A study of glass classification and identification based on K-means clustering and Euclidean distance

Yaxuan Kong [1, *], Zijun Liu [1], Wenyong Tan [1], Wenbo Wu [2]

[1] School of Civil Engineering, Lanzhou University of Technology, Lanzhou, Gansu, 730050, China

[2] School of Information and Electrical Engineering, Lanzhou University of Technology, Lanzhou, Gansu, 730050, China

* Corresponding author: yaxuankk@163.com

Abstract. Glass is buried in the process of weathering, resulting in a large number of exchanges of its chemical composition. To classify and identify the glass of unknown composition, firstly, the existing classification law of glass was analyzed, and the subclasses of glass types were divided based on principal component analysis and K-means cluster analysis, and the number of subclasses was determined according to the P value of cluster significance, and the high potassium glass was divided into 3 classes and the lead-barium glass was divided into 4 classes. Subsequently, a glass classification model was constructed for the unknown component glasses, and after obtaining the glass classification, the classification model of glass mean was constructed by Euclidean distance based on K-means clustering results, and the subclass classification results of all unknown samples were obtained.

Keywords: Principal component analysis, K-means cluster analysis, Glass mean classification model.

1. Introduction

Glass is a special class of silicate-based materials, and the creation, development, trade, and use of glass have an important place in the study of early human civilization and are important physical evidence of trade exchanges along the ancient Silk Road. The early glass was introduced to China through trade and other means in West Asia and Egypt, and its chemical composition was changed through improvement, evolving into our ancient glass.

Ancient glass is highly susceptible to weathering by the burial environment. Among the large number of ancient glass artifacts unearthed by archaeological work, a considerable portion of them have been subjected to different degrees of weathering, and a large number of internal elements have been exchanged with environmental elements, resulting in changes in their chemical composition and proportions, which affect the judgment of archaeologists on their categories and hinder the development of archaeological work. Therefore, the study of ancient glass archaeological work needs to establish a glass subclass relationship model and find its compositional correlation through the study of the chemical composition of excavated artifacts.

Before conducting the study, clustering methods are first investigated. Chaobo He [1] et al. proposed a fuzzy clustering-based attribute graph delineation method applied to protein complex discovery in protein interaction (PPI) networks, which outperforms traditional attribute graph delineation representative methods (e.g., SACluster and BAGC) in the case of small-scale data. Hong Xia [2] et al. proposed a service partitioning method based on particle swarm fuzzy clustering, which can better evaluate the interconnectivity among services, improve the accuracy of service clustering, and reasonably consider the service-to-service relationship. Ana Justel [3] et al. proposed a new split clustering method DivClusFD for the unsupervised classification of functional data, which will explore the function and its derivatives at multiple fixed points to find the sub-regions that can separate the most clusters, and estimate the number of clusters by the gap statistic to identify functions that are misclassified due to their atypical local behavior. Hongjie Jia [4] et al. proposed HMRF semi-supervised approximate kernel k-means algorithm, which has a shorter clustering time compared to the original HMRF semi-supervised clustering, good scalability, and is suitable for dealing with large-

scale Zhuqing Jiao [5] et al. proposed a multilayer functional brain network module partitioning method based on weighted clustering integration, which is aimed at solving the problem that the connections between nodes in the network change over time, and applied to analyze the differences between the corresponding modules of patients with Alzheimer's disease and normal people, which can better understand certain dynamic pathological changes in the brain connectome of patients. Yinghua Shen [6] et al. proposed an original algorithm called hyperplane partitioning to partition the entire data set into disjoint subsets, which is beneficial to improve the quality of clusters generated by clustering algorithms and has some superiority in some large data clustering. Shihua Zhang [7] proposed a new modularity function based on the generalized Q modularity and fuzzy C-mean clustering. Xiaochuan Pu [8] et al. introduced a local density value ρ and information entropy H to improve the K-means clustering algorithm, which was applied in the customer segmentation problem to fully reflect the customer value and its development space and improved the efficiency of the algorithm by improving the clustering algorithm. Yadi Zhang [9] et al. proposed an improved K-means based on quadtree space partitioning and density clustering the results showed that the QD-K-means algorithm is more accurate than other clustering algorithms and has better stability and robustness, which solves the problem of unstable algorithm efficiency due to randomly selected initial values of K-means algorithm. In addition, there are some other community clustering algorithms: among them, Yi Xu [10] et al. proposed a new community clustering algorithm based on simulated annealing and particle swarm optimization (SCBSP) for spectral clustering, which improves the global search ability, has better global convergence and has a better performance compared with the original spectral clustering.

