Scholar Publishing Group
Water Pollution Prevention and Control Project
https://doi.org/10.38007/WPPCP.2021.020101
ISSN 2790-2730 Vol. 2, Issue 1: 1-10

Assessment of Water Pollution Degree in Qingshui River Basin Based on Stochastic Forest Algorithm

Xianmin Ma[*]

Department of Information Engineering, Heilongjiang International University, Harbin 150025, China

maxianmin@hiu.net.cn

[*]*corresponding author*

Keywords: Water Pollution, Assessment of Pollution Degree, River Basin, Random Forest Algorithm

Abstract: Water is the active basis for the movement and transformation of various biological, physical and chemical substances and the flow of energy. On the one hand, the quality of water environment is very sensitive to changes in the external world; on the other hand, water quality affects the health of organisms and ecosystems through ecological support and other functions. The space-time distribution of water quality was analyzed according to the composition of basic substances and established water quality standards, which provided scientific basis for the rational development and utilization of water resources planning and management. In this paper, the stochastic forest algorithm was used to evaluate the water pollution degree of Qingshui River basin. This paper first analyzed the causes of Qingshui River pollution, and introduced the establishment method of random forest (RF) classification model. After that, this paper established the evaluation criteria of water pollution degree, and constructed the evaluation model of water pollution degree of Qingshui River basin according to the RF algorithm. In the experiment part, the quantity of decision trees (DT) of the RF model was set to 400 through sample training, and the classification accuracy of the RF classification model and the artificial neural network (ANN) model was compared. The experimental results showed that the RF classification model has a high classification accuracy rate of 97.33%, which can be used to assess the degree of water pollution. At the end of this paper, the RF classification model was used to evaluate the water pollution degree of the Qingshui River basin. The results showed that for the three stations in the Qingshui River basin, the water quality classification results are Class I, Class III, and Class IV. The water pollution degree of the Qingshui River basin is relatively serious.

1. Introduction

Water quality assessment mainly reflects the complex biological, chemical and physical conditions of water bodies, and is the basis for optimizing water quality management. In view of the multi-dimensional and holistic understanding of the water quality in the aquatic environment, scientists now tend to adopt a holistic approach of multiple indicators, and evaluate the water quality according to the content of different minerals or nutrients in the water, the physical properties of the water, such as hardness, salinity, oxygen content, and biological activity. The single attribute of water body state is taken as the basic evaluation scale, and the data processing is standardized to eliminate the differences between different attributes in the scale, so as to comprehensively evaluate the indicator scheme.

Many scholars have studied water pollution. Evans Alexandra EV reviewed the knowledge of selected areas of agricultural water pollution and identified future research needs [1]. Feng Xiaofang studied the progress, mechanism and trend of two-dimensional titanium carbide in the degradation of water pollutants [2]. Ji Mengzhi summarized and evaluated the role of bacteriophages in monitoring pathogens, tracking pollution sources, treating pathogens, infecting algal blooms and controlling sludge bulking and biofilm pollution in sewage treatment systems [3]. Sharma Rohit took different rivers in India as an example to analyze the water quality parameters from 2012 to 2016 [4]. Morin-Crini Nadia reviewed the persistent organic pollutants in lakes and marine ecosystems [5]. Wikurendra Edza Aria analyzed the water quality of the Pucang River in the Western Duo Arho Regency and controlled water pollution [6]. Although there are many studies on water pollution, further research is needed on how to evaluate water pollution.

Stochastic forest algorithm is a common evaluation and classification method. Iqbal Mudassir proposed the development of RF regression model to predict the tensile strength retention rate of reinforcement under laboratory conditions in alkaline environment [7]. Lu Shengfu used the RF regression model to evaluate the degree of depression symptoms [8]. Sadler J M used the RF algorithm to estimate the flood severity in urban coastal areas [9]. Wang Jing used RF algorithm to assess the risk of coronary heart disease [10]. Dharumarajan Subramanian used RF classification algorithm to predict soil texture categories [11]. Mobley William used RF classification to predict flood probability area [12]. Although the evaluation effect of RF algorithm is good, the application of RF algorithm in water pollution evaluation is not sufficient.

