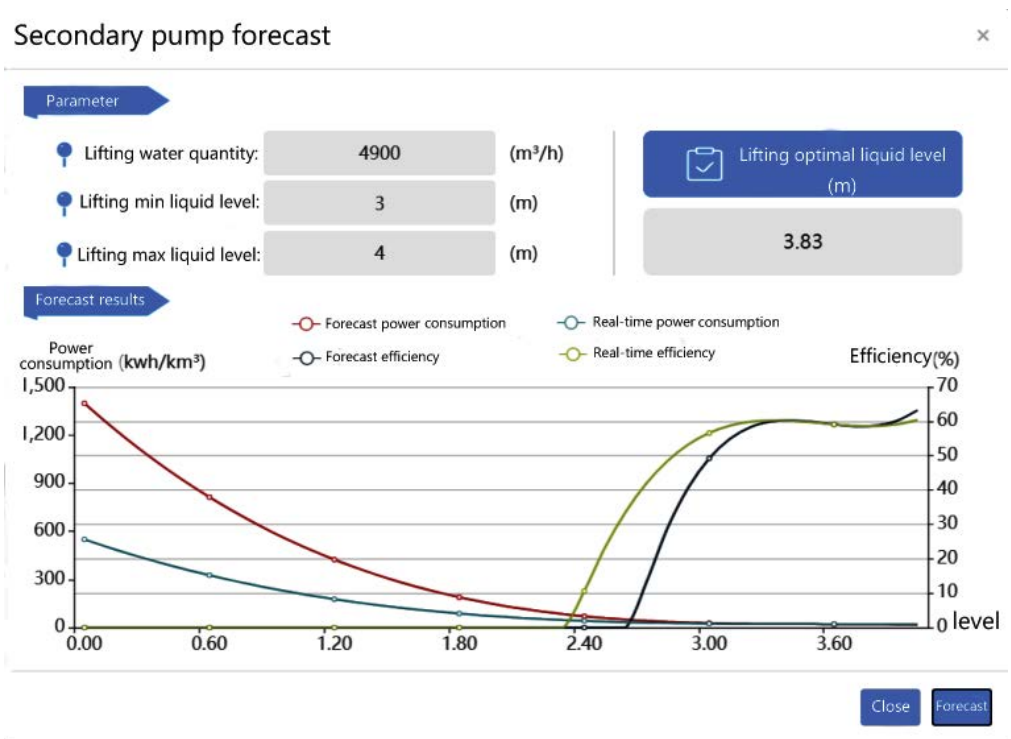

Fig. 15. Interaction design drawing of two-stage pump release prediction.

Table 5
Power consumption prediction

	Pump 1 frequency/(Hz)	Pump 2 frequency/(Hz)	Blower 1 frequency/(Hz)	Blower 2 frequency/(Hz)	Time/ (min)	Power consumption/ (kwh/km³)
Air flushing	\	\	30	30	30	16.07
Water flushing	30	30	\	\	30	25.02
Mixed flushing 1	40	\	33	33	30	25.09
Mixed flushing 2	30	35	35	\	30	35.35

Table 6
Historical data

	Lifting water quantity/(m^3/h)	Liquid level/(m)	Efficiency (%)	Power consumption (kwh/km^3)
1	5,659.72	3.92	58.86	20.05
2	5,512.15	3.96	58.54	19.97
3	5,529.51	3.72	59.15	20.86
4	5,546.88	3.73	60.53	20.32
5	6,038.77	3.53	59.79	21.51
6	5,949.07	3.55	59.90	21.35
7	5,989.58	3.46	59.92	21.78
8	5,966.44	3.48	60.35	21.53
9	6,001.16	3.40	60.95	21.66
10	5,821.76	3.42	59.92	21.96
11	5,914.35	3.33	60.30	22.20
12	5,873.84	3.34	60.39	22.13
13	6,021.41	3.35	62.15	21.45
14	5,879.63	3.37	59.86	22.18
20	5,714.70	3.41	60.87	21.66

the actual operation of the water plant. Through the prediction curve, the staff of the water plant can understand the data under the highest efficiency and the lowest energy consumption, and can intuitively compare the optimal operation parameters.