In summary, there are few studies on the classification of glass artifacts. In this paper, based on the existing data, we obtain the classification basis of glass types based on the basic information of two major categories of archaeological glass artifacts, and subclassify glass artifacts based on the principal component analysis and K-means clustering algorithm to build a model, and finally analyze and identify the types of unknown categories of glass artifacts.

2. Glass classification research

2.1. Research based on glass type classification

Firstly, the chemical composition of the two types of glass was obtained through data processing, and a scatter plot was drawn based on this data as shown in Figure 1.

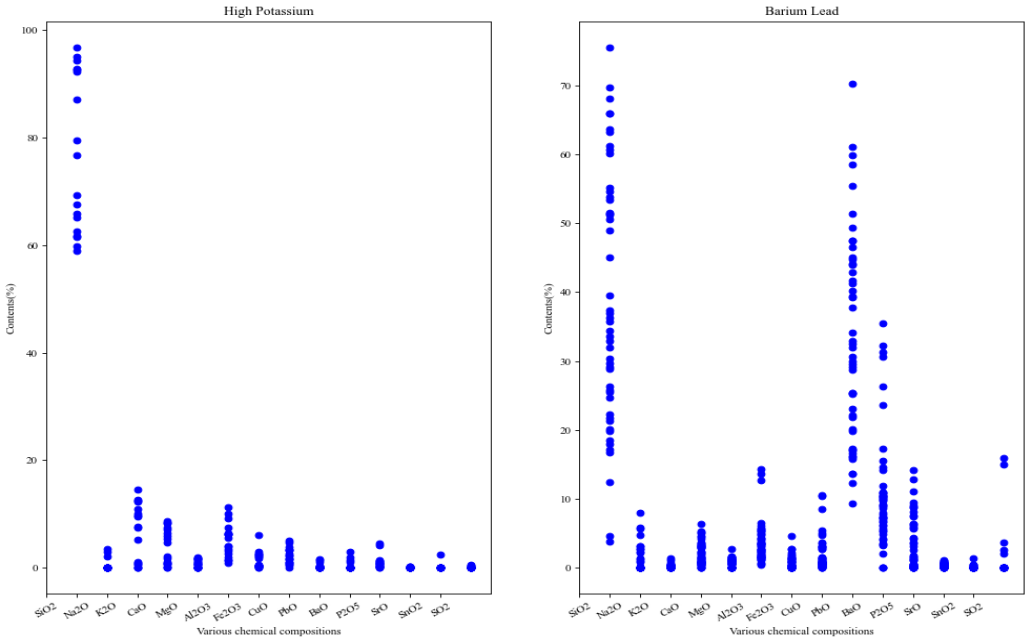

Figure 1. Scatter diagram of the percentage of the chemical composition of the two types of glass

The visualization of the scatter plot reveals the following findings:

In high potassium, the first three chemical components with high chemical content are SiO_2, K_2O, and Al_2O_3.

In the lead-barium glass, the first three chemical components with higher chemical content are SiO_2, PbO, and BaO.

Subsequently, since the average value reflects the general level of the phenomenon and the concentrated trend of the distribution, the average values of the chemical composition content of the two types of glasses were solved and ranked, and the results are shown in Tables 1 and 2.

