Chemical composition analysis of ancient glass products based on decision tree

Zhihao He [1,*,#], Weiduo Qin [2,#], Changping Duan [3,#]

[1] School of Food Science and Bioengineering, Tianjin Agriculture University, Tianjin, China, 300384

[2.]School of Humanities, Tianjin Agriculture University, Tianjin, China, 300384

[3.]School of Mechanical and Electrical Engineering, Lanzhou University of Technology, Lanzhou, Gansu, China, 730050

* Corresponding Author Email: hesecretarytrue@gmail.com

[#]These authors contributed equally.

Abstract. Due to the effects of prolonged burial, freshly unearthed ancient glass is often weathered to varying degrees, and it is difficult to identify the type of glass. We introduce machine learning into the composition analysis and type identification of ancient glass products. This objective is to build a reliable ancient glass classification model based on decision trees and two different k-means clustering algorithms. The performance of the decision tree is measured by the ROC curve. The performance of its clustering algorithm was evaluated by the Calinski-Harabasz index. The results show that the area of AUC in the decision tree is 1 and the highest Calinski-Harabasz index of the two clustering algorithms is 71.68. The predictive ability of the model was verified well.

Keywords: Machine learning, Decision tree, K-means clustering algorithm, Mini-batch k-means.

1. Introduction

1.1. Motivation

The protection of cultural relics is the responsibility of practicing the diversity of world and continuation of human heritage and the analysis of cultural relics is a very important field of cultural relics protection [1]. The development of machine learning provides a new direction for the analysis of cultural relics. Compared with traditional analysis, not only does the analysis of cultural through machine learning have higher accuracy rate, but also has the ability of continuous learning [2]. The study of ancient glass plays an important role in understanding the ancient Chinese culture. In addition, the classification and analysis of ancient glass is essential to distinguish authenticity and origin.

1.2. The related work

In the past few decades, the analysis of ancient glass has attracted the attention of a wider group of scholars [3]. Many studies on the composition analysis of ancient glass have emerged [4][5][6]. According to the chemical composition, five main types of ancient glass were basically determined [7]. Most of the analytical work on glass so far is to understand the basic principles of early glass technology [8], origin [9][10], type. In recent years, interdisciplinary technology has been widely used in the analysis of ancient glass [11]. Machine learning is also applied to glass-related fields [12][13], which is of great help in identifying glass types and restoring ancient glass.

1.3. Our contribution

The chemical composition of each item was analyzed before and after the weathering of this batch of cultural relics to study the relationship of glass type to surface weathering, ornamentation, and color in a collection of ancient glass. And the batch of glass was classified in detail through decision tree-kmeans clustering innovatively. The classification rules of high-potassium glass and lead-barium glass are explored and clustered through the decision tree, and the two glasses are clustered.

Comparing the two clustering methods, the k-means algorithm performs best in all evaluation indicators, k-mean is the most suitable algorithm for ancient glass classification, and the Calinski-Harabasz Index is 65 and 72 respectively.

1.4. The Arrangement

This paper is structured as follows: Section II, The type of glass makes a statistical analysis of the chemical composition content of weathering on the surface of cultural relics;Section III, construct decision tree-kmeans classification model to classify glass types in detail;Section IV, the conclusion.

2. Decision trees and k-means

As one of the fastest-growing technical fields today, machine learning can improve computer-like problems through existing experience [14], it shines in the problems of classification, regression, and clustering problems. Decision tree and k-means algorithms are two typical machine learning algorithms.

2.1. Decision trees

As a tree-based technology, the branches separated from the root are determined by data separation until the results appear at the leaf nodes [15]. The CART algorithm is a typical decision tree algorithm. It uses the Gini coefficient for performing feature selection of the explored samples. According to the feature selection standard, we need to minimize the Gini coefficient of each variable partition to determine the value of threshold under the variable. Threshold can be a splitting variable and splitting point.

$$Gini(S) = 1 - \sum_{n-1}^{N} \left(\frac{|C_n|}{|S|} \right) \tag{1}$$

The establishment process of the decision tree is to repeatedly calculate the influencing factors through this formula. When the Gini index is smaller, the division effect is better. Therefore, only the factors that have the greatest impact on the sample are selected as the optimal features to divide the left and right subtrees, and this operation process is repeated continuously. When the depth of the specified tree is reached, the operation stops, and an incomplete decision tree is generated iteratively.

