Research on glass classification prediction model based on machine learning

Xuehan Peng [1,*], Xinyan Tong [2], Xinyi Hu [1]

[1] School of Statistics, Capital University of Economics and Business, Beijing, China,100070

[2.]School of Accounting, Capital University of Economics and Business, Beijing, China,100070

* Corresponding Author Email: 2994784085@qq.com

Abstract. Glass is an important commodity on the ancient Silk Road, which is of great value to the study of ancient Chinese and Western trade. In this paper, aiming at the classification of glass cultural relics, the glass type assessment classification model and the random forest model are constructed by analyzing the proportion data of chemical composition content of existing glass samples. This paper summarizes the classification rules of two types of glass by drawing the pie chart of the proportion of each chemical component of high potassium glass and lead barium glass, and combining with a large number of literatures on glass weathering: the glass with high potassium oxide content is judged as high potassium glass; The probability of higher proportion of lead oxide and barium oxide is lead barium glass. Then, this paper selects several chemical components that occupy the main position in the composition as the standard, divides the lead barium glass and high potassium glass into three different subcategories, establishes the glass type evaluation classification model, uses C language, adopts multiple cycle logic nesting, and conducts the main division of glass categories, then introduces Di parameter to further deepen the division of glass categories, and runs each group of programs, Specific classification results can be obtained and brought into the data to obtain classification results. Finally, the sensitivity and rationality of classification results are analyzed by constructing two-dimensional descriptive statistical images of variances and influence parameters of high potassium glass and lead barium glass respectively.

Keywords: Glass type evaluation algorithm stochastic forest machine learning model spearman correlation coefficient

1. Research background

Glass is a valuable material evidence to witness the trade exchanges of the early Silk Road. Since the introduction of Western Asia and other regions into China, glass manufacturing technology has been localized in China, and its chemical composition has also changed greatly.

Because of the different cosolvents added in the refining process, the chemical composition of the glass produced is also different, such as high potassium glass and lead barium glass. At the same time, ancient glass is easily weathered by the influence of burial environment. Weathering will change the composition proportion of glass, thus affecting the judgment of glass category. Therefore, it is of great significance to study the factors that will affect the weathering of glass, find more detailed glass classification standards, and explore the methods to identify glass.

2. Introduction

Now there are a number of data related to ancient glass products in China. Archaeologists have divided these cultural relics into two types: high potassium glass and lead barium glass according to their chemical composition and other detection methods. Among them, the basic information of glass relics, the proportion of chemical composition of classified glass relics and the proportion of chemical composition of unclassified glass relics are studied respectively.

The given data is characterized by composition, that is, the cumulative sum of all components should be 100%. However, the cumulative sum of data may not be 100% due to factors such as detection means. Therefore, the data between 85%~105% and the accumulation of component proportions are regarded as valid data.

3. Analyze the statistical law of chemical composition content

This paper analyzes the statistical law of chemical composition content in combination with glass types. Therefore, this paper first classifies the data in Table 2 into two categories: high potassium glass and lead barium glass, and then compares the statistical law of chemical composition of weathering and non weathering data in high potassium glass data.

For high potassium glass, a pie chart of the proportion of weathered and unweathered chemical components is drawn respectively. Analysis shows that the proportion of silicon dioxide before weathering accounts for 70%, while after weathering, the content of other chemical components except silicon dioxide decreases a lot, and the silicon dioxide content reaches more than 80%. Therefore, it can be judged that the silicon dioxide content exceeds.