Table 1. High potassium glass chemical composition average descending order table

Ingredients	SiO_2	K_2O	Al_2O_3	CaO	CuO	Fe_2O_3	P_2O_5
Content	76.64389	6.401667	5.056667	3.845	2.155556	1.376111	1.028333
Ingredients	MgO	Na_2O	BaO	PbO	SnO_2	SO_2	SrO
Content	0.785	0.463333	0.398889	0.274444	0.131111	0.067778	0.027778

Table 2. Lead barium glass chemical composition average descending order table

Ingredients	SiO_2	PbO	BaO	Al_2O_3	P_2O_5	CaO	CuO
Content	38.87551	33.34918	10.49041	3.667551	3.292653	2.05	1.879592
Ingredients	Na_2O	SO_2	Fe_2O_3	MgO	SrO	K_2O	SnO_2
Content	0.90449	0.799592	0.655918	0.64551	0.347959	0.173469	0.058163

The descending order table of the two glasses above shows that the first three chemical compositions of high potassium glass and lead-barium glass are consistent with the results of the scatter diagram study. Since SiO_2 is the most important constituent element in glass, it is not used as a basis for classification. It is not difficult to find that these two glass types are classified based on the chemical composition content that accounts for a relatively high percentage. In the following subclass classification basis, this paper will also be based on this factor.

2.2. K-means clustering artifact classification model based on principal component analysis

In this paper, the data of 14 chemical components of glass artifacts, which are intricately related to each other and, through preliminary observations, certain chemical components account for a very low percentage and have minimal impact on the overall data, need to be studied, so a K-means clustering artifact classification model based on principal component analysis was developed.

The principal component analysis is for all the variables originally proposed, the duplicate variables (closely related variables) are deleted redundantly to reduce the data redundancy and create as few new variables as possible so that these new variables are two unrelated, and these new variables keep the original information as much as possible in terms of reflecting the information of the problem. It is a method used mathematically to reduce dimensionality. The process is as follows:

(1) First standardize the raw data.

(2) Calculating the correlation coefficient matrix.

(3) Calculation of eigenvalues and eigenvectors.

$$\begin{aligned} y_1 &= c_{11}A + c_{21}B + c_{31}C + c_{41}D + c_{51}E \\ y_2 &= c_{12}A + c_{22}B + c_{32}C + c_{42}D + c_{52}E \\ y_3 &= c_{13}A + c_{23}B + c_{33}C + c_{43}D + c_{53}E \cdots \\ y_4 &= c_{14}A + c_{24}B + c_{34}C + c_{44}D + c_{54}E \\ y_5 &= c_{15}A + c_{25}B + c_{35}C + c_{45}D + c_{55}E \\ &\vdots \end{aligned} \tag{1}$$

(4) Calculate the contribution rate of the eigenvalues (β_m).

$$\beta_m = \frac{\lambda_m}{\Sigma\lambda}, m = 1, 2, ...14 \tag{2}$$

(5) Summation of the contribution rate ranking to find the number of contribution rates that sum up to 80% or more.

In this model, the number of selected chemical components is mainly obtained by calculating the contribution ratio obtained from the principal component analysis. Firstly, the contribution rate was sorted in descending order, and the sum of the contribution rate was summed from the highest to the lowest, and the number of principal components with a sum of contribution rate over 80% was found to be $\boldsymbol{\alpha}$. Since the chemical composition ratio was found to be the main basis for the classification of glass types in the study, the group with the higher average chemical composition was extracted and then K-means cluster analysis was performed. Therefore, the data of $\boldsymbol{\alpha}$ groups with higher average chemical composition were extracted and subsequently subjected to K-means clustering analysis.

In clustering analysis, we set multiple K-values to compare and analyze the clustering results based on their significant differences, with stronger significance indicating stronger subclass variability and more reasonable models.

In the sensitivity analysis process, the data range was extended by taking $\alpha+1, \alpha+2, \alpha+3, \alpha+4$ four chemical composition data for sensitivity analysis. First, the custom indicator index - the center shift rate η_q, with the following equation.

$$\eta_q = \frac{\sum_{p=1}^{\alpha} |(x_p{}' - x_p)|}{\sum_{p=1}^{\alpha} x_p} \tag{3}$$

Where x'_p is the new clustering result center coordinate value and x_p is the original clustering result center coordinate value,q is taken as 1,2,3..., K.

The rationale is that the cluster centroids can better represent the degree of change in the cluster analysis, so the center shift rate was developed accordingly. Since the mean value of chemical composition share was previously ranked to make, the growth rate of its contribution decreases with each expansion of the data, and its change approximates a negative exponential relationship; the law of change of the two types of glass clustering centers is associated with the law of change of the chemical composition share, and the center offset rate should follow the negative exponential function trend as the data range is expanded in a similar logarithmic function form [11].

