Study on the Composition Analysis and Identification of Ancient Glass Products

Shirui Li*

College of physics, Sichuan University, Chengdu, China

*Corresponding author: shirui.li@foxmail.com

Abstract. The Silk Road was an important channel for cultural exchanges between ancient China and the West, and glass was a precious material evidence of early trade exchanges. Early glass was often made into ornaments and introduced into China. After absorbing its technology, ancient Chinese glass was made locally, similar to its appearance but different in chemical composition. Due to different fluxes added during refining, the glass can be divided into lead barium glass and potassium glass, ancient glass is easily weathered by burial environment. In the weathering process, due to a large amount of exchange between internal elements and external elements, the proportion of their components changes, which affects the correct judgment of their categories. The research on the relevant properties of ancient cultural relics is a major topic in the field of archaeology in China, which is of great significance. There are a number of relevant data on ancient glass products in China. Archaeologists have divided them into two types: high potassium glass and lead barium glass.In order to explore the relevant properties and chemical composition changes of these glass relics before and after weathering, this paper analyzes the classification rules of high potassium glass and lead barium glass; The classification results were analyzed for rationality and sensitivity.

Key words: Correlation analysis, Clustering algorithm, T test, Random forest.

1. Introduction

This paper analyzes the classification rules of high potassium glass and lead barium glass based on data, that is, analyzes the characteristics of grain decoration, color and chemical composition of high potassium glass and lead barium glass, and explores their classification rules. Random forest algorithm can be used to explore the importance of each component in classification and its rules. Each category is further classified by selecting appropriate chemical components for subclassing. The k-means algorithm in cluster analysis can be used for sub classification, and rationality and sensitivity analysis can be carried out.To identify the unknown glass types, we can first identify the glass types to which they belong by using the classification rules and the classification framework of subclasses, and then predict their subclasses. After classifying the original data, add category labels for reverse training, establish decision tree and random forest classification learning models respectively, classify the unknown category glass, and carry out sensitivity analysis on the model after adding noise to the given data.

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

2.1. theoretical analysis

The classification rules of high potassium glass and lead barium glass are analyzed by using the relevant data of chemical composition content. Firstly, the random forest algorithm is adopted. The random forest algorithm [1] is an integrated algorithm composed of multiple decision trees, which is used for classification prediction. It belongs to supervised learning, and its output category is determined by the mode of each tree output category.

2.2. Establishment of model

56 samples (80% of the data) were randomly selected for training, and the remaining 13 samples (20% of the data) were used for testing. With the component content as the variable and different types of glass as the dependent variable, the bootstrap self-help method is used to extract n training samples from the original data, and then n trees are established. These n trees constitute a random forest for comprehensive data discrimination and classification; During the tree building process, m variables are randomly selected from all variables at the nodes of each tree. With these m variables, the variables with the strongest classification ability are selected for data classification. The remaining unexampled data in Bootstrap is the test sample, which is mainly used to verify the performance of each tree, and to draw the ROC curve to verify the established model. The proportion of importance of each component in the classification is as shown in Fig 1:

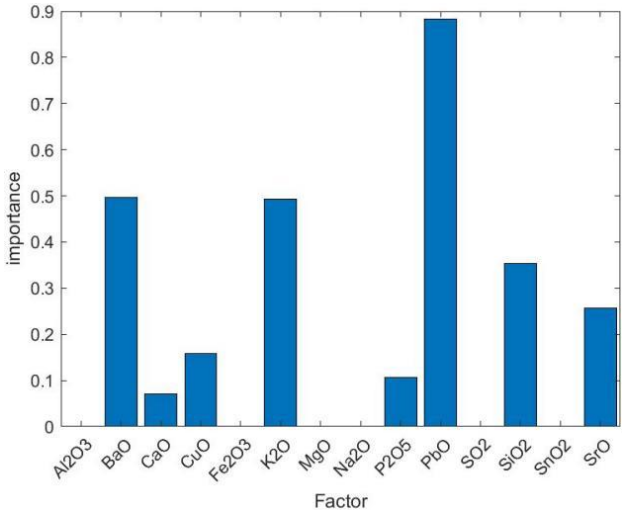

Fig. 1 Importance of each component

Run the model on matlab, import the test data into the model for prediction and classification, obtain the optimal number and number of leaves, and compare the prediction model with the actual probability as shown in Fig 2 and Fig 3 :

