Classification of ancient glassware based on K-means and decision trees

Jing Wang *, Kehan Chen

School of Communication, Dalian University of Technology, Dalian, China

* Corresponding author: wj122023@163.com

Abstract. In this paper, the relation between the chemical composition of different types of glass objects was analyzed, and the differences between the chemical compositions were used to classify the glass objects, mainly by K-means, decision tree, and Euclidean distance analysis. For the classification of glass artifacts, this article analyzed the chemical composition of different types of glass artifacts and the changes of chemical composition before and after weathering, and then obtained the differences of chemical composition of different types of glass artifacts to obtain the classification law. The model was tested for reasonableness by giving the classification criteria, and the sensitivity of the model was determined by the change in the number of classification species K.The sensitivity of the model was analyzed by calculating the Euclidean distances between the object and the cluster centers, and the variance of the model by calculating the Euclidean distances between the object and the cluster centers before and after the weathering of the glass objects.

Keywords: K-means, Decision Trees, Ancient Glassware, Chemical Composition.

1. Introduction

Ancient glassware is an important physical material for exploring the economic, technological and cultural exchanges between China and foreign countries along the Silk Road. The glassware of China has obvious characteristics of the times in terms of shape, production process, chemical composition, and distribution area. Although Chinese lead-barium glass and Western soda-lime glass" are similar in appearance, their main chemical compositions differed during refining because of the different fluxes added. The surface of glass and glass products came into contact with water and air, and then underwent a series of complex physical and chemical changes, resulting in changes in shape, color, chemical composition, etc., thus affect ing the correct judgment of archaeologists on their categories.

Archaeologists have classified these artifacts into two types of glass: high potassium glass and lead-barium glass, based on their chemical composition and other testing methods. In this paper, we use K-means elbow rule to determine the optimal number of subclasses for high potassium glass and lead-barium glass, and then use K-means to obtain the clustering results, based on which we use the decision tree algorithm to derive the ranking sign and classification tree for each chemical composition in the subclasses, and to justify the classification accuracy of the verification set. The analysis was carried out. The clustering centers of high potassium and lead-barium glass types before and after weathering were calculated according to K-means, and these were used as indicators for different types. The Euclidean distance is constructed between the target artifacts to be tested and the feature indicators of different categories, and the matching degree of the target artifacts to be tested is calculated, and the highest matching degree is the category of the target artifacts to be tested. According to the difference of matching degree between different target objects and different types of feature indicators, the sensitivity change of classification results is obtained.

2. Development and solution of a classification model for chemical composition of glass artifacts based on K-means and decision trees

2.1. Glass artifacts classification law

Based on the characteristic values of each chemical composition in different types of glass artifacts before and after weathering, it is assumed that the percentage of the ith chemical composition in different types of glass artifacts is yi′ The corresponding percentage of the ith chemical composition in another type of glass artifact is yi That is [1]:

$$ei=\frac{|yi'-yi|}{ymax} \tag{1}$$

Where (i=1, · · · ,n), ymax denotes the quantity that takes the largest value among yi and yi′, ei denotes the variability of the chemical element in the ith, and substituting the data into equation (1) yields

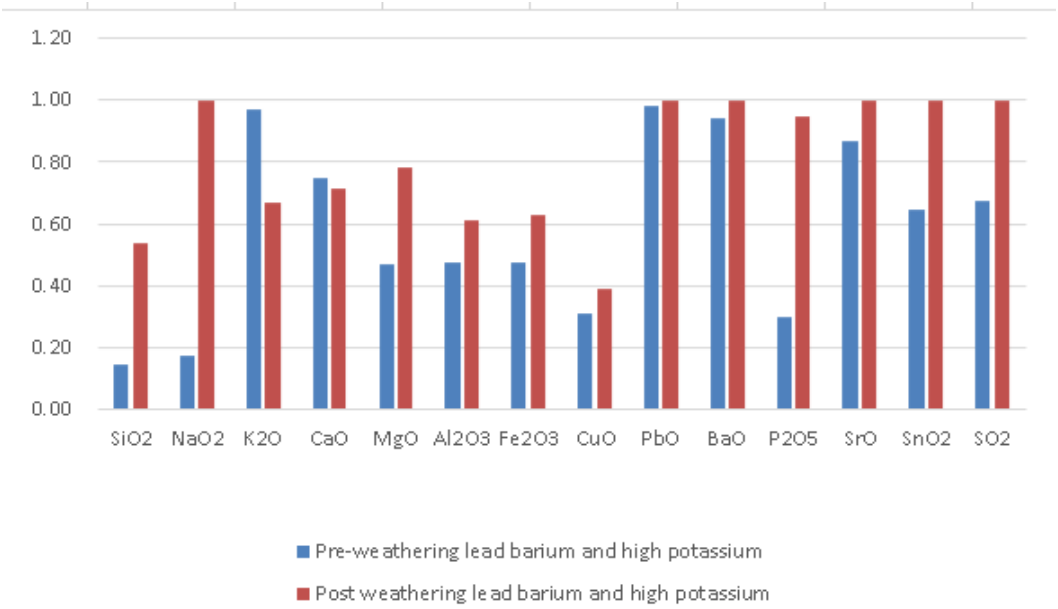

Figure 1. Difference of high potassium and Pb barium chemical elements before and after weathering

