Classification of Glass Products Based on Clustering Algorithm

Sixing Yan *, #, Keyu Zhang #, Jianxiang Sun #

Chongqing Jiaotong University, Chongqing, China, 400074

* Corresponding author: sihang0219@gmail.com

#These authors contributed equally.

Abstract. The analysis and identification of glass cultural relics is an important scientific basis for further research on the source and process analysis of glass. First, in order to clarify the classification basis for high-potassium and lead-barium glass, K-means cluster analysis was used to analyze and screen out suitable main related elements as the basis for glass classification. Finally, the accuracy and stability of the model were verified by comparing data, consulting literature, and analyzing sensitivity.

Keywords: Component identification, Cluster analysis, Data analysis.

1. Introduction

The study of glass objects has long been an important part of archaeological research. Ancient glass is extremely susceptible to weathering due to the influence of the burial environment. During the weathering process, a large number of internal elements and environmental elements exchanged, resulting in changes in its composition ratio, thereby affecting the correct judgment of its category. Classification and identification of ancient glassware is a challenging task, and various methods have been proposed to classify glass in recent years, including traditional chemical analysis and modern analytical techniques.

2. Literature review

The analysis and identification of glass cultural relics is of great significance for further research on its origin and craftsmanship. In recent years, mathematical modeling methods are increasingly used in the composition analysis and identification of substances. Mathematical modeling methods for ancient glass composition analysis and identification include principal component analysis (PCA), vector machine (SVM), error backpropagation neural network (BP), etc. These methods have been shown to improve the accuracy of ancient glass classification and identification, automate the process, and even enable multi-class classification. For example, Xian, Y. (2020) using PCA to analyze the composition and source identification process of turquoise cultural relics unearthed in Xinjiang [1]; Wen, D. (2021) combining BP and SVM for high-precision classification of ore samples [2]; Xiao, L. (2020) combining PCA and cluster analysis to classify ancient glass samples based on compositional data [3].

Despite these advances, there are still some limitations in the analysis and identification of ancient glass composition using mathematical modeling methods [4]. These include limited availability of high-quality compositional data, limited generalizability of models, and challenges in model validation. Despite their limitations, these methods have been shown to be effective in improving the accuracy and efficiency of ancient glass classification and identification.

3. Clustering Algorithm Model

3.1. K-means clustering model

The K-Means algorithm is a typical partition-based clustering algorithm and an unsupervised learning algorithm [5]. For a given sample set, according to the Euclidean distance between the samples, the sample set is divided into k specified clusters, so that the points in the cluster are connected as closely as possible[6], and the clusters and clusters are kept The Euclidean distance between them is as large as possible, recalculate the average value of each cluster point as the new cluster center, and continue to iterate this process cycle until the cluster center no longer changes, or the algorithm reaches the set The number of iterations, as shown in Figure 1 is the flow of the algorithm.

