Journal of Information and Communication Engineering (JICE), 3 (5): 178-182, December 29, 2017
e-ISSN: 2432-5465, p-ISSN: 2186-9162
Applied Science and Computer Science Publications

Water Bloom Warning Model Based on Random Forest

Yunxiang Liu and Hao Wu

Abstract: Based on the random forest classification algorithm, a warning model of water bloom is proposed. Using the collected data, Select the water quality, meteorological factors which like Chlorophyll a (Chl-a), water temperature (T), PH, nitrogen and phosphorus ratio (TN:TP), chemical oxygen demand (COD), total nitrogen (TN), total phosphorus (TP), dissolved oxygen Light (E) and so on as the impact factor and use them establish a warning model for Water bloom. And compared with the prediction accuracy of neural network model and SVM model. The results show that the water bloom warning model is established by using stochastic forest classification algorithm, the prediction accuracy is slightly higher than other algorithms. And the random forest algorithm has the characteristics of high robustness, China good performance, strong practicability can effectively carry out water bloom early warning.

Keywords: Random forest; Decision tree algorithm; water bloom warning; split attribute set.

I. INTRODUCTION

With the rapid growth of population, the rapid development of modern chemical and agricultural production, a large number of industrial and agricultural wastewater and domestic wastewater into the sea, lakes and urban reservoirs, causing serious water pollution. And most of them are untreated wastewater, and these untreated wastewater are discharged directly into the oceans, lakes and reservoirs, which exacerbates the eutrophication of water bodies. The eutrophication of water bodies is an important factor to induce water bloom [1-2] outbreak. The large area of water bloom makes the water environment worse and worse, and the pollution is aggravated gradually. In order to effectively control the water environment, scientific and effective prediction is essential. The forecast is conducive to the preparation of targeted preventive measures and early warning. In order to solve the problem of lake forecasting, domestic and foreign scholars have made some progress in carrying out early warning research, and different scholars have carried on different analysis from different angles. Multivariate statistical regression method, fuzzy mathematics, genetic algorithm and neural network method [3] are applied to the prediction of lake blooms and play a better role in water environment protection and management. However, these methods still exist a certain lack. For example, the problem of the local extremum of the neural network, the equation of the linear relationship obtained by the multivariate statistical regression method, is not obvious for the prediction of water bloom.

The key to solving the problem of water outbreak prediction is to determine whether the water blooms will break out according to the data of the influential factors collected by the monitoring station. This is a typical classification problem. Therefore, the application of decision tree algorithm can be a good prediction model. The decision tree algorithm has the characteristics of simple model and simple rule extraction, in which the CART algorithm is the classical algorithm in the decision tree algorithm. The CART algorithm uses the binary recursive splitting method to divide the sample set into two different sub-samples, so that each non-leaf node has two or so bifurcations, so the CART algorithm is formed by a binary tree Simple decision tree. However, based on the traditional CART algorithm to generate the water bloom warning model, in the judgment, there will still be running a long time, the accuracy is not high and other shortcomings. The random forest is a combination classifier which uses decision tree, which can improve the classification accuracy and solve the problem of over-fitting. Therefore, this paper will use the random forest [6-7] algorithm to improve the classification

Corresponding Author: Yunxiang Liu, School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai, China, E-mail: yxliu@sit.edu.cn

accuracy, and the use of random forest algorithm for water bloom warning problem provides a feasible method.

II. THE PRINCIPLE OF CART DECISION TREE ALGORITHM

Decision tree algorithm [7-8] is one of the classical algorithms in classification algorithm. The result of decision tree algorithm is very clear and easy to understand. Decision tree is a typical single classifier. To build a decision tree, First is to analyze and process the data in the sample set, and a series of rules and decision trees are generated by the inductive algorithm. Then, the new data is analyzed by using the decision tree. A decision tree is a process that uses a set of rules to separate data into different categories. CART is a binary tree algorithm in the decision tree algorithm. Its structure is relatively simple, it uses the binary segmentation method, which will be the current sample set is divided into two sample sub-set, and then use the binary sub-partition, the recursive repeat this action makes the final generation of each non-leaf nodes There are two left and right branches. Steps to generate a CART decision tree are as follows:

Step1 Calculate the GINI index of each attribute in the data set, the smaller the GINI index, the lower the impurity, the more "pure". Select the smallest GINI index attribute as the root node of the decision tree. For discrete properties, we examine all possible subsets of known sample sets, calculate these subsets GINI index, the minimum GINI index is chosen as the attribute splitting method and the minimum GINI index as the GINI index of the attribute. But for continuous attributes, it is necessary to do more step-by-step discretization of continuous attributes. Discretization needs to calculate the optimal segmentation threshold for each attribute, discretize it according to the segmentation threshold, and calculate its GINI index. The GINI index refers to the impure degree of the metric data partition and the impure level of the sample set E., and the smaller the GINI index, the more pure the sample is. Defined as follows:

$$\text{GINI（E）} = 1 - \sum_{i=1}^{m} p_i^2 \qquad (1)$$

the category set is{ C_1, C_2, ⋯ C_n }, | C_i | is the number of samples belonging to Ci in E, p_i=| C_i | / | E |is the probabilit that the sample in E belongs to class C_i.

When only the sample set is divided into two, the attribute A divides the training sample set E into two subsets E_1 and E_2, then the GINI index of E is divided:

$$\text{GINI}_A(E) = \frac{|E_1|}{|E|} GINI(E_1) + \frac{|E_2|}{|E|} GINI(E_2) \qquad (2)$$

among them, | E_k | / | E | is the probability that the sample set of samples belongs to k (k = 1, 2) subsets.

Step2 For continuous attributes segmentation, sample set to select attributes segmentation point GINI index minimum point threshold divide S into two parts <= *S* and >= *S*. If the splitting attribute is discrete, by calculating the GINI index of all subsets of the training sample set, the subset with the smallest GINI index is chosen as its splitting subset and split into two parts.

For a continuous attribute, in order to obtain the GINI index of the attribute, it is necessary to consider each possible split point. It is necessary to set the intermediate value of each of the adjacent values in ascending (or descending order) as a possible split point, using this splitting point to divide the attribute into two parts, we use the formula (2) to calculate the GINI index of each splitting point, and then select the splitting point of the smallest GINI index as the splitting point of this attribute.

Step3 The root nodes are split into two subsets E_1 and E_2 according to the attributes with the smallest GINI index, using Step1 the same method to recursively create the decision tree tree nodes. The categories of samples that are recycled to all leaf nodes are approximately the same, or the next subsequent splitting attribute is gone.

Step4 Using the cost complexity pruning algorithm for pruning to form a concise decision tree.

III. RANDOM FOREST ALGORITHM

Random forest is a combination of methods, its basic classification unit is the decision tree (here is the CART decision tree), a combination of a series of unbranched CART decision tree $\{h(\boldsymbol{X}, \Theta_k)\}$, $\{\Theta_k, k=1,2,\ldots,k\}$ is an independent and identically distributed random vector.And in the classification, each decision tree to vote, and finally return to the most votes in the class. It improves the shortcomings of a single decision tree in the classification, and is not easy to cause excessive fitting phenomenon. The flow chart of the random forest algorithm is shown in Fig 1.

(1) The K sub training sample sets ($D_1,D_2,\ldots,D_k$) is selected from the total training sample set D through Bootstrap sampling, establish K decision trees.

(2) Select m randomly from each of the N indexes on each node of the classification tree, according to the principle of minimum node purity, the best feature is selected from the M candidate indexes, and the nodes are classified, the decision tree is fully grown until the purity (GINI index) of each leaf node is minimal, and the decision tree is pruned.

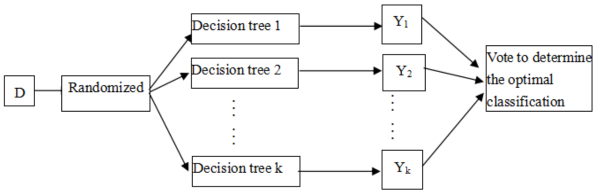

Fig. 1. Process of random forest algorithm

(3) Repeat step (2) to create k decision trees, from this k decision trees to create a random forest.