Figure 1 visually shows the transformation of the content of each chemical component of high potassium and lead-barium glass before and after weathering, and it can be concluded that there is a significant difference in the content of potassium oxide, lead oxide and barium oxide in high potassium glass and lead-barium glass before weathering, so we can distinguish two different kinds of glass by the three elements of potassium, barium and lead, and those containing more potassium are classified as high potassium glass, and those containing more lead and barium are classified as lead-barium glass; after weathering, high potassium glass will no longer contain chemical components such as lead oxide and barium oxide, so the two elements of lead and barium can still be used as the basis for classification.

2.2. K-means based analysis: sub-classification of glass types

On this basis the glass artifacts were divided into high potassium and lead-barium categories according to their types, and the magnitude of variability was calculated by substituting each chemical composition of the two types of glass artifacts before and after weathering into equation (1):

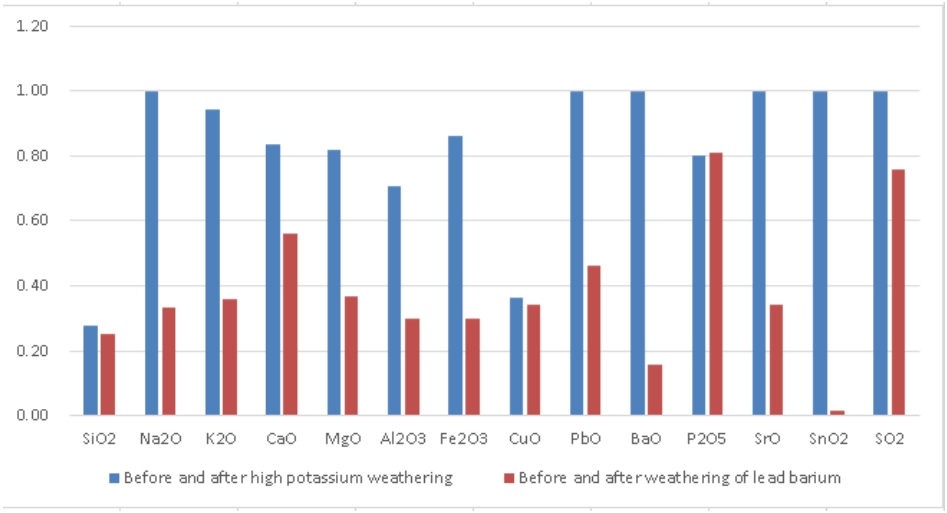

Figure 2. Difference of high potassium and Pb barium chemical elements before and after weathering

From the data in Figure 2, we can get the chemical composition of high potassium glass and lead-barium glass before and after weathering, and we can get the chemical composition with large variation, and when the difference of chemical composition is large, it means that the index has certain classification significance, so we can choose several elements with large variation to classify, and we can get different classification results.

In order to obtain the results of subclassing among glass artifacts of the same type, the K-means clustering algorithm was used to explore the results.According to[2] it is known that the K-means algorithm uses distance as a similarity index, so as to discover K classes in a given dataset and the center of each class is obtained based on the mean of all values in the class and each class is described by a cluster center.For a given dataset X containing n d-dimensional data points and the classes K to be divided, the Euclidean distance is chosen as the similarity metric and the clustering goal is to make the sum of squares of the clusters of each class minimum [3]:

$$J=\sum_{j=1}^{k}\sum_{i=1}^{n}|Xi-Uk| \tag{2}$$

The basic process is:

(1) K objects in the data space are selected as the initial centers, and each object represents a cluster center.

(2) For the data objects in the sample, according to their Euclidean distances from these clustering centers, they are assigned to the classes corresponding to the clustering centers closest to them according to the criterion of closest distance.

(3) updating the clustering centers, taking the mean value corresponding to all objects in each class as the clustering center of that class, and calculating the value of the objective function.

(4) judge whether the values of the clustering center and the objective function have changed, and output the result if they remain unchanged and return to step (2) if they have changed.