In order to study the current situation of water quality in the Qingshuihe River basin, this paper used the stochastic forest algorithm to evaluate the degree of water pollution in the Qingshuihe River basin. This paper discussed the problems of water shortage, serious pollution, backward wastewater treatment technology, and lack of water environment management in the upstream of Qingshui River basin, and then constructed a RF. For the Qingshui River basin, this paper constructed the water quality evaluation standard, and used the RF algorithm to build the evaluation model. The accuracy of the RF classification model was verified by experiments, and the current situation of water pollution in the Qingshui River basin was analyzed.

2. Current Situation of Water Pollution in Qingshui River Basin

The Qingshui River basin is completely located in Zhangjiakou City. Xigou River basin is sparsely vegetated, mainly composed of barren mountains and hills. The rocks are exposed all year round, causing serious soil erosion and serious damage in the dry season. The Donggou River and Zhenggou River basins have good vegetation, high forest coverage, strong biodiversity and strong water conservation capacity. In the past, the water of the Qingshui River was crystal clear. In recent years, with the rapid development of economy, the living conditions of residents have been greatly improved. Qingshuihe began to carry a large amount of industrial and domestic wastewater, causing many environmental problems, mainly in the following four aspects.

2.1. Reduction of Upstream Water Volume and Uneven Space-time Distribution

In recent years, the climate conditions in the upper reaches of the Qingshui River have changed. Due to the economic development, a large quantity of water resources have been mobilized, resulting in less precipitation, and the amount of water entering the river has decreased year by year.

Qingshui River is a seasonal mountain river, with uneven spatial and temporal distribution of precipitation. 80% of the water is concentrated in the flood period from June to September, and there is little or no irrigation in the other eight months, resulting in low flow efficiency in the dry period.

2.2. Coexistence of Urban Point Source Pollution and Rural Non-point Source Pollution

Improper industrial structure design is the main reason for urban point source pollution. Zhangjiakou is an old industrial city with few high-tech industries, and the pollution of energy-intensive industries accounts for a large part. The growth of urban population has also led to the increase of domestic sewage. Many industrial plants, large and small drainage pipes of life are continuously discharged into the Qingshui River, resulting in urban pollution. In recent years, sewage treatment has been improved to some extent, but it still needs to be strengthened. Organic and inorganic pollutants in mountain agriculture, such as fertilizers and pesticides, infiltrate through surface runoff and groundwater and pollute the catchment area. In recent years, the scale of cattle and poultry breeding has been expanding. More and more families raise pigs, cattle and sheep. Large farms have also begun to develop. They produce a large quantity of animal and poultry excreta, causing certain water pollution.

2.3. Backward Wastewater Treatment Technology

Although large industrial and mining enterprises in the Qingshuihe River basin have wastewater treatment facilities, most of the wastewater treatment equipment is outdated and backward, the wastewater treatment rate is difficult to keep up with the improvement of the enterprise’s production capacity, the amount of wastewater is increasing, the investment in urban treatment facilities is insufficient, and the treatment capacity is low, which generally cannot meet the needs of urban development.

2.4. Strengthening the Management of Water Environment in the Basin

A sound basin water environment management system is still under study, and the management mechanism from the whole basin area to the upper, middle and lower reaches is not specific. The lack of mechanism construction is also one of the main reasons for water pollution. Although the country has formulated corresponding laws to improve the water environment management of the river basin, the relevant laws and local regulations have not penetrated into all villages and families in the upstream mountain areas, and some mountain residents living in more remote areas have a weak awareness of the protection of water sources.

