Pak. j. life soc. Sci. (2024), 22(2): 9611-9622 E-ISSN: 2221-7630;P-ISSN: 1727-4915

Pakistan Journal of Life and Social Sciences

www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2024-22.2.00726

Water quality prediction based on High-Dimensional Dataset Integration Prediction Model

Jue Zhao[1, 2, a*], Vladimir Y. Mariano[1, b], Mideth Abisado[1, c]

[1]College of Computing and Information Technologies, National University, Manila, Philippines
[2]College of Computer Science, Hunan University of Technology and Business, Changsha, China

ARTICLE INFO

Received: Aug 19, 2024
Accepted: Oct 12, 2024

Keywords

High-dimensional Dataset
Water quality Prediction
MLP
SVR
XGboost
Cubist integrator
Machine learning

***Corresponding Author:**

yayagogo@126.com

ABSTRACT

River water quality has nonlinear and non-stationary characteristics, and the data set is huge and complex. To improve the accuracy of water quality prediction, a water quality prediction model based on ensemble machine learning technology was proposed, which mainly used multi-layer Perceptron (MLP), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost). Through dimensionality reduction and normalization of high-dimensional data sets, the key features extracted by different machine learning models are fused, and further optimized based on Cubist algorithm, the most accurate new prediction system for water quality prediction is developed. For this purpose, the water quality dataset of the Yellow River Basin from January 2021 to December 2022 based on 19 effective parameters is collected, and three important water quality indicators such as dissolved oxygen, ammonia nitrogen and water quality categories are selected to evaluate the model performance. The experimental results show that the performance of the proposed ensemble prediction model based on high-dimensional datasets is better than that of MLP, SVR and XGBoost models in R2, RMSE and MAE evaluation indicators. This study not only compares the performance of different models in water quality prediction, but also explores the significant improvement effect of the optimized Cubist algorithm on the prediction accuracy, which provides a valuable reference for the research of water quality prediction based on machine learning.

INTRODUCTION

Water is an important resource for human survival, and the quality of water quality directly affects human life and health and social and economic development. However, with the increasing level of social industrialization and urbanization, the water environment is polluted, leading to the deterioration of water quality and water disasters (Falconi et al., 2017; Peter et al., 2007) and pose a serious threat to human health and ecosystems (Vörösmarty et al., 2010). Water quality prediction is an important means of environmental protection and water resources management. Water quality prediction can also effectively help water ecosystem assessment and protection, and provide help and support for aquaculture, tourism, and other economic industries. In short, efficient, and accurate water quality prediction can not only help prevent and control water pollution and ensure the safety of drinking water, but also support the protection and restoration of the ecological environment and promote the sustainable use of water resources.

Traditional water quality prediction methods mainly use physical and chemical models. These models predict changes in water quality parameters by simulating physical and chemical processes in water bodies. Common models include 1D, 2D and 3D water quality models, such as CE-QUAL-W2, EFDC and MIKE series models, etc. These models can provide detailed water quality predictions by describing the dynamics of water flow, dissolved substances, temperature, and pollutants, etc., through systems of equations. However, these models usually require a large amount of field data and expertise for calibration and validation and have high computational complexity. Water quality prediction data usually contains multiple dimensions, such as physical indicators, chemical indicators, biological indicators, etc. Each index can be subdivided into multiple sub-indexes, such as total nitrogen can be subdivided into ammonia nitrogen, nitrous acid nitrogen, nitric acid nitrogen, etc. The change of water quality is affected by many factors, such as meteorological conditions, human activities and so on. These factors may interact with each other and have complex nonlinear relationships. Moreover, water quality changes over time. For example, the pollution status of rivers, lakes and other water bodies will change over time (Xu, 2020). Even due to instrument failure, human error and other reasons, water quality data may have missing values, reducing data quality, and affecting the accuracy of water quality prediction. Therefore, it is of great technical research value to use computer science and technology to improve the analysis and modeling of water resources data, and try to use advanced artificial intelligence technology, especially machine learning and computer vision technology to deal with the problems of high dimensionality, nonlinearity, dynamics, complexity, and absence of water quality data for intelligent water quality prediction.

RELATED WORKS

In recent years, scholars both domestically and internationally have extensively explored water quality prediction. Various methods have been utilized, including the ARIMA model (Luo et al., 2020), stepwise clustering analysis (Chang et al., 2015), multiple regression coupling model (Luo and Luo, 2016), grey fuzzy Markov chain (Yu et al., 2014), Bayesian networks (Graham et al., 2019) and artificial neural networks (Alizadeh and Kavianpour 2015). These methods can be broadly categorized into traditional prediction methods and artificial intelligence prediction methods. While traditional methods like regression analysis and time prediction are popular due to their well-established theoretical foundation, simplicity in calculation, and ease of implementation, they fall short in handling complex nonlinear data relationships, resulting in limited prediction accuracy. On the other hand, artificial intelligence prediction methods such as support vector machine (Zuo et al., 2018) and artificial neural network (Cheng et al., 2021) effectively overcome the limitations of traditional approaches by dealing with nonlinear relationships more efficiently and significantly improving prediction accuracy.