In order to better enable K-means to sub-classify high potassium and lead-barium glass, K we use the elbow rule to determine the K value for classification:

$$SSE=\sum_{i=1}^{k}\sum_{p\in C}|p-mi|^2 \tag{3}$$

Where, Ci is the ith class, p is all sample points in Ci class, and mi is the center of mass of Ci class (the mean of all samples in Ci). The core idea of the elbow method is that the larger the classification number K, the finer the sample division will be, the degree of aggregation of each class will gradually increase, and then the error squared, and SSE will naturally become smaller gradually[4]. Sensitivity analysis is performed on it, and the effect is:

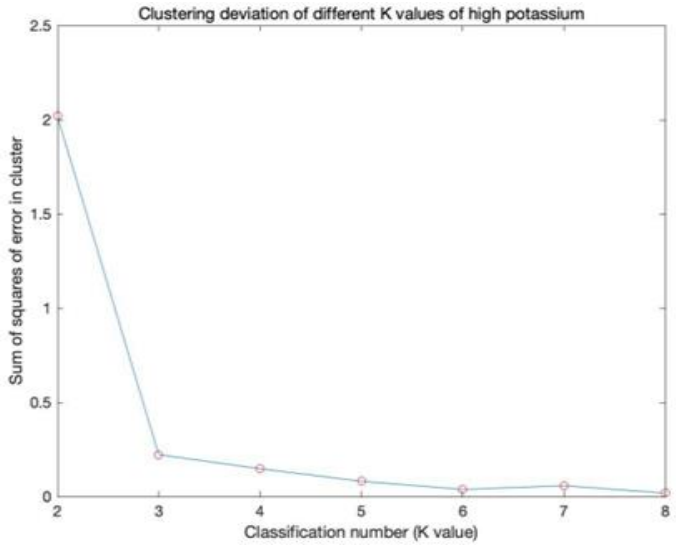

Figure 3. Clustering deviation of different K values of high potassium

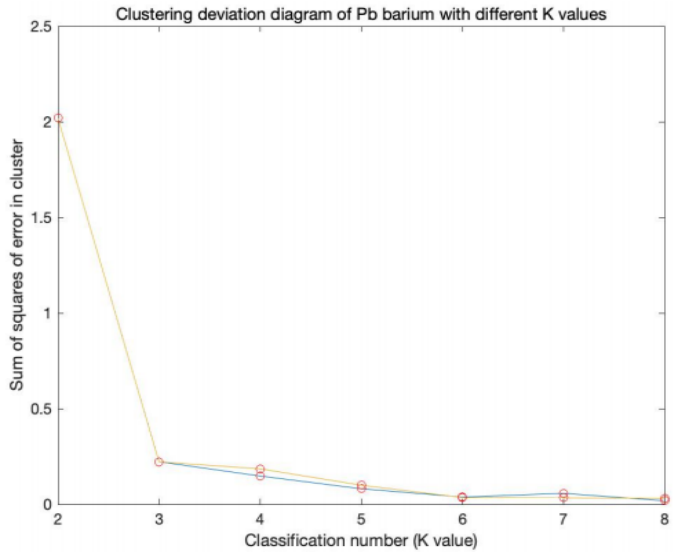

Figure 4. Clustering deviation diagram of Pb barium with different K value

In Figure 3, we can see that the SSE of high potassium starts to level off when the number of species K is 3. In Figure 4, the SSE starts to level off when the number of species K is 4. It is finally determined that the classification is better when high potassium is classified into 3 categories and lead-barium is classified into 4 categories. We used K-means to classify high potassium into 3 subclasses and lead-barium into 4 subclasses.

2.3. Decision tree based analysis

In order to derive the classification method for specific chemical elements and the rationality and sensitivity of the classification results, we used the decision tree algorithm of SPSS. According to [5] it is known that decision tree algorithm is a common classification prediction method in data mining techniques. Its decision process starts from the root node, tests the corresponding feature attributes in the item to be classified, and selects the output branches according to their values until it reaches the leaf nodes, and takes the category stored

in the leaf nodes as the decision result. In this paper, the ID3 algorithm is used to select the attribute with the largest value of information gain in the sample set as the test attribute, as each non-leaf node. Let S be the sample set, si ϵ S (i=1,2,3...,s), and the category attribute is Ci (i=1,2,3...,m). Assuming that si is the number of samples in category Ci, the information entropy contained in this set S is[6]:

$$\text{Entropy (S)} = -\sum_{i=1}^{m} \text{pi log2 pi} \tag{4}$$

Where pi=si/S, is the probability that any one data object belongs to Ci.Let the samples in set S be delimited by attribute A.There are k different values in attribute A.Then the conditional entropy of attribute A to sample set S is:

$$\text{Entropy A (S)} = \sum_{i=1}^{k} \left|\frac{Si}{S}\right| Enteopy(s) \tag{5}$$