3. RF Algorithm

The RF algorithm contains two main randomization ideas: random subspace idea and Bagging idea. The idea of random subspace is to randomly select a set of features with equal probability from the entire feature set of feature variables when dividing each DT node, and then select the best feature from this subset to divide the nodes. Bagging algorithm is one of the earliest supervised learning algorithms. It and RF algorithm are based on bootstrap method to resample to create multiple training sets. Resampling refers to randomly selecting a data set to reset N times and retrieving the same amount of data as the original data set to generate a sample set [13].

The RF algorithm only needs a few control parameters. It has high calculation efficiency and can process high size (multi-parameter) data. Moreover, it can learn quickly without special adjustment.

3.1. Construction of RF

RF is a combination classifier, which uses k DT as the basic classification group for learning control. It has a feature set of independent and uniformly distributed random vectors, and the DT is not associated. The growth of each tree does not need pruning. In the process of tree generation, the variable eigenvalues of each node are only generated from several randomly selected feature subsets [14]. The generation process of the DT is shown in Figure 1.

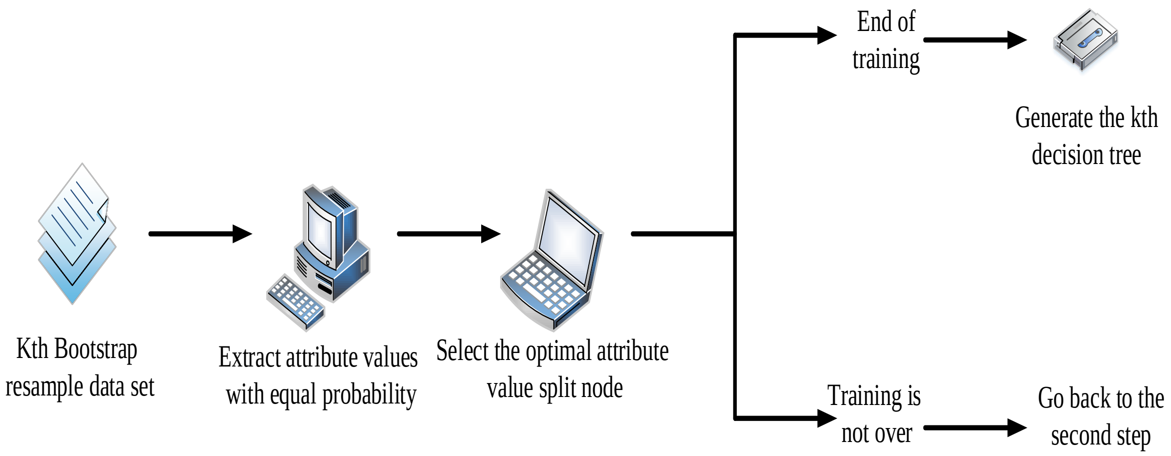

End of training
Generate the kth decision tree
Kth Bootstrap resample data set
Extract attribute values with equal probability
Select the optimal attribute value split node
Training is not over
Go back to the second step

Figure 1. Generation of DT

The randomness of data and attribute variable values allows the generation of a large quantity of trees, called "RFs". The generation of RFs is shown in Figure 2.

