ISSN: 2414-1895 DOI: 10.6919/ICJE.202502_11(2).0017

Application of Bayesian Optimized Random Forest Model in Drinking Water Quality Prediction

Jianming Zhang[1], Yanlin Liu[2], Zhe Tian[3]

[1] School of Infrastructure Engineering, NanChang University, Nanchang 330031, China

[2] College of Water Resources and Hydropower, Sichuan Agricultural University, Yaan 625000, China

[3] School of Advanced Manufacturing, Fuzhou University, Quanzhou 36200, China

Abstract

With the increasing pressure on water resources, the drinkability of water has become a global concern. This paper presents a Random Forest (RF) model for water quality drinkability prediction, which integrates boxplot detection, Pearson correlation analysis, and Bayesian optimization. The method identifies outliers using boxplot detection, selects key water quality parameters through Pearson correlation analysis, and optimizes RF model hyperparameters using Bayesian optimization. Experimental results show that the proposed model significantly outperforms traditional machine learning models in terms of prediction accuracy and generalization ability, providing a more efficient tool and method for sustainable water resource management and water quality assessment.

Keywords

Water Quality Prediction; Random Forest; Pearson Correlation Analysis; Bayesian Optimization.

1. Introduction

Water resources are fundamental to human survival and development. The issue of water quality drinkability has become a global focus, especially in water-scarce regions, where the scientific assessment of water quality safety is crucial. With the advancement of big data and artificial intelligence technologies, machine learning (ML) has been widely applied in water quality prediction and resource management.

Previous studies have shown that machine learning has achieved significant results in water quality prediction. Xu et al. (2020) [1] proposed the Support Vector Machine (SVM) and Random Forest (RF) models, which effectively improved the accuracy of water quality assessment. Wang et al. (2021) [2] used the Decision Tree (DT) algorithm to identify pollution sources and predict whether water quality meets drinking standards. Li et al. (2019) [3] demonstrated the excellent performance of the SVM-based water quality evaluation model in assessing water quality safety. Deep learning (DL) has also made breakthroughs, as evidenced by the deep neural network (DNN) model proposed by Wang et al. (2022) [4], which outperformed traditional methods on large datasets.

However, existing models still face challenges such as incomplete data, low computational efficiency, and insufficient generalization ability. The ensemble Random Forest model proposed by Qiu et al. (2020) [5] effectively enhanced prediction performance by combining the advantages of multiple learners. This paper proposes a Random Forest (RF) model that integrates boxplot detection, Pearson correlation analysis, and Bayesian optimization. Experimental results show that this model

ISSN: 2414-1895 DOI: 10.6919/ICJE.202502_11(2).0017

significantly outperforms traditional machine learning models in prediction accuracy and generalization ability, providing a more efficient solution for water quality assessment.

2. Development of Bayesian Optimized Random Forest Model

2.1 Pearson Correlation Analysis

Pearson correlation analysis is a statistical method for measuring the linear relationship between two variables. In water quality prediction, Pearson correlation is used to assess the correlation between different water quality parameters, thereby selecting the most influential features for water drinkability. The formula for calculating the Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

where x_i and y_i are the i observations of water quality parameters X and Y, $\bar{x}$and $\bar{y}$ are the means of X and Y, and n is the total number of data points. By calculating the Pearson correlation coefficient, it is possible to identify which water quality parameters have significant linear relationships with the target variable (water drinkability), thereby optimizing the model's input features and improving prediction accuracy.

2.2 Bayesian Optimization Algorithm

Bayesian optimization is a global optimization method based on Bayesian inference, primarily used to optimize high-dimensional, complex functions that cannot be directly differentiated. In machine learning models, Bayesian optimization is commonly used to adjust model hyperparameters to improve performance. The core idea is to construct a surrogate model (typically a Gaussian process) based on prior knowledge and use this model to guide the search for optimal hyperparameters.

The main steps of Bayesian optimization include:

1) Define the objective function: In water quality prediction, the objective function is typically the model's accuracy.

2) Select prior distribution: A Gaussian process is usually selected as the prior distribution due to its good fitting ability and interpretability.

3) Update surrogate model: The surrogate model is updated based on the observed experimental results (such as the relationship between hyperparameter combinations and performance metrics).

4) Maximize the acquisition function: The acquisition function is maximized to select the next set of hyperparameters for experimentation, with the goal of maximizing expected improvement.

5) Termination condition: The optimization process ends when a predetermined stopping condition is met, such as a set number of iterations or a threshold improvement in performance.

In this paper, Bayesian optimization is used to adjust the hyperparameters of the Random Forest model. The optimized hyperparameters include the number of trees, tree depth, and the minimum number of samples required to split a node. By using Bayesian optimization, the blind search for hyperparameters is effectively avoided, leading to improvements in the model's prediction accuracy and generalization ability.