Where |si| and |S| are the number of samples contained in Si and S, respectively.Using attribute A to delineate the sample set S, the information gain Gain(A) is:

$$\text{Gain (A)} = \text{Entropy (S)} - \text{Entropy A (S)} \tag{6}$$

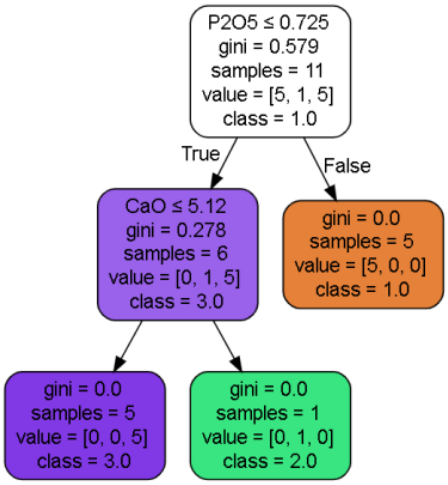

Figure 5. Classification structure of high potas sium decision tree

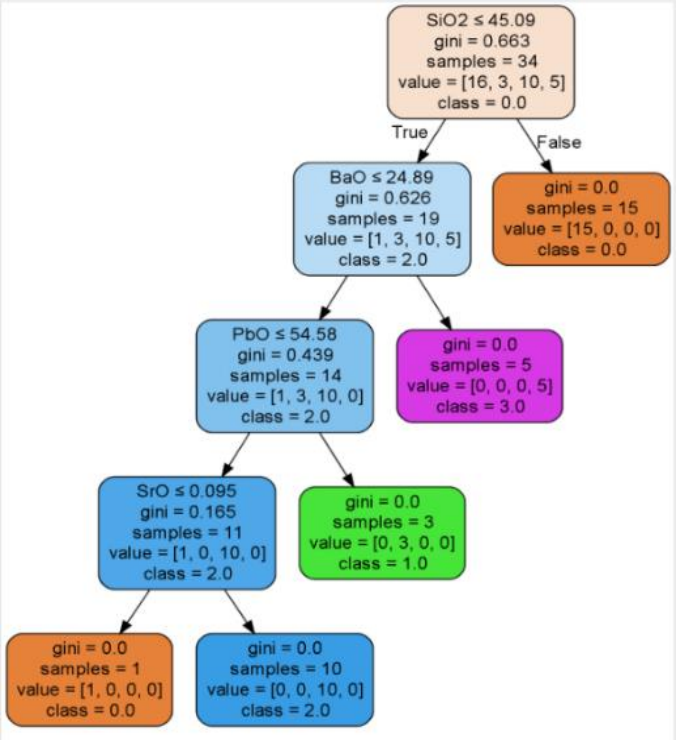

Figure 6. Classification structure diagram of decision tree for barium and lead

In Figure 5, when the content of P2O5 is greater than 0.725, the first classification result can be obtained, and then according to whether the content of CaO is greater than 5.12, it is divided into two categories, and in total, high potassium is divided into three categories; in Figure 6, when the content of Si02 is greater than 45.09, the first classification result can be obtained, and then according to whether the content of BaO is greater than 24.89, the second classification result can be obtained, and then according to whether the content of PbO is greater than 54.58, the third classification result can be obtained. The third classification result can be obtained when the content of PbO is greater than 54.58, and finally the total four classification results are obtained according to whether the content of SrO is greater than 0.095, which divides the lead barium into four categories. The internal nodes in the figure give the specific cut-offs of the branched features, the classification is based on a certain cut-off value of a feature.

• The gini information entropy is used to determine which feature to slice and dice [7].

• Sample type distribution is the number of samples belonging to each classification group in the node, e.g. [10,5,5] means that the three classification groups have 10, 5 and 5 samples respectively.

• Classification is the classification group to which the samples of this node are uniformly classified (this is determined by the group with the largest sample size).