2.3. Solving the model

2.3.1. Quantitative selection of chemical components:

Principal component analysis was performed on the glass artifact sample data by Matlab to calculate the eigenvalues λ_m of the correlation coefficients of each chemical component of the two glasses and the contribution β_m. The results are shown in Tables 3 and 4.

Table 3. Two types of glass characteristics root table

Constituent	C1	C2	C3	C4	C5	C6	C7
High Potassium	5.64	2.63	1.75	1.63	0.96	0.66	0.31
Lead Barium	3.58	2.95	1.66	1.08	0.91	0.84	0.75
Constituent	C8	C9	C10	C11	C12	C13	C14
High Potassium	0.20	0.12	0.07	0.02	0.01	0.01	0.00
Lead Barium	0.62	0.56	0.36	0.35	0.21	0.12	0.00

Table 4. Two types of glass contribution rate

Constituent	C1	C2	C3	C4	C5	C6	C7
High Potassium	40.30	18.77	12.49	11.65	6.85	4.73	2.23
Lead Barium	25.54	21.06	11.89	7.75	6.48	6.00	5.34
Constituent	C8	C9	C10	C11	C12	C13	C14
High Potassium	1.43	0.82	0.47	0.14	0.07	0.04	0.00
Lead Barium	4.42	4.03	2.61	2.50	1.47	0.88	0.03

After arranging the contribution rates in descending order, the top items with higher contribution rates were summed, and the number of principal components α was obtained when the contribution rate exceeded 80%. Subsequently, according to the mean size of chemical components, the α with a larger mean value was extracted items for K-means cluster analysis, and the results of the chemical component selection are shown in Figure 2.

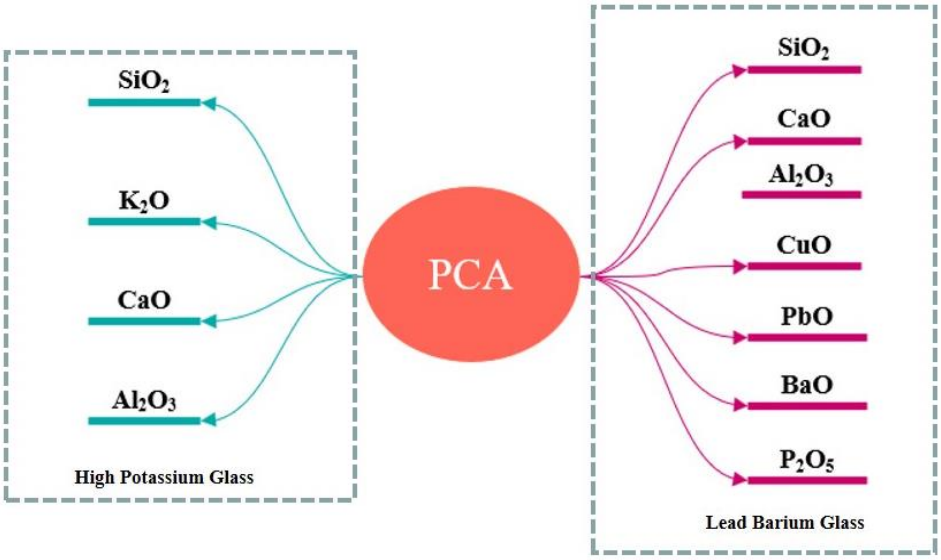

Figure 2. Two types of glass principal components selection chart

2.3.2. K-means clustering analysis:

After data processing based on the chemical components selected by principal component analysis, K-means clustering was performed using SPSS software, and the clustering groups were selected as K=2, 3, 4, 5, and 6, respectively.

When the number of clustering groups was K=2,3,4,5, the significance of each variable for a different number of groups was statistically processed and the results are shown in Tables 5 and 6.