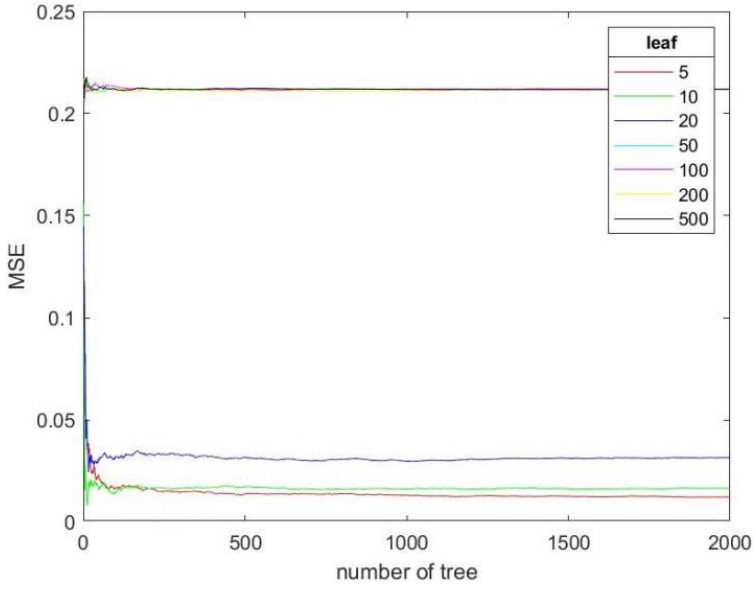

Fig. 2 Number of optimal numbers and number of leaves

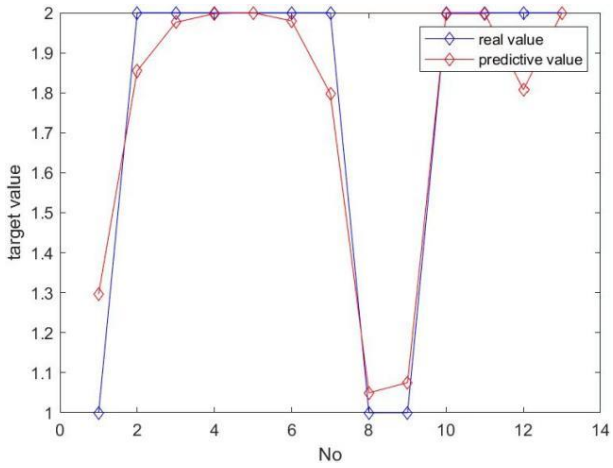

Fig. 3 Comparison between prediction model and actual probability

It can be seen from the above images that lead oxide is the most important when classifying high potassium glass and lead barium glass, followed by sodium oxide, potassium oxide, silicon dioxide and strontium oxide. After importing the prediction data, the prediction accuracy is high.

2.3. Sub classification of different categories

2.3.1 theoretical analysis

In order to select appropriate chemical components for the classification of the two categories, it is necessary to analyze the content rule of chemical components in each category to further classify them. Because it is necessary to further classify them according to their chemical composition, cluster analysis should be considered first. During cluster analysis, data objects are grouped according to the information describing objects and their relationships found in the data [2]. The greater the correlation among the groups, and the greater the gap between the groups, the better the clustering effect. In this problem, we want to sub divide the glass of the same type, and the description object is the data of the same type of glass. The purpose of grouping them is to divide them into new categories according to the content law of chemical components.

2.3.2 Model establishment

The k-means clustering algorithm is an iterative clustering analysis algorithm. Its steps are: pre divide the data into K groups, then randomly select K objects as the initial clustering center, then calculate the distance between each object and each sub cluster center, and assign each object to the nearest clustering center. Each time a sample is allocated, the cluster center of the cluster will be recalculated according to the cluster object, and the process will be repeated until the termination condition is met.

Input data are recorded as high potassium data set and lead barium data set respectively. Randomly select k groups of data as the center point O, and calculate the Euclidean distance from the remaining data to each center point O, whose expression is:

$$\mathrm{d}(x,y)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2} \tag{1}$$

For each group of data, compare and find the center point corresponding to the minimum Euclidean distance of the group of data, and classify the data into the group represented by the center point:

$$\mathrm{d}(x,c_i)=\frac{1}{2} \tag{2}$$

Where: c_i is the cluster center of the i th cluster, and x is the data object in the high potassium dataset or lead barium dataset.