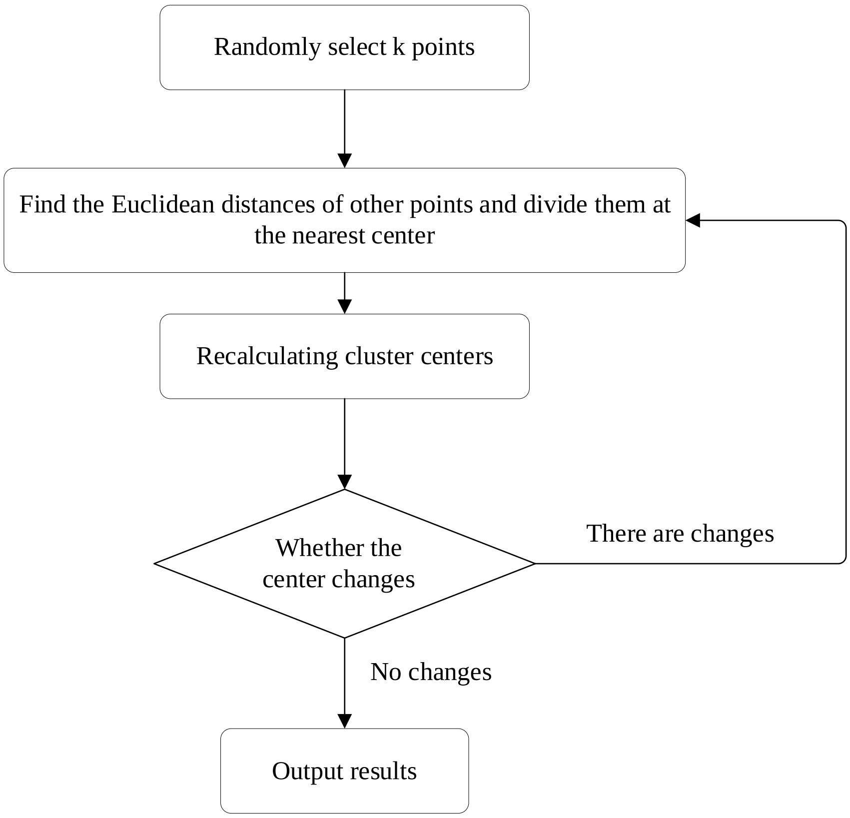

Figure 1. K-Means algorithm flow chart

Suppose $x_i(i=1,2,\cdots,n)$ is the data point, $\mu_j(j=1,2,\cdots,k)$ is the initialized data center, then the objective function is written as:

$$\min\sum_{i=1}^{n}\min_{j=1,2,\ldots,k}\|x_i-\mu_j\|^2 \tag{1}$$

3.2. Phylogenetic clustering model

Systematic clustering is a clustering analysis method that clusters data points into corresponding clusters by iteratively and continuously adjusting cluster centers [7]. The method calculates the distance between two types of point data, further combines the two groups with the smallest value, and then iterates this process until the calculation is aggregated into one type [8]. In the middle process, many cluster combinations will be generated, which can be distinguished by the coefficient draws an elbow diagram to select the optimal cut-off point to obtain a suitable clustering result. Figure 2 is the flow chart of the system clustering algorithm.