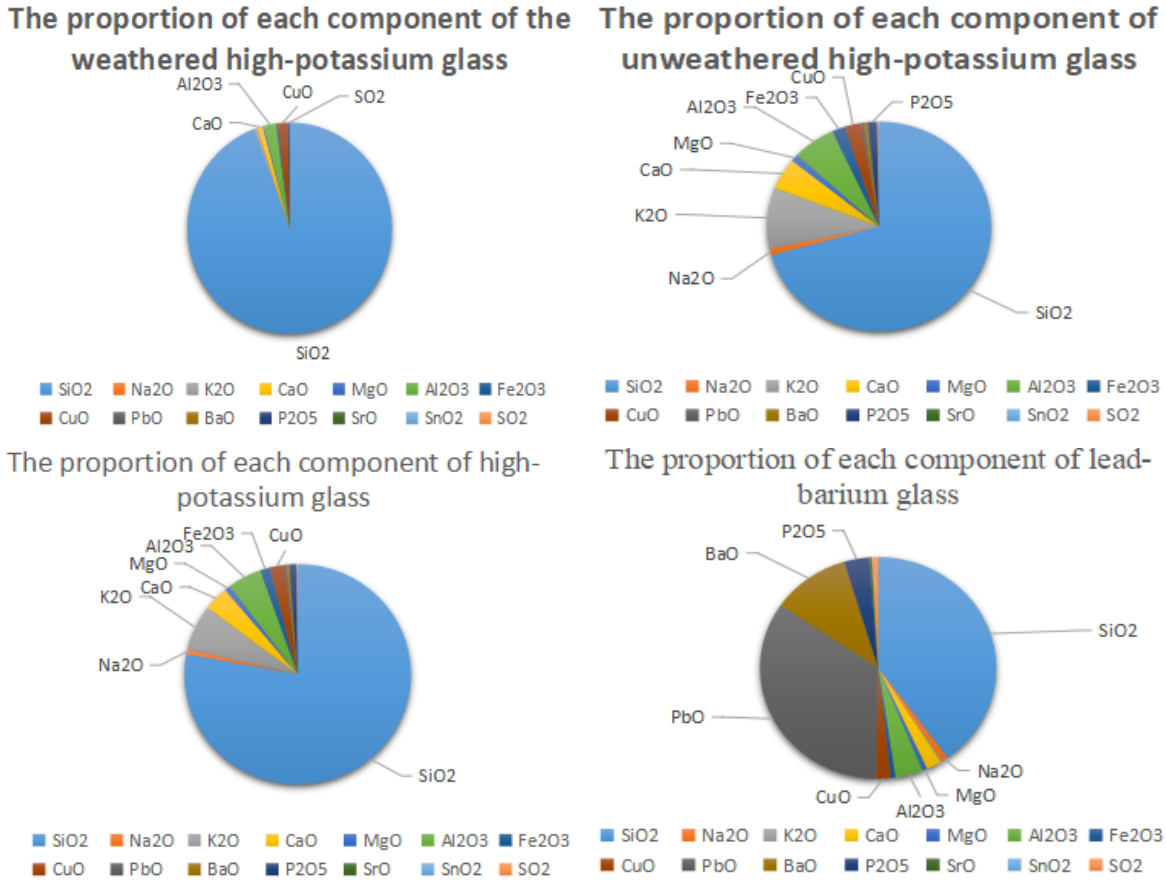


Fig 1. Classification of high potassium glass and lead barium glass

For high potassium glass and lead barium glass, the changes of each component before and after weathering were observed. Through observation, it can be found that the chemical composition of copper oxide is relatively stable for any type of glass. Therefore, take copper oxide as the observation point, compare the composition of copper oxide of glass before weathering with that of glass after weathering, find out the observation points with similar content of copper oxide before and after weathering, and match the data of similar observation points to predict the content of each chemical component of glass before weathering.

4. Analysis of the classification rules of high potassium glass and lead barium glass

This paper requires to analyze the classification rules of high potassium glass and lead barium glass. According to the chemical composition information of several known glass samples, the average proportions of each component of the two kinds of glass are calculated as the general proportions of such glass. Such data is representative and universal.

According to the data in this paper, pie charts of the proportion of each component of lead barium glass and high potassium glass are made, as shown in Figure 1 and Figure 2 below.

