Advances in Economics, Business and Management Research, volume 203
Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)

Simplification of Water Quality Classification in Beijing, Tianjin and Hebei Based on K-nearest Neighbor Algorithm

ZIYI ZHU[1]*

[1]*Statistics and Mathematics institute, Central University of Finance andEconomics, Yunnan, 102206*
（zhuziyi0904@163.com）

ABSTRACT
This paper repeated experiments for many times, and simplified the original five water quality categories into two categories, based on the samples of water quality monitoring in Beijing Tianjin Hebei Haihe River Basin. Based on water quality classification, the three most obvious variables of stratification: turbidity, total phosphorus and total nitrogen are selected. The k-nearest neighbor method is used to predict the classification results of the test set, and the accuracy is about 75%.

Keywords: *k-nearest neighbor, water quality classification, water quality predict*

1. INTRODUCTION

According to the differences in physical, chemical and biological characteristics, water is divided into five categories in the environmental quality standard for surface water. [1]One to three types of water quality can be used for domestic drinking water, and its water quality is good or only slightly polluted; Class IV to V water quality is generally used in industries, agriculture and landscapes that are not in direct contact with the human body. Water exceeding class V water quality standards basically no longer has the function of use.

In this report, the real-time measured water quality results of Beijing, Tianjin and Hebei Haihe River Basins on July 1, 2021 are selected and analyzed by using k-nearest neighbor algorithm (also abbreviated as KNN algorithm).

2. Data Source and Description

The report data comes from China environmental monitoring station（http://www.cnemc.cn/）, including the real-time monitoring data of 73 monitoring stations in Beijing Tianjin Hebei Haihe River Basin on July 1, 2021. The original data includes 9 indicators: water temperature, pH, dissolved oxygen, conductivity, turbidity, ammonia nitrogen content, total phosphorus, total nitrogen and water quality category.

The water quality category is divided into five categories class I - V in the original data. According to the data: class I - III water quality can be used for domestic water with light pollution; class IV and V water pollution is serious, and can only be used in non human direct contact occasions. In order to simplify the model, the five categories are divided into two categories, of which class I - III water quality is recorded as class 1; class IV and V water quality is recorded as class 2. After the original data are combined by water quality categories, there are 50 class 1 samples and 23 class 2 samples.[2]

3. DATA ANALYSIS

The final determination of water quality category can not accurately judge the water quality category of the samples given 8 indicators if a simple regression method is used because of many indicators, the correlation between indicators, insignificant individual indicators and so on. Therefore, next, we will select variables and select appropriate methods for modeling.

3.1 Data feature display of original data set and selection of modeling indicators, methods

3.1.1 Data feature display

The category of water quality is finally determined based on the comprehensive analysis of water

temperature, pH, dissolved oxygen, conductivity, turbidity, ammonia nitrogen content, total phosphorus and total nitrogen. However, due to the correlation between variables, not every variable is highly correlated with the final classification results. Therefore, when any two variables are selected for scatter diagram and different colors are marked according to whether the water quality is class 1 or class 2, there is not obvious aggregation of points of the same color.