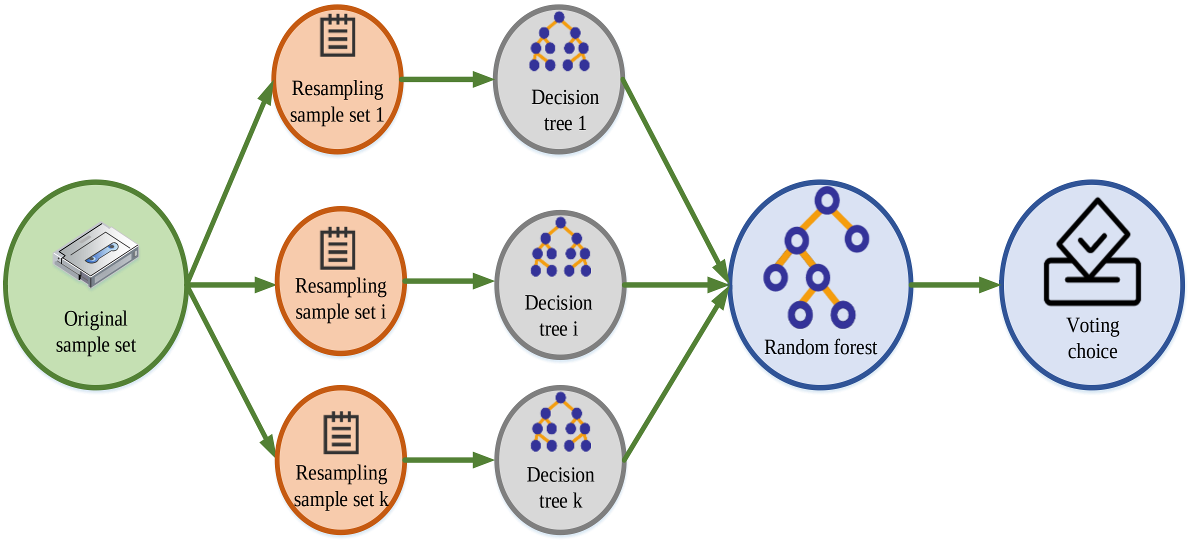


Figure 2. Generation of RF

When each DT is constructed, the parameter set is an independent and equally distributed random vector, because the random selection process of training sample set and attribute subset is independent, and the whole parameter set is the same.

The classification decision can be expressed as:

$$\mathrm{H(x)} = \arg\max_{\mathrm{r}} \sum\nolimits_{\mathrm{i}}^{\mathrm{k}} \mathrm{I} \tag{1}$$

$$\mathrm{h_i(x)} = \mathrm{Y} \tag{2}$$

Among them, H(x) represents the combined classification model; $\mathrm{h_i}$ is a single DT classification model; Y represents the target variable (or output variable).

3.2. Model Establishment of RF Classification Algorithm

The two most important parameters in the RF algorithm are ntree, which is the quantity of DT, and mtry, which is the quantity of attributes in the split attribute set. Here, mtry separates the quantity of attributes in the attribute set. For a given attribute value, the algorithm only selects the most effective attribute to split the nodes within its range. The attribute quantity mtry of the split attribute set is a user-configurable parameter, which is very sensitive to the classification efficiency of the RF [15]. The operation process of RF algorithm can be described as follows.

The original X data set is re-selected using Bootstrap method, and the learning sample $\theta_1^*, \theta_2^*, \ldots, \theta_K^*$ of group K is randomly generated, and K regression trees $(\mathrm{x}, \theta_\mathrm{i})$ are generated.

Without pruning the learning subgroup, m features are randomly selected from the M features of the node in the tree as the attribute of the node, and then separated from the node until the minimum purity of the node attribute is separated.

$$\omega_\mathrm{i}(\mathrm{x}, \theta) = \frac{1\{\mathrm{x_i} \in \mathrm{R_l}(\mathrm{x},\theta)\}}{\#\{\mathrm{j}:\mathrm{x_j} \in \mathrm{R_l}(\mathrm{x},\theta)\}} \tag{3}$$

$$\mu(\mathrm{x}) = \sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} \omega_\mathrm{i}(\mathrm{x}, \theta \mathrm{Y_i}) \tag{4}$$

ω_i represents weight. $\mathrm{Y_i}$ represents the dependent variable observation value.

The weight value of each observation value can be obtained by taking the average value of the DT weight $\omega_\mathrm{i}(\mathrm{x}, \theta_\mathrm{t})$:

$$\omega_\mathrm{i}(\mathrm{x}) = \frac{1}{\mathrm{k}} \sum\nolimits_{\mathrm{i}=1}^{\mathrm{k}} \omega_\mathrm{i}\,(\mathrm{x}, \theta_\mathrm{t})\mathrm{Y} \tag{5}$$

The final prediction result is:

$$\mu(\mathrm{x}) = \sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} \omega_\mathrm{i}(\mathrm{x}, \mathrm{Y_i}) \tag{6}$$

The advantages of RF classification algorithm are fast driving speed, simple algorithm and stable effect. The network only needs to be trained once to obtain ideal results. Therefore, it is necessary to identify and define an appropriate quantity of DT in advance, rather than merge repetitive tasks into the learning model to determine the quantity of DT. The original data set X is resampled to randomly generate K training samples and create K regression trees.