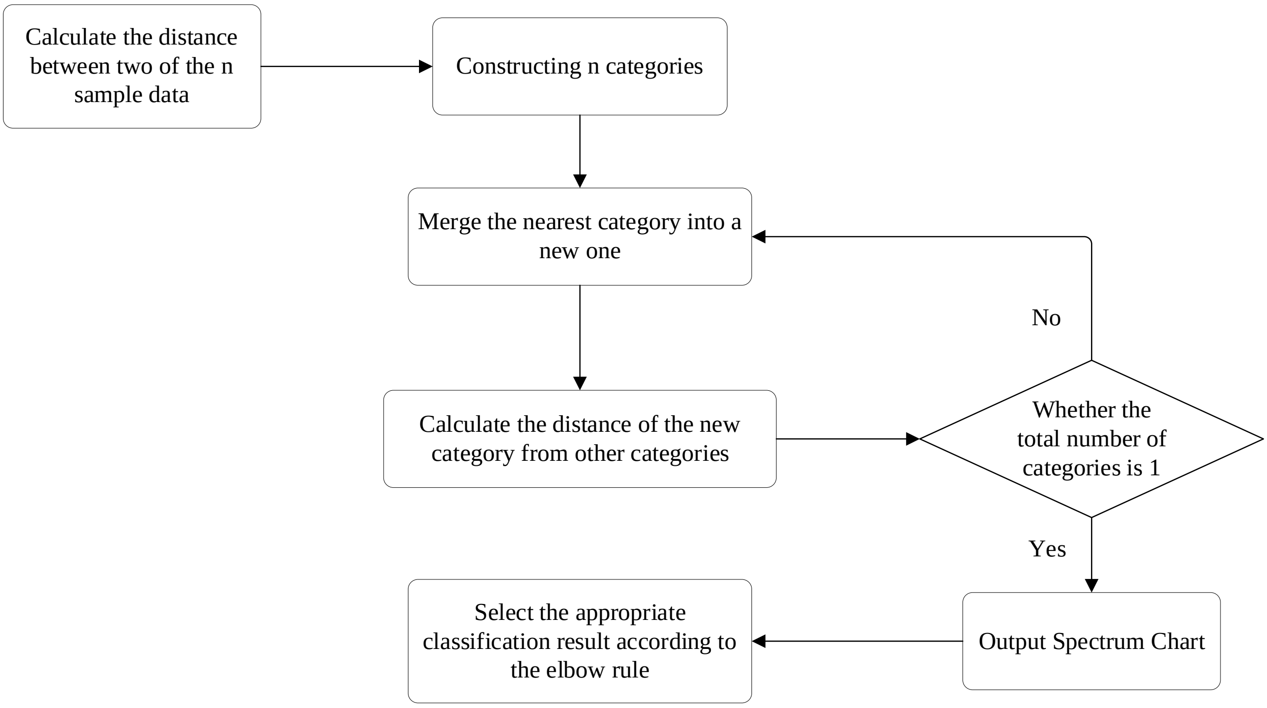

Figure 2. System clustering algorithm flow chart

3.3. Elbow rule

The sum of the degree of distortion of each class: the degree of distortion of each class is equal to the sum of the squares of the Euclidean distances between the center of the class and each point. Suppose a total of n samples are divided into K categories (K≤n-1, at least one category contains two points) [9].

Let C_k represent the kth class (k=1, 2…, K), and the position of the center of the class is recorded as u_k, then the degree of distortion of the kth class is:

$$\sum_{i\epsilon C_k} |x_i - u_k|^2 \tag{2}$$

Then define the total distortion degree of all categories:

$$J = \sum_{k=1}^{K} \sum_{i\epsilon C_k} |x_i - u_k|^2 \tag{3}$$

J is called the aggregation coefficient. The elbow rule is actually to find a more appropriate classification result by observing the line graph of the aggregation coefficient J in descending order of the clustering times.

Judgment criteria:

(1) The polymerization coefficient drops suddenly after reaching a certain polymerization;

(2) The aggregation coefficient tends to level off after reaching a certain aggregation and no longer changes significantly

4. Results

4.1. The main elements of the classification of large categories of glass

First, explore the classification rules of high-potassium and lead-barium glasses using the K-means classification of SPSS, substituting all the chemical composition data to obtain the following Table 1:

Table 1. K-means initial classification result table

Heritage sampling points	16	18	20	21	23	24	25	28	29
Known correct species	√	√	◯	√	◯	◯	◯	◯	◯
Substitute all elements	√	√	◯	√	◯	◯	◯	√	√
Heritage sampling points	30a	30b	31	32	33	35	37	42a	42b
Known correct species	◯	◯	◯	◯	◯	◯	◯	◯	◯
Substitute all elements	◯	◯	√	√	√	√	√	◯	◯

Tips: √ and ◯ are only used as classification representatives, there is no right or wrong, look down vertically, if the same symbol as the known category line, the classification is correct.

It is found that the clustering of species in the first half is relatively correct, but there is a large error in the second half, which shows that the chemical composition data of the independent variable cannot be fully substituted into the analysis, and a trade-off is required. The names of the two kinds of glasses are high potassium and lead-barium respectively. As the name suggests, the elements may be relatively large, so you can try to use them as the basis for clustering.

So, we changed the clustering basis variable to lead, barium and potassium, lead and barium respectively to compare with the known correct classification, as shown in Table 2:

Table 2. K-means last few classifications results table

Heritage sampling points	16	18	20	21	23	24	25	28	29
Known correct species	√	√	◯	√	◯	◯	◯	◯	◯
Substitute lead and barium	√	√	◯	√	◯	◯	◯	◯	√
Substitute potassium,lead and barium	√	√	◯	√	◯	◯	◯	◯	◯
Heritage sampling points	30a	30b	31	32	33	35	37	42a	42b
Known correct species	◯	◯	◯	◯	◯	◯	◯	◯	◯
Substitute lead and barium	◯	◯	◯	◯	◯	◯	◯	◯	◯
Substitute potassium,lead and barium	◯	◯	◯	◯	◯	◯	◯	◯	◯

It can be clearly seen that the second group that chooses lead and barium elements is inconsistent with the correct category only at the unweathered point of 29, while the third group that chooses potassium, lead, and barium elements is completely consistent with the correct category, so it can be analyzed that the classification of high-potassium and lead-barium glasses is based on potassium, lead, and barium elements. The experiment also tested other combinations, but the effect is not as good as the aggregation effect of the third group, which can verify the rationality of the characteristic variables selected by the third group.

