Research and Application of Bp Neural Network in Water Quality Testing

Lingxi Zeng

School of Computer Science, Wuhan University, Wuhan, Hubei, 430000, China
2020300004018@whu.edu.cn

Abstract. To ensure sustainable use of water resources and protect a robust ecological environment, developing effective and precise models for assessing water quality is essential. This research focuses on training detection models using water sample data collected in India. After preprocessing the data and constructing models using various methodologies, optimal model parameters were chosen by evaluating different hidden layers and neuron configurations to improve the model's learning capacity. Despite the suboptimal performance of the Back Propagation (BP) neural network with small-scale and weakly correlated data, parameter adjustments and suitable activation functions enhanced training effectiveness. Additionally, the training outcomes of alternative machine learning models on the same dataset were compared after training the BP neural network. The results demonstrate that gradient boosting trees exhibit superior performance under similar conditions, underscoring the critical importance of selecting appropriate models based on data characteristics. Specifically, when applied to small-scale datasets, experimental results using Gradient Boosting Decision Trees significantly outperform those obtained with BP neural networks, thereby effectively enhancing water quality detection models' accuracy.In utilizing a model recognized for its exceptional precision, this investigation revealed that relying on assessment criteria as markers for water quality analysis yielded less than optimal results.

Keywords: Back Propagation Neural Network; Gradient Boosting Decision Tree; Water Quality Detection.

1 Introduction

Water resources are fundamental to human survival and development. However, global challenges in water management are increasingly severe. Issues such as deteriorating water quality and the degradation of aquatic ecosystems have compromised ecosystem health, posing threats to both human health and ecological balance [1]. Therefore, it is crucial to strengthen water resource management, promote rational use of water resources, and ensure sustainable human health and the environment development [2].

Researchers worldwide have recently focused on developing efficient models for assessing water quality to ensure the rational utilization of water resources.

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_74

Traditional methods struggle to create unified water quality models due to aquatic environments' complexity and unknown variables. As a result, recent studies suggest selecting evaluation criteria suitable for water quality assessment by collecting, screening, and comparing information on various elements within aquatic environments [3,4]. These studies also propose constructing water quality detection models that adapt to environmental conditions by establishing mathematical relationships among these evaluation criteria.

Artificial neural network models have proven highly effective in addressing complex, nonlinear relationships in water quality detection. These models simplify the complexity of model development by requiring extensive data to establish relationships between inputs and outputs. Researchers have utilized artificial neural networks to streamline the computational complexity of model development processes. Back Propagation (BP) neural networks, a widely used model employing backpropagation techniques, excel in recognizing and classifying models. They enable learning patterns in water quality variations and have therefore been employed by researchers to propose various models for detecting water quality [5,6].

This article commences by addressing missing values in water sample data, followed by evaluating data correlations using a validation function to identify significant variables for input into a BP neural network. Throughout the neural network training process, techniques such as ten-fold cross-validation and early stopping are implemented, alongside the careful selection of activation functions to improve model fitting. After establishing the neural network, a BP neural network model is deployed to uncover nonlinear relationships between water quality and chosen evaluation factors, optimizing the model through parameter adjustments. Comparative analyses with alternative machine learning models are conducted to assess the dependability of BP neural networks in water quality detection models. Using models demonstrating superior performance, the research investigates whether the selected evaluation factors can reliably contribute to comprehensive assessments of water quality.

2 Data and Approaches

2.1 Data Origin

Past investigations have predominantly relied on water quality metrics, weather patterns, and environmental parameters to gauge water conditions comprehensively in evaluating water quality. Water quality metrics exhibit the strongest correlation with assessment outcomes and are frequently employed as primary benchmarks. Accordingly, this research process designated specific metrics from water quality data as evaluative criteria during experimentation. Utilizing around 2400 data entries from water samples in India sourced from the Kaggle database (Water Quality (kaggle.com)), the study concentrated on model training. The dataset encompasses nine influential factors, including pH levels, water hardness, total dissolved solids, chlorine levels, sulfate concentrations, conductivity, organic carbon content, trichloromethane levels, and turbidity, alongside portability as the key evaluation metric. The objective of this investigation revolves around forecasting water quality portability using these nine influential factors.

In light of incomplete entries within the dataset, about one-fifth of the data contained missing values for individual influencing factors. To uphold the integrity of model training accuracy, this study deemed these missing values as invalid data and consequently excluded them. As a result, approximately 2000 data points, encompassing all intact factors, were utilized for model training.