MLP has shown great potential in the field of water quality prediction because of its powerful nonlinear mapping ability and self-learning ability. Yang Weilun et al. (2023) and Liu Yanxin (2022) combined MLP with algorithms such as linear regression and PSO to improve the accuracy and stability of water quality prediction by fusing the advantages of different algorithms (Yang et al., 2022; Liu et al., 2022). Among them, linear regression method is used to deal with linear relationship, while MLP is used to deal with nonlinear relationship, and the combination of the two can better fit the change law of water quality data. The PSO algorithm is used to optimize the network parameters of MLP and improve the generalization ability of the model. Zhai et al. (2022) used ESN for water quality prediction and improved the prediction accuracy by optimizing the parameters through grid search (Zhai et al., 2022). Bai Yun et al. (2020) combined VMD and LSSVR to improve the accuracy of river water quality prediction (Bai and Li, 2020). Xiao Rongping et al. (2020) proposed a multi-algorithm combination method for river water quality prediction, which improved the accuracy of prediction through the combination of grey prediction, generalized autoregressive conditional heteroscedasedas model and discrete wavelet transform (Xiao et al., 2020).

As a powerful machine learning algorithm, SVR has shown significant advantages in the field of water quality prediction. Xue Tonglai (2020) used Genetic Algorithm (GA) to optimize the parameters of SVR model to improve the generalization ability and prediction accuracy of the model in water quality prediction (Xue et al., 2020). Zhou Peijun (2020) combined PLS, GWO and SVR to construct a new water quality prediction model to improve the prediction performance and stability of the model (Zhou, 2020). Luo et al. (2020) combined the ARIMA model with SVR to form a combined prediction method, to make full use of the advantages of the two algorithms and improve the accuracy and stability of water quality prediction (Luo et al., 2020). Cao Wenzhi et al. (2023) combined Ensemble Empirical Mode Decomposition (EEMD), Long Short-Term Memory (LSTM) and SVR. A multi-scale and multi-variable water quality prediction model is constructed (Cao, et al., 2023).

Although the above methods improve water quality prediction accuracy by fitting nonlinear relationships between variables, the model's generalization ability and robustness require further improvement due to the complexity of multi-dimensional data from numerous monitoring sites with various parameters. These high-dimensional data sets contain a lot of redundant information and noise, which also brings great difficulties to the construction and training of water quality prediction models. Aiming at the above problems, this paper proposes a water quality prediction method based on high-dimensional data set ensemble prediction model. By using MLP, SVR, XGboost and other machine learning technologies, combined with Cubist integrator for model optimization, a water quality prediction method based on high-dimensional data set integrated prediction model is constructed. This integrated method can effectively deal with high-dimensional data sets, improve the generalization ability and prediction accuracy of the model, and provide more reliable technical support for environmental protection and water resources management.

THE RESEARCH METHOD

A. MLP

Multilayer Perceptron (MLP) is a classical artificial neural network model, which was proposed by Frank Rosenblatt in 1958 (Rosenblatt and Frank, 1958). Neural networks can use training data for fitting and test data for performance evaluation. Hosseini et al. (2022) proposed that backpropagation algorithm can improve the generalization ability of data training and reduces the prediction error, or the difference between the estimated output and the actual output (Hosseini et al., 2022). The basic structure of the MLP model is shown in FIGURE 1(PP, 2021).

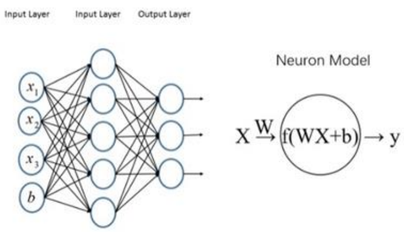

Figure.1 MLP Structure.

B. SVR

Support vector Machine regression (SVR) algorithm was first proposed by Vladimir Vapnik et al in 1995 (Vapnik and Vladimir, 1995). SVR algorithm can effectively deal with nonlinear data by finding the hyperplane that maximizes the margin for regression prediction. In 2001, Faruto et al. proposed a variety of kernel functions, which made SVR algorithm can be applied to high-dimensional data sets (Faruto, 2009). The SVR structure is shown in FIGURE 2 (dxw, 2019).

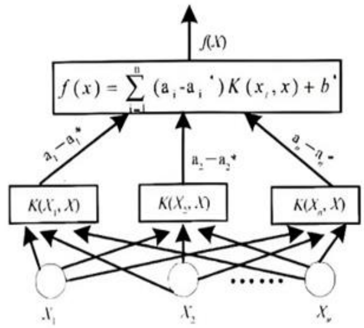

Figure.2 SVR Structure.

C. XGBOOST

In 2015, Chen et al. first proposed XGBoost model (Chen et al., 2015). XGBoost, short for "Extreme gradient boosting," is a tree-based machine learning algorithm that is known for its ability to handle complex datasets and its training efficiency. Multiple lifting trees can be generated in parallel at the same time in this technique (Nguyen et al., 2019). The XGBoost model can effectively solve many problems by using the gradient boosting technique (GB) (Duan et al., 2020). The general architecture of the XGBoost model is shown in FIGURE 3 (Zhao et al., 2023).

