ISSN 2616-5880 Vol. 4, Issue 5: 25-30, DOI: 10.25236/AJMC.2023.040504

Analysis and identification of the compositional correlations of glasses and the variability of different classes of compositions

Zhengyang Wang, Wenbo Shen*

University of Aeronautics and Astronautics, Nanjing, 210016, China
**Corresponding author: shenwenbo@nuaa.edu.cn*

Abstract: *The weathering process of ancient glass causes changes in the ratio of its chemical composition due to the influence of the external environment, which affects the correct determination of its category. In order to classify the composition of glass products and thus analyze the variability between the chemical compositions obtained. This paper establishes a glass classification model based on K-Means++ , and uses the unsupervised learning K-Means algorithm to find the seeds of the center of mass of K-means clustering by heuristic method, and determines the optimal number of classifications by C-H value, and analyzes the clustering of different types of glass before and after weathering and the changes of the categories to which each chemical component belongs before and after weathering according to the clustering results. The CHI index of high potassium glass increased significantly after weathering, and that of lead-barium glass decreased relatively after weathering; for the DBI index, the DBI index of high potassium glass increased and that of lead-barium glass decreased after weathering; for the contour coefficients, the analysis showed that the chemical composition correlations were stronger after weathering for lead-barium glass and weaker after weathering for high potassium glass.*

Keywords: *Glass classification model, K-Means++, Heuristic approach*

1. Introduction

Early glass was often made into ornaments imported into China, and our ancient glass was made locally after learning its techniques, so it is similar in appearance to foreign glass products, but not the same chemical composition. Ancient glass is highly susceptible to weathering by the burial environment [1]. During the weathering process, a large number of exchanges between elements occur, resulting in changes in the proportions of its composition, which in turn affects the judgment of its category. Finally, we analyzed the correlation between the chemical compositions of different types of glass artifacts and compared the differences in the chemical composition correlation between the different types of glass artifacts.

In the past few years, there have been many studies on the analysis of the composition of ancient artifacts, most of them using visual methods or optical inspection with specific instruments, which is time-consuming and laborious. The K-Means++ [2] algorithm is an optimization of the K-Means [3-4] method of randomly initializing the centers of mass, and the most essential difference between the K-Means algorithm and the K-Means algorithm [5] is the initialization process of the k clustering centers.

In this regard, we propose a K-means++-based [6-7] heuristic clustering model that can be used to classify artifacts for the analysis of their chemical composition.

2. Modeling and solution

2.1 Development of K-Means++-based model for glass type variability analysis

The basic principle of K-Means++ algorithm in the initialization process of cluster centers is to make the initial cluster centers as far away from each other as possible, so as to avoid the slow convergence problem mentioned above. The initialization process of K-Means++ algorithm is as follows: First, a sample point is selected randomly in the data set as the first initialized cluster center. Then the rest of the clustering centers are selected. Then calculate the distance between each sample point in the sample and the initialized clustering center, and choose the shortest distance among them, denoted as d_i. Then a new

ISSN 2616-5880 Vol. 4, Issue 5: 25-30, DOI: 10.25236/AJMC.2023.040504

data point is selected as the new cluster center. The principle of selection is that the point with larger distance has a higher probability of being selected as the cluster center. The above process is repeated until all *k* clustering centers are determined. For the *k* initialized clustering centers, the K-Means algorithm is used to calculate the final clustering centers.

2.2 Selecting the optimal number of clusters based on C-H values

The *Calinski-harabaz* index is essentially the ratio of inter-cluster distance to intra-cluster distance, and the overall calculation process is similar to the variance calculation, so it is also called the variance ratio criterion.

The data set X of capacity N is clustered into K classes, and the intra-class tightness (intra-class distance) is measured by calculating the sum of squares of the distances between each point in the class and the class center, and the separation of the data set (inter-class distance) is measured by the sum of squares of the distances between each class centroid and the data set centroid.

The formula for the C-H index is

$$s = \frac{\mathrm{tr}\left(B_k\right)(N-K)}{\mathrm{tr}\left(W_k\right)(K-1)}, \tag{1}$$

where B_k is the covariance matrix between classes and the covariance matrix of the data within class W_k. The detailed formula is as follows

$$B_k = \sum_{q=1}^{k} n_q \left(c_q - c_e\right)\left(c_q - c_e\right)^T, \tag{2}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} \left(x - c_q\right)\left(x - c_q\right)^T. \tag{3}$$

where c_q denotes the centroid of class q, c_B denotes the centroid of the data set, n_q denotes the number of data in class q, and c_q denotes the set of data in class q. The larger the score of the *Calinski-harabaz* index, the better it is (the smaller the intra-category covariance, the larger the inter-category covariance.