For the data in each cluster, select a group of data in the cluster in turn, calculate its Euclidean distance to the remaining data in the group, and calculate the Euclidean distance from the center point to all data in the cluster. Compare the two sizes. If the Euclidean distance of the data in the cluster is less, the data will be used as the new center point O. For the high potassium dataset and lead barium dataset of group m, calculate the Euclidean distance from the center point O, and substitute it into the cycle, until the center point does not change, stop iteration.

Since the K-means algorithm requires that the data be artificially divided into k working conditions, the optimal k value [3] is calculated by using the index of error square sum SSE. The calculation formula of SSE is:

$$SSE = \sum_{i=1}^{k} \sum_{n \subset c_i} \left| d\left(x, c_i\right) \right|^2 \tag{3}$$

Where: C_i is the ith cluster; C_i is the cluster center of cluster C_i; X is the data object of the high potassium data set or the lead barium data set; K is the number of clusters; SSE is the sum of Euclidean distance squares from the data in the cluster to the center point[4].

With the increase of cluster number k, the degree of aggregation of each cluster data gradually increases, and the SSE gradually decreases. When k is less than the true number of clusters, the increase of k has a greater impact on the degree of aggregation of each type, so the decline of SSE is very obvious. When k reaches the true number of clusters, the increase of k has a smaller impact on the degree of aggregation of each cluster, so the decline of SSE is not obvious. Therefore, the best cluster number k is selected according to the trend of the decline of SSE. The k values of high potassium and lead barium were determined by comparing the results of multiple fitting as shown in Fig. 4, 5, 6 and 7.