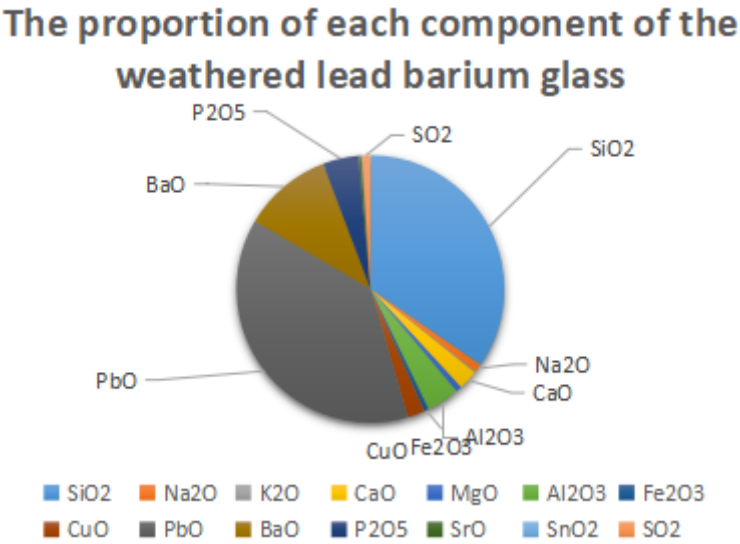


Fig 2. Proportion of each component of lead barium glass

The analysis shows that the main components of lead barium glass are silicon dioxide (40.41%), lead oxide (33.67%), barium oxide (10.63%), aluminum oxide (3.90%) and phosphorus pentoxide (3.41%).

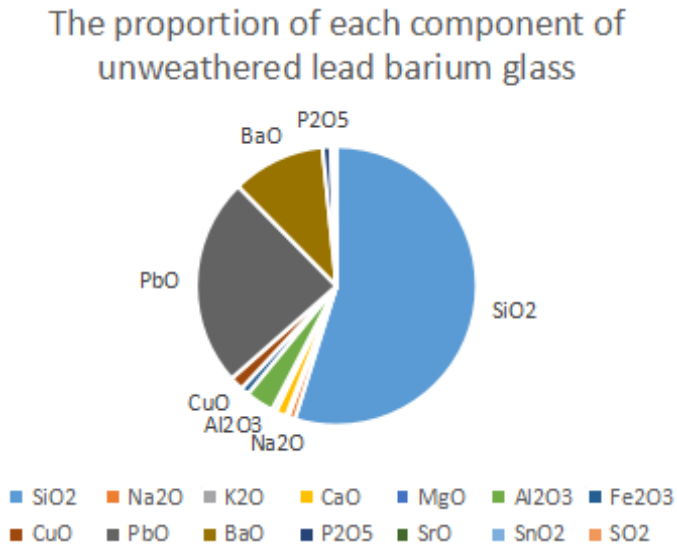


Fig 3. Proportion of components in high potassium glass

The main components of high potassium glass are silicon dioxide (77.95%), potassium oxide (6.66%), aluminum oxide (4.89%) and calcium oxide (3.59%).

By comparing the chemical composition proportion of lead barium glass and high potassium glass, the following analysis process can be obtained: as the main chemical composition of both types of glass (the silicon dioxide content of both types of glass exceeds 40%), silicon dioxide cannot be used to distinguish the type of glass. Therefore, continue to analyze other chemical components with high proportion. The average proportion of lead oxide (33.67%) and barium oxide (10.63%) in lead barium glass is very high, while the average proportion of lead oxide and barium oxide in high potassium glass is less than 1%.

The content of potassium oxide (6.66%) in high potassium glass is relatively high, while the average content of potassium oxide in lead barium glass is less than 0.5%. From this analysis, it can be seen that in the chemical composition, the glass with a high proportion of lead oxide and barium

oxide is probably lead barium glass, while the glass with a high proportion of potassium oxide is probably high potassium glass[1].