2.3 Solution of various evaluation metrics for K-means++ model

The *Calinski-harabaz* criterion, sometimes called the variance ratio criterion, can be used to determine the optimal *K* (number of clusters) value for clustering, which corresponds to having a large inter-cluster variance and a small intra-cluster variance, and the optimal number of clusters corresponds to having the highest.

Calinski-harabaz index value, set in *Matlab* in the range 2 to 5, where the larger the *C-H* value is, the better.

The scatter plot between each variable is drawn up and the cluster centers are represented by other shapes. The clustering effect is observed through the histogram, and the smaller the cross section is, the better the clustering effect is, and vice versa.

A visual analysis of high potassium glass compared to lead-barium glass before and after weathering is shown in Figure 1.

In addition to this we also programmed *Matlab* to solve for the clustering centroids with the classification of each chemical component, and the comparison of the categories to which the chemical components of different glass types belong before and after weathering is shown in the following table 1.

ISSN 2616-5880 Vol. 4, Issue 5: 25-30, DOI: 10.25236/AJMC.2023.040504

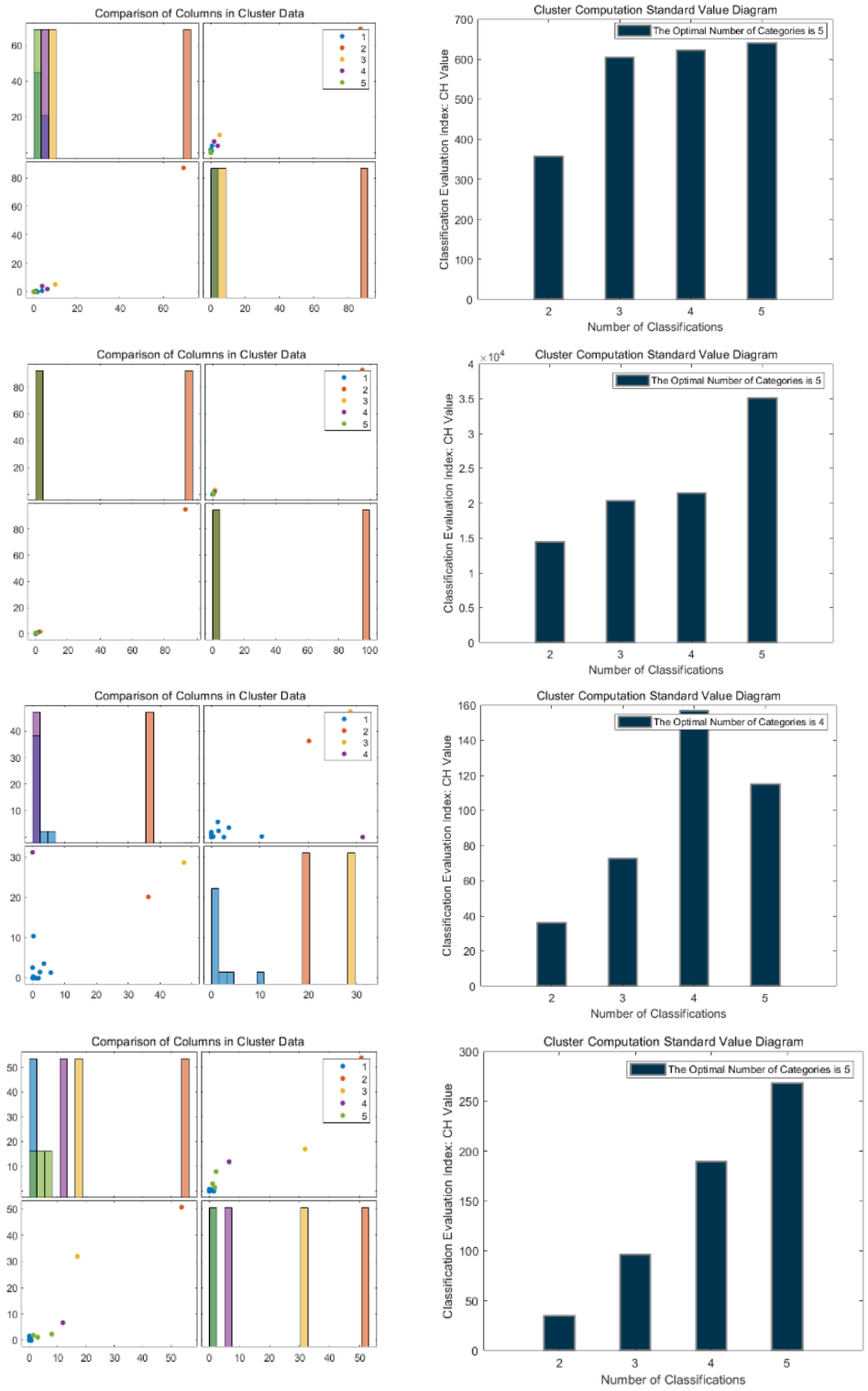

Figure 1: Histogram of clustering effect analysis before and after weathering

From the analysis of table 1, we can get: in the high potassium glass, the four chemical elements of silica (SiO_2), calcium oxide (CaO), aluminum oxide (Al_2O_3) and copper oxide (CuO) have changed before and after weathering, while all other chemical components belong to the same category before and after weathering; in the lead barium glass, only the chemical components of silica (SiO_2), aluminum oxide (Al_2O_3), lead oxide (PbO) and barium oxide (BaO) have changed before and after weathering, and all other chemical components belong to the same category before and after weathering.

ISSN 2616-5880 Vol. 4, Issue 5: 25-30, DOI: 10.25236/AJMC.2023.040504

Table 1: Comparison of the categories to which the chemical compositions of different glass types belong before and after weathering

Sort by Type	Classification			
	High Potassium non-weathering	High Potassium weathering	Lead-barium without weathering	Lead-barium weathering
SO_2	2	1	2	3
Na_2O	3	2	1	1
K_2O	1	2	1	1
CaO	4	5	1	1
MgO	3	2	1	1
Al_2O_3	4	3	4	1
Fe_2O_3	3	2	1	1
CuO	3	4	1	1
PbO	3	2	3	2
BaO	3	2	5	4
P_2O_5	3	2	1	1
SrO	3	2	1	1
SnO_2	3	2	1	1
SO_2	3	2	1	1