4. Assessment Model of Water Pollution Degree in Qingshui River Basin

4.1. Evaluation Criteria

The water quality evaluation standard in this study was based on the Groundwater Quality Standard, taking the grading standard of water quality factor content as the index value, and taking an upper limit on the random uniformity of the marginal function to generate training samples. There was no upper limit requirement for the water quality V standard. Considering the limitation of the measured factor content, the water quality V standard was selected with the upper limit of 2 times of the water quality IV standard, and the output values of the training samples at each level were 1, 2, 3, 4, and 5 respectively.

4.2. Construction of Evaluation Model

According to the above indicators, a model based on stochastic forest algorithm can be constructed to estimate the water pollution level of the Qingshui River basin.

(1) Sample data set establishment

According to the evaluation threshold level of each site, 100 sample sets were generated and randomly interpolated in 500 sample sets. After that, the evaluation levels and control points were listed as data sets as typical data sets.

(2) Sample settings

When collecting data, 500 data sets were randomly selected as learning samples and the rest as test samples.

(3) Construction of RF model

A kit for evaluating water pollution has been developed using MATLABR2016A software. The sample evaluation index was used as the input vector, and the evaluation degree was used as the output vector.

(4) Setting of model parameters

After the DT was created, the attributes were randomly selected, the branches were set and the decomposition was performed again. The quantity of DT and branches had a great impact on the results of the model. A repeated calculation model was created, with ntree increasing from 10 to 100 and mtry increasing from 1 to 10. The standard deviation was selected as the error evaluation standard, and the best value with the minimum standard deviation was selected.

(5) Threshold setting

According to step (4), the quantity of branches of the optimal DT was selected, and the five estimated level results of the test sample were set as the threshold.

