Glass Classification and Identification Based on K-Means Clustering Model

Yiting Lv[1], Hengyuan Tan[1], Feier Meng[1], Guanghu Shao[2,*], Sizhe Yang[1]

[1] School of Electrical and Automation Engineering, Hefei University of Technology, Hefei, Anhui, 230009

[2] Mechanics institute, Hefei University of Technology, Hefei, Anhui, 230009

* Corresponding Author Email: lights6782@gmail.com

Abstract. This paper proposes a method for the classification and identification of glass artefacts based on a K-mean clustering model. Using weathered and unweathered glass with high potassium and lead-barium as the main constituent as the main samples, a cluster analysis of their chemical composition was carried out to derive subclass subdivision and extract characteristic chemical components, based on which a sub-classification model of unknown glass based on the K-mean clustering model was established to provide a new method for the identification of glass artefact types. The results were analysed for reasonableness and sensitivity, and the model was proved to be of generalisation and application value.

Keywords: Entropy weighting, K-means clustering model, Spearman's correlation coefficient method, multi-factor variance method.

1. Introduction

The classification and identification of glass artefacts are of great importance to the study of ancient glass. The chemical composition of glass artefacts reflects their essential properties, and in-depth data analysis can provide an effective means of classifying and identifying glass artefacts. In recent years, there has been a growing interest in classifying and identifying ancient glass based on chemical composition analysis.

The K-means clustering model has been studied in great depth and its application is widespread. In the literature [1], K-means clustering algorithms were applied to categorise the intensity of low carbon travel intentions of residents, and in [2], K-means clustering algorithms were used to classify and identify sediment anomalies in aqueous systems, but this method has rarely been applied to the classification and identification of palaeoglass. The literature [3] points out that chemical composition analysis of glass artefacts is rarely carried out at this stage, and there is an obvious gap in the study of glass artefacts relying on chemical composition data analysis. This paper, therefore, proposes a K-mean clustering model-based method for the classification and identification of glass artefacts, in which glass is subdivided into subclasses by K-mean clustering of characteristic chemical composition data.

2. Glass classification model building and solving

The literature [4] states that high potassium and lead-barium glass are the most common. Therefore, this paper combines the data of high potassium and lead-barium glass artefacts from online databases and the literature [5] as a sample for the study and is labelled...The selected 58 groups of sample data have the following characteristics:

The only two types are high potassium and lead-barium; their surface weathering is known; their main chemical content is known and the main chemical constituents include: silica (SiO_2), sodium oxide (NaO), potassium oxide (K_2O), calcium oxide (CaO), magnesium oxide (MgO), aluminium oxide (Al_2O_3), iron oxide (Fe_2O_3), copper oxide (CuO), lead oxide (PbO), barium oxide (BaO), phosphorus pentoxide (P_2O_5), strontium oxide (SrO), tin oxide (SnO), sulphur dioxide (SO_2).

The classification pattern of high potassium glass and lead-barium glass was described by relying on the chemical composition data. Subsequently, the number of subclass categories was determined according to the aggregation coefficient, and the K-means clustering method could be used to discover the structure implied by the data [6], and the data were automatically categorised [7], and finally, the subclass subdivision results were derived by using SPSS software. The significance test value in the analysis results was used as a criterion to screen the characteristic chemical components, and the category segmentation was based on the minimum distance between the content of this component in the sample data and the central value of the clustering. Subsequently, this classification result was compared with the classification type in the sample, and the rationality of the clustering model was described by solving for the error rate magnitude. The raw data were processed at ±5% and ±10% fluctuation rates, and the theoretical classification was compared to the results to describe the sensitivity of the clustering model by analysing the rate of change.