2.2. K-means

K-means clustering is to iteratively divide K clusters as the initial cluster center, then calculate the distance between each sample and each K-value cluster center, and assign each sample to the cluster with the shortest distance to it center K objects are randomly selected as initialization clustering centers. The distance from each object to K clustering centers is calculated by Euclidean algorithm, and those close to the centroid of K value are grouped into the same category, and the initial clustering results are obtained. Sort according to the order of increasing distance, select the initial clustering center of each category, and find out the new clustering center of each category after classification.

$$y_k = \frac{1}{|C_k|} \sum_{x \in C_k} x \tag{2}$$

2.3. Mini Batch k-means

Mini Batch K-Means clustering algorithm randomly selects a part of data from the data set as a data subset to train the algorithm, which reduces the computing time compared with the K-Means algorithm [16], and contributes to the optimization of the objective function. The algorithm is based on randomly extracting a small subset of data, and iteratively calculates the center point of each cluster according to the data points contained in the subset until the center point is stable [17].

2.4. Calinski-Harabasz index

As an index widely used in cluster evaluation, The Calinski-Harabasz Index is based on clusters and between clusters, which is defined as follows.

$$s(k)=\frac{tr(B_k)m-k}{tr(W_k)k-1} \tag{3}$$

The larger B_k is, the higher the degree of dispersion between clusters, and the smaller W_k is, the closer the relationship between clusters is. With the increase of the ratio, the Calinski-Harabasz index value increases, that is, the better the clustering effect is [18].

2.5. Dataset and symbol description

Table 1. The symbol description

Symbol	Demonstrations
S	Total number of samples
$\lvert C_n \rvert$	The number of samples in S belonging to the n class
K	the corresponding number of clusters
B_k	Covariance matrix between categories
tr	Trace of matrix
W_k	Covariance matrix of data within categories

The dataset contains detailed data of 58 ancient glass. For the code and data required in the paper, see https://github.com/heshuji/paper/tree/main/ancient_glass

3. Experiments and analysis

3.1. Statistical law analysis of component content

Through descriptive statistical analysis, the contents of chemical components on the surface of high-potassium lead-barium glass before and after weathering were analyzed, and the statistical rules were revealed by box plots.