Table 5. Significant P-value for the number of different groups of high potassium glass

	K=2	K=3	K=4	K=5	K=6
SiO_2	0.000***	0.000***	0.000***	0.032*	0.001***
K_2O	0.000***	0.000***	0.041*	0.001***	0.002***
CaO	0.000***	0.001***	0.001***	0.002***	0.001***
Al_2O_3	0.000***	0.000***	0.001***	0.000***	0.002***

Table 6. Significant P-value for the number of different groups of lead barium glass

	K=2	K=3	K=4	K=5	K=6
SiO_2	0.000***	0.000***	0.000***	0.000***	0.000***
CaO	0.000***	0.002***	0.003***	0.019**	0.012**
Al_2O_3	0.004***	0.019*	0.016**	0.007***	0.095*
CuO	0.079*	0.000**	0.000***	0.000***	0.000***
PbO	0.000***	0.000***	0.000***	0.000***	0.000***
BaO	0.067*	0.000***	0.000***	0.000***	0.000***
P_2O_5	0.000***	0.000***	0.006***	0.036*	0.013**

Since fewer categories at K=2 would lead to rougher subcategories, the subcategories at K=2 were not considered. According to the analysis of the significance P-value results in Table, it was found that the significance P-value of some chemical components in high potassium glass at K=4 tended to increase faster, which would reduce the variability of such classification categories, so the classification of clustering analysis at K=3 was used. The first three chemical components in lead-barium glass do not change significantly in significance, and the remaining four chemical components show a trough at K=4, at which time the significance increases, increasing the variability between the divided categories. Therefore, high potassium was divided by clustering with K=3, and lead barium was divided by clustering with K=4. The clustering results are shown in Tables 7 and 8.

Table 7. High potassium glass clustering centroid coordinates

Clustering type	Center point coordinates				Number of divisions
	SiO_2	K_2O	CaO	Al_2O_3	
Medium Weathering	81.06	4.87	2.24	4.43	9
High Weathering	63.62	10.81	6.36	7.34	7
Low weathering	93.96	0.54	0.87	1.93	2

Table 8. Lead barium glass clustering centroid coordinates

Clustering type	Center point coordinates							Number of divisions
	SiO_2	CaO	Al_2O_3	CuO	PbO	BaO	P_2O_5	
Silicon Lead	29.48	2.96	3.17	1.31	41.79	9.56	5.12	21
Silicon-Aluminum	57.49	1.14	5.06	1.16	19.88	7.97	1.12	17
Copper Barium	16.04	1.91	1.18	7.23	29.94	31.15	4.09	6
High Lead	19.35	2.75	2.25	1.52	59.39	4.70	5.08	5

Through the analysis of Table 4 Center point coordinate analysis, it is found that the changes of the three elements of high potassium glass except silica show some similarity, and the trend of changes of these three chemical components are inversely proportional to silica, which is more in line with the definition of weathering, so it is considered that these three subspecies are the description of the degree of weathering, and named according to the degree of weathering. The division of lead-barium glass, is named according to the percentage of its chemical composition.

2.3.3. Sensitivity Analysis

The two types of glasses were polymerized four times, respectively, and their chemical compositions were selected $\alpha+1,\alpha+2,\alpha+3,\alpha+4$. The centroids were solved to calculate the center shift rate η each time, and the results are shown in Figure 3.