4. Conclusion

This paper proposes a waterworks optimization control system based on machine learning, through the historical data of plant water treatment process. The main advantage of this article can be summarized as: first, although machine learning technology has been widely used, the application and research towards waterworks optimization control aspect is insufficient. This paper proposed a waterworks optimization control system based on machine learning, offered a new way for optimal control of water, at the same time confirmed the effectiveness of the machine learning in the application of water energy consumption prediction and configuration optimization direction. Second, this paper designs an optimization control system for water plant optimization, consisting of a concise and convenient control interface for the above subsystems, and combines the theory and the use of the water plant optimization control system.

Acknowledgments

This work is supported by National Natural Science Foundation of China (51708299), Major Science and Technology Program for Water Pollution Control and Treatment (2017ZX07201001).

References

[1] J. Hakanen, K. Miettinen, K. Sahlstedt, Wastewater treatment: new insight provided by interactive multiobjective optimization, Decis. Support Syst., 51 (2010) 328–337.
[2] H. Han, L. Zhang, Y. Hou, J. Qiao, Nonlinear model predictive control based on a self-organizing recurrent neural network, IEEE Trans. Neur. Networks Learn., 27 (2016) 402–415.
[3] A. Saptoro, State of the art in the development of adaptive soft sensors based on just-in-time models, Procedia Chem., 9 (2014) 226–234.
[4] B. Szelag, K. Barbusinski, J. Studzinski, L. Bartkiewicz, Prediction of wastewater quality indicators at the inflow to the wastewater treatment plant using data mining methods, E3S Web Conf., (2017) 1–8.
[5] P. Yongeun, L. Mayzonee, M.K. Young, Development of enhanced groundwater arsenic prediction model using machine learning approaches in Southeast Asian countries, Desal. Water Treat., 57 (2016) 12227–12236.
[6] K. Ding, J. Zhang, X. Zhu, The model of pump head data mining based on SVM, Appl. Mech. Mater., 2685 (2013) 3263–3268.
[7] A. Sharafati, S. Asadollah, M. Hosseinzadeh, The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty, Process Saf. Environ., 140 (2020) 68–78.
[8] V. Sousa, J. Matos, N. Matias, I. Meireles, Statistical comparison of the performance of data-based models for sewer condition modeling, Struct. Infrastruct. Eng., 15 (2019) 1680–1693.
[9] S. Włodzimierz, K. Wojciech, Determination of the optimal operational parameters for a three-phase fluidised bed bioreactor with a light biomass support when used in treatment of phenolic wastewaters, Biochem. Eng. J., 20 (2014) 49–56.
[10] C. Song, H. Wang, P. Li, A Receding Optimization Control Policy for Production Systems with Quadratic Inventory Costs, IFAC Proceedings Volumes, 2004, pp. 713–717.
[11] W. Ang, W.M Abdul, H. Nidal, P.L Choe, A review on the applicability of integrated/hybrid membrane processes in water treatment and desalination plants, Desalination, 363 (2015) 2–18.
[12] S. Meng, Y.R. Shen, E. Wang, Basic science of water: challenges and current status towards a molecular picture, Nano Res., 8 (2015) 3085–3110.
[13] E.S. Rigobello, A.D. Dantas, L.D. Bernardo, Removal of diclofenac by conventional drinking water treatment processes and granular activated carbon filtration, Chemosphere, 92 (2013) 184–191.
[14] M. Vliet, J. Yearsley, W. Franssen, F. Ludwig, Coupled daily streamflow and water temperature modelling in large river basins, Hydrol. Earth Syst. Sci., 16 (2012) 4303–4321.

[15] L. Zhang, C. Chen, J. Bu, D. Cai, X. He, Active Learning based on locally linear reconstruction, IEEE Trans. Pattern Anal., 33 (2011) 2026–2038.
[16] C. Chen, L. Zhang, J. Bu, Constrained Laplacian Eigenmap for dimensionality reduction, Neurocomputing, 73 (2010) 951–958.
[17] T.K. Ho, Random Decision Forest, Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278–282.
[18] X. Wang, L. Wang, N. Li, An application of decision tree based on ID3, Phys. Procedia, 25 (2012) 1017–1021
[19] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Network, 10 (1999) 988–999.
[20] G. Hong, J. Kwan, J. Lim, J. Jo, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, J. Environ. Sci., 32 (2015) 90–101.
[21] H. Yoon, S.C. Jun, Y. Hyun, A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, J. Hydrol., 396 (2011) 128–138.
[22] Q.J. Wang, Using genetic algorithms to optimise model parameters, Environ. Model. Softw., 12 (1997) 27–34.