2.4 Differential judgments of chemical composition association relationships between different categories

After clustering analysis using the K-Means++ algorithm, how to evaluate the clustering results, how to determine the advantages and disadvantages of the results of each clustering algorithm, and how to determine the values of the parameters of the clustering algorithm can be illustrated from one side by the clustering performance metrics and the selection of the parameters of the clustering algorithm and algorithm. Clustering performance metrics are usually divided into external and internal metrics. The internal metrics are unsupervised and do not require a benchmark data set or an external reference model, and the distance between the sample points and the cluster center in the sample data set is used to measure the merit of the clustering results.

1) CHI Indices: Kalinsky-Harabas Index

$$s(k) = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \frac{N-k}{k-1}, \tag{4}$$

where N is the sample size in the data set, k is the number of clusters (i.e., the number of categories), is the inter-cluster dispersion matrix, i.e., the covariance matrix between different clusters, is the intra-cluster dispersion matrix, i.e., the covariance matrix of the data within a cluster, and tr denotes the trace of the matrix. In linear algebra, the sum of the elements on the main diagonal (the diagonal from the top left to the bottom right) of a matrix A of order n is called the trace (or number of traces) of the matrix A. The higher the dispersion between the data, the larger the trace of the covariance matrix will be. The lower the dispersion within the group, the smaller the trace of the covariance will be, and the smaller it will be, while the greater the dispersion between the groups, the larger the trace of the covariance will be, which is exactly what we want, so the higher the *Calinski-harabaz* index, the better. The evaluation criterion is, in summary, the larger the better.

2) DBI index: It is a clustering algorithm that evaluates a metric *DB* calculates the average sum of intra-class distances of any two classes divided by the distance between the centers of the two classes and finds the maximum value. A smaller *DB* means a smaller intra-class distance and a larger inter-class distance.

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j, i, j \in [1,K]} \frac{s_i + s_j}{M_{ij}}, \tag{5}$$

Where denotes the dispersion of sample points in a class, then is the distance between class *i* and the center of class *j*. The calculation formula is shown below.

