Original Research Paper | Computer Science | Volume : 4 ¦ Issue : 9 ¦ Sep 2019 ¦ e-ISSN : 2456-5040

THE APPLICATION OF GENETIC ALGORITHM TO OPTIMIZE XGBOOST ALGORITHM IN MBR SIMULATION

Ziming liu

Tian Gong University (TGU), Chaina.

ABSTRACT

Membrane bio-reactor (MBR) is one of the mainstream technologies of modern wasted-water treatment, but membrane fouling is a key factor which restricts the development of MBR. The formation of membrane fouling directly leads to a decrease in membrane flux. As one of the important parameters to measure membrane fouling, membrane flux is the focus and difficulties of membrane fouling research. In this paper, XGBoost algorithm is used to simulate and predict MBR membrane flux. In view of the shortcomings of XGBoost with many parameters, combined with the characteristics of genetic algorithm (GA), such as strong global search ability and fast convergence speed, the parameters of XGBoost are optimized and adjusted. By analyzing the prediction results of GA-XGBoost algorithm, and comparing with the experimental data, the results show that GA-XGBoost neural network prediction model is superior to XGBoost algorithm prediction model in the prediction of MBR membrane flux, with higher accuracy.

KEYWORDS: Genetic Algorithm; XGBoost; Membrane Flux; Membrane Fouling.

Introduction:

Membrane Bio-reactor (MBR) is a unique wasted-water treatment process that combines biological treatment process of activated sludge with an innovative membrane filtration system. Due to its high-quality water production and compact footprint, it is an effective tool for industrial water treatment and water reuse, while MBR is also the preferred wasted-water treatment technology for activated sludge processes. MBR is one of the most important innovations in wasted-water treatment technology nowadays. It overcomes the shortcomings of traditional activated sludge processes, including space requirement for secondary sedimentation tanks, liquid–solid separation problems, generation of excess sludge, and restrictions on removal of stubborn things. [1]MBR is a mixture of a traditional biological treatment system and physical liquid–solid separation using membrane filtration in one system. The MBR technology provides the following advantages, for example, traditional sewage treatment technology, high-quality effluent, higher volumetric loading rates, shorter hydraulic retention times (HRT), longer solid retention times (SRT), and lower sludge production, etc[2]. However, some problems have arisen in the process of MBR wasted-water treatment, such as the need to control membrane fouling issues, higher energy costs, and the potential high costs of periodic membrane replacement. Research shows that membrane fouling remains the most challenging problem in membrane bio-reactor operation[3]. Membrane fouling is considered to be a major disadvantage of MBR because it reduces membrane flux of MBR, which increases the MBR's operating costs in the meantime. So the size of membrane flux can be used as a measure of membrane fouling[4,5].

At the present days, the method based on machine learning is widely used in the analysis and prediction of MBR membrane flux, which has the characteristics of large-scale parallel processing, strong learning ability and high nonlinearity. XGBoost algorithm is an enhancement algorithm based on the regression tree model of gradient lifting decision tree for classification[6], and it is one of the machine learning algorithms with good application effect at present. The main features are to use many strategies to prevent over fitting, to add sparse data processing, to use cross validation and EarlyStop, to prevent too deep tree building, to support parallelization of nodes at the same level can be parallelized, specifically for a node, the best split point is selected within the node, and the candidate split point calculation gain uses multi-threaded parallel to accelerate the training speed. The main disadvantage of this method is presorting. Before iteration, presorting is performed on the characteristics of nodes, and the optimal segmentation points are selected through traversal. When the amount of data is large, greedy method is time-consuming, and the number of adjustable parameters is large, which affects the prediction efficiency. Therefore, genetic algorithm (GA) is selected to optimize XGBoost algorithm to achieve better prediction effect.

1. Theory:

1.1. XGBoost algorithm:

XGBoost (Extreme Gradient Boosting) is proposed by Dr. Chen Tianqi of Washington University[7], and it is used in the Higgs signal recognition competition of Kaggle. It has attracted wide attention due to its outstanding efficiency and high prediction accuracy. XGBoost algorithm is an enhancement algorithm for classification based on the regression tree model of the gradient lifting decision tree. Newton method is used to solve the extreme value of the loss function, and the loss function is extended to the second derivative through the second-order Taylor derivative to obtain the analytical solution as the gain of the establishment tree, so as to optimize the loss function. Figure 1 shows the general flow of boosting algorithm based on classification regression tree. First, learn a tree from the sample to get the first set of estimation results Y1, then learn the second tree according to the results obtained by Y1, compare the difference between the real label and the prediction label in the previous step, and through the above analogy, algorithm errors can be effectively reduced.[8]