5. Study on the classification of glass subclasses

As for the classification of glass subcategories, if the number of glass in each category is not different, it means that the classification effect is ideal. Therefore, this paper uses the variance of the number of cultural relics sampling points (respectively set as d, e, f) that are classified into three categories as a way to evaluate the advantages and disadvantages of the classification of glass subcategories. Therefore, the classification results of this question can be analyzed for their sensitivity and rationality by using excel to construct two-dimensional descriptive statistical images of the order variance of sampling points and influence parameters of high potassium glass and lead barium glass respectively[2].

Taking high potassium glass as an example, three mean value parameters a, b, c are calculated. When the constant a changes, the classification of glass cultural relics will also change. When the number of glass cultural relics in the three categories d, e, f is not different, it is considered that this classification is relatively successful. Since variance is a variable that reflects the volatility and stability of data, this paper considers building a two-dimensional relationship between array variance and a, b, c variables. Since a, b, c are fixed values, it is not possible to directly build a two-dimensional relationship model as independent variables. Therefore, the control variable method was used to construct three groups.

Case 1: array b and c are fixed values, a (i, 1) is independent variable, the initial value of a is taken as the center point, 0.1 is taken as the distance between groups, and 10 sample points are selected. Then, the fixed value of a is extended to the independent variable interval of [a-0.1 * 5, a+0.1 * 5] centered on a, that is, a increases in length by 0.1 from a-0.5 to a+0.5. On this basis, calculate the variance of each row of data according to the new array a, b, c, as shown in the following formula.

$$S_i^2 = \frac{\left(a(i,1)-M_i\right)^2+\left(b-M_i\right)^2+\left(c-M_i\right)^2}{3}$$
$$M_i = \frac{a(i,1)+b+c}{3} \tag{1}$$

Where, a (i, j) represents the first column of data in row i, b (i, 2) represents the second column of data in row i, c (i, 3) represents the third column of data in row i, and Mi represents the mean value of a (i, j), b, and c.

After the above calculation, establish a two-dimensional line chart of variance array and array a.

Case 2: array a and c are fixed values, b is independent variable, initial b value is taken as the center point, 0.1 is taken as the distance between groups, and 10 sample points are selected. Then, the fixed value of b is extended to the independent variable interval of [b-0.1 * 5, b+0.1 * 5] centered on b, that is, b increases in length by 0.1 from b-0.5 to b+0.5. On this basis, the variance of each row of data is calculated according to the new array a, b, c, and the process of variance calculation is completely consistent with that of case I. After the above calculation, establish a two-dimensional line chart of variance array and b array.

$$S_i^2 = \frac{\left(a-M_i\right)^2+\left(b(i,2)-M_i\right)^2+\left(c-M_i\right)^2}{3}$$
$$M_i = \frac{a+b(i,2)+c}{3} \tag{2}$$

Case 3: array a and b are fixed values, c is independent variable, initial c value is taken as the center point, 0.1 is taken as the distance between groups, and 10 sample points are selected. Then, the

constant value of c is extended to the independent variable interval of [c-0.1 * 5, c+0.1 * 5] centered on c, that is, c increases in length by 0.1 from c-0.5 to c+0.5. On this basis, the variance of each row of data is calculated according to the new array a, b, c, and the process of variance calculation is completely consistent with that of case I. After the above calculation, a two-dimensional line chart of variance array and c array is established.

$$S_i^2 = \frac{(a-M_i)^2 + (b-M_i)^2 + (c(i,3)-M_i)^2}{3}$$
$$M_i = \frac{a+b+c(i,3)}{3} \tag{3}$$

Based on the above situation, the above three two-dimensional broken line diagrams of high potassium glass are summarized, and the sensitivity analysis diagram is as follows:

