ISSN 2616-5872 Vol.4, Issue 8: 48-52, DOI: 10.25236/AJEE.2022.040809

Research on the Water Pollution Control Mechanism of Xiangjiang River Basin Based on Machine Learning

Yu Yuan, Dazhi Xu[*]

College of Economics and Management, Hunan University of Arts and Science, Changde, China
[*]*Corresponding author*

Abstract: *Due to the complexity of water pollution treatment and the impact of external factors, precise regulation of water pollution has always been one of the problems in the environmental field. Traditional methods cannot meet the increasingly complex engineering project requirements. Machine learning methods developed in recent years provide a series of effective solutions for such problems. Taking the Xiangjiang River basin as an example, this paper introduces the characteristics of machine learning methods such as artificial neural network, support vector machine and random forest, and expounds the application of machine learning methods in the field of water pollution control in the Xiangjiang River basin from three aspects of water quality prediction and early warning, fault diagnosis of sewage treatment system and intelligent control, and analyzes the advantages of machine learning methods compared with traditional methods and the problems existing in the application of machine learning methods in water pollution control in the Xiangjiang River basin, the prospect and trend of the application of machine learning methods in the field of water pollution control in the Xiangjiang River basin are forecasted.*

Keywords: *machine learning, water pollution, intelligent control, governance, mechanism*

1. Introduction

In recent years, with the increasing investment of Hunan Province in water pollution control of the Xiangjiang River basin, the renewal and iteration of water treatment technology and the continuous expansion of the scale of wastewater treatment facilities, the treatment process has become more and more complex; In addition, because the quality and quantity of the sewage in the Xiangjiang River basin fluctuate greatly and are subject to more external interference, the sewage treatment system is characterized by strong connectivity and hysteresis, and each part of the system is related to each other. A change in a condition may affect the whole system, which has a strong uncertainty [1]. However, traditional means, such as relying on experience or simple control systems, can no longer meet the needs of the current sewage treatment system in the Xiangjiang River basin, resulting in poor operation quality, low treatment efficiency and low resource utilization during the construction and operation of sewage treatment facilities in the Xiangjiang River basin, which has attracted extensive attention.

Machine learning methods can make full use of big data to perform nonlinear regression, classification and prediction, diagnose abnormal data points and find optimal decision-making methods for multi-objective systems. It is one of the important means to solve the problems of complex engineering systems in recent years [2]. At the same time, the machine learning method has a high fault tolerance, can adapt to large changes in input data, can make good use of the data sets generated in the process of sewage treatment in the Xiangjiang River basin, and can achieve better learning effects through continuous optimization. Because of these characteristics, in recent years, in the field of sewage treatment, many researchers have applied machine learning methods to solve complex practical engineering problems, and made a series of progress in solving the low accuracy of water quality prediction, difficulties in fault diagnosis and delayed regulation in sewage treatment[3-4]. This paper will discuss and analyze the application of machine learning in the field of wastewater treatment in Xiangjiang River basin in recent years, with a view to providing reference for researchers of wastewater treatment in Xiangjiang River basin.

2. Classification and introduction of machine learning methods

The machine learning method mainly uses the selected model to learn the input data, extract valuable features or information from the complex data set, summarize the reasonable change trend, and then make data prediction[5]. It is a method that can readjust the parameters or structure in the model to improve the accuracy and reliability of prediction after comparing the deviation between the predicted value and the actual value. Machine learning can be divided into supervised learning, unsupervised learning and semi supervised learning according to the different ways in which the model marks the input data.