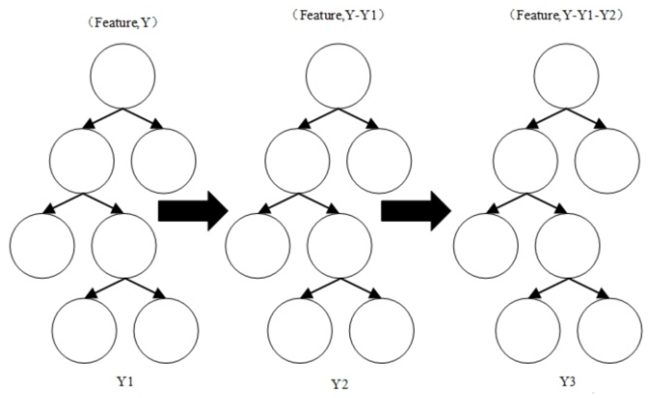

Figure 1: General flow of boosting algorithm

1.2 Genetic algorithm(GA):

Genetic algorithm (GA) was first proposed by Professor Holland of the University of Michigan in the 1970s, based on selection and genetics. It is used for reference an adaptive, highly parallel, random global search algorithm generating from the biological evolution theory of nature. Genetic algorithm simulates breeding, mating and mutation in natural selection and genetic evolution, which has three operations, namely selection, crossover and mutation[9].

(1) Selection:

The selection operation means selecting the individuals with high adaptability from the initial population to form new groups. The selection of GA ordinarily uses the roulette wheel method, which is a playback random sampling method. The probability that each individual enters the next generation is equal to the ratio of its fitness value to the sum of individual fitness value in the entire population. Its mathematical expression is as follows:

$$P_i = \frac{F_i}{\sum_{i=1}^{n} F_i})$$

Among them, P_i is the probability of selecting the individual, F_i is the fitness of the individual, n means the entire number of individuals in the population.

(2) Crossover:

Crossover refers to the exchange of two parts of a pair of chromosomes in a certain way to form two new individuals.

(3) Mutation:
Mutation is a sort of the genetic mutation and the individual chromosome has mutated to be a new individual. GA has also a mutation probability and that is very small.

2. Simulation prediction model of membrane flux based on GA-XGBoost:
When XGBoost algorithm is used alone, there are many kinds of parameters, different combinations of parameters will get different evaluation scores, resulting in large deviation of results. Genetic algorithm uses probabilistic optimization method and probabilistic transfer rules instead of deterministic rules, which can adjust the search direction adaptively and find the optimal solution. Therefore, based on the defects of XGBoost algorithm and the advantages of genetic algorithm, the combination of the two algorithms can enhance the global optimization ability, speed up the evolution of the algorithm, and improve its convergence accuracy. So this paper combines the selection, crossover and mutation operations of genetic algorithm with XGBoost algorithm to construct GA-XGBoost hybrid algorithm. The specific algorithm implementation steps are as follows.

Step 1: generate a set of default parameters according to XGBoost algorithm.

Step 2: initialize the genetic population.

Step 3: calculate the fitness function, which is obtained from the output score of XGBoost algorithm.

Step 4: rank the top ten population individuals with the lowest fitness.

Step 5: introduce genetic algorithm. Select, cross and mutate the new population.

Step 6: extract the population that has been operated by genetic algorithm, and recombine the population with high adaptability that has not been selected before.

Step 7: update individual extremum and group extremum.

Step 8: end condition. Recalculate the fitness function to determine whether it meets the cycle termination condition. If a better fitness value is obtained, proceed to the next step, otherwise return to step 4.

Step 9: output the prediction results of XGBoost algorithm.

Step 10: calculation error.

3. Simulation results of XGBoost optimization based on GA:
3.1 Experimental method:
In this experiment, the actual MBR system operation data of a sewage treatment plant is selected for prediction and evaluation. In the process of analysis, first of all, the known data is sorted out, individual blank data is filled with pandas data, and the processed data is divided into training data set and test data set, the allocation ratio is set to 9:1, the training set is used to train XGBoost model, the test set is used to test the final experimental results, in which XGBoost parameters use the original default parameters. The comparison results show that the predicted results of XGBoost membrane permeability prediction model with default parameters are basically consistent with the data trend of the actual MBR system operation results, as shown in Figure 3-1, the basic trend prediction can be achieved, but the results of relative error are poor, as shown in table 3-1 and figure 3-2, and the results of relative error smaller can be obtained by optimizing and adjusting its parameters after need.

3.2 Analysis of results:
In order to study whether the algorithm effectively improves the prediction accuracy, this paper uses the initial XGBoost algorithm model to predict the experimental results under the same conditions, and compares the prediction results with the XGBoost algorithm optimized by GA. The comparison results are shown in Figure 3.