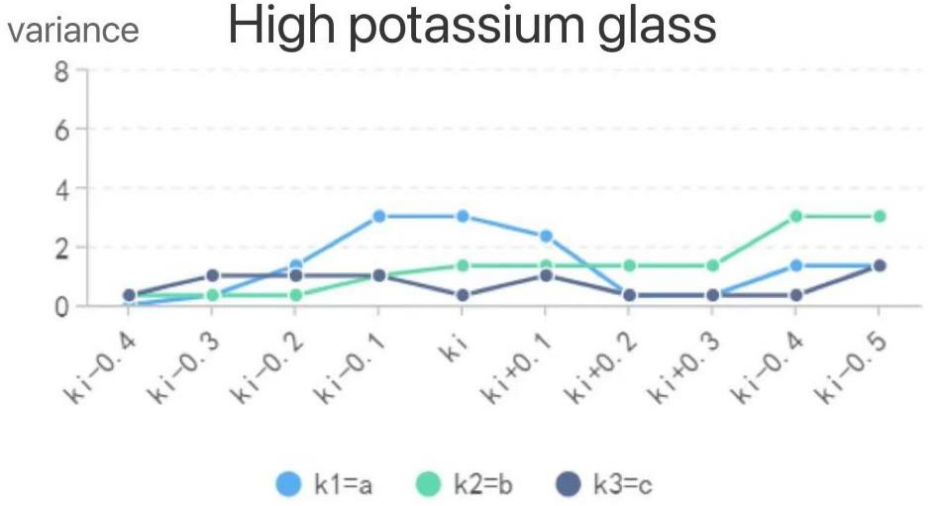


Fig 4. Sensitivity analysis of high potassium glass to parameters a, b and c

Similarly, a two-dimensional line chart of lead barium glass based on a, b, c parameter changes can also be made. The sensitivity analysis chart of lead barium glass can be obtained by summarizing the three two-dimensional line charts as follows:

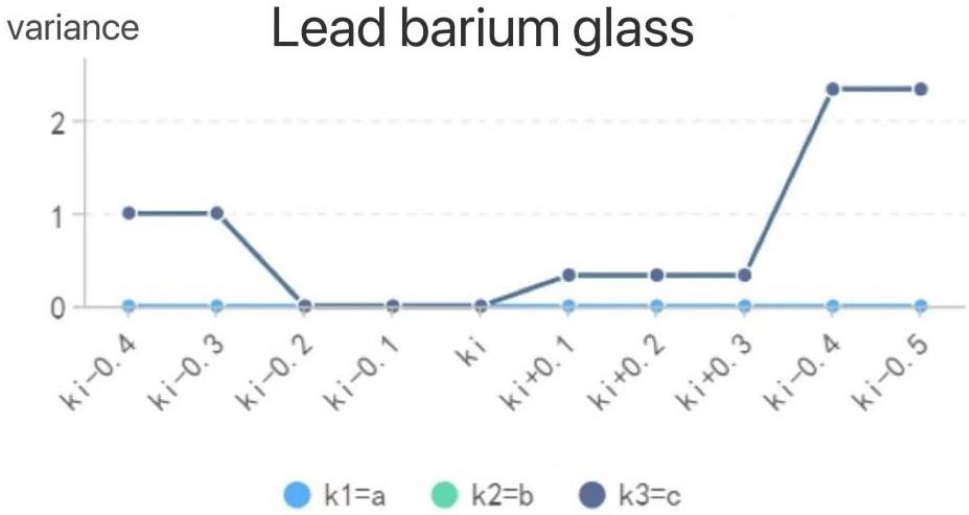


Fig 5. Sensitivity analysis of lead barium glass to parameters a, b and c

From the analysis of the above two figures, it can be seen that the change of the broken line chart is relatively stable, and the variance is close to 0. When parameter a,b. When c changes, the variance of the number of glasses classified into the three major categories is still very small, indicating that

when the parameters change, the classification model is not greatly affected, and the model is still widely used.