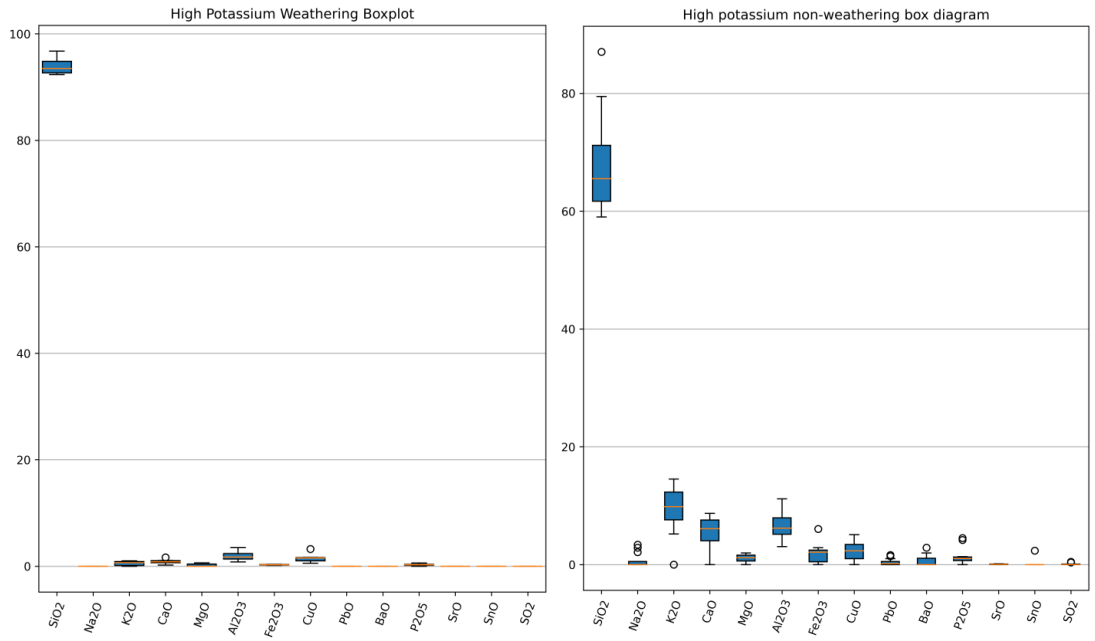

Fig 1. High potassium weathered and non-weathered chemical content

Because other oxides are greatly affected by weathering, the proportion of other oxides in the total proportion decreases, while silica has strong stability and less material flow, so when the sample surface is weathered, the proportion of the substance in the sample increases.

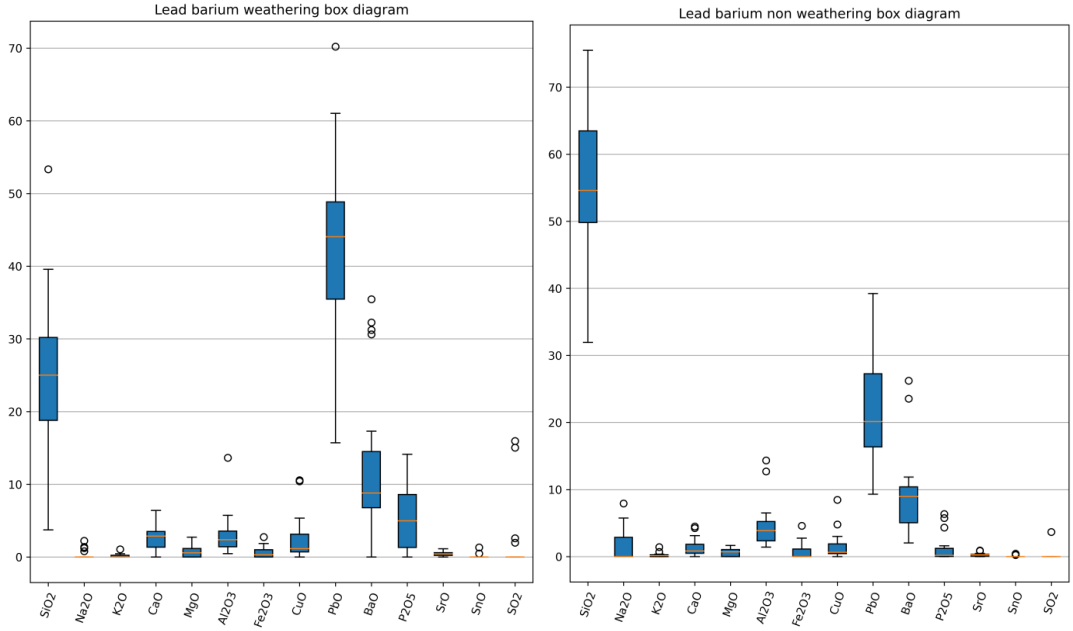

Fig 2. Lead barium weathered and non-weathered chemical content

The properties of glass samples using lead ore as cosolvent are relatively stable compared with high potassium glass products. Except for silica and iron oxide, the chemical content of the weathered surface increased in the total proportion.