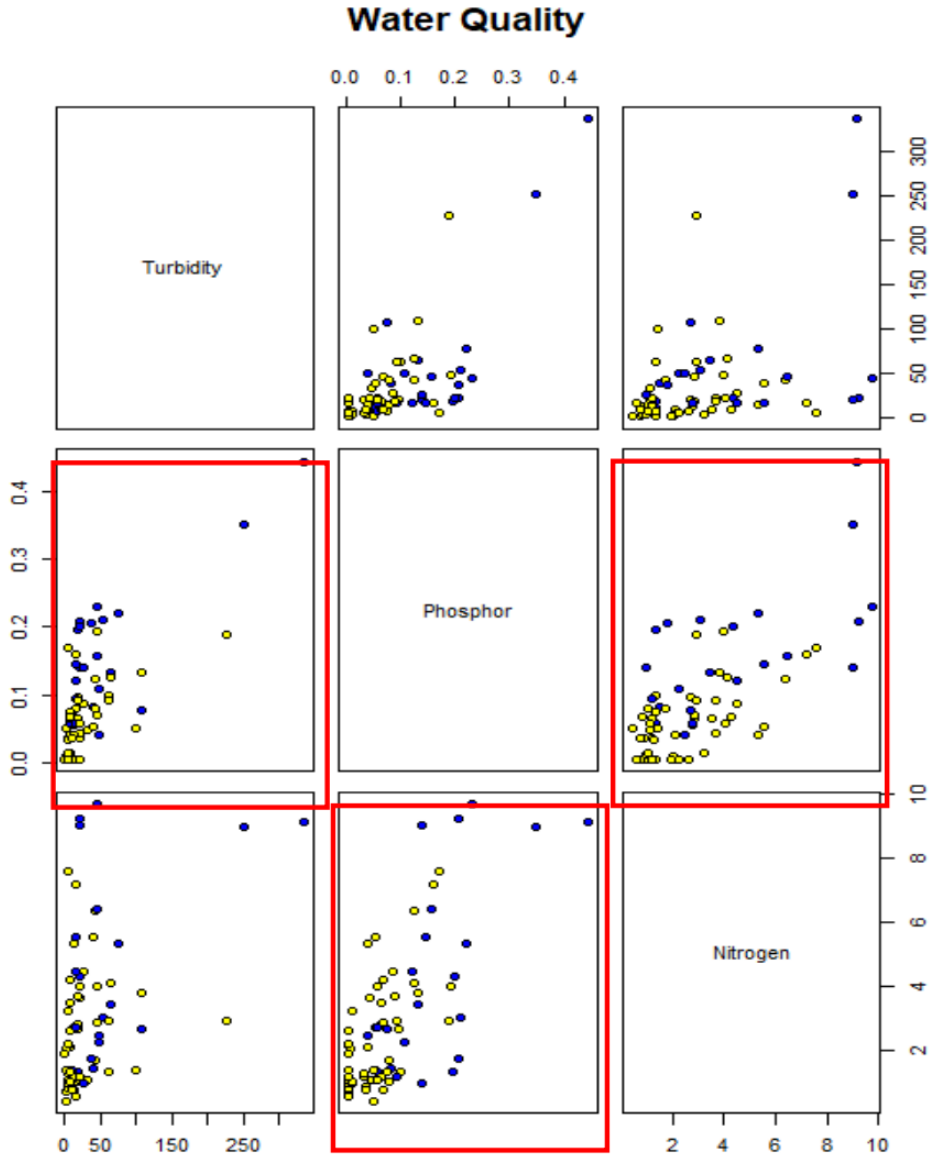

Figure 1 Scatter matrix of turbidity, total phosphorus and total nitrogen according to water quality category

Because the graph matrix of all variables is too large, and not all scatter diagrams between variables have obvious stratification, according to the observation, the three variables with the most obvious stratification are selected as the graph matrix. Figure 1 above shows the scatter diagram matrix of the three most obvious variables in the total variable diagram matrix -- turbidity, total phosphorus and total nitrogen. (yellow dots are class 1 and blue dots are class 2)

The scatter diagram with obvious stratification has been circled in red circle. It can be seen that in the scatter diagram with variable total phosphorus and total nitrogen as the horizontal and vertical axis and the scatter diagram with turbidity and total phosphorus as the horizontal and vertical axis, the sample points of class 2 water quality are gathered above or on the right side of class 1 water quality sample points, the stratification phenomenon is obvious, and there are obvious characteristics of cluster distribution.

3.1.2 Selection of modeling index and modeling method

Based on the observation of scatter matrix of all variables, the stratification phenomenon is obvious for total phosphorus, total nitrogen and turbidity. Therefore, in order to make the final classification result more accurate, only the three variables of total phosphorus,

total nitrogen and turbidity in the original data set are retained.