2.2 The Back Propagation Neural Network

The Back Propagation Neural Network (BPNN) is a type of artificial neural network (ANN) that emulates biological neural networks to execute computational tasks. It comprises interconnected artificial neurons that mimic how biological neurons transmit and process information. These networks can learn autonomously, demonstrating strong adaptability to data correlations and effectively handling nonlinear data in parallel [7,8]. Currently, numerous neural network models exist, with BPNN, Counter Propagation Network (CPN), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) being prominent and widely applied. In this research, the BPNN introduced by Rumelhart and Hinton is employed for model development and training [9,10].

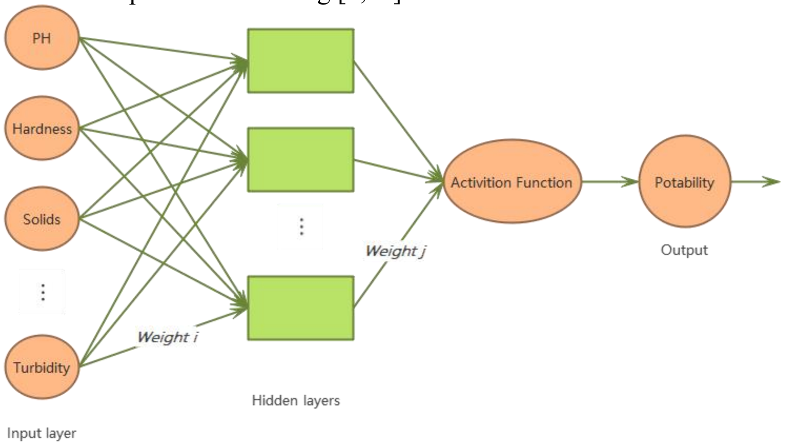

Fig. 1. Depicts a BP Neural Network model trained on data from Indian water quality monitoring(Photo/Picture credit : Original)

Fig. 1 showcases the structure and functionality of the backpropagation (BP) neural network. This neural network comprises an initial layer for input, intermediate hidden layers, and a concluding output layer. The input layer accepts data features, with the hidden layers' numerous neurons forming intricate connections via sophisticated feature blends and nonlinear alterations. Following this, the output layer generates outcomes derived from acquired data patterns. Post computation, the neural network employs backpropagation to refine data connections within the hidden layers, enhancing model efficiency by adjusting from output to input.

2.3 Evaluation Indices

This research utilizes the Pearson correlation coefficient, widely employed in correlation analysis, to assess relationships between variables in the dataset. Based on the correlation findings, the strongest associations were identified between hardness, trichloromethane levels, and potability, with coefficients of 0.700 and 0.679 respectively. Conversely, turbidity and chlorine levels showed weaker correlations with potability, registering only 0.309 and 0.352, respectively, among the nine factors examined. Moreover, organic carbon content, conductivity, and sulfate content exhibited notable negative correlations with potability, suggesting that lower levels of these constituents in water are linked to higher potability. After comparing organic carbon content, turbidity and chlorine levels, which displayed weaker associations, were excluded, leaving seven influencing factors as input data for the model.

3 Results Examination and Debate

3.1 Synopsis

To improve the performance of the BP neural network, this research adopts cross-validation, a widely recognized technique for training BP neural network models. Specifically, ten-fold cross-validation is employed to divide the data into ten subsets for model validation and parameter tuning, thereby enhancing the model's fitting efficiency.

Visual representations illustrating the distribution of influential factors and assessment metrics are provided in the examination of data distribution. Findings reveal a minimal significance in linear relationships between influential factors and assessment metrics. Consequently, the model integrates the rectified linear unit (ReLU) function as its activation function to capture non-linear relationships better.

Additionally, early stopping mechanisms are introduced to counteract model overfitting, mitigating the risk of rapid overfitting to training data associated with a higher number of neurons. By analyzing validation outcomes, early cessation of model training is implemented to prevent decreased accuracy on the validation set resulting from model overfitting.

3.2 The Experimental Results Analysis

This study adjusted the performance of the BP neural network by varying the number of neurons per layer and the number of hidden layers. According to the data presented in Table 1, increasing the number of hidden layers and neurons in the BP neural network marginally enhances the model's performance. However, the BP neural network did not achieve satisfactory results in terms of data prediction for the dataset utilized in this study. Following adjustments, the accuracy stabilized around 0.65.

Table 1. The relationship between the number of hidden layers and validation set accuracy in the BP neural network

Hidden Layers (Neurons per layer)	Accuracy
3 layers（50，100，50)	0.623
4 layers（50，100，100，50)	0.627
4 layers（50，100，100，50)	0.656
5 layers（50，100，100，100，50)	0.661
5 layers（100，200，200，200，100)	0.657

3.3 Discussion

In addition, this research conducted experiments with various datasets to train alternative machine learning models, identifying optimal performance. The study compared the BP neural network against popular classifiers like support vector machines, logistic regression, and gradient boosting trees. Results in Table 2 indicate that gradient boosting trees achieved superior performance with an accuracy of 0.72 on validation data. Conversely, support vector machines and logistic regression showed learning results comparable to those of the BP neural network, plateauing at around 0.65 accuracy.