Table 2. Glass type and chemical composition

cultural relic sampling point	01	02	03	04	05	…
SiO2	69.33	36.28	87.05	65.88	61.58	…
Na2O	0	0	0	0	0	…
K2O	9.99	1.05	5.19	9.67	10.95	…
CaO	6.32	2.34	2.01	7.12	7.35	…
MgO	0.87	1.18	0	1.56	1.77	…
Al2O3	3.93	5.73	4.06	6.44	7.50	…
Fe2O3	1.74	1.86	0	2.06	2.62	…
CuO	3.87	0.26	0.78	2.18	3.27	…
PbO	0	47.43	0.25	0	0	…
BaO	0	0	0	0	0	…
P2O5	1.17	3.57	0.66	0.79	0.94	…
SrO	0	0.19	0	0	0.06	…
SnO2	0	0	0	0	0	…
SO2	0.39	0	0	0.36	0.47	…
surface weathering	unweathered	weathered	unweathered	unweathered	unweathered	…
type	high potassium	lead barium	high potassium	high potassium	high potassium	…

3.2. Decision tree-kmeans

Divide Glass type and chemical composition into 80% training set and 20% test set. The decision tree composed of factors affecting the classification of high potassium glass and lead-barium glass was trained. The accuracy of the model is the highest when the maximum depth is 2 and the cotyledon node contains at least sample 2. The evaluation index that can be obtained by ROC curve can reduce the interference caused by different test sets, and measure the performance of the model more objectively on the basis of previous test sets. The ROC curve reflects the trend of sensitivity (FPR) and accuracy (TPR) of decision tree affecting the classification of high potassium glass and lead-barium glass, as shown in figure 3.

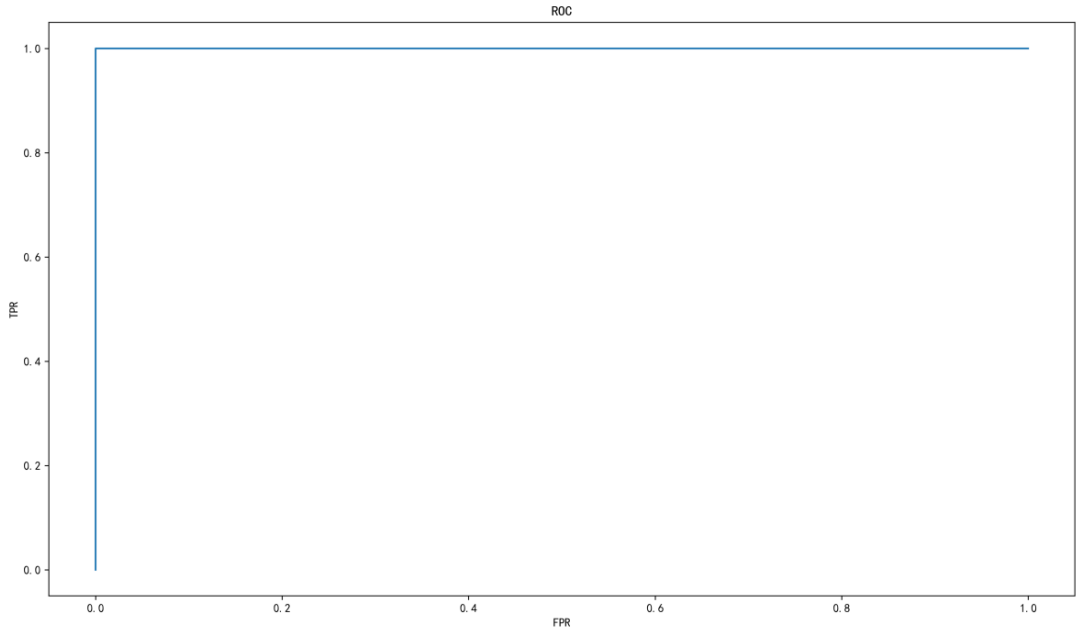

Fig 3. ROC curve of decision tree of factors affecting different glass classification

AUC area of 1This is a perfect classifier.