4.2. Results of further subclassification of glass

Select the appropriate chemical composition for the two major categories of high-potassium glass and lead-barium glass to divide them into subcategories. Use SPSS to import two sets of data of all unweathered high-potassium glass and lead-barium glass respectively, substitute all the chemical composition data, and use the system clustering method to further divide the two types of glasses into subcategories [10], as shown in Figure 3 and Figure 4. Pedigree chart:

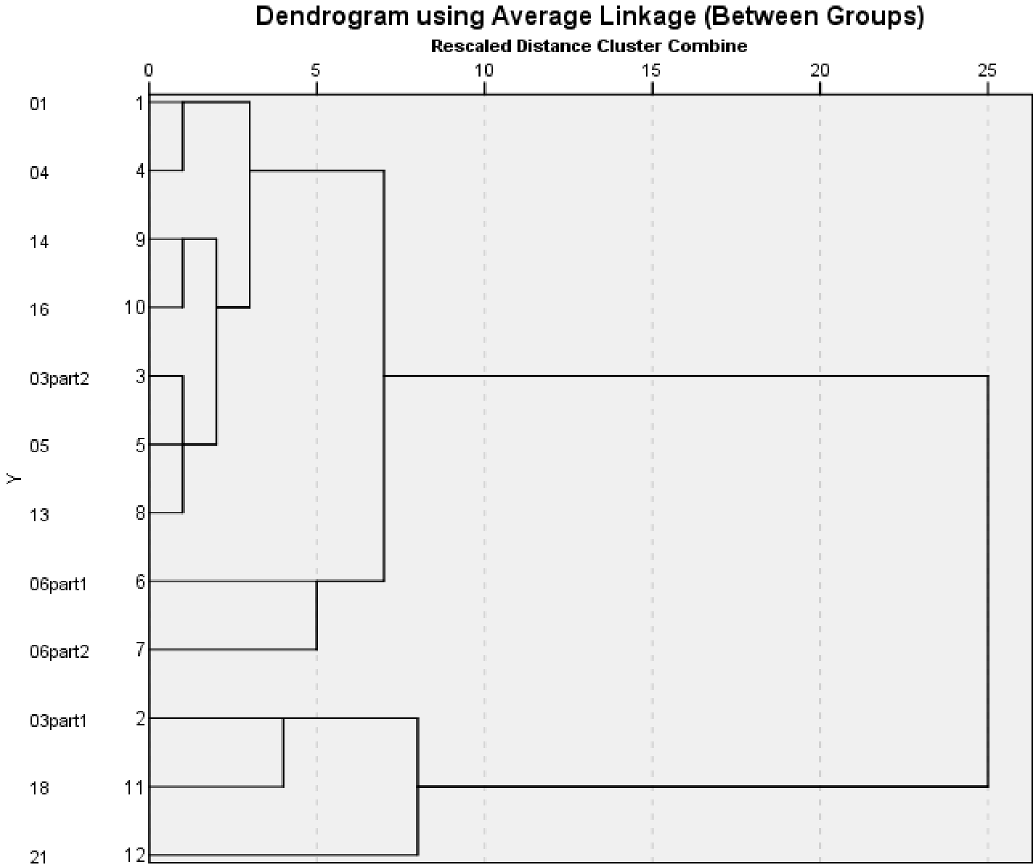

Figure 3. Cluster pedigree diagram of high potassium glass

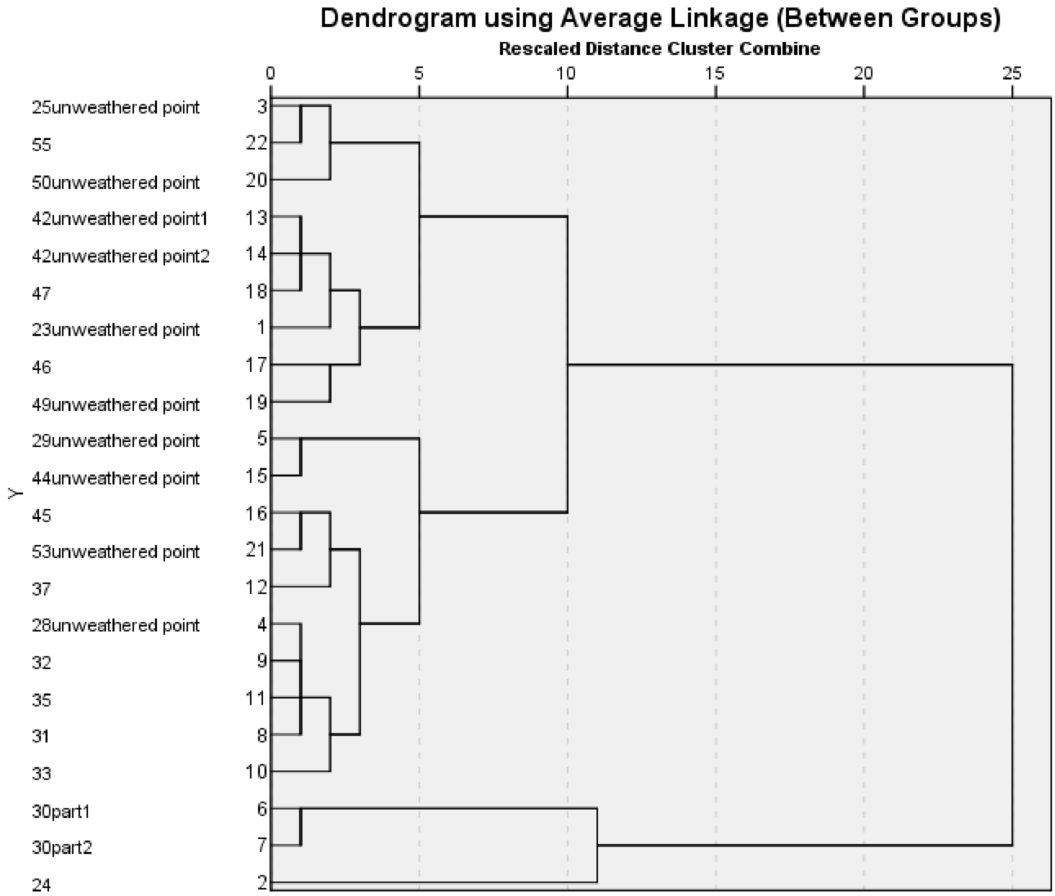

Figure 4. Cluster pedigree diagram of lead-barium glass

Then record the polyline (elbow diagram) of the aggregation coefficient and the aggregation coefficient of the category in the clustering process of the two types of glass systems, as shown in Figure 5 and Figure6:

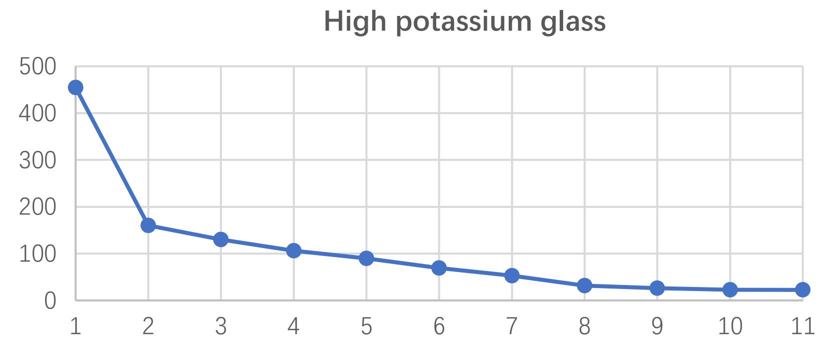

Figure 5. The iterative line graph of aggregation coefficient J of high potassium glass during the clustering process

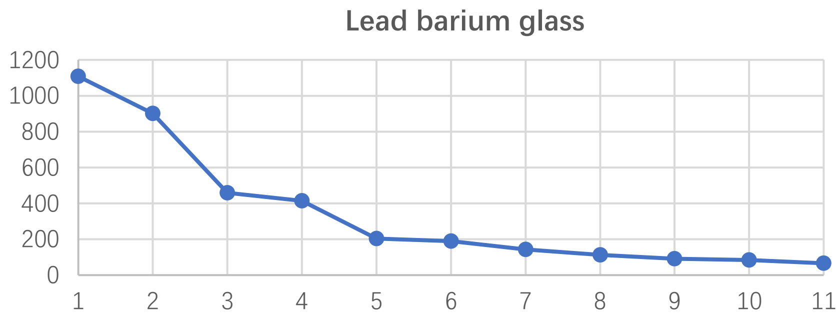

Figure 6. The iterative line graph of the aggregation coefficient J of lead-barium glass in the clustering process

As shown in Figure 5 and Figure 6, in the elbow diagram of high-potassium glass, when the aggregation category is 2, the aggregation coefficient drops sharply, and then the change tends to be gentle, so we set the number of categories of high-potassium glass to 2 ;Similarly, in the elbow diagram of lead-barium glass, when the polymerization coefficient is 3, it drops sharply, but it does not tend to level off until it is 5, so the number of categories of lead-barium glass is set to 5.