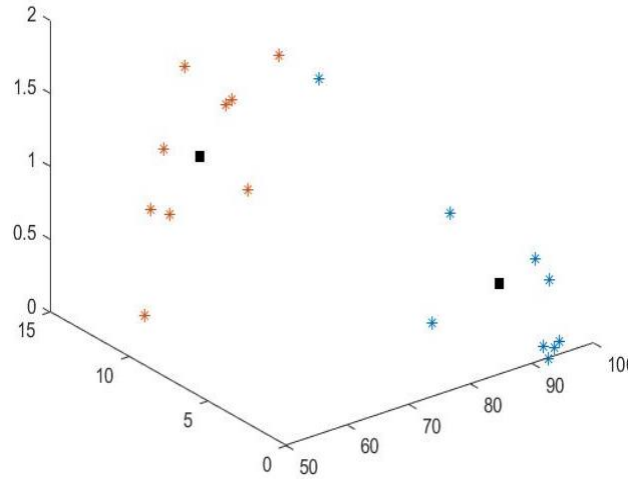

Fig. 4 The k value of high potassium is 2

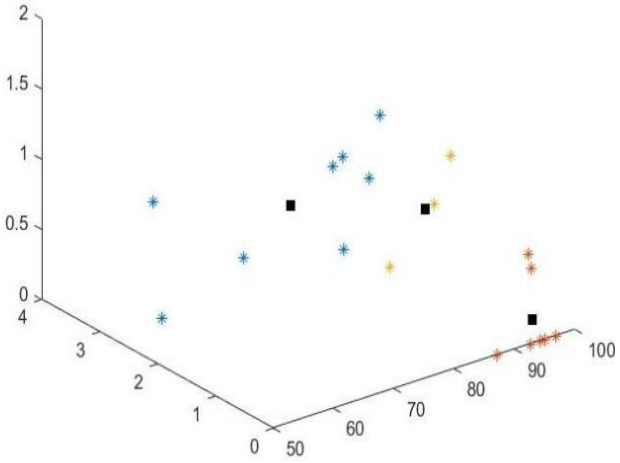

Fig. 5 The k value of high potassium is 3

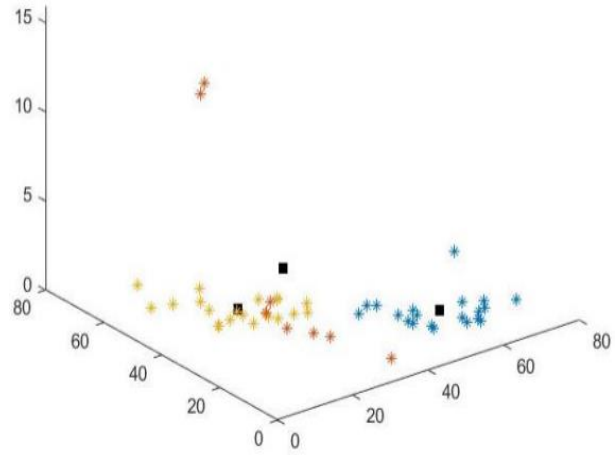

Fig. 6 K value of lead barium is 3

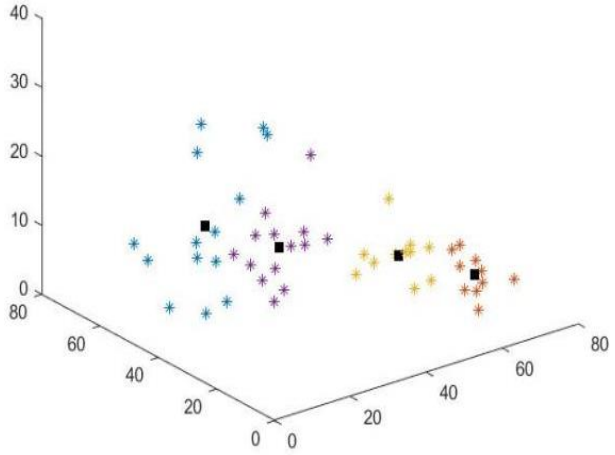

Fig. 7 K value of lead barium is 4

The above images were observed through multiple test experiments. Points of the same color represent the same kind, the same color points gather, and different colors disperse, so the correlation is large. It can be seen that for high potassium glass, when the k value is 2, the correlation between the groups is large, but the gap between groups is large. For lead barium glass, when the k value is 4, the correlation between groups is large, but the gap between groups is large. Therefore, the k value of high potassium is determined as 2, and that of lead barium is determined as 4. That is to say, the high potassium category is divided into two categories, and the lead barium category is divided into four categories.

2.3.3 Solution of model

In matlab, write K-means model code. All the components of high potassium glass were introduced to cluster them, and the k value was set as 2. The significance p value of each component and cluster is calculated by spss, and the following table is obtained to observe the table and compare the p value of each chemical component and cluster category[5]. The smaller the p value, the greater the correlation between the component and the clustering category. Therefore, we selected the three chemical components (silica, calcium oxide, aluminum oxide) with the lowest p value to cluster again. The table is as shown in Table 1:

Table 1. Re clustering results of high potassium

Clustering category	silicon dioxide	calcium oxide	alumina
1	69.33	6.32	3.93
2	87.05	2.01	4.06
1	61.71	5.87	5.5
1	65.88	7.12	6.44
1	61.58	7.35	7.5
1	67.65	0	11.15
1	59.81	5.41	10.05
2	92.63	1.07	1.98
2	95.02	0.62	1.32
2	96.77	0.21	0.81
2	94.29	0.72	1.46
1	59.01	8.7	6.16
1	62.47	8.23	9.23
1	65.18	8.27	6.18
2	79.46	0	3.05

By observing the content comparison of each chemical component of the two cluster species, they are divided into two categories. Type 1 refers to the relatively low content of silicon dioxide and

relatively high content of calcium oxide and aluminum oxide, so it is named as high silicon potassium. Type 2 refers to the relatively high content of silicon dioxide and relatively low content of calcium oxide and aluminum oxide, named as high calcium aluminum potassium.

Similarly, all the components of lead barium glass are introduced to cluster them, and the k value is set as 4. The significance p value of each component and cluster is calculated by spss, and the following table (see appendix) is obtained to observe the table and compare the p value of each chemical component and cluster category[6]. The smaller the p value, the greater the correlation between the component and the clustering category. Therefore, we selected the three chemical components (silicon dioxide, lead oxide and barium oxide) with the lowest p value to cluster again. The table (part) is as shown in Table 2:

Table 2. Re clustering results of lead and barium

Clustering category	Q6-(SiO2)	Q14-(PbO)	Q15-(BaO)
3	36.28	47.43	0
4	20.14	28.68	31.23
4	4.61	32.45	30.62
3	33.59	25.39	14.61
4	31.94	29.14	26.23
2	60.12	17.24	10.34
3	32.93	49.31	9.79
1	26.25	61.03	7.22
1	16.71	70.21	6.69
3	18.46	44.12	9.76
2	51.26	21.88	10.47
2	51.33	20.12	10.88
1	12.41	59.85	7.29
3	21.7	44.75	3.26

By observing the content comparison of each chemical component of the two cluster species, they are divided into four categories. Type 1 shows relatively high content of lead oxide, type 2 shows relatively high content of silicon dioxide, type 3 shows relatively low content of silicon dioxide, relatively high content of lead oxide, and type 4 shows relatively high content of barium oxide. These four types are named as high lead, high silicon, high silicon, low lead and high barium in the lead barium category.