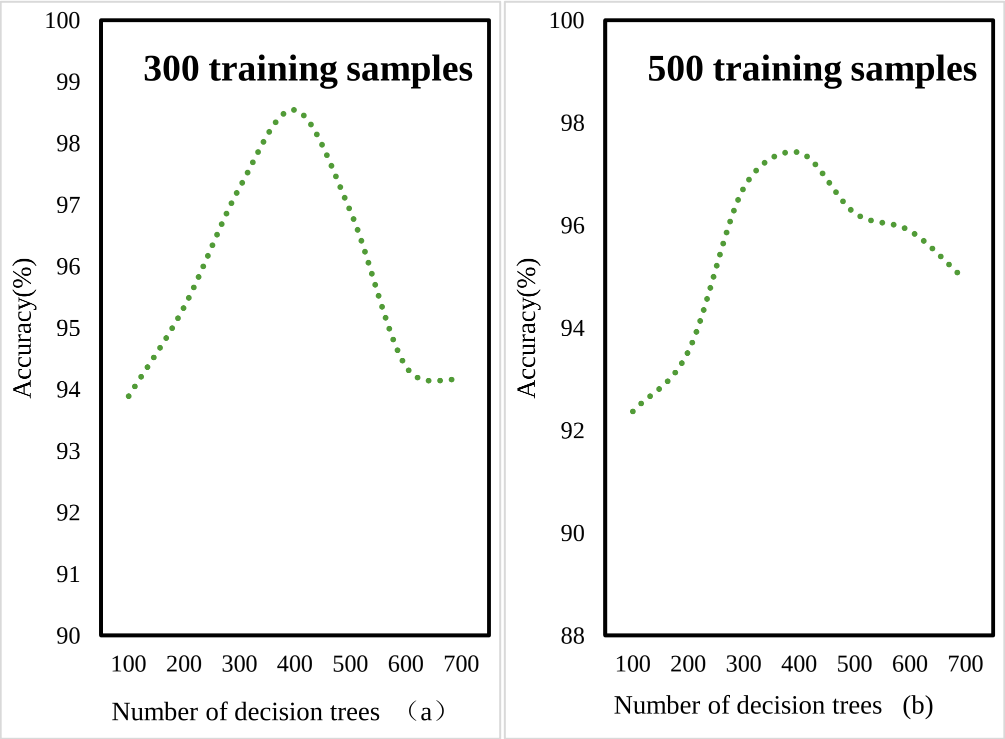
(6) Model result output

According to the new threshold, the output result of the test sample was compared with the threshold, and the final evaluation result was obtained.

5. Assessment Experiment of Water Pollution Degree in Qingshui River Basin

5.1. Determination of the Quantity of RF DT

For data with the same complexity and sample quantity, the quantity of DT to be generated was basically the same, so the quantity of trees can be determined in advance by training samples. 300 groups of training samples and 500 groups of training samples were selected for sample training. The results of sample training are shown in Figure 3.



(a) Sample training operation results for 300 sets of training samples
(b) Sample training operation results for 500 sets of training samples

Figure 3. Sample training operation results

Figure 3 (a) shows the sample training operation results for 300 sets of training samples, and Figure 3 (b) shows the sample training operation results for 500 sets of training samples.

According to Figure 3 (a), when the quantity of training samples was 300, with the increase of the quantity of DT, the classification accuracy of RF showed a trend of increasing first and then decreasing. When the quantity of DT was 100, the sample training accuracy was 93.89%; when the quantity of DT was 300, the sample training accuracy was 97.28%; when the quantity of DT was 400, the sample training accuracy was 98.54%. When the quantity of DT was 400, the sample training accuracy rate was the highest. When the quantity of DT exceeded 400, the sample training accuracy rate would gradually decrease.

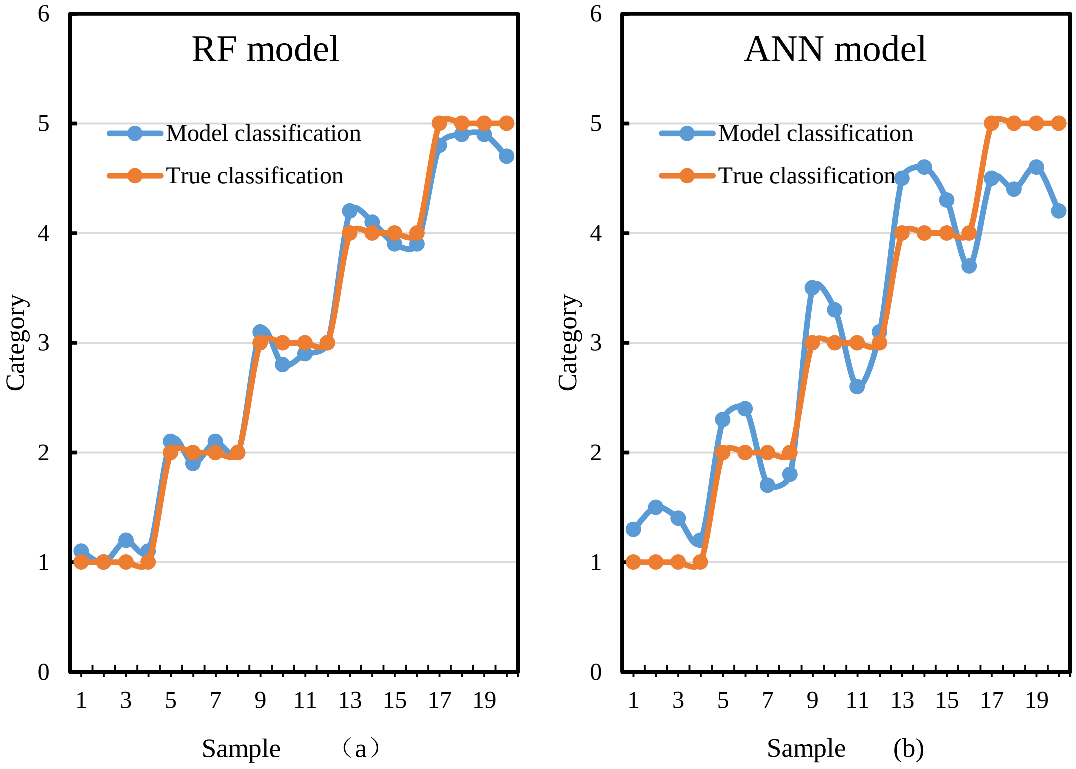
According to Figure 3 (b), when the quantity of training samples was 500, with the increase of the quantity of DT, the classification accuracy of RF showed a trend of increasing first and then decreasing. When the quantity of DT was 100, the accuracy of sample training was 92.37%; when the quantity of DT was 300, the accuracy of sample training was 96.73%; when the quantity of DT was 400, the sample training accuracy was 97.42%. When the quantity of DT was 400, the sample training accuracy rate was the highest. When the quantity of DT exceeded 400, the sample training accuracy rate would gradually decrease.