ISSN 2616-5880 Vol. 4, Issue 5: 25-30, DOI: 10.25236/AJMC.2023.040504

$$s_i = \left\{ \frac{1}{n} \sum_{j=1}^{n} \left| X_{ij} - A_i \right|^q \right\}^{\frac{1}{q}}, \tag{6}$$

where denotes the *j* th data point in class *i*; denotes the center of class *i*; n denotes the number of data points in class *i*; q takes 1 to denote: the mean of the distance from each point to the center, and *q* takes *2* to denote: the standard deviation of the distance from each point to the center, both of which can be used to measure the degree of dispersion.

$$M_{ij} = \left\{ \sum_{k=1}^{K} \left| a_{ki} - a_{kj} \right|^q \right\}^{\frac{1}{q}}, \tag{7}$$

where denotes the value of the *K*th attribute of the centroid of the *i*-th class.

3) Contour coefficient: For a sample set, its contour coefficient is the average of all sample contour coefficients. The range of the contour coefficients is [-1,1], and the more similar the distance of samples in the same category and the more distant the distance of samples in different categories, the higher the score.

The profile coefficient *s* for a given sample is:

$$s = \frac{b - a}{\text{sum}(a, b)}, \tag{8}$$

where a denotes the average distance of a sample from other samples in its cluster and b denotes the average distance of a sample from other samples in the cluster.

The total contour coefficient of clustering *SC* is

$$SC = \frac{1}{N} \sum_{i=1}^{N} s_i, \tag{9}$$

The final solution was programmed in Matlab to derive the changes in the indicators before and after weathering for different glass types, which were analyzed as shown in the following table.

Table 2: Comparative analysis of K-Means++ clustering evaluation indexes

Category	Optimal number of classifications	CHI	DBI	Profile factor
High Potassium non-weathering	4	622.815	0.39238	0.84211
High Potassium weathering	5	21478.3469	0.42253	0.7886
Lead-barium non-weathering	5	387.5223	0.12902	0.78246
Lead-barium weathering	4	157.051	0.10074	0.93844

From the above table 2, we can get that the number of classification of both high potassium glass and lead-barium glass will decrease after weathering. The *CHI* index of high potassium glass increases significantly after weathering, while the *CHI* index of lead-barium glass decreases somewhat after weathering, and we can identify whether the clustering effect of each glass type is good or not based on the *CHI* index. For the *DBI* index, there is a certain increase of high potassium glass after weathering, while there is a slight decrease of lead-barium glass; besides, the contour coefficient of high potassium glass is farther from 1 after weathering, while the contour coefficient of lead-barium glass is close to 1 after weathering. The trend is that the correlation between the chemical components of high potassium glass is weaker after weathering, while the correlation between the components of lead-barium glass is increasing.

3. Results

The choice of the location of the k initialized centers of mass has a significant impact on the final clustering results and the running time, so the choice of the appropriate k centers of mass is required. The K-Means++ algorithm is an optimization of the K-Means method of randomly initializing the center of

ISSN 2616-5880 Vol. 4, Issue 5: 25-30, DOI: 10.25236/AJMC.2023.040504

masses. The CHI, DBI, and contour coefficient are representative evaluation indicators in the clustering algorithm, and the introduction of these parameters can show the correlation between the chemical components and the clustering effect more intuitively.

References

[1] Mo Dingyuan. Research on the relationship between GDP and industrial structure in Baise City based on gray correlation analysis [J]. Mall Modernization, 2022(12): 124-126.
[2] Zhu Yingpei. Research on the composition system and production process of glass beads excavated from Yanghai Cemetery, Shanshan County, Xinjiang [D]. Northwestern University, 2018.
[3] Huang ZY, Zhou Lingke. Fault classification based on K-means Bayes and AdaBoost-SVM [J]. Computer System Applications, 2022, 31(07): 239-246.
[4] Kang Wei, Cao Wengeng, Xu Lixia, Nan Tian, Gao Yuanyuan, Nie Ziyi. A representative groundwater level calculation method based on k-means clustering and Tyson declustering [J/OL]. South-North Water Diversion and Water Science and Technology (in English and Chinese): 1-12 [2022-09-18]. http://kns.cnki.net/kcms/detail/13.1430.TV.20220726.1029.002.html
[5] Pu Huizhong. K-means cluster analysis algorithm in artificial intelligence + personalized learning system [J]. Intelligent Computers and Applications, 2022, 12(08): 152-156.
[6] Lin E Y. Starbucks as the Third Place: Glimpses into Taiwan's Consumer Culture and Lifestyles[J]. Journal of International Consumer Marketing, 2012, 24(1-2):119-128.
[7] Hu Yongkai. Big data visualization implementation under cloud platform [D]. University of Electronic Science and Technology, 2022.