3. Rationality and sensitivity analysis

3.1. Rationality analysis

The K-means algorithm of clustering division is used in the sub classification of each type. Firstly, the k-value of each category is determined by comparing the clustering effects for many times, and then all components of the two types are clustered, and the significant P-value of each chemical component on the clustering effect is analyzed to determine the chemical components that have a major impact on the clustering effect. Then the principal components are clustered to get the clustering results, so as to divide the subtypes of each category. In the division of subclasses, we have reasonable theoretical basis and reliable mathematical calculation. The scatter plots of the final clustering results are shown in Figure 5 and Figure 10, with large differences between groups and high similarity within groups. The classification of multiple results is consistent, indicating that the clustering results are more reasonable.

3.2. sensitivity analysis

Change the influence of clustering parameters on the results, add small noise to the original data to obtain a new partition result and compare it with the original result, and analyze its sensitivity[7].

First, the two types of original data are processed with random noise in a small range, and then the noise processing sensitivity analysis is conducted for the subcategory of the two types. The results are visualized and the following line graph is obtained as shown in Fig 8 and Fig 9:

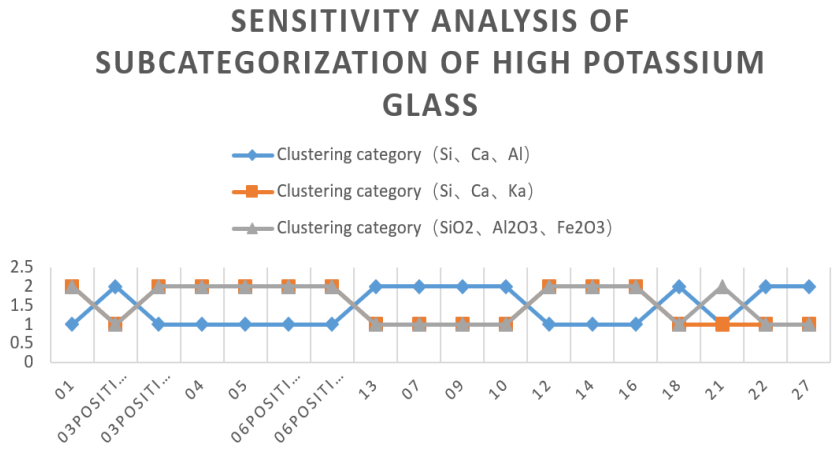

Fig. 8 Sensitivity Analysis of High Potassium Glass Subcategorization

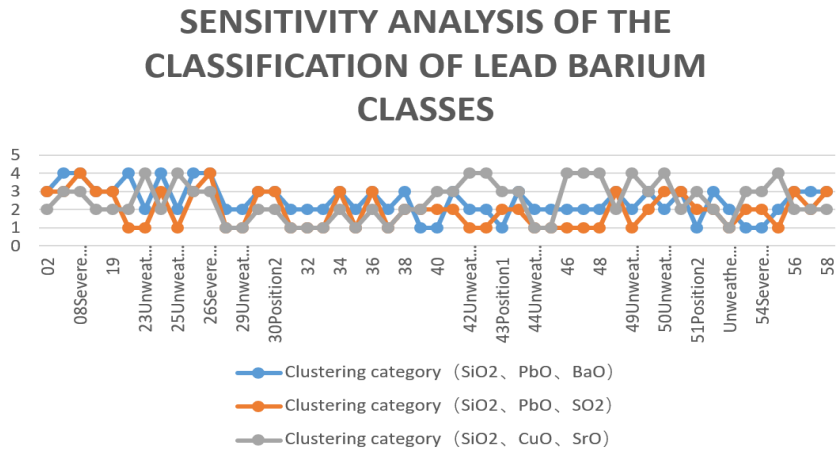

Fig. 9 Sensitivity analysis of lead barium glass subcategory

According to the above broken line diagram, the classification sensitivity of lead barium and high potassium is good. Moreover, the sensitivity of classification of high potassium glass and lead barium glass is higher.