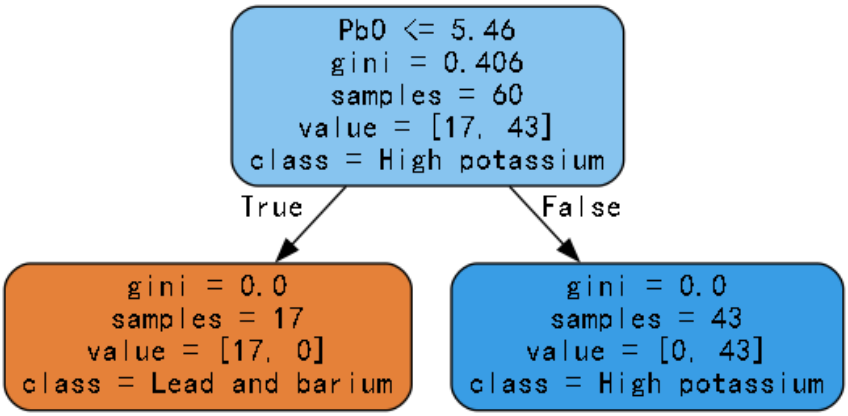

Fig 4. decision tree of factors affecting different glass classification

From figure 4, we analyze the classification law of high-potassium glass and lead-barium glass. When the content of lead oxide on the surface of cultural relics is less than or equal to 5.46%, the glass is high-potassium glass; on the contrary, when the surface lead oxide of the cultural relic is more than 5.46%, the glass is lead-barium glass.

We chose two different clustering categories, K-Means and MiniBatchK-Means, and carried out a comparative experiment. The Calinski-Harabasz score was used as an index to evaluate the clustering effect. In order to directly reflect the clustering of the model in space, we reduce the dimension of the model and draw the clustering distribution map in two dimensions.

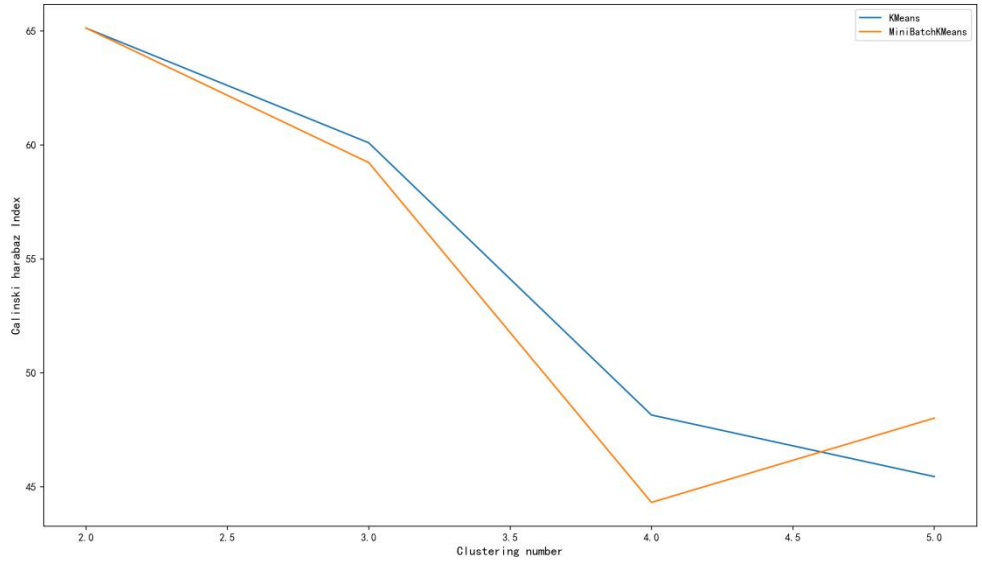

Fig 5. Calinski-Harabasz index of High potassium Glass under different clustering algorithms and K values