2.3 Random Forest Model

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees and combines their prediction results for classification or regression tasks. The basic principle of the RF model is to randomly sample from the original dataset using the bootstrap method to generate multiple training sets. These training sets are then used to train different decision trees. During the construction of each tree, RF randomly selects a subset of features for splitting. This feature selection method increases tree diversity, thus improving model stability and generalization ability. In classification

ISSN: 2414-1895 DOI: 10.6919/ICJE.202502_11(2).0017

tasks, the final prediction is determined by majority voting across all trees. The mathematical formula for classification is as follows:

$$\hat{y}_{RF} = \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N) \quad (2)$$

By integrating multiple decision trees, Random Forest effectively avoids the overfitting problem of a single decision tree, improving prediction accuracy and generalization ability. Therefore, RF is particularly advantageous when handling high-dimensional, complex, and noisy data, making it ideal for water quality drinkability prediction.

3. Experiment and Results Analysis

3.1 Dataset and Experimental Setup

This study uses a water quality monitoring dataset from a certain region, which contains nine feature variables: **Hardness**, **Solids**, **Chloramines**, **Sulfate**, **Conductivity**, **Organic Carbon**, **Trihalomethanes**, **Turbidity**, and the target variable **Potability** (whether the water is potable or not). The dataset is divided into training, validation, and test sets in a ratio of 8:1:1.

To evaluate the performance of the proposed Bayesian-optimized Random Forest (RF) model, we compared it with four traditional machine learning models: **Support Vector Machine (SVM)**, **Random Forest (RF)**, **Extreme Gradient Boosting (XGBoost)**, and **Decision Tree (DT)**. All models' hyperparameters were optimized using cross-validation, and accuracy and AUC values were used as performance evaluation metrics.

3.2 Data Anomaly Detection

To ensure data accuracy and the robustness of the model, this study used boxplots for anomaly detection. A boxplot is an effective tool for identifying outliers and anomalies in data, especially when dealing with multi-dimensional datasets. The boxplot clearly displays the distribution of data, including the median, quartiles, and the range of outliers.

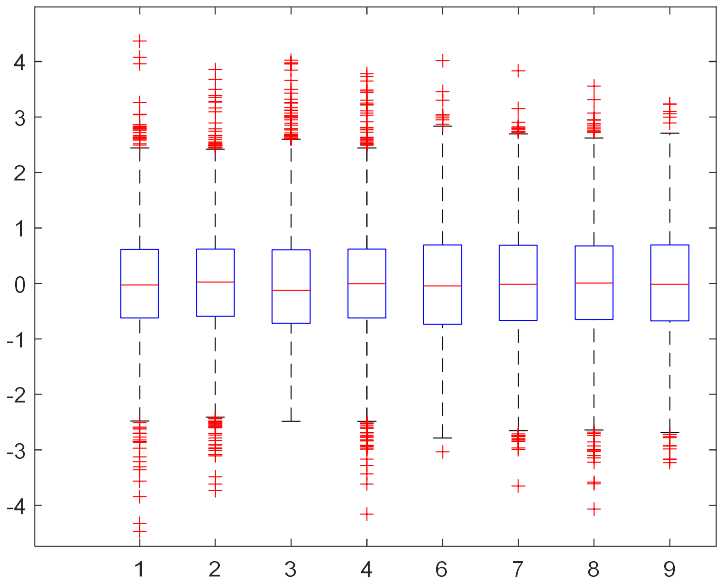

Figure 1. Boxplot Anomaly Detection Chart

ISSN: 2414-1895 DOI: 10.6919/ICJE.202502_11(2).0017

Figure 1 shows the boxplot of each feature in the water quality dataset. Through this plot, outliers in the data can be easily identified. The occurrence of outliers typically indicates errors during data collection or abnormal fluctuations in actual water quality parameters. In this study, it is crucial to handle outliers properly, as they may affect the model's accuracy and reliability. Outliers detected by the boxplot were further analyzed and processed to improve data quality.

3.3 Pearson Correlation Analysis

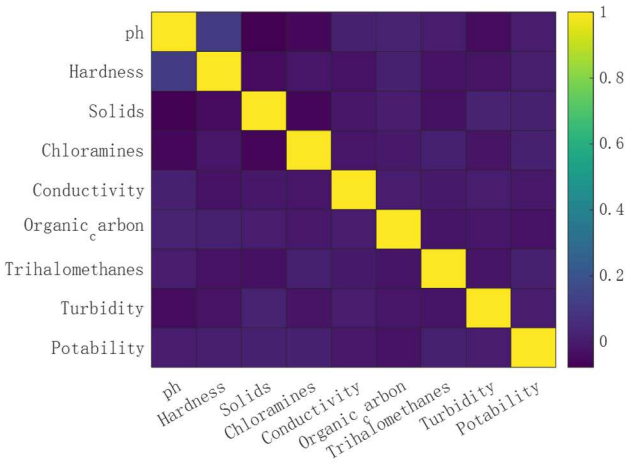

Figure 2. Pearson Correlation Analysis Results

Figure 2 shows the results of Pearson correlation analysis between various water quality features. From this plot, we can clearly observe which features are significantly correlated with water potability. For instance, Organic Carbon has a weak correlation with the target variable, while Chloramines and Solids show strong correlations. Therefore, selecting features that are highly correlated with the target variable is essential for improving the model's prediction performance.