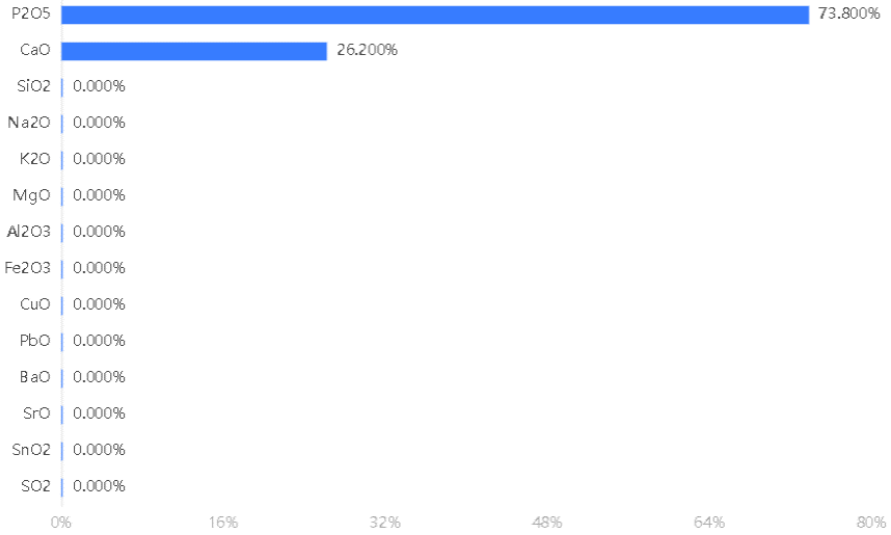

Figure 7. Importance diagram of high potassium classification index

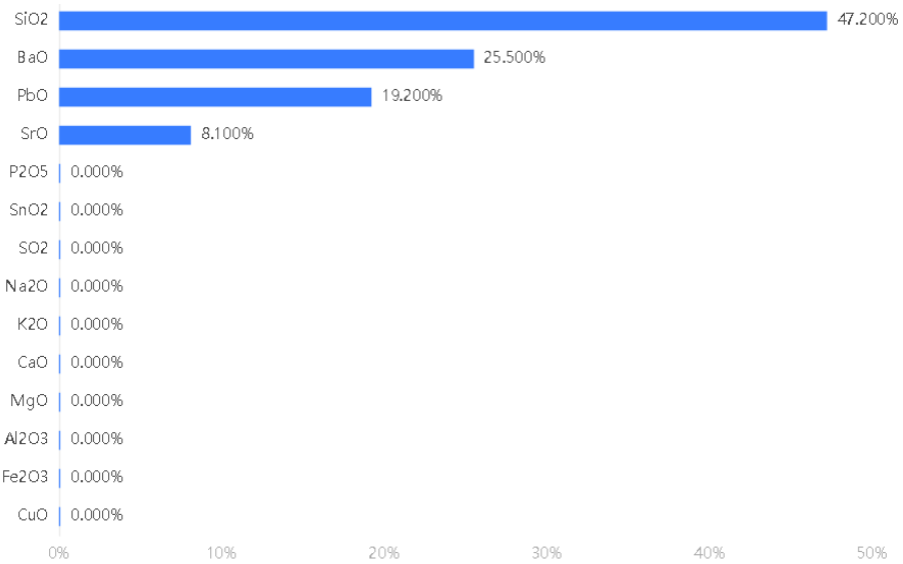

Figure 8. Importance diagram of lead barium classification index

The important indicators of P2O5 and CaO in the classification of high potassium are obtained in Figure 7, and the important indicators of SiO2, BaO, PbO and SrO in the classification of

lead and barium are obtained in Figure 8.