2.1. Description of the classification pattern of high-potassium glass and lead-barium glass

After qualitative analysis of the sample data, the classification pattern between high-potassium glass and lead-barium glass is elaborated: for high-potassium glass, weathering causes SiO_2 a sharp increase in content, while a more significant decrease in the content of K_2O, CaO,Al_2O_3 and P_2O_5.In the case of lead-barium glass, weathering causes a more significant increase in PbO, a slight increase in the content of P_2O_5,a more significant decrease in the content of SiO_2

2.2. Subclass subdivision based on K-means clustering

The number of categories determined.

The number of categories in the systematic clustering model was determined using aggregation coefficients. The study samples can be initially divided into four categories: high potassium unweathered, high potassium weathered, lead-barium unweathered and lead-barium weathered, with each of the major categories of m. The samples were divided into K The number of samples in each of these categories was divided into four subcategories, and the number of samples in each subcategory was guaranteed to be $K < (m-1)$. The samples were divided into four subclasses, one for each major class and one for each subclass. If A_{nk} denotes then of the major class subcategory, the u_{nk} denotes the first k the central position of the subclass, then the corresponding coefficient of aggregation J_n is obtained from equation (1).

$$J_n = \sum_{k=1}^{K} \sum_{i \in A_{nk}} |x_{ni} - u_{nk}|^2 \quad (1)$$

Accordingly a line graph of the aggregation coefficients for the four broad categories was made, as shown in Figures 1 to 4.

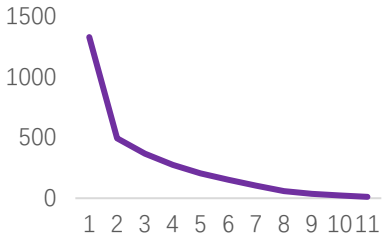

Figure. 1 High Potassium weathered

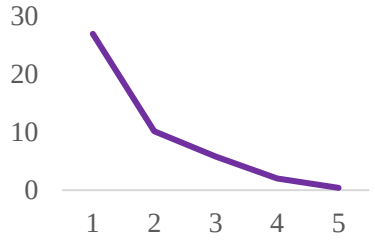

Figure. 2 High potassium unweathered

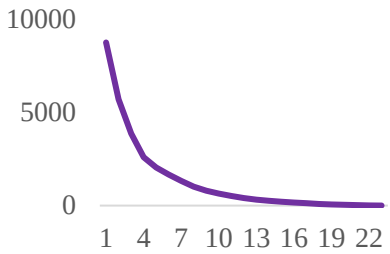

Figure. 3 Lead and barium weathered

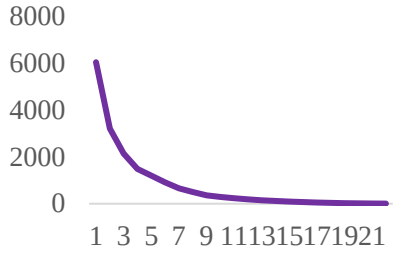

Figure. 4 Lead and barium unweathered

Analysis shows that there is a clear inflexion point in the line graph, whose horizontal coordinate corresponds to a value. Before this category, the J_n there is a clear downward trend, after which the J_n the inflexion point can be used as the elbow of the coefficient of aggregation graph. According to the elbow principle, when $K = a$ The number of subcategories is optimal at that time, and the number of subcategories for the four main categories can be derived as shown in Table 1.

Table 1 Results of the determination of the number of subcategories

Broad Category Divisions	High potassium unweathered	High Potassium weathered	Lead and barium unweathered	Lead and barium weathered
Number of subcategories	2	2	2	3