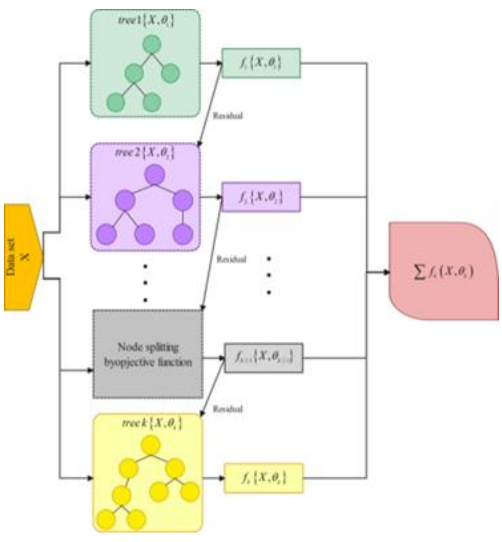

Figure.3 XGBOOST Structure.

D. CUBIST

Robert Quinlan developed Cubist in 2004 (Quinlan), creator of the C4.5 decision tree algorithm, as an extension of the M5 architecture tree (Quinlan, 1992). It is a rule-based machine learning algorithm for classification and regression tasks that builds decision trees by recursively partitioning data based on feature values. Cubist works well when dealing with noisy or high-dimensional regression problems. The algorithmic results of prediction models are more accurate than simple regression models and simpler than artificial neural networks.

DATASET

The data set used in this experiment is the surface water quality monitoring data of China from 2021 to 2023. The data came from the monitoring data of the water quality automatic monitoring station of the National surface water quality automatic monitoring network. The monitoring data mainly included water temperature, pH, dissolved oxygen, conductivity and turbidity, ammonia nitrogen, permanganate index, total nitrogen and total phosphorus, and some water stations measured total organic carbon, chlorophyll a, algae density, VOCs, biological toxicity, fecal coliform, and heavy metals. In this study, 41,464 surface water data of the Yellow River Basin from January 1, 2021 to December 12, 2022 were selected. The dataset is shown in FIGURE 4.