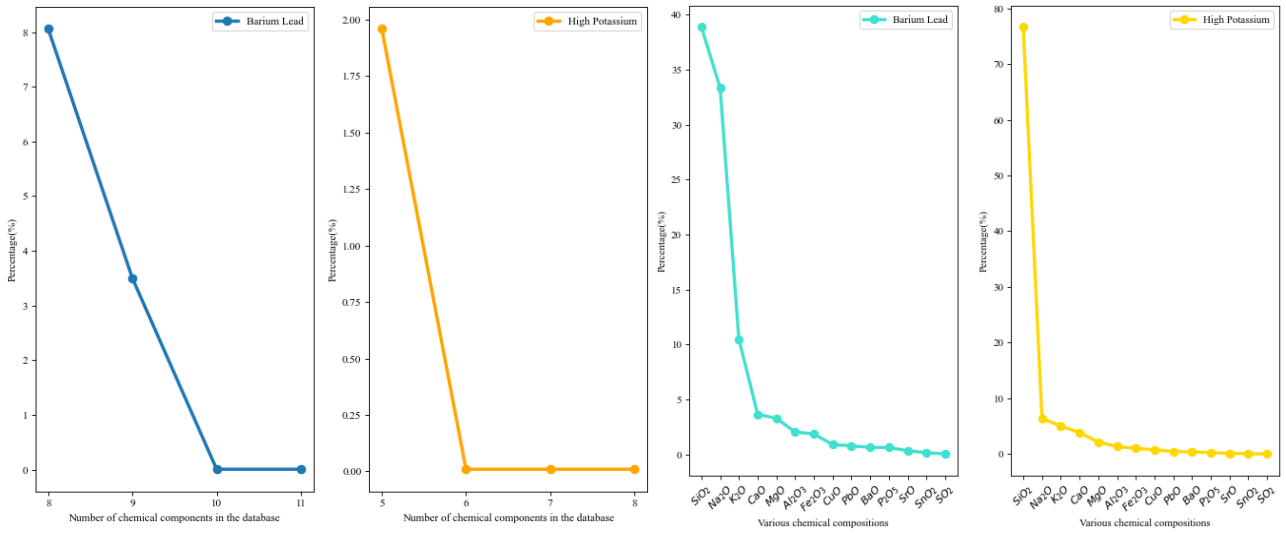

Figure 3. Results of significant differences between the two types of glass

From Figure 3, it can be obtained that the cluster analysis center shift rate was exponentially distributed during the increase of the number of chemical components, which was similar to the distribution of the percentage of chemical components, all of which were exponentially distributed, indicating that the rate of change of the model changed in parallel with the increase of the database, indicating the good sensitivity of the model.

3. Glass identification study

3.1. Modeling of major categories of high potassium glass and lead-barium glass

If the model established in the previous section is to be used to discriminate subclasses of unknown samples in the glass samples to be tested, the first step is to make a broad class judgment of the unknown samples. By observing the data of the glass artifact samples, it can be found that high potassium glass and lead-barium glass have some variability: high potassium glass has a high content of potassium oxide and a low content of lead oxide and barium oxide; lead-barium glass has a low content of potassium oxide and a high content of lead oxide and barium oxide. Therefore, it can be determined qualitatively that the unknown samples belong to high potassium glass or lead-barium glass initially, but some cases cannot be determined qualitatively because some samples have similar content of relevant components. Qualitative judgment can no longer meet the needs of classification of high potassium glass and lead-barium glass, then it is necessary to establish classification criteria to quantitatively discriminate the unknown samples in the glass samples to be tested by the available data, and the model establishment steps are as follows.

(1) Preliminary quantitative judgment and data rejection

The sample with zero potassium oxide content is high potassium glass, and the sample with zero-sum of lead oxide and barium oxide is lead-barium glass. If the content of both is zero, the known glass artifact sample and the glass sample to be tested are combined in a way that is judged to be high potassium glass. Subsequently, the data in the glass artifact samples were processed according to the glass type after the above-mentioned zero values were removed.

(2) Calculate the quantitative classification interval of glass types

Calculate the ratio of the sum of lead oxide and barium oxide content to potassium oxide for high potassium glass and lead-barium glass samples, respectively. (t)

$$t = \frac{(B_i + C_i)}{A_i} (i = 1, 2, \cdots, n) \tag{4}$$

Where A_i is the potassium oxide content of this sample, B_i is the lead oxide content of this sample, and C_i is the barium oxide content of this sample The magnitude of the ratio t for each sample is shown in the lower panel of Figure 4.