2.1 Supervised learning

Supervised learning is one of the important branches of machine learning methods. It is mainly used to learn and train data sets of known categories, find the relationship between data features and categories through computational models, and predict according to the training results. Supervised learning is a machine learning method that is widely used in various fields, mainly including linear regression, SVM, BC, ANN, RF and logistic regression algorithms. Among them, regression algorithms such as linear regression and logical regression are mainly used to study the relationship between simple independent variables and dependent variables[6]. However, due to the complexity of sewage treatment process, conventional regression calculation cannot meet the requirements of system prediction, early warning and monitoring. In contrast, SVM and ANN are more suitable for solving such complex problems. SVM is often used for classification. Its essence is to project the sample data into a higher dimensional space in the form of vectors and establish a hyperplane. The purpose is to find a hyperplane that is the smallest of all samples[7]. This method can minimize the empirical error and model complexity to improve the classification effect or solve the general regression problem. However, SVM is only suitable for processing small-scale data. If the sample size is too large, it will make the calculation process too complex to ensure the accuracy of classification; The ANN algorithm is to transfer the input signal from one neuron to another in the form of activation function, input the signal value in the activation function, and then input it to the next layer after a certain nonlinear calculation until the output results. Common activation functions include Sigmaid, Tanh, ReLU and ELU functions[8]. These two kinds of machine learning algorithms are widely used in the related fields of environmental prediction, including environmental ecology, water treatment and data modeling for water quality monitoring.

2.2 Unsupervised learning

Unsupervised learning is another important category of machine learning methods. Unlike supervised learning methods, unsupervised learning is mainly used to analyze unclassified data, extract potential relationships or features from data sets, and then classify them into clusters. At present, the mainstream unsupervised learning algorithms mainly include PCA, K-means clustering, CNN and SOM. Among them, PCA and K-means clustering are the two most basic unsupervised learning methods. PCA is widely used for data dimensionality reduction. It can extract low dimensional subspaces from high-dimensional data and preserve the diversity of data as much as possible. The specific method is to map high-dimensional (n-dimensional) features to low dimensional (k-dimensional) features. k-dimensional is a new orthogonal feature, also known as the main component, which is a k-dimensional feature reconstructed from the original n-dimensional features, Therefore, it is suitable for dealing with complex and multi-dimensional wastewater process problems. However, due to its poor adaptive ability, it is not practical for wastewater projects with large changes in water quality and quantity; K-means clustering algorithm is a commonly used clustering algorithm. The given data object is divided into k different clusters by iteration and converges to the local minimum to complete the clustering process. This method runs fast and is suitable for processing large data sets. However, since the algorithm output depends on random seeds (the selection of k-value and cluster center depends on random seeds), repeated operations are required to optimize the k-value and cluster center.

2.3 Other machine learning methods

In addition to supervised learning and unsupervised learning, there are some other types of algorithms, such as semi supervised learning algorithm that can learn from a small amount of labeled data and reinforcement learning algorithm for online modeling. The former can only rely on a small number of labeled data and a large number of unlabeled data for learning, and the training model has been applied in practical projects; The latter is to use the information transmitted to the system by the changing environmental state to judge whether the change brings corresponding benefits, store the change and benefits, and then constantly seek the next decision that can achieve the maximum benefits. The typical one is the Q-learning algorithm, which stores the environment variables in the system and the benefits that the variables can bring by building a Q table, and then selects the actions that can obtain the maximum benefits according to the Q table. In the field of wastewater treatment, Q learning algorithm can be used to optimize the hydraulic retention time (HRT) of reactors.

3. Governance mechanism of machine learning in Xiangjiang River basin sewage treatment

3.1 Fault diagnosis during sewage treatment in Xiangjiang River basin

There are many traditional fault diagnosis methods. The fault diagnosis methods based on historical data (quantitative data, qualitative data and process data) are generally only applicable to simple or linear mechanical problems. When dealing with high-dimensional and nonlinear sewage treatment problems in the Xiangjiang River basin, the traditional fault diagnosis methods pay more attention to the microstructure and emphasize the characteristics of timely changes, and cannot fully describe the change laws of complex systems[9].

Fault diagnosis of sewage treatment facilities in Xiangjiang River basin by machine learning can be transformed into a problem of state classification based on historical data. Typical supervised learning methods (such as SVM, ANN, Bayesian network, etc.) can expand the problem of fault diagnosis from binary classification to multi category classification, so as to achieve a more reliable fault detection effect. At the same time, in unsupervised learning, such as K-means clustering, PCA and expectation maximization clustering methods, abnormal problems can be determined as separate clusters or points away from normal clusters, so as to achieve the effect of fault diagnosis[10].