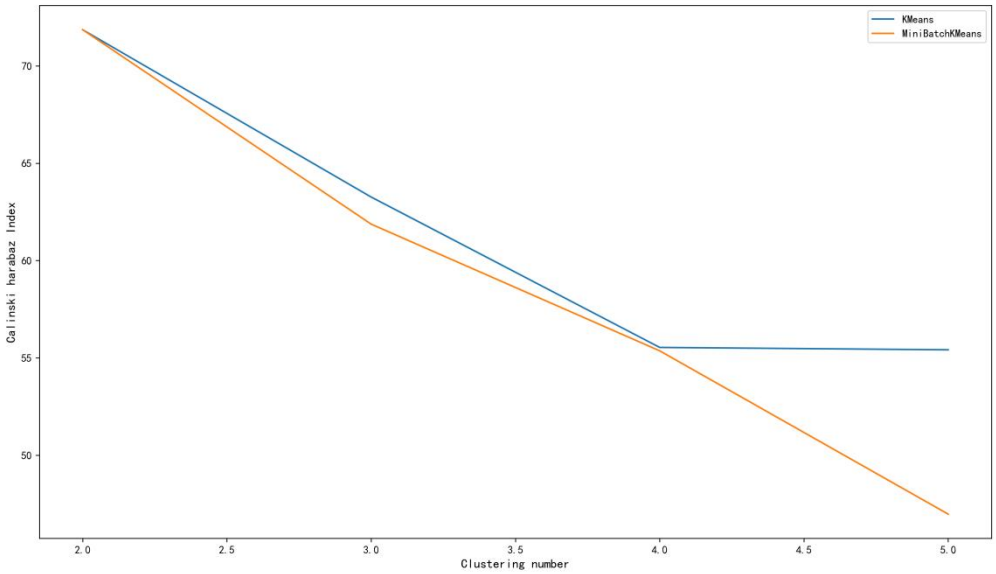

Fig 6. Calinski-Harabasz index of lead-Barium Glass under different clustering algorithms and K values

We can find that the unified results are obtained by using the two methods, that is, the highest score of Calinski-Harabasz index is 71.86 and 65.12 respectively, indicating that the clustering effect corresponding to K value is the best at this time.

We divide high potassium glass and lead-barium glass into two categories respectively. The statistical relationship between different chemical composition and different types of glass products was constructed by scatter plot, and the categories were analyzed.