1	Provinces	City	river	River basin	Name of section	Monitoring time	Water quality class	Temperature of water	pH	pH category	Dissolved oxygen	Dissolved oxygen category	Potassium permanganate	Potassium permanganate category	Ammonia nitrogen	Ammonia nitrogen class	Total phosphorus	Total nitrogen	Electrical conductivity	Degree of turbidity	Chlorophyll	Density of algae	Site status
2	省份	城市	河流	流域	断面名称	监测时间	水质类别	水温	pH	pH类别	溶解氧	溶解氧类别	高锰酸钾	高锰酸钾	氨氮	氨氮类别	总磷	总氮	电导率	浊度	叶绿素	藻密度	站点情况
3	山西省	太原市	汾河	黄河流域	上兰	2021/1/1 0:00	I	0.2	8.34	-	10.99	-	1.51	-	0.142	-	0.005	2.92	904.4	4.1	-2	-2	正常
4	陕西省	宝鸡市	千河	黄河流域	千河公路桥	2021/1/1 0:00	II	1	8.19	-	11.57	-	1.67	-	0.138	-	0.025	5.96	564.9	22.6	-2	-2	正常
5	山东省	济南市	玉符河	黄河流域	卧虎山水库	2021/1/1 0:00	II	2.6	7.97	-	11.86	-	0.82	-	0.16	-	0.013	3.59	660	3.8	-2	-2	正常
6	陕西省	宝鸡市	渭河	黄河流域	卧龙寺桥	2021/1/1 0:00	II	2.3	8.04	-	11.59	-	1.07	-	0.253	-	0.048	7.4	918.7	10.8	-2	-2	正常
7	山西省	太原市	汾河	黄河流域	汾河水库出口	2021/1/1 0:00	I	2.7	8.34	-	12.38	-	1.57	-	0.025	-	0.005	1.66	724.8	1.5	0.006	2350865	正常
8	山东省	济南市	黄河	黄河流域	泺口	2021/1/1 0:00	II	1.8	8.45	-	14.26	-	3.46	-	0.136	-	0.038	3.83	853.9	382.9	-2	-2	正常
9	河南省	洛阳市	洛河	黄河流域	洛宁长水	2021/1/1 0:00	I	7	8.35	-	10.47	-	1.59	-	0.025	-	0.01	2.43	458.8	2.3	-2	-2	正常
10	山西省	太原市	汾河	黄河流域	温南社	2021/1/1 0:00	IV	1.8	7.85	-	10.39	-	5.17	-	1.089	-	0.17	7.72	1274.5	13.7	-2	-2	正常
11	河南省	洛阳市	伊河	黄河流域	潭头	2021/1/1 0:00	I	2.5	8.29	-	12.76	-	1.14	-	0.025	-	0.009	3.41	533.2	4.2	-2	-2	正常
12	河南省	洛阳市	洛河	黄河流域	白马寺	2021/1/1 0:00	III	6.8	8.34	-	12	-	4.6	-	0.437	-	0.073	5.99	835.5	13	-2	-2	正常
13	河南省	洛阳市	洛河	黄河流域	高崖寨	2021/1/1 0:00	II	4.9	8.33	-	12.02	-	1.72	-	0.072	-	0.047	3.58	438.9	11.6	-2	-2	正常
14	陕西省	宝鸡市	渭河	黄河流域	魏家堡	2021/1/1 0:00	III	4.5	8.3	-	11.46	-	1.16	-	0.623	-	0.103	5.13	807.9	16.2	-2	-2	正常
15	河南省	洛阳市	伊河	黄河流域	龙门大桥	2021/1/1 0:00	II	2.9	8.02	-	14.39	-	2.22	-	0.058	-	0.024	3.95	808.9	8.1	-2	-2	正常
16	山西省	太原市	汾河	黄河流域	上兰	2021/1/1 4:00	II	0.2	8.3	-	10.96	-	1.5	-	0.151	-	0.005	2.97	918.8	4	-2	-2	正常
17	陕西省	宝鸡市	千河	黄河流域	千河公路桥	2021/1/1 4:00	II	0.8	8.12	-	11.37	-	1.55	-	0.139	-	0.024	5.83	566.3	18.6	-2	-2	正常
18	山东省	济南市	玉符河	黄河流域	卧虎山水库	2021/1/1 4:00	I	2.1	8.01	-	12.17	-	0.86	-	0.149	-	0.012	3.68	662.1	3.8	-2	-2	正常
19	陕西省	宝鸡市	渭河	黄河流域	卧龙寺桥	2021/1/1 4:00	II	2	8.02	-	11.55	-	1.39	-	0.263	-	0.048	7.78	924	9.5	-2	-2	正常
20	山西省	太原市	汾河	黄河流域	汾河水库出口	2021/1/1 4:00	I	2.7	8.34	-	12.39	-	1.67	-	0.025	-	0.005	1.71	724.8	1.6	0.007	2453965	正常
21	山东省	济南市	黄河	黄河流域	泺口	2021/1/1 4:00	II	1.7	8.45	-	14.22	-	3.08	-	0.153	-	0.024	3.33	857.9	391.1	-2	-2	正常
22	河南省	洛阳市	洛河	黄河流域	洛宁长水	2021/1/1 4:00	I	6.9	8.35	-	10.46	-	1.69	-	0.025	-	0.01	2.3	458	3	-2	-2	正常
23	山西省	太原市	汾河	黄河流域	温南社	2021/1/1 4:00	IV	0.8	7.8	-	10.11	-	5.38	-	1.228	-	0.191	7.89	1286.4	14.5	-2	-2	正常
24	河南省	洛阳市	伊河	黄河流域	潭头	2021/1/1 4:00	I	2	8.1	-	11.72	-	1.16	-	0.025	-	0.009	3.4	542.5	3.8	-2	-2	正常
25	河南省	洛阳市	洛河	黄河流域	白马寺	2021/1/1 4:00	III	6.5	8.32	-	11.96	-	4.65	-	0.413	-	0.077	5.98	841.4	13.4	-2	-2	正常
26	河南省	洛阳市	洛河	黄河流域	高崖寨	2021/1/1 4:00	II	4.7	8.32	-	11.9	-	1.64	-	0.083	-	0.043	3.56	438.7	11.4	-2	-2	正常
27	陕西省	宝鸡市	渭河	黄河流域	魏家堡	2021/1/1 4:00	III	4.2	8.27	-	11.24	-	1.15	-	0.599	-	0.098	5.15	800.5	13.6	-2	-2	正常
28	河南省	洛阳市	伊河	黄河流域	龙门大桥	2021/1/1 4:00	II	2.7	8.01	-	14.08	-	2.38	-	0.073	-	0.023	3.99	799.3	7.2	-2	-2	正常
29	山西省	太原市	汾河	黄河流域	上兰	2021/1/1 8:00	II	0.2	8.28	-	11.05	-	2.02	-	0.149	-	0.005	3.01	926.5	3.5	-2	-2	正常

Yellow River Water quality data

Figure.4 Screenshot of the Yellow River water quality dataset.

EXPERIMENTAL DESIGN

A. THE WORKFLOW OF EXPERIMENTAL

The purpose of this study is to improve the accuracy and generalization ability of water quality prediction. In recent years, using machine learning methods to predict ecological environments such as water quality has become a research hotspot. This experiment aims to study and compare the performance of MLP, SVR, and XGBoost models in water quality prediction and design respective optimization algorithms as inputs to the Cubist algorithm to further improve the accuracy of water quality prediction. The experimental process is described as follows:

1. Data preprocessing: Collect historical water quality data, including water temperature, pH value, dissolved oxygen, ammonia nitrogen, and total phosphorus. Clean the data, handle missing values and outliers. Missing values are filled in using interpolation or mean filling, and outliers are removed or corrected through statistical analysis methods. Normalize or standardize the data to eliminate the influence of different feature scales on model training. Divide the data into a training set and a test set according to an 8:2 ratio. Among them, the training set has 33171 data points, and the remaining 20% of the data (8292 data points) are used for testing.