The machine learning method is based on a huge database to continuously monitor and predict the system, so as to find problems and effectively achieve remote and local maintenance[11]. When the machine learning method is used for fault diagnosis, it is not necessary to pay attention to the operation mode of each part of the sewage treatment in the Xiangjiang River basin and the various biochemical reaction processes involved, but to collect and process the water quality indicators, operation status, environmental factors and other data of the entire system, monitor and diagnose from a global perspective, effectively making up for the defects of traditional methods in fault diagnosis of sewage treatment in the Xiangjiang River basin. Its basic principle is to compare the current monitoring data or system status with the normal or abnormal historical data accumulated previously, find the similarities between the current data and historical data through classification or clustering, analyze the changes and exceptions of the inherent background, and timely diagnose whether the operation of the sewage treatment system in the Xiangjiang River basin is normal or not. Therefore, machine learning needs to collect a large amount of raw data, combine appropriate data analysis techniques, convert these data into valuable information, and make positive decisions based on this information to optimize the overall performance.

3.2 Application of machine learning in sewage fault diagnosis and early warning in Xiangjiang River basin

The application of machine learning method in fault diagnosis and early warning of sewage treatment system in Xiangjiang River basin mainly includes timely detection of sensor failure, sudden water pollution, pipeline leakage and large fluctuation of system operating parameters.

In the application examples of Xiangjiang Community Sewage Treatment Plant and Sequencing Batch Reactor (SBR) pilot plant, the researchers used cross validation to determine several principal components based on training set[12]. The former used PCA model to analyze the cause of failure, and the latter used classification to identify the failure of dissolved oxygen sensor or liquid level sensor, both of which achieved fault diagnosis effects applicable to engineering applications, However, this multivariate statistical method needs to assume that the environmental conditions will not change significantly during the water treatment process, so the latter is only applicable to the situation where the water quality changes little.

In addition, with the continuous development of machine learning technology, more and more researchers found that a single machine learning model could not better analyze the entire wastewater treatment process. If only SVM is used for fault diagnosis of sewage treatment plants in the Xiangjiang River basin, the error rate will be high, reaching about 30%. However, in case of misdiagnosis of the sewage treatment system, it may cause great losses. Therefore, some hybrid machine learning models are applied to the field of fault diagnosis of sewage treatment in the Xiangjiang River basin. For example, the improved genetic algorithm and K-means clustering algorithm can be integrated to analyze the historical data of the sewage plant, establish the rules for fault diagnosis of the sewage plant process, and use machine learning to complete fault diagnosis and early warning of the sewage plant in the Xiangjiang River basin. In the application example of Changsha No.2 Sewage Treatment Center, multi class SVM is used, and GA algorithm is used to optimize SVM. After more than 180 generations of optimization and evolution the misclassification rate can be reduced to 1.86%, basically meeting the requirements of fault diagnosis in sewage treatment system. Because there are many and varied faults that may occur in the process of sewage treatment in the Xiangjiang River basin, and these problems are likely to exist in the same sewage treatment project together, a single machine learning algorithm cannot achieve better results when seeking machine learning for fault diagnosis. Based on the above analysis, it can be seen that fault diagnosis and prediction using multiple algorithm hybrid machine learning methods is expected to become the mainstream trend of future technology development.

4. Discussion

Although machine learning is more and more widely used in the field of sewage treatment, and can effectively ensure the normal operation of sewage treatment system under different circumstances, to achieve energy conservation and consumption reduction, there are still problems to be solved.

(1) Machine learning model has black box property, and its own accessibility and interpretability are poor, which may affect the stability of the system to a certain extent.

(2) In the Xiangjiang River basin sewage treatment system, many parameters cannot be obtained by current sensors and other hardware equipment, such as technical characteristics, environmental conditions, meteorological conditions, social conditions, process design direction, etc. These parameters are an indispensable part of the control process, but because they are difficult to quantify and evaluate, machine learning cannot bring them into the category of learning and training.

(3) In terms of prediction, early warning and fault diagnosis, there is often a problem of data imbalance in machine learning, that is, most of the data collected in the system are normal samples, and the number of fault and abnormal samples collected is very small, so the gap between the two is large. The classical classification and recognition technology requires that all types of samples be classified as equally as possible, which makes the application of machine learning methods have certain limitations. In terms of system control, compared with using traditional mathematical models or traditional PID control models, machine learning models have higher accuracy, but they are very dependent on historical data, requiring a large number of background values as reference. Once severe water quality changes occur, the reliability of models built on the basis of large amounts of historical data will decline, and they cannot cope with changes in extreme conditions, However, large fluctuations in water quality and quantity are common in Xiangjiang River basin sewage treatment, so relying on machine learning for early warning, decision-making and control cannot completely replace the traditional system.