K-means clustering analysis model building

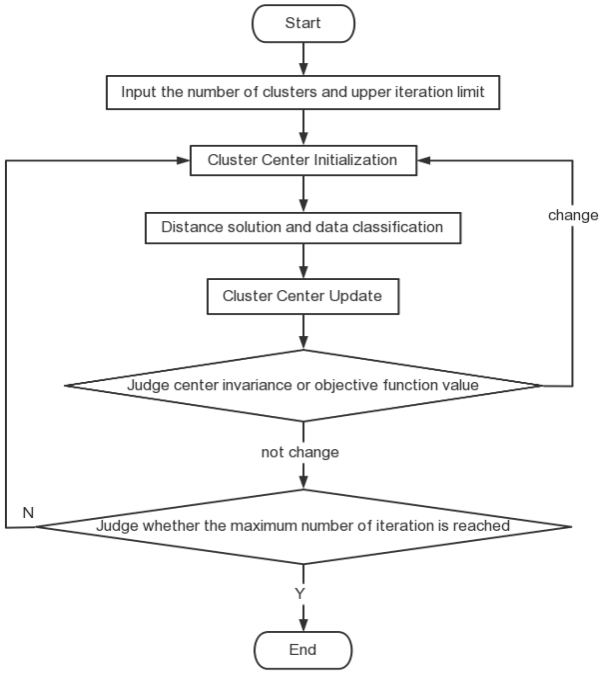

Figure. 5 K-means clustering model building flowchart

The K-means clustering model building process is shown in Figure 5 and its specific steps are divided into:

(1)Input parameter setting: Based on the result determined by the number of subclassifications, select an as the number of clusters, and set the maximum number of iterations b.

(2)Cluster centre initialisation: from the original data set, i.e. the data of each chemical component content of the four major classes, randomly select a. The internal objects are used as the initial clustering centroid.

(3)Distance solving and data assignment: for the remaining data, calculate the first i data value e_i with the selected first j cluster centre data value E_j Euclidean distance between d_{ij}.

$$d_{ij} = \sqrt{\sum_{i=1}^{S-a} \sum_{j=1}^{a} (e_i - E_j)^2} \quad (2)$$

This reflects the proximity of the source data to the cluster centres: the smaller the Euclidean distance, the greater the proximity of the data to the cluster centres and the higher the degree of aggregation. Therefore, the Euclidean distance can be used as a classification indicator to determine the cluster centre with the highest degree of aggregation with any of the component data values. e_i The clustering centre with the highest degree of aggregation will be assigned to the sub-category represented by this cluster. A total of a subclasses can be divided, each containing f data.

(4)Cluster centre update: take the first j of all data values in the subclasse$_i$of the first subclass $\frac{1}{f}\sum_{i=1}^{f} e_i$ as the cluster centre of the corresponding category, and solve for the objective function value.

(5)Conditional judgement: the clustering results are output when the central value or the objective function value remains unchanged and the upper limit of the iteration is reached, which is the classification result of further subdivision of the four major classes of artefacts.

2.3. Breakdown of results

The partial significance values, and cluster centroids, of the data on the content of the chemical components contained in the major classes are shown in Table 2. From this, the characteristic components for continuing the subclassification of each major category can be derived as the basis for subclassification. The specific division method and results are shown in Figure 6, where feature component 1 refers to K_2O,CaO and Al_2O_3,feature component 2 refers toSiO_2 and Al_2O_3,feature component 3 refers toAl_2O_3,PbO and BaO,feature component 4 refers to SiO_2,Al_2O_3,CuO,PbO,BaO and SO_2.Some of the clustering results are shown in Table 3.

Table. 2 Significance values of chemical composition content, cluster centroids

Data type	Major categories	SiO_2	K_2O	CaO	MgO	Al_2O_3
Significance value	High potassium unweathered	0.00	0.01	0.04	0.65	0.08
	High Potassium weathered	0.02	0.10	0.06	0.12	0.04
Clustering Central value	High potassium unweathered 2	81.06	4.87	2.24	0.92	4.43
	High Potassium weathered 2	95.36	0.84	0.52	0.00	1.20

Table. 3 Clustering results

Broad Category Divisions	High potassium unweathered		High Potassium weathered		Lead and barium unweathered		Lead and barium weathered		
Subcategories	1	2	1	2	1	2	1	2	3
Number of subcategories	9	3	3	3	7	16	17	3	4
Individual number	01 03	03 18	09 10	07 22	20 24	23 28	02 11	08 19	26 39