2. MLP model construction: To select the optimal model for predicting water quality, this study tries to design and compare several MLP models with different learning algorithms, transfer functions, and hidden neurons. The training process uses cross-validation method to adjust the learning rate, batch size, etc.

3. SVR model construction: In this study, we selected linear kernel, RBF kernel, polynomial kernel, and Sigmoid kernel functions to try and develop several SVR models. We set appropriate regularization parameters C and penalty parameters ε, used appropriate solvers, trained the models, and then selected the optimal model from them.

4. XGBoost model construction: We used the XGBoost library to try and build linear regression models, gradient boosting tree models, and distributed random tree models, setting different learning rates, tree numbers, and maximum depths for training. Then we selected the optimal model from them.

5. Cubist model integration optimization: In this study, the Cubist algorithm will serve as a secondary optimizer for the MLP, SVR, and XGBoost models to process high-dimensional water quality data to further improve prediction accuracy. Train the parameters of three models as inputs for the Cubist algorithm, and use regression tree-based rule generation methods, cost complexity pruning, and information gain pruning to optimize the parameters of the Cubist algorithm and ensure the prediction effect of the model.

6. Model evaluation: We used R^2, RMSE, and MAE as three statistical indicators to evaluate the model performance and analyze the accuracy of the prediction level.

7. Results Analysis: Through the experimental process described above, compare the performance of the three models in water quality prediction and analyze their respective advantages and disadvantages. Explain the effect of optimizing the Cubist algorithm on improving water quality prediction accuracy. Provide reference for further research on machine learning-based water quality prediction.

B. The Result of Experimental

Since dissolved oxygen, ammonia nitrogen concentration and water quality category can best reflect the water quality condition, this paper mainly selects three indicators of dissolved oxygen, ammonia nitrogen and water quality category for prediction. Through experiments, MLP, SVR, XGBoost three models have discovered better prediction models and achieved good prediction effects. After the second optimization of the integrated Cubist model, the effect is even better.

1.MLP model

In this paper, four MLP models (as shown in TABLE 1) were successively designed for experiments, and it was finally found that the MLP-S model had the best effect in predicting ammonia nitrogen, dissolved oxygen and water quality types, as shown in FIGURE 5.

TABLE 1: MLP model

MLP-S	MLP-D	MLP-B	MLP-E
Model architecture: Three hidden layers with 100, 100, 50 neurons, respectively Activation function: ReLU Maximum number of iterations: 500 Optimization algorithm:	Model architecture: Three hidden layers with 100, 100, 50 neurons, respectively A Dropout layer is added after each hidden layer with a drop rate of 0.5 Activation function: ReLU Optimization algorithm: Adam Optimizer	Model architecture: Three hidden layers with 100, 100, 50 neurons, respectively A Batch Normalization layer is added after each hidden layer Activation function: ReLU Optimization algorithm: Adam Optimizer	Model architecture: Three hidden layers with 100, 100, 50 neurons, respectively Activation function: ReLU Optimization algorithm: Adam Optimizer

Stochastic Gradient Descent (SGD)			

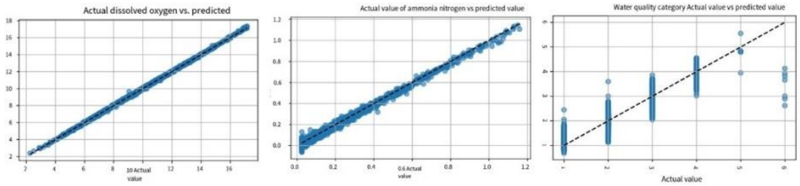

Figure.5 Results of MLP-S model prediction of ammonia nitrogen, dissolved oxygen, and water quality type.

2.SVR model

In this paper, four SVR models are successively designed for experiments as shown in TABLE 2. Finally, it is found that the SVR-R model has the best effect on predicting ammonia nitrogen, dissolved oxygen and water quality categories, as shown in FIGURE 6.

TABLE 2: SVR model

SVR-L	SVR-R	SVR-P	SVR-S
Kernel function expression: $K(x, x') = x \cdot x'$ Optimization algorithm: The LIBSVM solver is used Model parameters: The regularization parameter C: 1.0 The penalty parameter ε : 0.1	Kernel function expression: $K(x, x') = \exp(-\gamma \|x - x'\|^2)$ Optimization algorithm: The LIBSVM solver is used Model parameters: The regularization parameter C: 1.0 The penalty parameter ε : 0.1 γ (kernel width parameter) : 1 / (number of features)	Kernel function expression: $K(x, x') = (x \cdot x' + r)^d$ Optimization algorithm: The LIBSVM solver is used Model parameters: The regularization parameter C: 1.0 The penalty parameter ε : 0.1 The order of the polynomial is d: 3 r (offset) : 0.0	Kernel function expression: $K(x, x') = \tanh(\gamma x \cdot x' + r)$ Optimization algorithm: The LIBSVM solver is used Model parameters: The regularization parameter C: 1.0 Penalty parameter ε : 0.1γ (kernel width parameter) : 0.1 r (offset) : 0.0