Due to the obvious hierarchical nature of data, it is very suitable to use k-nearest neighbor algorithm to analyze data. The idea of k-nearest neighbor algorithm is: after determining the number of nearest neighbors K to be compared, for a sample point, find the nearest K samples around it, observe the classification of the nearest K samples, and take the most categories as the categories of samples to be determined.

3.2 Basic assumptions of the model

Basic assumption: Based on the assumption of "congeneric attraction", the k-nearest neighbor algorithm believes that most of the sample points closest to a sample point belong to the category of the sample point. According to the stratification shown in the figure above, this assumption is applicable to this data set.

3.3 Model setting and analysis of model test results

3.3.1 Sample set extraction and parameter setting of model

Re select the data set of variables, the sample points with water quality category of 1 or 2 are scattered, and the original samples are randomly divided according to the sample size of test set and training set of 7:3. The final training set has 51 samples and the test set has 22 samples.[3]

According to the experience of k nearest neighbor theory, the value of K is generally selected according to the rounding of $K = \sqrt{n}$ (n is the sample size). Due to the large difference in value range among data variables, in order to make the influence of each variable on distance the same, the values of turbidity, total phosphorus and total nitrogen are standardized according to the formula: $(x - mean(x))/(\sqrt{var(x)})$.

In addition, in this model, the distance between samples uses the Euclidean distance.[4]

3.3.2 Analysis of prediction results of repeated extraction test set

Because the division of different training sets and test sets will lead to different accuracy of final prediction[5], I repeatedly extracted the test set and training set for 100 times to calculate the prediction efficiency of k-nearest neighbor model on the test set, and draw the broken line diagram as shown in Figure 2 below.

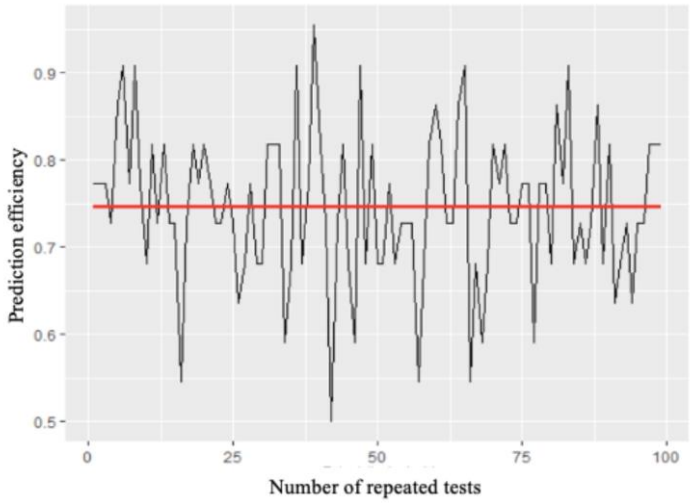

Figure 2 Line chart of predicted efficiency of repeated test

After calculation, the average prediction efficiency of the 100 experiments is 0.7470, that is, the position of the red line in Figure 2.

In general, the prediction efficiency of the classification prediction results of the selected turbidity, total phosphorus and total nitrogen by k-nearest neighbor method is more than 50%, and according to the above repeated tests, the probability prediction of nearly 75% is consistent with the actual classification results. Figure 3 below is a scatter diagram of the predicted value and the real value in the test set in an independent experiment. It can be seen that most of the predicted values are equal to the real value.