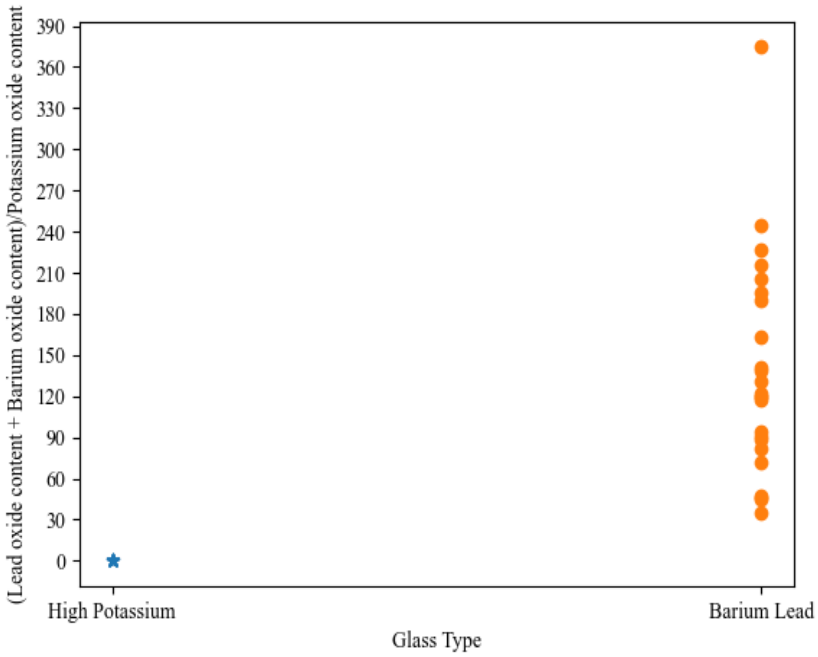

Figure 4. High potassium, lead barium t-value size to scatter plot

It can be seen visually in Figure 4 that there is a large difference between the t-values of the high potassium glass and the lead-barium glass samples. The range of t-values for high potassium glass is [0.01,0.35] and for lead-barium glass is [35.11,374.57]. The t-values of all high potassium glasses were not greater than 1, and the t-values of the lead-barium glass samples were not less than 10. The difference in classification was distinguished significantly, so the model can be used as a basis for the classification of high potassium and lead-barium glasses.

3.2. Classification of glass subclasses to discriminate

Firstly, the categories of the unknown samples in Form 3 were derived based on the large class judgment model.

Secondly, the central value corresponding to each chemical composition indicator obtained by k-means clustering in Problem 2 and the corresponding value of the chemical composition in the attached glass samples to be tested are used as the central indicator and the pending indicator, respectively, and the glass means classification model is constructed by calculating the Euclidean distance between the central indicator and the pending indicator of each category, to quantify the matching degree between the chemical composition of the unknown artifact and each characteristic indicator. The Euclidean distance formula is as follows:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5}$$

Where x_i is the central indicator, y_i is the indicator to be determined, and d is the Euclidean distance (sample matching degree)

Finally, the magnitude of the Euclidean distance is compared. The glass subclass corresponding to the feature indicator with the lowest category match (shortest distance) is taken as the best category for this sample to be tested.

3.3. Model solving

First, the unknown samples of the glass samples to be tested were analyzed for preliminary classification results as shown in Table 9 below:(High potassium glass is abbreviated as HP, and barium lead glass is abbreviated as LB in the table.)

Table 9. Preliminary classification results table

Artifact Number	A1	A2	A3	A4	A5	A6	A7	A8
Type	HP	LB	LB	LB	LB	HP	HP	LB

After analyzing the broad categories to which each unknown sample in the glass sample to be tested belongs, the results shown in Tables 10 and 11 were obtained after performing Euclidean distance calculations for each chemical composition index corresponding to the index to be determined using the central values obtained by solving in the previous section for the different glass categories.

Table 10. Euclidean distance between the unknown sample and high potassium glass center value

High potassium glass	Medium Weathering	High Weathering	Low weathering
A1	7.29	20	17.21
A6	33.93	1.23	93.19
A7	4.46	90.98	61.53

Table 11. Unknown sample and lead barium glass center value Euclidean distance

lead barium glass	Silicon Lead	Silicon Aluminum	Copper barium	High lead
A2	18	29.75	40.63	33.18
A3	33.1	33.59	24.06	51.74
A4	32.02	39.25	45.25	45.25
A5	66.23	69.24	85.36	79.11
A8	57.24	77.29	67.68	87.41

The above results were analyzed to compare the respective matching degrees. Combining the subcategory categories established in the previous section, the categories belonging to the glass samples to be tested can be obtained as shown in Table 12: (The name in the table will be indicated by the initials of the category, e.g., high potassium and high weathering category - HP-HW)