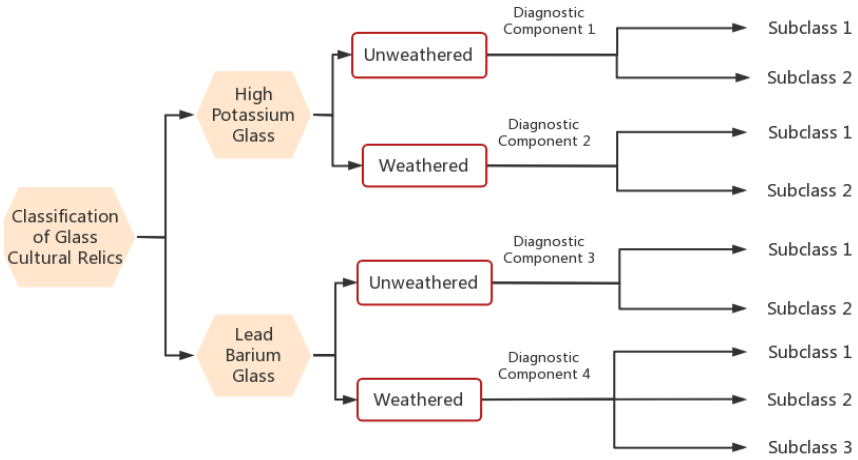

Figure. 6 Graph of specific division methods and results

2.4. Reasonableness of results, sensitivity analysis

Reasonableness analysis

Analysis of the complete data corresponding to Table 2 shows that when the significance value $P < 0.05$ when a significant correlation between the chemical component and the category can be considered to exist. Therefore, the chemical component that meets the significance test is selected as the characteristic indicator for subcategory classification, and the $j = 1,2,\cdots,q$ denotes the marker of the selected characteristic indicator. After that, the first k The clustering centre of the first$j$ The cluster centroids of the chemical components of the first subclass were extracted H_{kj} .

Extract the first artifact from the original data r of an object j,The content data of the first characteristic chemical component of the object G_{rj}.The distance between the content data and the cluster centroid is calculated according to Equation 3.

$$d_{rk} = \sqrt{\sum_{j=1}^{q}(G_{rj} - H_{kj})^2} \tag{3}$$

This distance reflects the proximity between the chemical content of the object and the subclass range; the smaller the distance, the closer the chemical composition of the object is to the characteristics of the subclass, and therefore the minimum distance can be $\min\{d_{rk}\}$ as the basis for subcategorisation of cultural objects.

Accordingly, the results of the theoretical analysis of the sub-classification of artefacts were derived by classifying each artefact according to its chemical composition only. The results were compared with the K-means clustering results of the artefacts and their error rates were calculated based on equation (4) :

$$\tau = \frac{\sum_{r=1}^{s} L_r}{s} \tag{4}$$

L_r is a logical variable, taking 1 when the control results are inconsistent.s indicates the total number of artefacts for which controls were carried out.

Some of the results of the rationality analysis are shown in Table 4.

Table. 4 Results of the rationality analysis

Heritage number	Lead barium weathered subclass 1 distance	Lead and barium unweathered subclass 2 distance	High Potassium weathered Subclass 1 Distance	Minimum Distance	Classification results
01	66.438	21.700	23.275	5.770	High potassium unweathered subclass 1
11	25.997	28.957	58.980	8.727	Lead and barium unweathered subclass 1
22	83.505	37.361	0.868	0.868	High potassium weathered subclass 1
26	42.694	64.635	88.862	15.098	Lead and barium weathered subclass 2

According to the results of the rationality analysis, the theoretical results of the subclassification of the chemical components of each artefact agreed with the results of the K-means clustering model with a number of 60 and an error rate of 0.1045, indicating that the K-means clustering model obtained a high accuracy rate of subclassification and a more significant rationality.