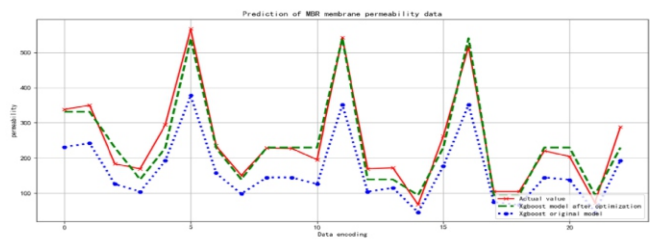

Figure 2: Experimental prediction results

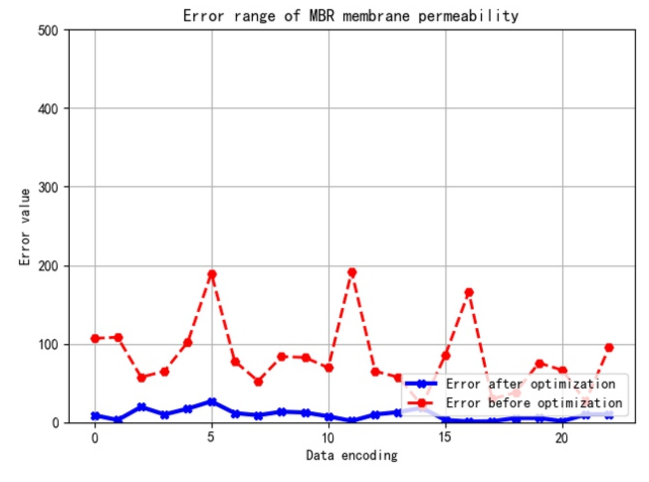

Figure 3: Experimental relative error

It can be seen from Figure 1 and Figure 2 that under the premise of selecting the same test sample, the prediction accuracy of MBR membrane pollution prediction model after GA optimization XGBoost is significantly higher than that before optimization, and the error reduction is large. Therefore, the model optimized by GA-XGBoost improves the accuracy of MBR membrane pollution prediction.

4. Conclusion:
In this paper, XGBoost algorithm is used as a starting point to simulate and predict MBR membrane flux. Aiming at the problem that XGBoost has too many parameters and it is difficult to use global search to optimize, genetic algorithm (GA) is combined with XGBoost algorithm, and the better parallel computing ability of GA algorithm is integrated. XGBoost algorithm is optimized based on GA algorithm to quickly find the optimal XGBoost under MBR membrane pollution prediction model. Parameters to improve the training efficiency. The results show that the prediction performance of XGBoost algorithm optimized by GA is improved compared with that before optimization, and the reliability and accuracy are greatly improved. Therefore, the MBR prediction method based on GA optimization XGBoost can more effectively and accurately predict MBR membrane flux, accurately judge the degree of membrane pollution, and timely clean it to keep its membrane flux at a higher level, so as to improve the production efficiency, which has a certain guiding significance for the actual MBR engineering.

References:

[1] Wei Li, Zhigang Cheng, Kaitao Wu. Industrial Application and Discussion of Membrane Bioreactor Technology [J]. Guangdong Chemical Industry, 2012, 39(07):112-113+143.
[2] Noor Sabrina Ahmad Mutamim,Zainura Zainon Noor,Mohd Ariffin Abu Hassan,Adhi Yuniarto,Gustaf Olsson. Membrane bioreactor: Applications and limitations in treating high strength industrial wastewater[J]. Chemical Engineering Journal,2013,225.
[3] Meng F , Zhang S , Oh Y , et al. Fouling in membrane bioreactors: An updated review[J]. Water Research, 2017, 114:151-180.
[4] Haiyan Mu. New progress in research on membrane fouling in membrane bioreactor [J]. Urban road bridge and flood control,2018(08):169-172+19-20.
[5] Haowen Wang. Causes and Control of Membrane Pollution in Hollow Fiber Membrane for Advanced Treatment of Wastewater [J]. Chemical and pharmaceutical engineering,2016,37(03):48-53.
[6] S. Paisitkriangkrai, C. Shen, and J. Zhang, "Fast pedestrian detection using a cascade of boosted covariance features, "IEEETrans. CircuitsSyst. Video Technol., vol.18, no.8, pp.1140–1151,Aug.2008.
[7] T. Chenand C. Guestrin, "XGBoost: A scalable tree boosting system, "in Proc. 22nd ACMSIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug.2016,pp.785–794.
[8] FanJ, WangX, WuL,etal. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China [J]. Energy Conversion and Management,2018,164:102-111.
[9] Hui Zhi,Sanyang Liu. Face recognition based on genetic algorithm[J]. Journal of Visual Communication and Image Representation,2018.