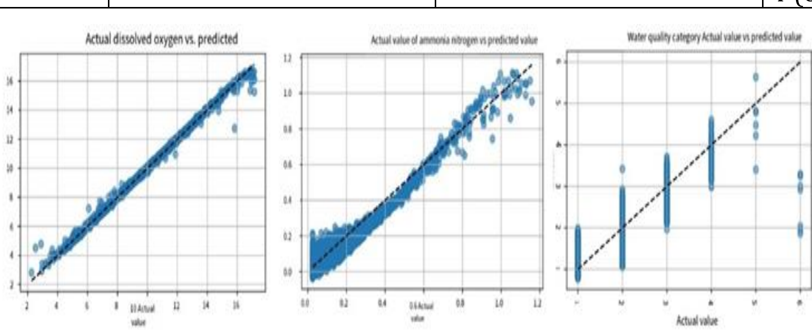

Figure.6 Results of SVR-R model prediction of ammonia nitrogen, dissolved oxygen, and water quality type.

3.The XGoost model

In this paper, four XGBoost models are designed for experiments as shown in TABLE 3. Finally, it is found that the XGBoost-S model has the best effect on predicting ammonia nitrogen, dissolved oxygen and water quality categories, as shown in FIGURE 7.

TABLE 3: XGBOOST model

XGBoost-S	**XGBoost-GBT**	**XGBoost-GBL**	**XGBoost-D**
Multiple decision trees are constructed and weighted voting is performed by the lifting algorithm. Objective function: Squared error	gbtree (Tree-based lifting method) : Multiple decision trees are built and weighted voting is performed by a lifting algorithm. Objective function: Squared error Model parameters: booster: gbtree	gblinear (Lifting Based on Linear Models) : This method builds multiple linear models and uses a lifting algorithm for weighted voting. Objective function: Squared error Model parameters: booster: gblinear	dart (Decision Tree-based Dropout Boosting) : Randomly drops some trees in each boosting round to reduce overfitting. Objective function: Squared error Model parameters: booster: dart

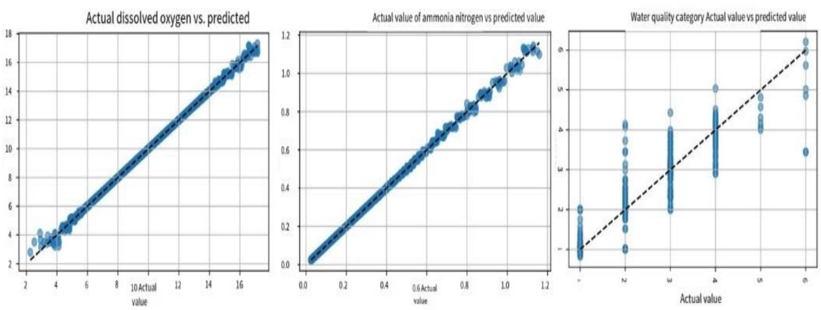

Figure.7 Results of XGboost-s model prediction of ammonia nitrogen, dissolved oxygen, and water quality type.

4.Cubist model optimization

The architecture shown in FIGURE 8 was used to create the Integration prediction model for calculating the water quality prediction. The optimal prediction results of MLP, SVR, and XGBoost models are used as input features of the Cubist algorithm to construct a new training set. Depth, etc. Cubist employs a regression tree-based rule generation approach using cost-complexity pruning and information gain pruning to avoid overfitting. Through verification, the prediction effects of Cubist model on ammonia nitrogen, dissolved oxygen and water quality types are shown in FIGURE 9-11.

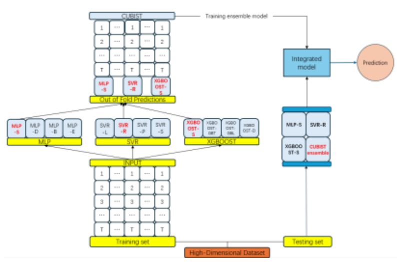

Figure.8 Flowchart of the Integration prediction approach.

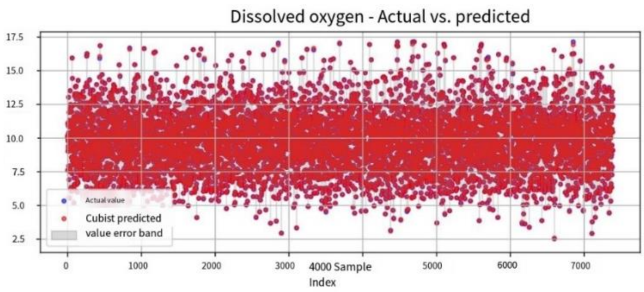

Figure.9 Results of Cubist model prediction of dissolved oxygen.