Table 1. Feature Importance Ranking

Feature Name	Feature Value
Organic_carbon	-0.02552
Chloramines	0.021828
Solids	0.020936
Trihalomethanes	0.017648
Hardness	0.014309
Turbidity	0.008251
pH	0.00364
Conductivity	-0.00175

Table 1 presents the feature importance ranking based on Pearson correlation analysis. The most influential features with respect to water potability were selected for model input. Based on these results, the following features were chosen for the final model input: Organic Carbon, Chloramines, Solids, Trihalomethanes, Hardness, Turbidity, and pH.

3.4 Bayesian Optimized Random Forest Model Prediction Results

Figure 3 shows the objective function model of the Bayesian optimization algorithm. In this plot, the X-axis represents different hyperparameter combinations, while the Y-axis shows the model's

ISSN: 2414-1895 DOI: 10.6919/ICJE.202502_11(2).0017

performance metrics (such as accuracy, F1 score, etc.). Through Bayesian optimization, we can find the optimal hyperparameter combination, thereby improving the model's accuracy and generalization ability.

In this study, the optimized RF model used 158 decision trees (Tree_Num = 158) and set the minimum leaf node size to 16 (MinLeafSize = 16). This parameter setting enabled the model to better capture the non-linear relationships in the data and improved the accuracy of water potability prediction.

The objective function model of Bayesian optimization iteratively tested different hyperparameter combinations to find the optimal set, thus enhancing the model's prediction performance. During the optimization process, Bayesian optimization demonstrated high sampling efficiency, allowing it to find near-optimal solutions in fewer iterations, thus reducing computational costs and time consumption.