6. Analysis of chemical composition of unknown glass relics

To analyze the chemical composition of unknown glass relics It can identify the type of glass according to the proportion of chemical composition. Form 2 provides the chemical composition proportion of the classified glass relics. We can use this as a learning template to build a classification model based on machine learning methods to judge the glass type in Form 3.

In this paper, the random forest model is mainly used for classification prediction, and the decision tree and XGBoost integrated learning models are compared to verify the accuracy of the random forest model to the glass type prediction results in Table 3, as well as the sensitivity of the random forest model. The following article will introduce the basic principles of the three models used in this question.

Decision tree: A basic method for classification and regression. For different types of response variables, different decision trees can be established to solve the problem. For discrete response variables, the decision tree is also called the classification tree, and the splitting rule uses the Gini index. For continuous response variables, the decision tree can be called the regression tree, and the splitting rule used is the square error. This paper mainly discusses the classification of glass types, using a classification tree.

Stochastic forest: a natural nonlinear modeling tool of statistical learning theory. It uses Bootstrap resampling method to extract multiple samples from the original samples, model the decision tree for each Bootstrap sample, then combine the predictions of multiple decision trees, and finally obtain the prediction results through voting. Random forest can significantly improve the accuracy of individual decision tree, because the composite model reduces the variance of individual model. Therefore, this question selects the random forest model as the main model of classification prediction, and the decision tree and XGBoost as the comparison model.

6.1. Establishment of stochastic forest model

In this paper, we use python programming to establish a random forest model[3], randomly select 70% of the data as the training set, and the remaining 30% of the data as the test set. Since the random forest is composed of multiple decision trees, the final classification results are also voted after independent judgment of multiple trees. The random sampling with placement can ensure that the training sets of each decision tree are different, and avoid that the classification results of the finally trained trees are identical[4].

After training multiple decision trees with randomly selected training sets, the established random forest classification model is applied to the test set to obtain the classification evaluation results of the model and calculate the accuracy of the model. The results are shown in Figure 2 below. There are four evaluation indexes for random forest model.

Accuracy: predict the proportion of correct samples in the total samples. The greater the accuracy, the better.

Recall rate: in the results that are actually positive samples, the predicted proportion is positive samples. The larger the recall rate, the better.

Accuracy: the proportion of predicted positive sample results that are actually positive samples. The greater the accuracy, the better.

F1: The harmonic average of the accuracy rate and the recall rate. The accuracy rate and the recall rate affect each other. Although both are high, it is an expected ideal situation. However, in practice, the accuracy rate is often high, the recall rate is low, or the recall rate is low, but the accuracy rate is high. If you need to give consideration to both, you can use the F1 indicator.

Table 1. Evaluation Results of Random Forest Model

	Accuracy	recall	Accuracy	F1
Training Set	1	1	1	1
Test Set	1	1	1	1

It can be seen from Table 1 that the accuracy, recall, accuracy and F1 of the constructed random forest model in the test set are all 1, indicating that the model has a high accuracy and consistency in judging the problem. Therefore, the chemical composition proportion of the unknown glass relics in Table 3 is brought into this random forest model to judge the type.

6.2. Model solution

Table 2. Prediction and Evaluation Results of Random Forest Model Test Data

Cultural relic number	Forecast results_ Y	Probability of prediction results_ Lead barium	Probability of prediction results_ hyperkalemia
A1	hyperkalemia	0.32	0.68
A2	Lead barium	0.84	0.16
A3	Lead barium	0.88	0.12
A4	Lead barium	0.85	0.15
A5	Lead barium	0.69	0.31
A6	hyperkalemia	0.25	0.75
A7	hyperkalemia	0.31	0.69
A8	Lead barium	0.88	0.12

6.3. Sensitivity analysis

This question requires an analysis of the sensitivity of the classification results, and the purpose of the sensitivity analysis is to evaluate the accuracy of the analysis results. Other research methods can be used to analyze the same data to see whether the results change and the extent of change, so as to explain the stability of the research results.