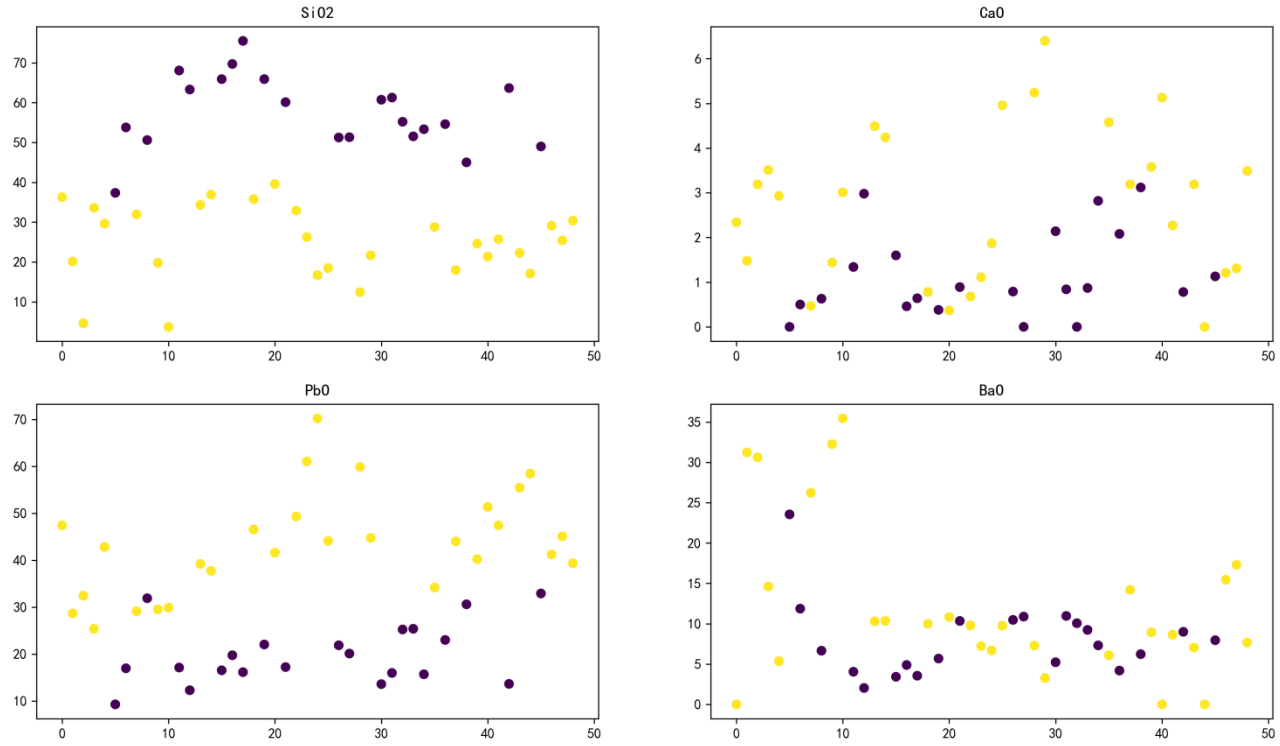

Fig 7. Statistical relationship between lead-Barium Glass and Chemical composition

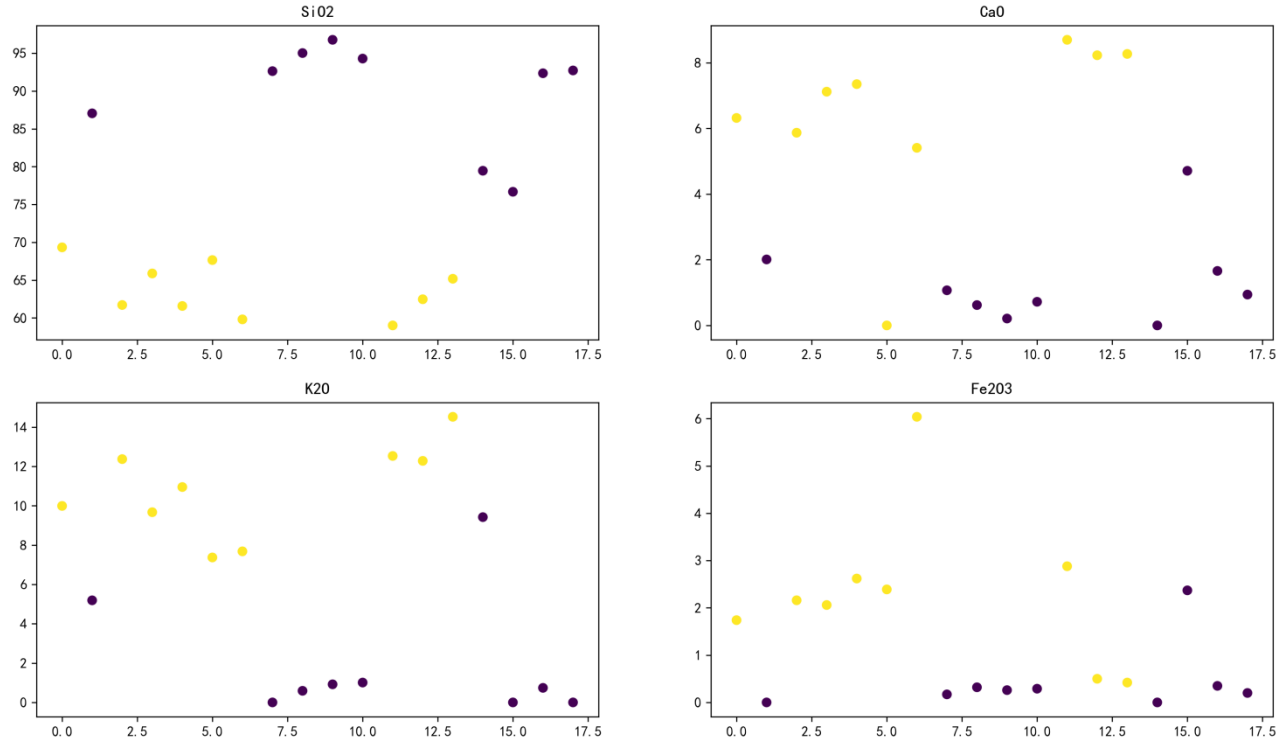

Fig 8. Statistical relationship between High potassium Glass and Chemical composition

(1) Two classifications of High potassium glass:

The first category: when the surface of the glass product contains more silicon dioxide, most of the calcium oxide, potassium oxide and iron oxide are less.