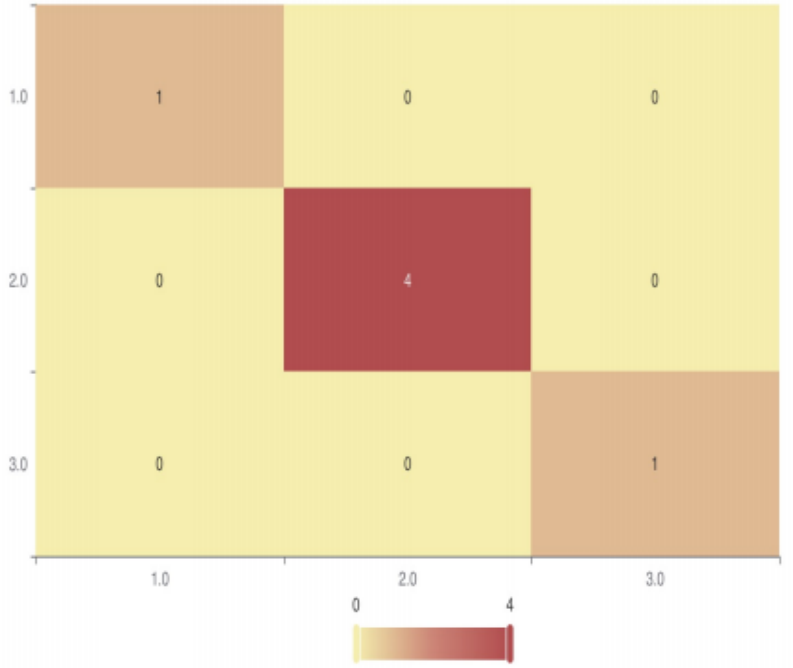

Figure 9. High potassium confusion matrix heat map

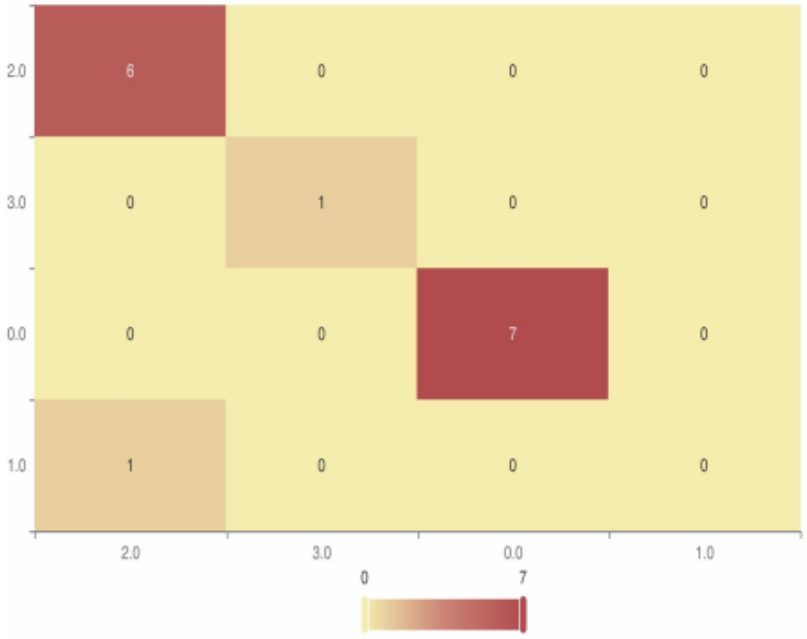

Figure 10. Lead barium confusion matrix heat map

The confusion matrix heat map illustrates: the vertical coordinate is the type of K-means classification, and the horizontal coordinate is the prediction result of the decision tree test set. The prediction results are correct if they are the same as K-means, and are distributed on the main diagonal, and not on the non-main diagonal. In Figure 9,we can see that the classification results are correct, and in Figure 10,there is an error in the classification results, and the overall

training effect is more satisfactory.

Table 1. High potassium class model assessment results

	Accuracy	Recall Rate	Accuracy rate		F1
Training set	1	1	1	1	1
Test set	1	1	1	1	1

Table 2. Lead-barium model evaluation results

	Accuracy	Recall Rate	Accuracy rate	F1
Training set	1	1	1	1
Test set	0.933	0.933	0.876	0.903

From the above table, it can be concluded that the model has good prediction effect. Table 1 shows that the classification results of the high potassium classification model have good results in terms of recall, accuracy, precision, and F1,and Table 2 shows that the classification results of the lead-barium classification model have errors, but the recall, accuracy, precision, and F1 are all within a good range. Table 2 shows that although there are errors in the classification results of the lead-barium classification model, the recall, accuracy, precision, and F1 are all within a more desirable range, and the overall effect is more desirable.

• Accuracy: The proportion of correctly predicted samples to the total samples, the larger the accuracy, the better[8].

• Recall rate: the proportion of results that are actually positive samples that are predicted to be positive samples, the larger the recall rate, the better.

• Precision rate: the proportion of results predicted to be positive samples that are actually positive samples, the larger the precision rate, the better.

• F1: The summation average of precision and recall. Precision and recall affect each other, although high both is a desired ideal situation, but in practice, it is often the case that the precision rate is high and the recall rate is low, or the recall rate is low, but the precision rate is high. If a balance of both is needed, then the F1 metric can be used. From the above evaluation results, it is clear that the subcategorization basis meets the expected results and the model stability is good.

3. Development and solution of an identification model for unknown artifacts based on Euclidean distance [9]

3.1. Computation of Euclidean distance based on clustering centers

The clustering centers of different classes before and after weathering can be obtained by substituting the data of glass artifacts before and after weathering into the K-means algorithm calculation, as shown below.