For the research classification model, this paper plans to use two machine learning models, decision tree[5] and XGBoost[6], to analyze and predict the data, and then verify the accuracy of the model. After that, it is actually applied to the type identification of the unknown glass relics in Form 3, and compared with the classification results of the previously constructed random forest model, to investigate the differences in the final classification results, so as to complete the sensitivity analysis.

According to the classification prediction of Table 3 of the decision tree and XGBoost models, it is found that the predicted classification results are completely consistent with the random forest. The important chemical components calculated by the random forest model are silicon dioxide, potassium oxide, lead oxide and barium oxide, as shown in the figure below. The importance characteristics calculated by the decision tree[7] and XGBoost[8] are also similar to the random forest. Therefore, we should prove that the random forest model used in this question has a very high prediction accuracy and accurate calculation.

7. Conclusion

This paper has independently created a glass category evaluation and classification model. Using C language as the programming language, it can not only be applicable to the screening of a small amount of data, but also has significant advantages in processing a large amount of data. It is universal and innovative, and it can solve the operation of classification by relatively simple means

This paper innovatively applies the random forest method of machine learning to the field of identification of glass cultural relics, and uses decision trees and XGBoost model as the assistant of accuracy identification, which enhances the persuasiveness of the results. At the same time, in the practical application as a means of identification, it does not need too much data processing, directly input relevant data, reduce the interference of human factors, and directly judge the category by the model. This paper chooses random forest model instead of decision tree as the main classification model, because random forest has higher prediction accuracy and accuracy than decision tree, which is also one of the advantages of the model.

References

[1] Yang Xin Glass defect detection and classification based on machine vision [D]. Fujian Institute of Engineering, 2021. DOI: 10.27865/d.cnki.gfgxy.2021.000021

[2] Wang Yumei Classification definition and code of glass fiber [C]//. Collection of propaganda and implementation meeting of glass fiber product standards and industrial inorganic fiber technology seminar, 2019:9-20. DOI: 10.26914/c.cnkihy.2019.068339

[3] Cheng Shaochi, You Guangrong. Research on evaluation index reduction method based on improved random forest algorithm [J/OL]. System Engineering and Electronic Technology: 1-9 [2022-09-18] http://kns.cnki.net/kcms/detail/11.2422.TN.20220907.1648.004.h tml

[4] Zhang Ziyue. Research on the Credit Problem of Small, Medium and Micro Enterprises Based on Random Forest [J]. China Collective Economy, 2022 (24): 104-106

[5] Wang Jingxiang. Principle Research and Practical Application of Decision Tree Algorithm [J]. Computer Programming Skills and Maintenance, 2022 (08): 54-56+72. DOI: 10.16184/j.cnki.comprg.2022.08.043

[6] Xu Yi, Zhen Jianing, Jiang Xiapeng, Wang Junjie. UAV remote sensing and XGBoost mangrove species classification [J]. Journal of Remote Sensing, 2021,25 (03): 737-752

[7] Liu Yang. Inductive classification and application of decision tree in hierarchical teaching[J]. Journal of Huaibei Vocational and Technical College,2022,21(04):70-72+112.DOI: 10.16279/j.cnki.cn34-1214/z.2022.04.012.

[8] Wang Kunzhang, Jiang Shubo, Zhang Hao, Chao Zheng. Regression Classification Regression Life Prediction Model Based on XGBoost[J/OL]. Testing in China:1-8[2022-10-17]. http://kns.cnki.net/kcms/detail/51.1714.TB.20220616.1537.002.html

[9] Si Shoukui, Sun Xijing. Mathematical modeling algorithm and application [M] Beijing: National Defense Industry Press, 2016-2

[10] Cui Rumai, Si Shoukui. Mathematical Model of Data Visualization Processing [J]. Journal of Naval Academy of Aeronautical Engineering, 2010, 25 (01): 105-108+112