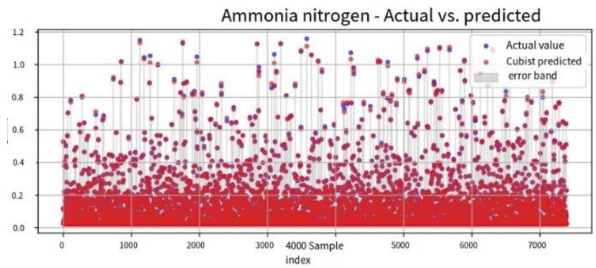

Figure.10 Results of Cubist model prediction of ammonia nitrogen.

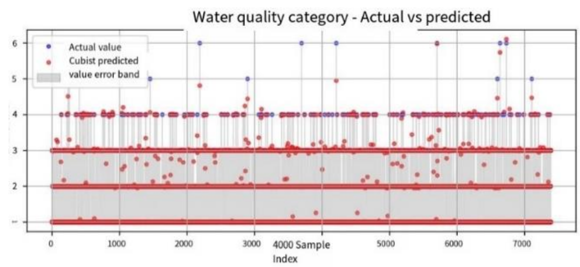

Figure.11 Results of Cubist model prediction of water quality type.

5. RESULT ANALYSIS

After establishing the prediction model, R^2, RMSE and MAE are used to evaluate the accuracy and performance of the model. The comparison results of the analysis indicators of different prediction models are shown in Table 4.

TABLE 4: Comparison of analysis indicators of different prediction models

MODEL	TARGET	R2	RMSE	MAE
MLP-S	Dissolved Oxygen	0.999586	0.04329	0.031562
SVR-R	Dissolved Oxygen	0.998124	0.092136	0.048288
XGBoost-S	Dissolved Oxygen	0.999992	0.005945	0.004366

Cubist	Dissolved Oxygen	0.999993	0.005743	0.004081
MLP-S	Water Quality Category	0.877954	0.254337	0.174255
SVR-R	Water Quality Category	0.859363	0.273023	0.171231
XGBoost-S	Water Quality Category	0.990138	0.072297	0.01034
Cubist	Water Quality Category	0.991561	0.06688	0.00447
MLP-S	Ammonia Nitrogen	0.99065	0.014366	0.01058
SVR-R	Ammonia Nitrogen	0.915097	0.043291	0.035219
XGBoost-S	Ammonia Nitrogen	0.999984	0.000586	0.000263
Cubist	Ammonia Nitrogen	0.999984	0.000586	0.000262

According to the experimental results, Cubist model is better than other models because:

(1) Cubist can fuse information from different machine learning models to obtain more comprehensive information and improve prediction accuracy. In the Cubist model, taking the trained parameters of MLP, SVR, and XGBoost models as input is equivalent to fusing the prediction results from different models, which can make up for the shortcomings of a single model and improve the robustness and accuracy of the prediction.

(2) MLP, SVR, and XGBoost models automatically perform feature extraction during the training process, and extract key features related to water quality prediction from the original data. Using these extracted features as input to Cubist can avoid Cubist to repeat feature extraction, improve efficiency, and possibly obtain a more effective feature representation. Cubist model takes the optimal models of MLP, SVR, and XGBoost as input, which can play their respective advantages and compensate for each other's shortcomings, to obtain better prediction results.

CONCLUSION

In the study of water quality prediction, we design and develop the optimal prediction models of MLP, SVR and XGBoost, and then aggregate the results and import them into the super Learner model Cubist to form a new integrated prediction model for high-dimensional data: HD-Cubist-Integration Learner model. This is a high-quality water quality prediction technology, which combines information fusion, feature extraction advantages, model complementary advantages and its own rule learning ability. It synthetically uses information and features from different models to learn more effective rules, to obtain better prediction results than a single model. The experimental results highlight its high accuracy and performance.

Funding

This work was supported in part by Teaching reform Research Project of Hunan Province under Grant HNJG-2022-0878; in part by Hunan Education Science Planning Project under Grant XJK21BXX003 ND210887.

REFERENCES

Alizadeh, M. J., and Kavianpour, M. R, "Development of wavelet GANN models to predict water quality parameters in Hilo Bay, Pacific Ocean." Marine Pollution Bulletin 60, no. 1-2 ,2015: 171-178.

Bai, Y., & Li, Y, "Prediction of river water quality based on VMD-LSSVR." Journal of Safety and Environment 03 ,2020: 1162-1168. DOI: 10.13637/j.issn.1009-6094.2019.0727.

Chen, Tianqi, et al., "Xgboost: Extreme Gradient Boosting." R Package, version 0.4-2, 1, 2015, pp. 1-4.

Chang, C., et al., "Prediction of Phytoplankton Growth Based on Stepwise Clustering Analysis." China Environmental Science 9.9 ,2015: 2805-2812.

Cheng, C., et al., "Personalized sleep seizure prediction based on deep neural network." Pattern Recognition and Artificial Intelligence 4, no. 4 ,2021: 333-342.

Cao, Wenzhi, et al., "Water Quality Prediction Model Based on EEMD-LSTM-SVR." Systems Engineering 04 ,2023: 1-12.