Sensitivity analysis

(1) Sensitivity interpretation and choice of volatility ratio

Sensitivity refers to the degree of variation in the output when the input data fluctuates compared to the baseline data, the greater the variation, the more unstable the output and the more sensitive the model.

When the data fluctuates within a smaller interval, the model's resistance and sensitivity both have a greater impact on the output results, so it is not possible to use the smaller fluctuation interval for sensitivity analysis; when a larger fluctuation interval is chosen, the output results are mainly affected by sensitivity, so a larger fluctuation interval should be chosen for sensitivity analysis of data fluctuation processing.

The increased volatility of the input content data can be reflected in the sensitivity of the model by the magnitude of the change in the output indicators. Therefore the selected volatility ratio μ_{rj} for the content data G_{rj} The volatility update is carried out and the updated data is $G_{rj'}$

$$G_{rj'} = (1 + \mu_{rj})G_{rj} \quad (\mu_{rj} = 5\%、 -5\%、10\%、 -10\%) \tag{5}$$

The data is updated and then passed into the model for the reasonableness analysis for the rate of change solution, and the final bar chart reflects the distribution of the rate of change under different volatility.

Some of the results of the sensitivity analysis are shown in Table 5. The rates of change are shown in bar chart 7.

Table. 5 Results of sensitivity analysis

Fluctuation ratio	Heritage number	Lead barium weathered subclass 1 distance	High Potassium weathered Subclass 1 Distance	Minimum Distance	Classification results
5%	01	68.80	19.83	9.19	High potassium unweathered subclass 1
-5%	11	26.47	60.66	10.56	Lead and barium unweathered subclass 1
10%	22	91.11	9.09	6.77	High potassium weathered subclass 1
-10%	26	32.53	74.79	7.86	Lead and barium weathered subclass 2

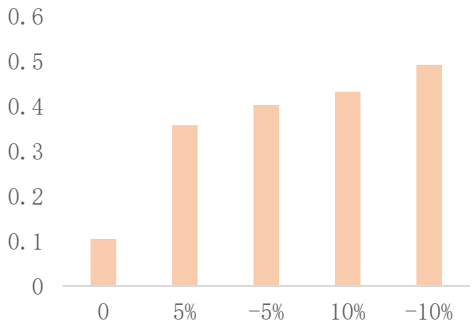

Figure. 7 Rate of change histogram

Based on the rate of change at different volatility ratios, it can be concluded that an increase in data volatility leads to a significant increase in the rate of change and therefore a higher sensitivity of the clustering segmentation method. At the same time, the rate of change increases significantly more significantly in the stages where volatility increases from 0 to 5% or decreases to -5% than in the stages where it increases from 5% to 10% or decreases from -5% to 10%, i.e. as the volatility of the data increases, the instability of the model gradually increases and the more sensitive the model becomes.

3. Development and solution of glass identification models

The compositional data for the unknown chemical major group (i.e. unknown whether it is high potassium or lead-barium) was identified in the web database as a test sample for use as the test data here. The degree of surface weathering of the glass artefacts in this data for the unknown chemical major category is known and the chemical composition type is consistent with that in the 2 data.

Based on the above subclass subdivision of glass artefact samples of known broad classes, chemical compositions with high significance of differences were selected as generic feature indicators and a generic subclass delineation model based on K-means clustering was established. The compositional data of unknown chemical macroclasses (i.e. unknown whether they are high

potassium or lead-barium) were identified in the web database as test samples for subclassification. Their sensitivities were analysed using similar ideas, but given the small amount of data in the test sample compared to the study sample, the results were only analysed for individual controls.