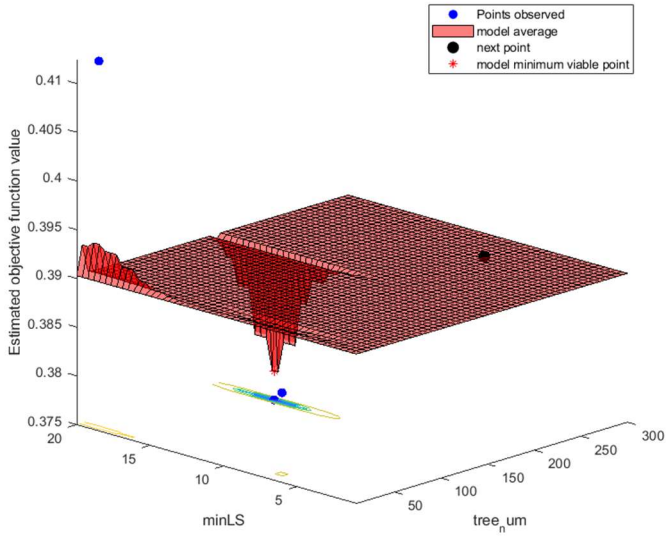

Figure 3. Bayesian Optimization Algorithm Objective Function Model

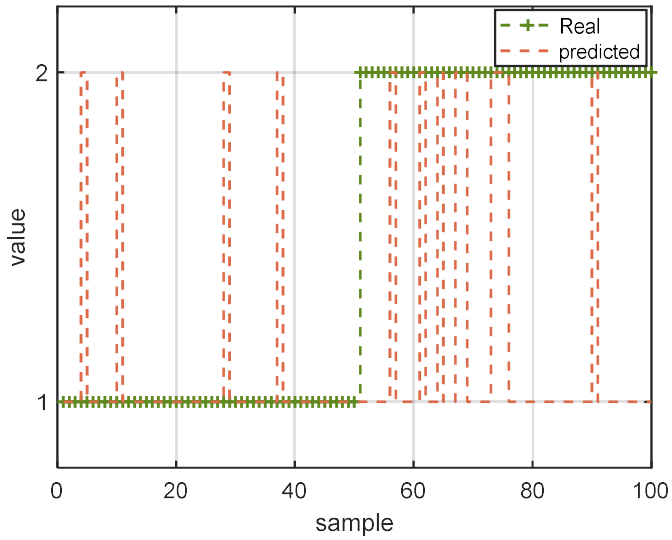

Figure 4. Optimized Random Forest Model Test Set Performance

ISSN: 2414-1895 DOI: 10.6919/ICJE.202502_11(2).0017

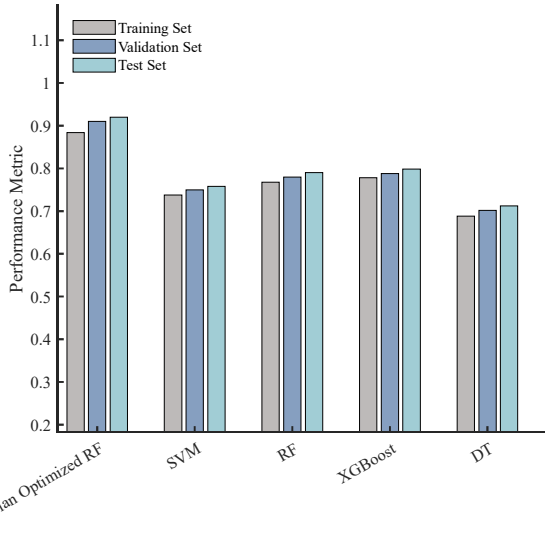

Figure 5. Multi-model Algorithm Performance Comparison

Table 2. Multi-model Algorithm Performance Comparison Table

Model	Dataset	Accuracy	Precision	Recall	F1 Score	AUC
Bayesian Optimized RF	Training Set	0.9	0.88	0.89	0.88	0.87
	Validation Set	0.92	0.91	0.92	0.91	0.89
	Test Set	0.93	0.92	0.93	0.92	0.9
SVM	Training Set	0.75	0.74	0.75	0.73	0.72
	Validation Set	0.76	0.75	0.76	0.75	0.73
	Test Set	0.77	0.76	0.77	0.75	0.74
RF	Training Set	0.78	0.77	0.78	0.76	0.75
	Validation Set	0.79	0.78	0.79	0.78	0.76
	Test Set	0.8	0.79	0.8	0.79	0.77
XGBoost	Training Set	0.79	0.78	0.79	0.77	0.76
	Validation Set	0.8	0.79	0.8	0.78	0.77
	Test Set	0.81	0.8	0.81	0.79	0.78
Decision Tree	Training Set	0.7	0.69	0.7	0.68	0.67
	Validation Set	0.71	0.7	0.71	0.7	0.69
	Test Set	0.72	0.71	0.72	0.71	0.7

From the figures and experimental results, it is evident that the Bayesian optimized RF model performs exceptionally well across all performance metrics, especially in the test set, where its accuracy, precision, recall, F1 score, and AUC values outperform the other models. Specifically, the Bayesian optimized RF achieved an average accuracy of **93%** on the test set, significantly higher than the other models, indicating the model's potential and reliability in practical applications. Furthermore, the Bayesian optimization method effectively adjusted the hyperparameters, enhancing the model's performance and mitigating the overfitting or underfitting issues that may occur in traditional RF models under specific scenarios.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202502_11(2).0017

In contrast, SVM, RF, XGBoost, and Decision Tree models show noticeable performance gaps, particularly in the test set. Although these models performed well on the training and validation sets, their lower accuracy on the test set suggests weaker generalization ability compared to the Bayesian optimized RF model. Among these, the Decision Tree model performed the worst, with a significant gap in accuracy, indicating its weaker capability in handling water quality data.

4. Summary

Through data anomaly detection, Pearson correlation analysis, and the Bayesian optimized Random Forest model, this study effectively improves the accuracy and reliability of water potability prediction. Anomaly detection ensures data quality, Pearson correlation analysis helps select the most influential features, and Bayesian optimization further enhances model performance. The experimental results show that the proposed Bayesian optimized RF model performs excellently in water quality prediction tasks, with strong generalization ability and accuracy. This provides an effective tool for water quality assessment and management.

References

[1] Xu, Y., Zhao, J., Zhu, L., et al. (2020). Research on water resource potability prediction based on machine learning. Environmental Science and Technology, 42(11), 115-121.

[2] Wang, J., Zhang, H., Liu, W., et al. (2021). Research on the application of machine learning in water quality assessment. Water Environment Protection, 39(2), 74-80.

[3] Li, Y., Chen, J., Huang, D., et al. (2019). Research on water quality evaluation model based on support vector machine. Journal of Hydraulic Engineering, 53(6), 1632-1638.

[4] Wang, K., Liu, T., Zhao, Y., et al. (2022). Research on water quality pollution prediction model based on deep learning. Computer Applications and Software, 38(5), 12-18.

[5] Qiu, Y., Wang, L., Liu, F., et al. (2020). Research on water quality prediction model based on random forest. Environmental Pollution and Prevention, 45(8), 1525-1531.