Duan, J., et al., "A Novel Artificial Intelligence Technique to Predict Compressive Strength of Recycled Aggregate Concrete Using ICA-XGBoost Model." Engineering Computations, 2020, pp. 1-18.

Falconi, T. M. A., et al., "Quantifying tap-to-household water quality deterioration in urban communities in Vellore, India: The impact of spatial assumptions." International Journal of Hygiene and Environmental Health 220.1,2017: 29-36.

Faruto, Y. L, LIBSVM-farutoUltimateVersion-a toolbox with implements for support vector machines based on libsvm. Software available at http://www.ilovematlab.cn, 2009.

Graham, S. E., et al., "Using Bayesian Networks to Predict Risk to Estuary Water Quality and Patterns of Benthic Environmental DNA." Integrated Environmental Assessment and Management 1.1 ,2019: 93-111.

Hosseini, Seyed, et al., "Application of Reliability-Based Back-Propagation Causality-Weighted Neural Networks to Estimate Air-Overpressure Due to Mine Blasting." Engineering Applications of Artificial Intelligence 115 ,2022: 105281.

Luo, Ke, et al., "Application of ARIMA-SVR Combination Method in Water Quality Prediction." Journal of Yangtze River Research Institute 10 ,2020: 21-27.

Luo, K., et al., "Application of ARIMA-SVR Combination Method in Water Quality Prediction." Journal of Yangtze River Research Institute 10.10 ,2020: 21-27.

Luo, H., and Luo, Y, "Application of Multiple Regression Coupled Model in Prediction of Cyanobacterial Biomass in Dianchi Lake." In Proceedings of the Annual Conference of

the Chinese Society of Environmental Sciences, edited by Chinese Society of Environmental Sciences, 1200-1207. Beijing: China Environment Press, 2016.
Liu, Y. X., Jia, J. H., Li, H. B., et al., "Effects of rainfall on ammonia nitrogen in Fuhe River using PSO-MLP model." Science, Technology and Engineering 32 ,2022: 14511-14517.
Lonely you go in, "MLP-Mixer: MLP is all you need?" PP flying paddle AI Studio. aistudio.baidu.com, 2021.
Nguyen, H., et al., "Developing an XGBoost Model to Predict Blast-Induced Peak Particle Velocity in an Open-Pit Mine: A Case Study." Acta Geophysica, vol. 67, 2019, pp. 477-490.
Peter, L. B., et al., "Comparison of Green Algal Bloom Intensity and Related Water Quality Parameters at Paired 'Bloom' and 'Non-Bloom' Sites." Journal of Phycology 43, December 2007: 33-34.
Quinlan, J. R, Data Mining Tools See5 and C5.0. http://www.rulequest.com/see5-info.html.
Quinlan, J. R, "Learning with Continuous Classes." 5th Australian Joint Conference on Artificial Intelligence, World Scientific, 1992, pp. 343-348.
Rosenblatt, Frank, "The Perceptron: A Probabilistic Model for Information Storage and Organization." Psychological Review 65.6 ,1958: 386-388.
Vapnik, Vladimir N, The Nature of Statistical Learning Theory. Springer Science & Business Media, 1995.
Vörösmarty, C. J., et al., "Global Threats to Human Water Security and River Biodiversity." Nature 467.7315,2010: 555-561.
Xiao, R. P., Ge, H. B., & Sun, X, "River water quality prediction based on multi-algorithm combination." Water Conservancy Technical Supervision 02 ,2020: 136-139.
Xue, T. L., Zhao, D. H., & Han, F, "Research on SVR water quality prediction model based on GA optimization." Environmental Engineering 03 ,2020: 123-127. DOI: 10.13205/j.hjgc.202003021.
Xu, Z. B, "Research and Realization of Water Quality Prediction System Based on Hybrid Optimized BP Neural Network." Master's Dissertation, Beijing University of Technology, 2020.
Yu, H., et al., "Application of Grey Fuzzy Markov Chain to Predict the Change Trend of Haihe River Water Quality." China Environmental Science 3.8 ,2014: 810-816.
Yang, W. L., Gao, Y. X., & Cao, L, "Linear regression method and MLP combined to predict the comprehensive water quality index of Lijiahe Reservoir." Shaanxi Water Resources 06 ,202): 19-21+25.
Zhai, X. A., et al., "Water quality prediction method based on echo state network." Journal of Hebei Normal University of Science and Technology 02 ,2022: 82-88.
Zhou, P J, Construction and Application of Water Quality Prediction Model Based on PLS-GWO-SVR. Master's thesis, Yanshan University, 2020.
Zhao, J., et al., "Super Learner Ensemble Model: A Novel Approach for Predicting Monthly Copper Price in Future." Resources Policy, vol. 85, 2023, p. 103903.
Zuo, J., et al., "Haze prediction method based on Co-evolutionary artificial fish swarm algorithm and SVM." Pattern Recognition and Artificial Intelligence 8, no. 8 ,2018: 725-739.
51dxw, "Design of Data Prediction Module Based on SVR Model." 51dxw, 2019. https://www.51dzw.com/embed/embed_128812.html.