3.1. Subclass classification modeling

The subclass classification model is shown in Figure 8. The data from the test samples were analysed and processed to obtain a preliminary classification of the artefacts: weathered and unweathered classes were distinguished by the index of weathering or non-weathering. Secondly, the chemical components with significance values within the range of 0.05 were selected as the characteristic indicators for subclassification of weathered and unweathered categories, and clustering analysis and subclassification were carried out.

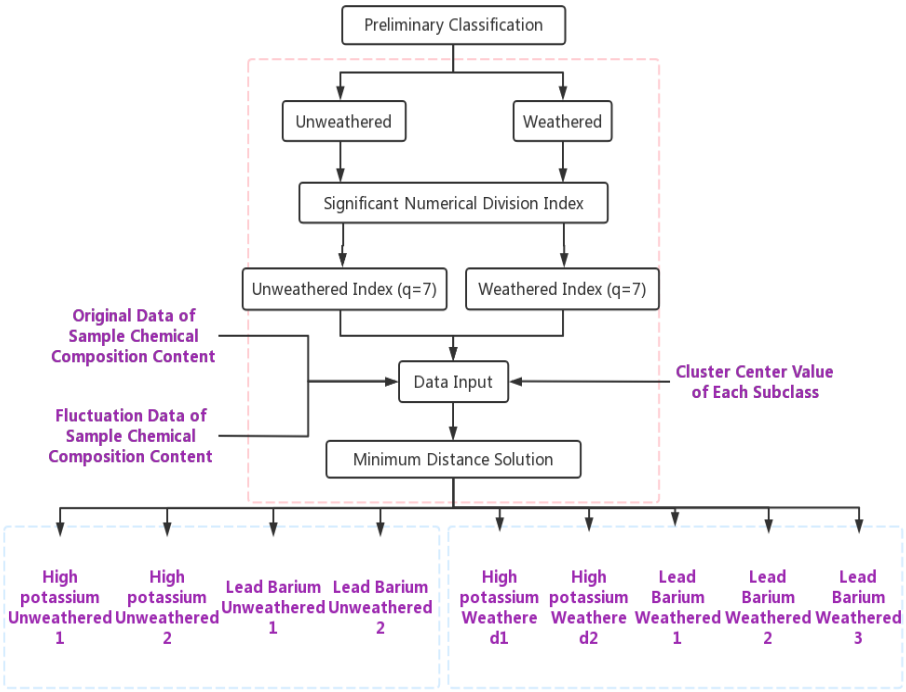

Figure 8 Model of heritage sub-categorisation

Table. 6 Results of the heritage category breakdown

Known categories	Unweathered				Weathering			
Heritage number	A1	A3	A4	A8	A2	A5	A6	A7
Shortest distance	6.73	15.38	9.21	10.83	11.66	30.03	1.29	2.96
Breakdown by category	High Potassium 2	Lead Barium 1	Lead Barium 1	Lead Barium 2	Lead Barium 3	High Potassium 1	High Potassium 1	High Potassium 1

3.2. Sensitivity analysis of results

Similar to the sensitivity analysis method in 2, the test data for the chemical composition of the artefacts were I_{rj} were processed as follows.

$$I_{rj}' = (1 + \mu_{rj})I_{rj} \tag{6}$$

The processed data was fed into the subclass classification model to obtain the classification results, as shown in Table 7.

Table. 7 Results of sub-categorisation of artefacts after treatment

Fluctuation ratio	Known categories	Unweathered				Weathering			
Heritage number		A1	A3	A4	A8	A2	A5	A6	A7
-5%	Shortest distance	8.77	14.73	12.16	12.59	11.06	30.31	4.26	6.56
	Break down by category	High Potassium 2	Lead Barium 1	Lead Barium 1	Lead Barium 2	Lead Barium 3	Lead Barium 3	High Potassium 1	High Potassium 1
5%	Shortest distance	6.65	15.83	7.06	9.16	12.57	26.96	3.01	4.17
	Break down by category	High Potassium 2	Lead Barium 1	Lead Barium 1	Lead Barium 2	Lead Barium 3	High Potassium 1	High Potassium 2	High Potassium 2
10%	Shortest distance	8.42	16.54	5.52	19.26	21.47	24.63	7.14	6.31
	Break down by category	High Potassium 2	Lead Barium 1	Lead Barium 1	Lead Barium 2	Lead Barium 1	High Potassium 1	High Potassium 2	High Potassium 2