(4) A well-grown k decision trees are formed by a random forest combinatorial classifier to predict new unknown samples. The final classification of the sample waiting for the test is based on the majority of the voting results of the random forest combination classifier. The formula for its classification is:

$$f(x_t) = majority \quad vote\{h_i(x)\} \text{ (i=1,2,…,k)} \qquad (3)$$

Where: Majority vote is the result of a majority vote.

IV. EXAMPLE APPLICATION

A. Establishment of early warning model of bloom based on random forest

The outbreak of water blooms is a result of a multi-factor integrated effect, and its influencing factors are very numerous and complex. The water quality data selected by the Institute of Science and Technology from the Chaohu Lake half of the blue algae blooms online monitoring base station. Cited indicators include chlorophyll a (Chl-a) [10], water temperature (T), PH, nitrogen phosphorus ratio (TN:TP), chemical oxygen demand (COD), total nitrogen (TN), total phosphorus (TP), dissolved oxygen (DO), light (E) and other water quality, meteorological factors. According to the international and national literature defined Chl-a concentration threshold of 0.003 mg / L, when the Chl-a concentration of more than 0.003 mg / L, it means that the outbreak of bloom may need to make preventive work, less than 0.003 mg / L when the water environment in good condition, water bloom is unlikely to break out.

Based on the above-mentioned influencing factors, we can construct the water bloom warning model with random forest. Enter the measured data set of influencing factors of bloom, and then set the number of random forest trees [11] and the number of variables to train the model to form the corresponding random forest. Here the number of trees to select the first 100, the number of variables 3. Water bloom warning model of the establishment of the process shown in Figure 2. The number of decision trees is also the construction of the decision tree in the forest is the core of the model establishment, each decision tree is the maximum growth, no pruning. Which affects the operation speed and classification effect of stochastic forest classification algorithm. Therefore, the number of decision trees is very important for the establishment of water bloom warning model.

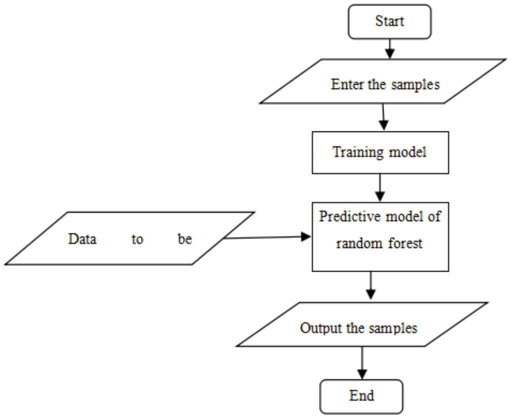

Fig. 2. The operational process of random forest forecasting model

B. Evaluation of model performance

In order to evaluate the classification performance of the random forest model comprehensively and comprehensively, this paper chooses the overall classification accuracy (ACC) index to evaluate the classification performance of the model. Since the overall classification accuracy is likely due to the occasional nature of the selected training samples and test samples. In order to minimize the uncertainty of the effect of the selected training samples on the results, ten sub-cross validation methods were used to analyze the neural network model, SVM model and RF model classification ability.

(1) The ACC index indicates the degree of difference between the predicted value and the true value of the model. The larger the value, the stronger the classification ability of the model used, the formula is as follows:

$$A_{CC} = T_{P} / T_{N} \tag{4}$$

T_P is the number of samples correctly classified; T_N is the total number of samples.

(2) Ten ten cross-validation is the idea of all the data randomly divided into 10 equal parts, respectively, D_1, D_2 ... D_{10}, rotate 9 of them as training data, leaving 1 as the test data to calculate the classification accuracy of the model. Repeat the test, taking the average of the accuracy of 10 results as a result of a cross validation test. Repeat the above steps 10 times, and then seek the mean to comprehensively evaluate the classification of the model. The 108 sets of data were selected from all the data as input samples to construct the stochastic forest prediction model, and the remaining 12 sets of data were used to verify the accuracy of the prediction model. Respectively, through the random forest, neural network, SVM vector machine to establish water blooms warning model and through the ten times the cross validation to get specific predictions in Table 1.