Lead-barium glass: Glass No. 20 has a chemical composition content of only 88.41%, which leads to deviations in clustering. It is a separate group and will not be counted as a sub-category. The subspecies can be divided into 4 types: Pb-Ba-Na glass, Pb-Ba-Ca-Fe glass, Pb-Ba-Cu glass, Pb-Ba glass.

High-potassium glass: There is no obvious deviation in each sample, and the characteristic elements can be screened out as calcium and aluminum. Therefore, high-potassium glass subspecies can be divided into two categories: K-Ca-Al glass and K-Al glass. Analyze the rationality of the classification results. We first screened through Excel to observe whether the sub-categories of glasses have the characteristics of specific components, as shown in Table 3:

Table 3. Specific compositional characteristics of glass

	High potassium glass			Lead barium glass		
Composition	K_2O	PbO	BaO	K_2O	PbO	BaO
Unweathered	0-14.52	0-1.62	0-2.86	0-1.41	9.3-39.22	2.03-26.23
Weathering	0-1.01	0	0	0-1.05	15.71-70.21	0-32.25

Consult recent archaeological documents on the detailed classification of glass cultural relics [11], as shown in Figure 7 and Figure 8:

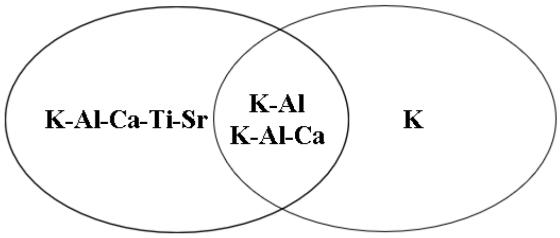

Figure 7. Subclassification of high potassium glasses.

Note: The big circle on the left is the subcategory of high potassium glass in the literature, and the big circle on the right is the subcategory of high potassium glass in the cluster analysis.

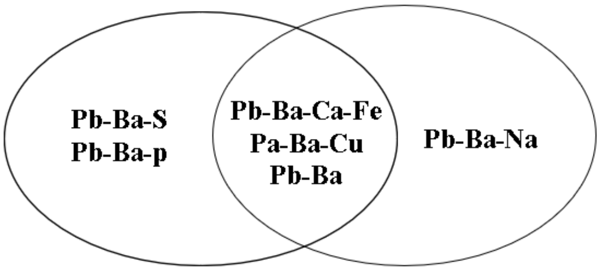

Figure 8. Subclassification of lead-barium glass

Note: The large circle on the left is the subcategory of lead-barium glass in the literature, and the large circle on the right is the subcategory of lead-barium glass in the cluster analysis.

It can be seen that there are many overlaps with the subcategories of glass in archaeological literature, which verifies the rationality and reliability of the results.

4.3. Sensitivity Analysis of Model Results

Sensitivity is a test of an established model. With other parameters unchanged, we can use the perturbation method to simply verify the reliability of the model. Change the important reference composition data in the glass calculated by the model, and then perform cluster analysis to observe whether the classification result of the glass changes significantly [12].

In cluster analysis, the high sensitivity of a variable means that a small change in the value of the variable may lead to a large change in the cluster assignment, indicating that the variable has a high proportion in distinguishing different clusters, so it should be given more weight in the analysis. High weight, while ensuring the reliability of the verification results [13].

For the subcategories of Pb-Ba-Ca-Fe glass and Pb-Ba-Cu glass in lead-barium glass, samples are randomly selected in the subcategory, and when other data remain unchanged, the important elements that determine the classification are changed to average value to see if the subclass changes.