Table 2. Machine Learning Models and Validation Set Accuracy

Model	Accuracy
Support Vector Machine	0.647
Gradient Boosting Tree	0.721
BP Neural Network	0.661
Logistic Regression	0.650

The experimental investigation reveals that the BP neural network fails to attain satisfactory outcomes when trained on data exhibiting weak correlations, as utilized within this research. Support vector machines and logistic regression similarly exhibit subpar performance in fitting functions to such weakly correlated datasets. In contrast, gradient boosting trees excel by building and refining decision trees, showcasing superior efficacy in managing such data and leading to markedly enhanced accuracy.

4 Conclusion

This research develops a BP neural network using pertinent water quality data from India. Adjusting network parameters contrasts the BP neural network's performance with alternative machine learning models. The following insights emerge from the experimental outcomes:

1) Training the dataset with a BP neural network can significantly suffer from poorly correlated data. Weak correlations among data points can impede the network's ability to learn relationships, thereby leading to degraded training results effectively.Utilizing high-performing models, this research reveals that the precision attained in evaluating water quality with assessment criteria remains disappointingly minimal, failing to effectively represent the true status of water quality. In contrast to

conventional indicators like Chemical Oxygen Demand (COD) and Particulate Organic Carbon (POC), there exists a conspicuous disparity in their efficacy. Thus, there remains ample room for enhancing the utilization of these criteria in assessing water quality.

2) Increasing the number of hidden layers and neurons in the BP neural network enhances its parameter count and complexity, thereby improving its learning capability. Additional hidden layers and neurons enable better capture of intricate patterns and relationships within data, improving fitting performance, particularly with small-scale datasets. Furthermore, employing suitable data augmentation techniques or cross-validation methods further boosts the BP neural network's efficacy.Yet, owing to the extensive array of structural options accessible for BP neural network models, pinpointing the optimal configuration proves challenging, thus constraining the attainment of its theoretical effectiveness.

3) In the realm of intricate nonlinear challenges, the backpropagation (BP) neural network modifies its weights gradually towards local enhancements. This iterative refining process may steer the algorithm towards local minima, where weights stabilize at minimal local points. Moreover, BP neural networks are susceptible to either overfitting or underfitting when trained on scant datasets, thus affecting their ability to generalize. Hence, when handling small-scale datasets with weak correlation to evaluation metrics, the performance gap between BP neural networks and gradient boosting trees can manifest distinctly under these circumstances.

References

1. Mi, Y.P., Wang, X.P., & Jin, X.: Machine learning-based method for predicting water quality COD. Journal of Zhejiang University (Engineering Science), 42(5),790-793 (2008).
2. Zhang, Y., & Gao, Q.Q. :Nest Lake water quality assessment based on random forest classification algorithm. Journal of Environmental Engineering, 10(2), 992-998(2016). .
3. Xu, H.M., & Yang, T.H.: Lake water quality assessment research based on support vector machine classification algorithm. Journal of Jilin University (Earth Science Edition), 36(4), 570-573 (2006).
4. Wu, G.Z.: Application research of support vector machine in eutrophication assessment and water quality prediction of lakes (Doctoral dissertation). Inner Mongolia Agricultural University (2008).
5. Ni, S.H., & Bai, Y.H.: Application of BP neural network model in groundwater quality assessment. Systems Engineering Theory and Practice, 20(8), 124-127 (2000).
6. Chen, S.Y., & Li, Y.W.: Water quality evaluation model based on fuzzy artificial neural network recognition. Advances in Water Science, 16(1), 88-91 (2005).
7. Guo, J.S.: Research on water quality evaluation and simulation based on artificial neural network (ANN) (Doctoral dissertation). Chongqing University. (2002).
8. Li, Q., Zhang, J.H., Lin, L.Y., et al.: A review of water environmental quality assessment methods. Modern Agricultural Science and Technology, (19), 285-287, 290 (2011).
9. Li, R.Z: Research progress and trend analysis of water quality prediction theoretical model. Journal of Hefei University of Technology (NaturalScienceEdition), 29(1),26-30 (2006).

10.Nagy, H. M.: Prediction of sedi-ment load concentration in rivers using artifical neural net-work mod, Journal of Hydraulic Engineering,128(6):588-595 (2002).