TABLEI. STATISTICS OF CROSS VALIDATION'S RESULT

Subset	Predictive accuracy Of RF model(%)	Predictive accuracy of SVM model （%）	Predictive accuracy of neural network model（%）
D1	91.67	91.67	83.33
D2	83.33	75.00	83.33
D3	91.67	83.33	91.67
D4	83.33	83.33	75.00
D5	91.67	91.67	91.67
D6	100	83.33	75.00
D7	91.67	83.33	91.67
D8	83.33	91.67	83.33
D9	75.00	91.67	66.67
D10	83.33	83.33	66.67
Average	87.48	85.83	71.66

The results of cross validation show that the accuracy of the predictive model of random forest blooms reaches 87.48%, the highest among the three models. The accuracy of SVM bloom prediction model is 85.83%. The accuracy of neural network prediction model is only 71.66%.

V. CONCLUSION

The results of the final test show that the predictive model established by stochastic forest algorithm solves the problems of insufficient robustness and learning of other algorithms, but also can ensure the accuracy of prediction and can also find an important factor affecting the outbreak of water bloom for the prediction of water outbreaks proposed an effective way for the hydrological monitoring department to prevent water outbreaks to provide an effective scientific basis.

REFERENCES

Adopting this sort of perspective will allow us to perceive living systems holistically. Put another way, living things are not the union of elements that they had been thought to be in a reductionist sense in conventional natural sciences. Rather, they are complex systems in which interactions occur between various elements at all levels from the micro to macro, and thus they display chaotic behavior on the whole. If we assume this, then the possibility arises that describing and controlling chaos dynamics as a universal principle that prescribes the chaotic nature of pulse waves and respiration could contribute to determining and controlling the status of living systems as a whole.

REFERENCES

[1] Wang Li, Gao Chong, Wang Xiaoyi and so on. Nonlinear dynamics analysis and water bloom prediction of cyanobacteria growth time variation system [J]. Journal of Chemical Industry and Engineering, 2017, 68 (3) :1065-1072.

[2] QIN B. The changing environment of Lake Taihu and itsecosystem responses[J]. Journal of Freshwater Ecology, 2015,30(1): 1-3.

[3] Jiao Licheng, Yang Shuyuan, Liu Fang. Seventy Years Beyond Neural Networks : Retrospect and Prospect [J]. Chinese Journal of Computers, 2016, 39 (8) :1697-1716.

[4] Li Haixin. Using"random forest"for classification and regression[J]. Chinese Journal of Applied Entomology, 2013, 50 (4) : 1190-1197.

[5] Dong Shishi, Huang Zhexue. A Brief Theoretical Overview of Random Forests[J]. Journal of Integration Technology, 2013, 2 (1) :1-7.

[6] Hu Tianyi, Dai Bo, He Qi. Slope Stability Forecasting Model Based on Radom Forest Classification Algorithm[J]. Yellow river, 2017, 39 (5) :115-118.

[7] Pan Dasheng, Qu Wanwen. An Improved ID3 Decision Tree Mining Algorithm[J].Journal of Huaqiao University(Natural Science) ,2016, 37 (1) :71-73.

[8] Xie Niuniu. A Survey of Decision Tree Algorithms[J].Software Guide,2015, 14 (11) :63-65.

[9] Zhang liang,Ning Qian. Two improvements on CART decision tree and its applocetion[J]. Computer Engineering and Design,2015, 36 (5): 1209-1213.

[10] Qi Lingyan, Huang Jicong. Spatial-temporal variation characteristics of chlorophyll-a concentration in Lake Hongze[J]. Journal of Lake Sciences, 2016, 28 (3):583-591.

[11] LIU Min, LANG Rongling, CAO Yongbin. Number of trees in random forest[J]. Computer Engineering and Applica-tions, 2015, 51 (5):126-131.