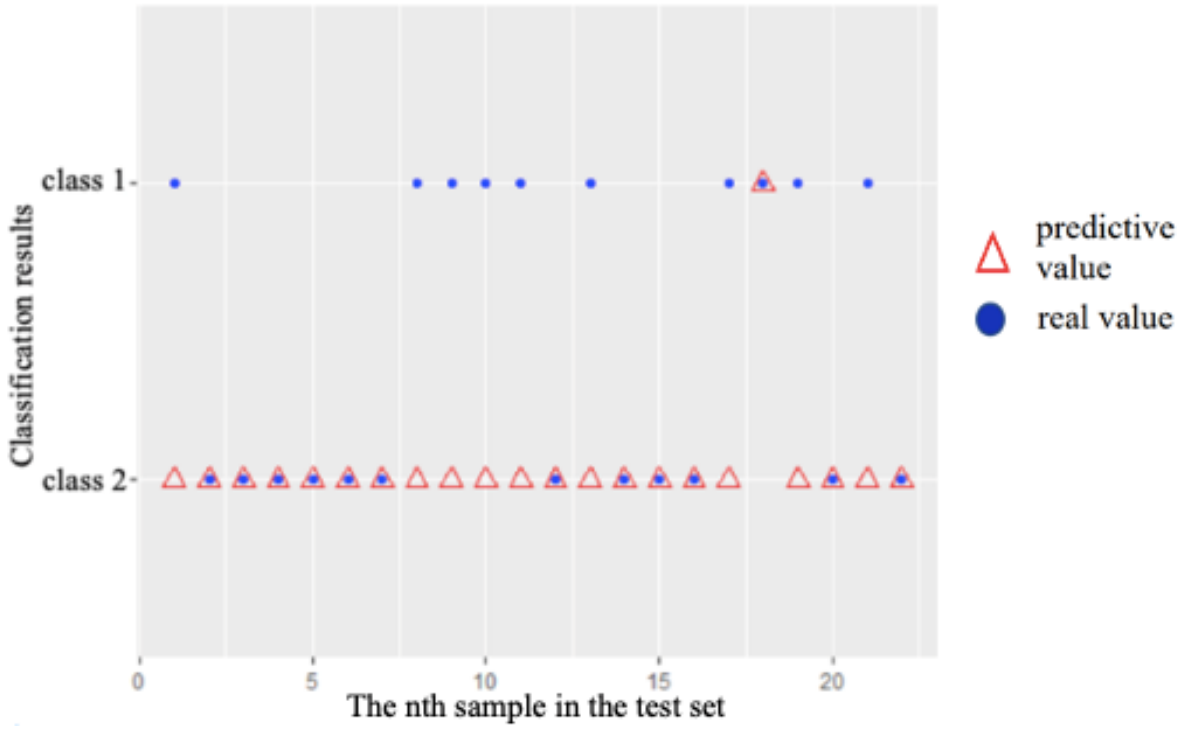

Figure 3 Scatter plot of predicted value and real value in an experiment

4. CONCLUSION

From the water quality monitoring samples of Beijing Tianjin Hebei Haihe River Basin on July 1, 2021, according to the results of the above repeated experiments, after merging and simplifying the original five water quality categories into two categories (recorded as class 1 and class 2) and selecting the three most obvious variables: turbidity, total phosphorus and total nitrogen, the k-nearest neighbor method is used to predict the classification results of the test set, and the accuracy is about 75%, The prediction is good.

However, there are still deficiencies in the establishment of this model: if the water quality classification method according to the five categories of the original data is only 73, due to the limited number of samples, the prediction effect of using k-nearest neighbor method is bound to be much worse than that of classifying it into two categories. In order to achieve the prediction effect divided into two categories in this paper, more samples are necessary.

In a word, compared with the eight variables in the original data set, only three variables of turbidity, total phosphorus and total nitrogen are used in this model to obtain good prediction results, which greatly simplifies the water quality evaluation method of the original data set. Therefore, when the water quality does not need to be carefully divided into five categories and can be simplified into two categories I - III and IV - V, the k-nearest neighbor algorithm is very simple and efficient.

REFERENCES

[1] Environmental quality standard for surface water(GB3838-2002),2002-4-26

[2] General Administration of Quality Supervision, Inspection and Quarantine Quality standard for ground water GB/T 14848-93 ,1993-12-30

[3] Tian Shuguang, song Yaolian. Research on improved KNN classification algorithm based on Gaussian function [J]. Data communication, 2021 (03): 39-43.

[4] Principle and implementation of KNN algorithm,https://www.cnblogs.com/sxron/p/5451923.html

[5] Machine learning - KNN algorithm https://www.cnblogs.com/ybjourney/p/4702562.html