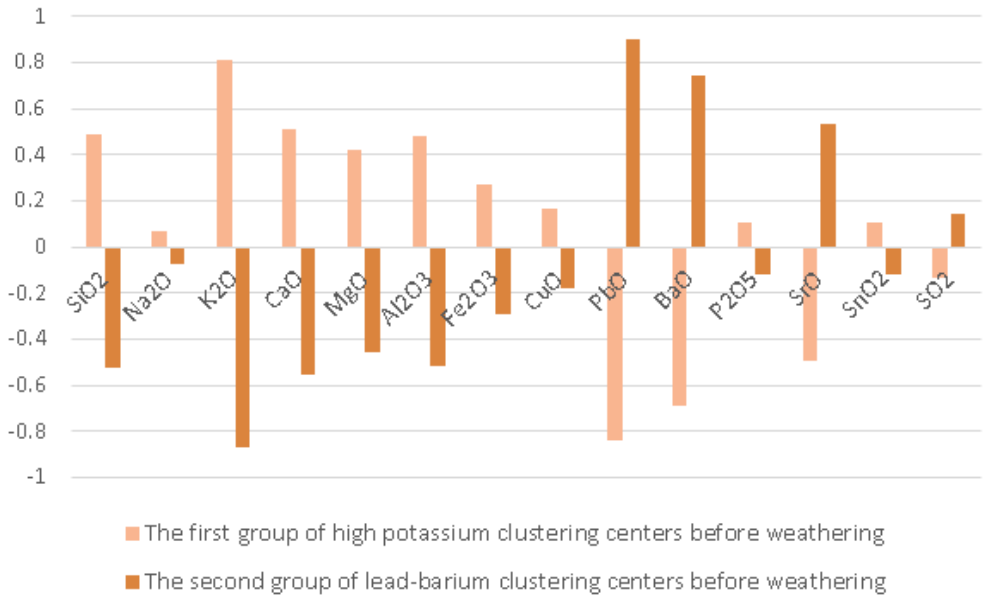

Figure 11. Comparison of cluster centers before weathering

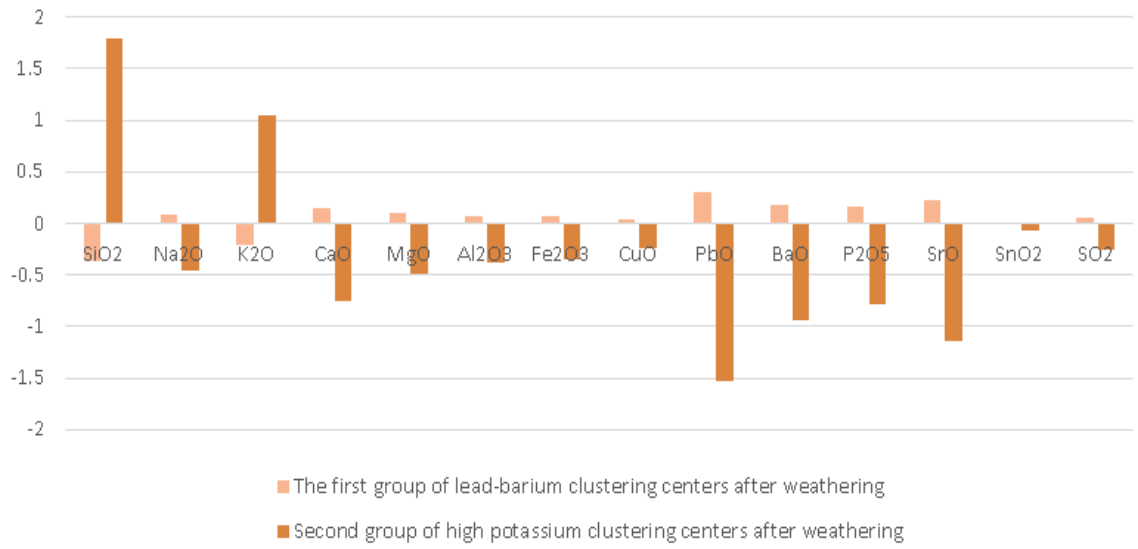

Figure 12. Comparison of cluster centers after weathering

Figure 11 shows the clustering centers of high potassium and lead barium before weathering as the eigenvalues of each class, and the two different glass products have some differences in the clustering centers of each eigenvalue, and Figure 12 shows the clustering centers of high potassium and lead barium after weathering as the eigenvalues of each class, and the two dif ferent [10] glass products still have some differences in the clustering centers of each eigenvalue, so it is meaningful to determine the class to which the glass products belong according to the Euclidean distance from the clustering center of each eigenvalue Therefore, it is meaningful to determine the types of glass products based on the Euclidean distance from the cluster center of each eigenvalue.First,the validity of the data was judged, and the sum of each chemical com- position was between 85%and105%, and there was no abnormal data. Then, we used Matlab to measure the matching degree of the chemical composition of the data with the eigenvalues of xi and the chemical composition of class i in the artifacts to be tested as yi , using the Euclidean distance.