Given the small sample size, the calculation of the rate of change was not performed and only individual cases of the results were analysed against each other. When the percentage of fluctuation is 5% or -5%, the number of inconsistencies between the category breakdown results and the original data analysis is within 1, which means that the prediction results after the fluctuation treatment are in good agreement with the prediction results of the standard data, and the model is not too sensitive at this time. Very sensitive.

4. Conclusion

The K-mean clustering model-based method for classifying and identifying glass artefacts proposed in this paper has the advantages of simple principles, easy implementation and better clustering results [8]; the ANVOA discrepancy test is applied to improve the efficiency of the model; the model is built based on known data and verified using other known data, which has high accuracy and rigour. The analysis of the results shows that the model has a strong resistance to interference for small fluctuation intervals, and a strong sensitivity for classification of large intervals. The model is general enough to be used in other compositional-based classification and identification of substances.

Ancient times have left behind a large number of valuable glass objects, both of our own manufacture and those imported from the West, which are an important cultural heritage. However, they have all suffered from corrosion and damage to varying degrees, and some have been damaged quite severely due to their instability [9]. The aim of this paper is not only to explore a model for the analysis of the composition of ancient glass, but also to raise awareness of the conservation of ancient cultural relics through the advancement of analytical techniques, so that these precious relics can be properly protected and the study of Chinese ancient glass can go global [10].

References

[1] Wu Wen Jing, Peng Jia HongFei, Low carbon travel intention data mining based on K-means clustering and random forest algorithm [J]. Journal of South China University of Technology (Natural Science Edition),2019,47(7):105-111. DIO:10.12141/j.issn.1000-565X.180116.

[2] Tian Mi,Wang Xueqiu,Hao Libo. Research on the extraction method of geochemical anomalies in aqueous sediments based on K-mean clustering[C]. // Proceedings of the 2017 Annual Academic Conference of the Geological Society of China. 2017:70-74.

[3] Gan Fuxi. Some views on the study of ancient glass in China [J]. Journal of Silicates,2004,32(2):182-188. DIO:10.3321/j.issn:0454-5648.2004.02.015.

[4] Ma Yanru. The ancient glass of Bozhou [J]. Cultural Identification and Appreciation, 2012(12):94-97. DIO:10.3969/j.issn.1674-8697.2012.12.019.

[5] Fu Xiufeng, Gan Fuxi. A study on the composition of a group of ancient glasses from southern and southwestern China based on multivariate statistical analysis[J]. Conservation and Archaeological Science,2006,18(4):6-13. DOI:10.3969/j.issn.1005-1538.2006.04.002.

[6] Pan Junliang,Shi Yuexiang,Li Pingting. A new particle swarm optimization clustering method [J]. Computer Engineering and Applications,2012,48(8):179-181.

[7] Tao Y,Yang F,Liu Y,Dai B. Research and optimization of K-means clustering algorithm [J]. Computer Technology and Development,2018,28(06):90-92.

[8] Tang Zhe. A k-means clustering analysis based on genetic algorithm[D]. Changsha University of Technology,2014.

[9] Lu Shoulin. Ancient glass and its conservation [C]. Technology of heritage conservation (1981-1991).2010:299-309.

[10] Wan Fu-Bin,Gan Fu-Xi. Let Chinese ancient glass research goes to the world--Interview with academician Gan Fu Xi [J]. Journal of Guangxi University for Nationalities (Natural Science Edition),2009,15(04): 25-41.DOI:10.16177