It can be seen from the analysis of the data that if the quantity of DT was determined as 400, a higher sample classification accuracy can be obtained.

5.2. Classification Accuracy

In this study, in order to reduce the random impact of the quantity of DT, 100 RF classification models were generated after determining the quantity of DT. According to the RF classification categories, the average value of the water quality classification results was used as the classification results.

This paper used the RF classification model and ANN model to classify water quality, and compared them with the real water quality types. The results of water quality classification are shown in Figure 4.



(a) Water quality classification results of stochastic forest classification model
(b) ANN model classification model water quality classification results

Figure 4. Water quality classification results

Figure 4 (a) shows the water quality classification results of the stochastic forest classification model, and Figure 4 (b) shows the water quality classification results of the ANN model classification model.

According to Figure 4 (a), when the stochastic forest classification model was used for water quality classification, the average values of the water quality classification results for four Class 1 water were 1.1, 1, 1.2 and 1.1 respectively. For 4 water samples of Class 2, the average values of water quality classification results were 2.1, 1.9, 2.1 and 2 respectively. For 4 samples of Class 3 water, the average values of water quality classification results were 3.1, 2.8, 2.9 and 3 respectively. For 4 samples of Class 4 water, the average water quality classification results were 4.2, 4.1, 3.9 and 3.9 respectively. For 4 samples of Class 5 water, the average water quality classification results were 4.8, 4.9, 4.9 and 4.7 respectively. It can be seen from the comparison data that the water quality classification results of the RF classification model were close to the real results.

According to Figure 4 (b),when the ANN model was used to classify the water quality, the average values of the water quality classification results were 1.3, 1.5, 1.4 and 1.2 respectively for four parts of Class 1 water. For 4 water samples of Class 2, the average values of water quality classification results were 2.3, 2.4, 1.7 and 1.8 respectively. For 4 samples of Class 3 water, the average water quality classification results were 3.5, 3.3, 2.6 and 3.1 respectively. For four water samples of Class 4, the average values of water quality classification results were 4.5, 4.6, 4.3 and

3.7 respectively. For 4 samples of Class 5 water, the average water quality classification results were 4.5, 4.4, 4.6 and 4.2 respectively. It can be seen from the comparison data that the water quality classification result of the ANN model was far from the real result.

The accuracy of water quality classification of the two classification models was calculated. The accuracy of water quality classification of the stochastic forest classification model was 97.33%, and the accuracy of water quality classification of the ANN model was 89.74%. The stochastic forest classification model had a higher accuracy of water quality classification.