The second category: when the content of silicon dioxide on the surface of the glass product is less, most of the calcium oxide, potassium oxide and iron oxide are more.

(2) two classifications of lead-barium glass:

The first category: when the surface of the glass product contains more silicon dioxide, most of the calcium oxide, lead oxide and barium oxide are less.

The second category: when the content of silicon dioxide on the surface of the glass product is less, the content of calcium oxide is affected by the relevant environment, while the content of lead oxide and barium oxide is more.

4. Conclusion

In this work, the decision tree-clustering algorithm is used to subclassify ancient Chinese glass. Through the decision tree to determine the key conditions of the first classification, whether the content of lead oxide is more than 5.46%. Different clustering numbers and the performance of clustering algorithms are compared. The results show that the clustering of 2 is the optimal classification, and the performance of the two algorithms is the same when the clustering is 2, and the score of Calinski-Harabasz index is the highest, which is 71.86 and 65.12 respectively. This model plays a guiding role in the classification of ancient glass. Next, we will use the weathered glass to predict its chemical composition before weathering, hoping to provide some help to the restoration of ancient glass.

References

[1] Liu, B., Mu, K., Ye, F., Deng, J., & Wang, J. (2020). Immovable cultural relics disease prediction based on relevance vector machine. Mathematical Problems in Engineering, 2020.

[2] Sun, H., Liu, M., Li, L., Yan, L., Zhou, Y., & Feng, X. (2020). A new classification method of ancient Chinese ceramics based on machine learning and component analysis. Ceramics International, 46(6), 8104-8110.

[3] Rehren, T., & Freestone, I. C. (2015). Ancient glass: from kaleidoscope to crystal ball. Journal of Archaeological Science, 56, 233-241.

[4] Beck, H. C., & Seligman, C. G. (1934). Barium in ancient glass. Nature, 133 (3374), 982-982.

[5] Laubengayer, A. W. (1931). THE WEATHERING AND IRIDESCENCE OF SOME ANCIENT ROMAN GLASS FOUND IN CYPRUS 1. Journal of the American Ceramic Society, 14 (11), 833-836.

[6] Akmeemana, A., Weis, P., Corzo, R., Ramos, D., Zoon, P., Trejos, T., ... & Almirall, J. (2021). Interpretation of chemical data from glass analysis for forensic purposes. Journal of Chemometrics, 35(1), e3267.

[7] Sayre, E. V., & Smith, R. W. (1961). Compositional categories of ancient glass. Science, 133 (3467), 1824-1826.

[8] Freestone, I.C. (2004). The Provenance of Ancient Glass through Compositional Analysis. MRS Proceedings, 852.

[9] Fuxi, G. (2009). Origin and evolution of ancient Chinese glass. Ancient glass research along the Silk Road, 1-40.

[10] Sayre, E. V., & Smith, R. W. (1973). Analytical studies of ancient Egyptian glass (No. BNL--18562). Brookhaven National Lab.

[11] Henderson, J. (2013). Ancient glass: an interdisciplinary exploration. Cambridge University Press.

[12] Xiong, J., Zhang, T. Y., & Shi, S. Q. (2019). Machine learning prediction of elastic properties and glass-forming ability of bulk metallic glasses. MRS Communications, 9(2), 576-585.

[13] Alcobaca, E., Mastelini, S. M., Botari, T., Pimentel, B. A., Cassar, D. R., de Leon Ferreira, A. C. P., & Zanotto, E. D. (2020). Explainable machine learning algorithms for predicting glass transition temperatures. Acta Materialia, 188, 92-100.

[14] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[15] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.

[16] Béjar Alonso, J. (2013). K-means vs Mini Batch K-means: a comparison.

[17] Fitriyani, S. R., & Murfi, H. (2016, May). The K-means with mini batch algorithm for topics detection on online news. In 2016 4th International Conference on Information and Communication Technology (ICoICT) (pp. 1-5). IEEE.

[18] Wang, X., & Xu, Y. (2019, July). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In IOP Conference Series: Materials Science and Engineering (Vol. 569, No. 5, p. 052024). IOP Publishing.