$$d(x, y) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + \cdots + (xn - yn)^2} = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (7)$$

Where (i=1, · · · ,n), calculated by Matlab programming, gives:

Table 3. European distance matching table for artifacts to be tested

Artifact Serial Number	Matching Clustering	European distance
A1	Lead and barium before weathering	2.2904
	High potassium class before weathering	1.6835
	After weathering lead barium class	0.8444
A2	High potassium class after weathering	3.3455
	Lead and barium before weathering	1.9094
A3	High potassium class before weathering	1.9125

The results are shown in Table 3, and it can be seen that for a known pre- or post-weathering glass product, the use of Euclidean distances can reflect the degree of matching with both categories, and there is some difference between the two calculated Euclidean distances.

3.2. Identifying models for unknown artifacts

Divide the object to be tested into two categories, respectively before weathering and after weathering, and calculate the Euclidean distance between the object to be tested and the class center of high potassium and lead barium glass before weathering and the Euclidean distance between the object to be tested and the class center of high potassium and lead barium glass after weathering, respectively, and classify the object to be tested into the category with smaller Euclidean distance by the result of the calculated Euclidean distance, so it can be concluded that:

Table 4. Table of glass artifact types to be tested

High potassium type glass	Lead and barium type glass
A1, A6, A7	A2, A3, A4, A5, A8

Table 4 classifies the unknown artifacts A1, A6, and A7 as high potassium glass types; A2, A3, A4, A5, and A8 as lead-barium glass types.

4. Conclusion

According to Figure 1, there is a significant difference in the content of potassium oxide, lead oxide and barium oxide in high potassium glass and lead-barium glass before weathering, so we can distinguish two different kinds of glass by potassium, barium and lead, and those containing more potassium are classified as high potassium glass, and those containing more lead and barium are classified as lead-barium glass; after weathering, high potassium glass will no longer contain chemical components such as lead oxide and barium oxide, so the two elements of lead and barium can still be used as the basis for classification.

According to K-means, the optimal effect was determined when the K value was equal to 3 and 4, i.e., high potassium was classified into 3 subcategories and lead-barium into 4 subcategories. After sensitivity analysis and evaluation of the subclassification basis, it was found that the results were as expected, and the stability of the model was good.

The analysis and identification of the composition of ancient glass objects is a very serious problem, and a complete, effective, and reliable model is necessary to solve such problems, especially for the analysis of correlation and classification of species, which can be better and more widely used, and can be applied to many similar problems. In this paper, we use K-means algorithm to classify glass artifacts, which is more efficient and scalable for dealing with large data. The decision tree can handle multiple outputs and is more visualized and completely in- dependent of data scaling.

References

[1] Wichao. A review of correlation coefficient research [J]. Journal of Guangdong University of Technology, 2012, 29 (3): 12 - 17.

[2] Yu Lianmin. Improvement and Application of K-means Algorithm in Cluster Analysis [D]. Shandong University of Science and Technology, 2019.

[3] Zeng Yimiao. Improved K-means clustering algorithm based on cyclic dataset [J]. Software, 2021, 42 (11): 74 - 76.

[4] Wang Zhehui, Cheng Ji, Qu an etc. Application of machine learning algorithms in 5G network shunt enhancement [J]. Digital Communications World, 2022, No.216 (12): 73 - 76+80.

[5] Jiang, Xingli, Wang, Jianhui. A decision tree algorithm for disease classification prediction [J]. Information and Computer (Theoretical Edition), 2021, 33 (11): 51 - 53.

[6] Xu Yajie, Liang Jinghan. Construction of student performance evaluation model based on K-means fusion decision tree classification algorithm [J]. Wireless Internet Technology, 2022, 19 (22): 134 - 137.

[7] Sun Pei, Wang Quanfang, Yi Jiewei etc. Extraction of wheat and rape planting range based on decision tree classification method [J]. Agriculture and Technology, 2022, 42 (24): 7 - 11.DOI: 10.19754/j.nyyjs.20221230002.

[8] Yan Hongli, Luo Yonglian. Classification of breaking news based on decision tree method [J]. Electronic Technology and Software Engineering,2020, No.172 (02): 194 - 195.

[9] Wang Hui, Zhang Wenjie, Liu Jie etc. Flight delay prediction model based on classification regression decision tree algorithm [J]. Journal of Civil Aviation University of China, 2022, 40 (03): 35 - 40.

[10] Wang Weishuai, Gao Yue, Liu Hongyuan etc. Evaluation of agricultural sustainability in Chengdu based on the Euclidean distance method [J]. China Agricultural Resources and Zoning,2019, 40 (07): 209 - 215.