5.3. Water Pollution Degree of Qingshui River Basin

The water quality in the Qingshui River basin was classified by using the stochastic forest algorithm model, and the water quality was evaluated at three water quality monitoring stations, namely, Wailaimiao Station, Zhangjiakou Station and Gaojiatun Station. The water quality results of Beilaimiao Station were classified as Class I, Zhangjiakou Station as Class III, and Gaozhatun Station as Class IV. In view of the good water quality of Wallamio Station and the serious water pollution of Gaojiatun Station, it was found that the water pollution mainly occurs in the lower reaches of Qinghai River.

6. Conclusion

In this paper, the stochastic forest algorithm was used to analyze the degree of water pollution in the Qingshui River basin. In this paper, the RF was constructed firstly, and then the evaluation standard of water pollution degree was determined. After that, the sample data set was established, and the RF model was constructed. The parameters and thresholds of the model were set, and the evaluation model was finally obtained. In this paper, the quantity of DT in RF was determined by sample training. The classification accuracy of the RF assessment model was verified by experiments, and the water quality of the Qingshui River basin was classified. The research showed that the water quality in the upper reaches of Qingshui River is better, and the water pollution in the lower reaches is more serious as it is closer to the lower reaches. The water quality pollution in Qinghai River is mainly concentrated in the lower reaches.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Evans, Alexandra EV. "Agricultural water pollution: key knowledge gaps and research needs." Current opinion in environmental sustainability 36.2 (2019): 20-27.

[2] Feng, Xiaofang. "Review MXenes as a new type of nanomaterial for environmental applications in the photocatalytic degradation of water pollutants." Ceramics International 47.6 (2020): 7321-7343.

[3] Ji, Mengzhi. "Bacteriophages in water pollution control: Advantages and limitations." Frontiers of Environmental Science & Engineering 15.84 (2020): 1-15.

[4] Sharma, Rohit. "Water pollution examination through quality analysis of different rivers: a case study in India." Environment, Development and Sustainability 24.6 (2020): 7471-7492.

[5] Morin-Crini, Nadia. "Worldwide cases of water pollution by emerging contaminants: a review." Environmental Chemistry Letters 20.4 (2020): 2311-2338.

[6] Wikurendra, Edza Aria. "Water quality analysis of pucang river, sidoarjo regency to control water pollution." Environmental Quality Management 32.1 (2020): 133-144.

[7] Iqbal, Mudassir, Daxu Zhang, and Fazal E. Jalal. "Durability evaluation of GFRP rebars in harsh alkaline environment using optimized tree-based random forest model." Journal of Ocean Engineering and Science 7.6 (2020): 596-606.

[8] Lu, Shengfu. "Semi-supervised random forest regression model based on co-training and grouping with information entropy for evaluation of depression symptoms severity." Math. Biosci. Eng 18.4 (2020): 4586-4602.

[9] Sadler, J. M. "Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest." Journal of hydrology 559.4 (2018): 43-55..

[10] Wang, Jing. "Risk assessment of coronary heart disease based on cloud-random forest." Artificial Intelligence Review 56.1 (2020): 203-232.

[11] Dharumarajan, Subramanian, and Rajendra Hegde. "Digital mapping of soil texture classes using Random Forest classification algorithm." Soil Use and Management 38.1 (2020): 135-149.

[12] Mobley, William. "Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: a pilot study in southeast Texas." Natural Hazards and Earth System Sciences 21.2 (2020): 807-822.

[13] Alnahit, Ali O., Ashok K. Mishra, and Abdul A. Khan. "Stream water quality prediction using boosted regression tree and random forest models." Stochastic Environmental Research and Risk Assessment 36.9 (2020): 2661-2680.

[14] Liu, Yang. "Risk prediction and diagnosis of water seepage in operational shield tunnels based on random forest." Journal of Civil Engineering and Management 27.7 (2020): 539-552.

[15] Li, Xing, Zhiping Wen, and Huaizhi Su. "An approach using random forest intelligent algorithm to construct a monitoring model for dam safety." Engineering with Computers 37.6 (2020): 39-56.