For the data of part 1 of Pb-Ba-Ca-Fe glass sample 30, only the calcium and iron elements were adjusted to the average value, and the cluster iteration was performed, and new changes were found, as shown in Figure 9:

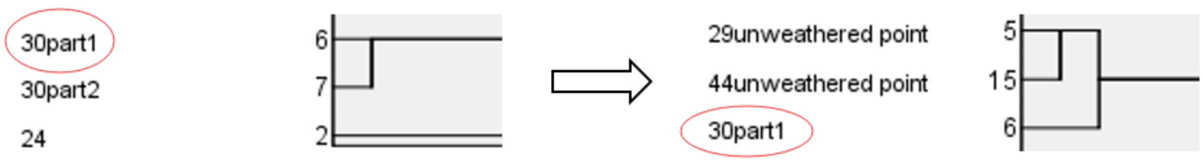

Figure 9. Clustering iterative partition difference

Similarly, changing the Pb-Ba-Cu glass sample data for systematic clustering cluster analysis has similar results, which can verify that the key classification elements determined by the model are accurate.

5. Conclusion

In conclusion, the study aimed at classifying ancient glass compositions using K-means clustering analysis. The results of the model and multiple experiments determined the main elements used for classifying high potassium and lead-barium glass into larger categories. Further clustering analysis was conducted to subcategorize the two main glass categories. Finally, sensitivity analysis was carried out to verify the accuracy of the model. The study provides valuable insights into the classification of ancient glass compositions and has potential implications for the field of archaeology and materials science. It is hoped that this research will contribute to a deeper understanding of ancient glass production and usage.

References

[1] Xian Yiheng, Li Xintong, Zhou Xueqi, et al. Composition analysis and origin identification of turquoise cultural relics unearthed from two sites in Xinjiang [J]. Spectroscopy and Spectral Analysis, 2020, 40 (3): 4.

[2] Wen Dapeng, Liang Xiyin, Su Maogen, et al. Classification and identification of ores by laser-induced breakdown spectroscopy combined with PCA-PSO-SVM [J]. Progress in Lasers and Optoelectronics, 2021 (023): 058.

[3] Xiao Liang, Yong Yuanyuan, Wang Xing, et al. Quantitative classification of Miocene Rhizoma fruit fossils in eastern Zhejiang [J]. Frontiers of Earth Science, 2020 (004): 027.

[4] Ioulia Papageorgiou. Ceramic investigation: how to perform statistical analyses [J]. Archaeological and Anthropological Sciences, 2020, 12 (9): 150 - 165.

[5] Mohiuddin Ahmed, Raihan Seraj, Syed Mohammed Shamsul Islam, et al. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation [J]. Electronics, 2020, 9 (8).

[6] Siwei Wang, Miaomiao Li, Ning Hu, et al. K-Means Clustering with Incomplete Data. [J]. IEEE Access, 2019, 7: 69162 - 69171.

[7] Enrique H. Ruspini, James C. Bezdek, James M. Keller, et al. Fuzzy Clustering: A Historical Perspective [J]. IEEE Computational Intelligence Magazine, 2019, 14 (1): 45 - 55.

[8] Mohammad Jafarzadegan, Faramarz Safi-Esfahani, Zahra Beheshti, et al. Combining hierarchical clustering approaches using the PCA method [J]. Expert Systems with Applications, 2019, 137: 1 - 10.

[9] Emrah Hancer, Bing Xue, Mengjie Zhang, et al. A survey on feature selection approaches for clustering[J]. Artificial Intelligence Review: An International Science and Engineering Journal, 2020, 53 (6): 4519 - 4545.

[10] Kaminskyy Roman, Shakhovska Nataliya, Kryvinska Natalia, et al. Dendrograms-based disclosure method for evaluating cluster analysis in the IoT domain [J]. Computers & Industrial Engineering, 2021, 158.

[11] Adolfo Molada-Tebar, Ángel Marqués-Mateu, José Luis Lerma, et al. Dominant Color Extraction with K-Means for Camera Characterization in Cultural Heritage Documentation [J]. Remote Sensing, 2020, 12 (3): 520 - 520.

[12] Razavi Saman, Jakeman Anthony, Saltelli Andrea, et al. The Future of Sensitivity Analysis: An Essential Discipline for Systems Modeling and Policy Support [J]. Environmental Modelling & Software, 2020, 104954-.

[13] Felix Riede, Christian Hoggard, Stephen Shennan, et al. Reconciling material cultures in archaeology with genetic data requires robust cultural evolutionary taxonomies [J]. Palgrave Communications, 2019, 5 (1): 1 - 9.