4. Predict the type of glass relics of unknown category

To identify the types of glass relics of unknown categories, it is necessary to analyze their chemical composition. First, predict whether the unknown cultural relics belong to high barium or lead barium, and then classify them into subclasses. In 1.1, the random forest algorithm is used to analyze the chemical composition of the two glass types and explore their classification rules[8]. After importing the test data, the discrimination accuracy is high, so when identifying the type of unknown glass, the data in Table 3 is directly imported into the random forest model trained in 1.2, and the prediction results of its type are as follows:

Table 3. Identification Results of Unknown Cultural Relics

Cultural relic number	Forecast results_ Y	Probability of prediction results_ Lead barium	Probability of prediction results_ hyperkalemia
A1	hyperkalemia	0.15	0.85
A2	Lead barium	0.89	0.11
A3	Lead barium	0.89	0.11
A4	Lead barium	0.9	0.1
A5	Lead barium	0.82	0.18
A6	hyperkalemia	0.03	0.97
A7	hyperkalemia	0.05	0.95
A8	Lead barium	0.99	0.01

It can be seen from the above table that cultural relics A1, A6 and A7 are identified as high potassium type, while other cultural relics are identified as lead barium type. The probabilities of the two prediction results are greater than 0.8, indicating that the reliability of the prediction results is high and the accuracy of type prediction is high.

4.1. Further classify the types of unknown cultural relics

Through the identification of unknown cultural relics in the previous section, we can divide these cultural relics into high potassium and lead barium[9]. In this paper, K-means clustering algorithm is used to sub divide high potassium and lead barium. In order to further divide the unknown cultural relics, the chemical composition data of cultural relics identified as high potassium and lead barium are imported into the model trained in 1.3.3 to identify their sub categories.

Import high-K cultural relics into the high-K clustering algorithm model to obtain the following table 4 :

Table 4. Prediction of Subcategory of High Potassium Cultural Relics

Cultural relic number	Forecast results_ Y	Probability of prediction results_ 1	Probability of prediction results_ 2	silicon dioxide	alumina	calcium oxide
A1	1	0.9	0.1	78.45	7.23	6.08
A6	2	0	1	93.17	1.52	0.64
A7	2	0.11	0.89	90.83	5.06	1.12

It can be seen from the above table that cultural relic A1 belongs to the type with low silica content and high alumina and calcium oxide content. Cultural relic A6 and A7 belong to the type with high silica content and low alumina and calcium oxide content[10]. Moreover, the probability difference of prediction results is large, and the prediction accuracy is high.

Lead barium cultural relics are introduced into the high potassium clustering algorithm model, and the following tables are obtained through reverse training of classification labels and decision trees as shown in Table 5:

Table 5. Prediction of sub categories of lead barium cultural relics

Cultural relic number	Forecast results_ Y	silicon dioxide	Lead oxide	Barium oxide
A2	4	37.75	34.3	0
A3	3	31.95	39.58	4.69
A4	4	35.47	24.28	8.31
A5	2	64.29	12.23	2.16
A8	2	51.12	21.24	11.34

According to the above table, cultural relics A2 and A4 belong to type 4 of lead barium, cultural relics A3 belong to type 3 of lead barium, and cultural relics A5 and A8 belong to type 2 of lead barium.

4.2. sensitivity analysis

Use the same method in 1.3 to process the noise of data, and obtain the following line chart:

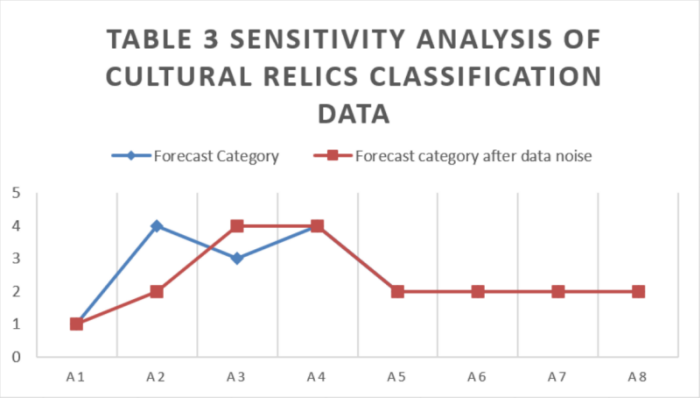

Fig. 10 Sensitivity Analysis of Unknown Cultural Relics Classification Data

It can be seen from the above figure that the prediction results are highly sensitive.