Table 12. Unknown cultural relics sub-class classification results table

Artifact Number	A1	A2	A3	A4
Types of artifacts	HP-MW	LB-SL	LB-CB	LB-SL
Artifact Number	A5	A6	A7	A8
Types of artifacts	LB-SL	HP-HW	HP-MW	LB-SL

4. Conclusion

For the glass classification problem, the data were first processed, and the processing results such as the scattering and mean ranking of the chemical components of the two types of glasses were analyzed, and the division of the two types of glasses was mainly based on the size of the chemical component share. Subsequently, the number of chemical components was determined based on the contribution of each component in the principal component analysis, and α chemical components were selected for K-means cluster analysis from the highest to the lowest mean of the components, and the number of subclasses was determined according to the P value of cluster significance, and the high potassium glass was classified into 3 classes and the lead-barium glass was classified into 4 classes. Finally, the clustering results at K=2 were analyzed to verify the rationality of the model. The

sensitivity analysis of the model was found to be more sensitive because of the correlation between the chemical composition types and the cluster centroid shift.

For the problem of glass identification with unknown composition, since the qualitative judgment cannot meet the classification needs of high potassium glass and lead-barium glass, we first constructed a model for determining the broad categories of glass, and after calculating the quantitative classification intervals of glass types for the data in the glass artifact samples, we obtained the t-value intervals of [0.01,0.35] for high potassium glass and [35.11,374.57] for lead-barium glass, with significant differentiation of classification methods. Second, the classification model of glass mean was constructed by calculating the central indicator of each category (the central value obtained by k-means clustering in Topic II) and the Euclidean distance of the indicators to be determined, which was used to discriminate the subclasses of all unknown samples one by one, and the classification results are shown in Table 14.

References

[1] Chaobo He, Shuangyin Liu, Lei Zhang, Jianhua Zheng. A fuzzy clustering-based method for attributed graph partitioning [J]. Journal of Ambient Intelligence and Humanized Computing, 2019, 10 (9).

[2] Xia Hong, Dong Qingyi, Gao Hui, Chen Yanping, Wang ZhongMin. Service Partition Method Based on Particle Swarm Fuzzy Clustering [J]. Wireless Communications and Mobile Computing, 2021, 2021.

[3] Ana Justel, Marcela Svarc. A divisive clustering method for functional data with special consideration of outliers. [J]. Adv. Data Analysis and Classification, 2018, 12 (3).

[4] Hongjie Jia, Liangjun Wang, Heping Song. HMRF semi-supervised approximate kernel k-means algorithm [J]. Computer Science, 2019, 46 (12): 31 - 37.

[5] Zhuqing Jiao, Xuelian Ming, Yin Cao, Chun Cheng,Shui-Hua Wang. Module partitioning for multilayer brain functional network using weighted clustering ensemble [J]. Journal of Ambient Intelligence and Humanized Computing, 2019 (prepublish).

[6] Shen Yinghua, Pedrycz Witold, Chen Yuan, Wang Xianmin, Gacek Adam. Hyperplane Division in Fuzzy C-Means: Clustering Big Data [J]. IEEE Transactions on Fuzzy Systems, 2019.

[7] Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang. Identification of overlapping community structure in complex networks using fuzzy c -means clustering [J]. Physica A: Statistical Mechanics and its Applications, 2006, 374 (1).

[8] Pu Xiaochuan, Huang Junli, Qi Ning, Song Changsong. Application of K-Means algorithm based on density information entropy in customer segmentation [J]. Journal of Jilin University (Science Edition), 2021, 59 (05): 1245 - 1251. DOI: 10. 13413/j.cnki.jdxblxb.2020190.

[9] Zhang Yadi. Improved K-means algorithm based on quadtree space partitioning and its validity verification [D]. Anhui University, 2021. DOI: 10.26917/d.cnki.ganhu.2021.000319.

[10] Yi Xu, Zhi Zhuang, Weimin Li, Xiaokang Zhou. Effective community division based on improved spectral clustering [J]. Neurocomputing, 2018, 279.

[11] Lei Mingjun, Chen Jian, Chen Xinzhao, Wang Jiannan. Application of sensitivity analysis method in body lightweighting [J]. Automotive Engineering, 2009, 31 (07): 682 – 685.