5. Conclusion

At present, machine learning is widely used in the field of sewage treatment, involving monitoring, prediction, early warning, fault diagnosis, intelligent control and other technical links, and has an extremely broad development prospect. However, due to the characteristics of machine learning, its current application has limitations. First of all, the sewage system in the Xiangjiang River basin is very complex, involving a variety of physical, chemical and biological reactions, and there is still a lot of information that has not been extracted can be used, such as pollutant types, toxicity, microbial community structure and functions. Therefore, it is necessary to develop new detection methods, improve the machine learning data system, and provide higher dimensional, more valuable and more representative data for machine learning methods. Secondly, in terms of specific machine learning model application, a single machine model cannot adapt to the sewage treatment problem in the Xiangjiang River basin due to its inherent shortcomings and problems, so the joint use of multiple models or the use of mixed models to deal with sewage problems has gradually become the mainstream trend. Thirdly, because machine learning itself has black box nature, poor interpretability and accessibility, and machine learning cannot cope with extreme situations, it is of great significance to carry out basic research, clarify the basic principle of machine learning algorithm and improve its practicability. Finally, scientific researchers and engineering technicians in the field of sewage treatment in the Xiangjiang River basin lack the understanding of machine learning algorithms and related theories, so multidisciplinary cross research is needed to develop machine learning methods more suitable for sewage treatment systems.

Acknowledgements

This work was supported by the Social Science Foundation of Hunan Province of China (Grant No.18JD54).

References

[1] Qian W.J., He C.F. China's regional difference of water resource use efficiency and influencing factors[J]. China Population, Resources and Environment, 2011, 21(2):54-60.
[2] Sun Z.L., Sun H., Su Y. Water use efficiency and its influencing factors in arid areas of Northwest China[J]. Journal of Ecology and Rural Environment, 2017, 33(11):961-967.
[3] Xu X.Y., Wang H.R. Liu H.J., et. Al. Evaluation report on water resources utilization efficiency in China[M]. Beijing: Beijing Normal University Press, 2010
[4] Song G.J., He W. Benchmarking of city water resource utilization efficiency in China[J]. Resources Science, 2014, 36(12): 2569-2577.
[5] Jamaluddin M Y, David H. The efficiency of the National Electricity Board in Malaysia: An intercountry comparison using DEA[J]. Energy Economics,1997,19(1):255-269 .
[6] Mai Y.Z., Sun F.L., Shi L., et.al. Evaluation of China's industrial water efficiency based on DEA model[J].Journal of Arid Land Resources and Environment, 2014,28(11):42-47.
[7] Zhao L.S., Sun C.Z., Liu F.C. Two-stage utilization efficiency of the interprovincial water resources under environmental constraint and its influence factors in China[J]. China Population, Resources and Environment, 2017, 27(5):27-36.
[8] Dong Z.F., Yu E.Y., Qiu L., et al. Water efficiency evaluation of the provincial regions in China based on DEA model[J]. Ecological Economy, 2012(10):43-47.
[9] Hu L.K. Dynamic evaluation of water resources utilization efficiency in Yunnan province based on WCA-MEPP model[J]. Journal of Water Resources & Water Engineering, 2017,28(4):75-87+81.
[10] Tony A. Productive Efficiency and Allocative Efficiency: Why Better Water Management Might not Solve the Problem [J]. Agricultural Water Management, 1999, 40(3):71-75.
[11] Anwandter L. Can Public Sector Reforms Improve the Efficiency of Public Water Utilities an Empirical Analysis of the Water Sector in Mexico Using Data Envelopment Analysis[D]. Maryland: University of Maryland, 2000.
[12] Wang X.Y., Zhao L.G. Agricultural water efficiency and the causal factors——A stochastic frontier analysis based on Chinese provincial panel data: 1997-2006[J]. Issues in Agricultural Economy, 2008, 29(3): 10–18.