5. Conclusion

In the classification of high potassium glass and lead barium glass, lead oxide is the most important, followed by sodium oxide, potassium oxide, silicon dioxide and strontium oxide. After importing the prediction data, the prediction accuracy is high. Therefore, the glass can be divided into high potassium glass or lead barium glass according to the content of lead oxide, sodium oxide, potassium oxide, silicon dioxide and strontium oxide. Through many tests and experiments, observe Fig 8-Fig 10. Points of the same color represent the same kind, the same color points gather, and different colors disperse, so the correlation is large. It can be seen that for high potassium glass, when the k value is 2, the correlation between the groups is large, but the gap between groups is large. For lead barium glass, when the k value is 4, the correlation between groups is large, but the gap between groups is large. Therefore, the k value of high potassium is determined as 2, and that of lead barium is determined as 4. That is to say, the high potassium category is divided into two categories, and the lead barium category is divided into four categories. According to the results of cluster analysis, the classification of this subclass is reasonable and sensitive.

Import the data into the random forest model. Table 5 shows that the identification results of cultural relics A1, A6 and A7 are high potassium type, while other cultural relics are lead barium type. The probabilities of the two prediction results are greater than 0.8, indicating that the reliability of the prediction results is high and the accuracy of type prediction is high. Then classify them into subcategories and introduce K-means clustering algorithm. We can get that cultural relic A1 belongs to the type with low silica content and high alumina and calcium oxide content. Cultural relic A6 and A7 belong to the type with high silica content and low alumina and calcium oxide content. Moreover, the probability difference of prediction results is large, and the prediction accuracy is high. Cultural relics A2 and A4 belong to lead barium type 4, cultural relics A3 belong to lead barium type 3, and cultural relics A5 and A8 belong to lead barium type 2.

References

[1] Wang Bo, Zhuang Jijun, Xiong Jun, Luo Xiaochen. Research on stochastic forest algorithm based on high-dimensional feature clustering optimization [J]. Journal of Jinggangshan University (Natural Science Edition), 2022,43 (05): 52-56

[2] Sui Yong. Comprehensive evaluation of wheat inferior flour quality based on principal component and cluster analysis [J]. Modern Flour Industry, 2022,36 (02): 56

[3] Spring flowers K-means clustering research based on swarm intelligence algorithm [D]. Dalian University of Technology, 2019. DOI: 10.26991/d.cnki.gdllu.2019.003511

[4] Gong Yunqing, Tong Lichen, Yu Ze, Zhang Xu. Research on the method of rotating machinery misalignment fault diagnosis based on Pearson correlation coefficient [J]. China New Technology and New Product, 2022 (05): 48-50. DOI: 10.13612/j.cnki.cntp.2022.05.013

[5] Li Jiakang, Tao Zhilin, Xu Bo, et al Quantitative analysis of tobacco leaf texture based on random forest [J] Hubei Agricultural Science, 2022,61 (14): 155-159. DOI: 10.14088/j.cnki.issn0439-8114.2022.14.028

[6] Li Ruiguang, Duan Pengyu, Shen Meng, et al Traffic classification algorithm of Internet of Things devices based on random forest [J] Journal of Beijing University of Aeronautics and Astronautics, 2022,48 (2): 233-239. DOI: 10.13700/j.bh.1001-5965.2020.0383

[7] Wang Bo, Zhuang Jijun, Xiong Jun, Luo Xiaochen. Research on stochastic forest algorithm based on high-dimensional feature clustering optimization [J]. Journal of Jinggangshan University (Natural Science Edition), 2022,43 (05): 52-56

[8] Spring flowers K-means clustering research based on swarm intelligence algorithm [D]. Dalian University of Technology, 2019. DOI: 10.26991/d.cnki.gdllu.2019.003511

[9] Gong Yunqing, Tong Lichen, Yu Ze, Zhang Xu. Research on the method of rotating machinery misalignment fault diagnosis based on Pearson correlation coefficient [J]. China New Technology and New Product, 2022 (05): 48-50. DOI: 10.13612/j.cnki.cntp.2022.05.013

[10] Sui Yong. Comprehensive evaluation of wheat inferior flour quality based on principal component and cluster analysis [J]. Modern Flour Industry, 2022,36 (02): 56