ISSN 2616-5880 Vol. 4, Issue 3: 39-46, DOI: 10.25236/AJMC.2023.040306

Identification of ancient glass products based on K-means composition analysis

Yaowen Zhang[1,*], Jie Guo[1], Qiuling Deng[2]

[1]*Department of Civil Engineering, Guilin University of Technology, Guilin, China*
[2]*Department of Electronic Information Engineering, Guilin University of Technology, Guilin, China*
[*]*Corresponding author:2163757536@qq.com*

Abstract: *In order to study the classification laws of glass types, the data were first divided into two categories of weathered and unweathered points, and then k-means cluster analysis was used to subdivide each category of data into two categories. It was found that the artifacts in these two categories corresponded to high potassium glass and lead-barium glass, respectively, indicating that k-means cluster analysis could be used as a classification law for high potassium glass and lead-barium glass. Since there are 14 chemical components in each of the four categories, it is more difficult and complicated to use them as the basis for subcategory classification, so principal component analysis was applied to reduce the dimensionality, and the 14 chemical components were replaced by comprehensive indicators (principal components) filtered by the cumulative contribution of eigenvalues over 80%. Then, the sample glass was classified into 15 classes by applying SPSS software to classify the principal components of each class as variables, respectively, and the samples as one event for clustering. In order to verify whether the classification method established by this model is realistic, the results of the division of each category into classes were analyzed separately using ROC curves for reasonableness and sensitivity in this paper, and the final reasonableness and sensitivity were both good.*

Keywords: *Glass type, Classification law, K-means cluster analysis, Principal component analysis method, Cluster division*

1. Introduction

The main chemical composition of quartz sand is silicon dioxide (SiO2), which is the most important raw material for glass making. However, because the refining process in ancient times was not high enough to reach the melting point of pure quartz sand, the ancients would add fluxes to pure quartz sand to lower the melting temperature of quartz sand in glass making. In ancient times, people often used grass ash, natural soda ash, saltpeter and lead ore as fluxes, and limestone as stabilizer. Different fluxes contain different chemical compositions, resulting in different chemical compositions in the manufactured glass. For example, the ancient Chinese invented lead-barium glass, adding lead ore as a flux to quartz sand in the refining and firing process, and eventually obtaining lead-barium glass, which contains more lead oxide (PbO) and barium oxide (BaO). There is also potassium glass, which is popular in Lingnan, China, as well as in Southeast Asia and India, and other regions, is manufactured by melting fluxes that have a high content of potassium [1].

Ancient glass is easily affected by the environment and weathered. In the process of weathering, elements inside the glass will exchange with elements outside, which in turn will change the original composition of the glass, so that the ratio of various chemical components contained inside the glass will change, creating a great disturbance in the judgment of glass categories. To find the influence of different types and differentiation on the chemical composition of glass, and to establish a model to distinguish the nature of glass by its chemical composition, it is important to study the ancient glass "culture"[2].

This study intends to investigate the classification laws of high potassium glass and lead-barium glass. And a few more major chemical compositions were selected as the discriminatory criteria for a second and more detailed classification of different categories of glasses, and the results of the classification were analyzed for reasonableness and sensitivity.

ISSN 2616-5880 Vol. 4, Issue 3: 39-46, DOI: 10.25236/AJMC.2023.040306

2. Model building and analysis

2.1 Classification rules for two types of glass

In the study of high potassium glass and lead-barium glass, we first divided the data into two categories: weathered and unweathered, and then used k-means cluster analysis to subdivide the data in each category into two categories.

(1) Principle of clustering analysis

K-means cluster analysis [3] is a simple and efficient analysis method that is often applied to cluster analysis of large-scale data and has resulted in many other kinds of cluster analysis algorithms.

The k in k-means cluster analysis represents the data to be divided into k classes, and it is necessary to first determine the mean vector of the samples in the cluster selected as the center of the cluster during the iteration, and the data in each cluster should satisfy that the sum of squares to the center of that cluster is the smallest, and thus the data to be classified into k clusters, each of which.

(2) The main steps of k-means clustering analysis

Step 1: Select any k objects from the input data as the initial clustering centers.

Step 2: According to the mean value of different clustering objects, go to calculate the distance from the target object to these clustering centers, and according to the minimum distance is used to re-classify the corresponding objects.

Step 3: Calculate the average value of each cluster.

Step 4: Repeat Step 2 and Step 3 until each cluster no longer changes.

Since the question is to determine the classification law of high potassium glass and lead-barium glass, the k value here is taken as 2, that is, the target objects are divided into two classes, corresponding to high potassium glass and lead-barium glass respectively.

(3) Clustering results

Table 1: Comparison of predicted and actual types of artifacts

Artifact Number	Actual Type	Clustering Type	Artifact Number	Actual Type	Clustering Type
07	High Potassium	High Potassium	38	Lead Barium	Lead Barium
09	High Potassium	High Potassium	39	Lead Barium	Lead Barium
10	High Potassium	High Potassium	40	Lead Barium	Lead Barium
12	High Potassium	High Potassium	41	Lead Barium	Lead Barium
22	High Potassium	High Potassium	43	Lead Barium	Lead Barium
27	High Potassium	High Potassium	43	Lead Barium	Lead Barium
48	**Lead Barium**	**High Potassium**	49	Lead Barium	Lead Barium
02	Lead Barium	Lead Barium	50	Lead Barium	Lead Barium
08	Lead Barium	Lead Barium	51	Lead Barium	Lead Barium
08	Lead Barium	Lead Barium	51	Lead Barium	Lead Barium
11	Lead Barium	Lead Barium	52	Lead Barium	Lead Barium
19	Lead Barium	Lead Barium	54	Lead Barium	Lead Barium
26	Lead Barium	Lead Barium	54	Lead Barium	Lead Barium
26	Lead Barium	Lead Barium	56	Lead Barium	Lead Barium
34	Lead Barium	Lead Barium	57	Lead Barium	Lead Barium
36	Lead Barium	Lead Barium	58	Lead Barium	Lead Barium

Note: The bolded ones are artifacts with wrong predictions.

After dividing the detection points into weathering points and unweathered points, the data were finally classified into two categories by cluster analysis. For the classification of weathering points, we got 7 artifacts belonging to high potassium glass and 25 glasses belonging to lead-barium glass. Compared with the actual 6 artifacts of high potassium glass and 26 artifacts of lead-barium glass, there is a little prediction error, but the accuracy rate reaches 96.25%, which can meet the requirements of use. For the classification of unweathered points, we also divided them into two types of glass, high potassium and lead-barium, and finally obtained 13 high potassium glasses and 22 lead-barium glasses, which still have an error with the actual value type, but the accuracy rate still meets the requirements of use. Among them, the classification results of undifferentiated glass are shown in Table Details see Table 1.

(4) Classification law

The above classification of glass using k-means cluster analysis was divided into high potassium

ISSN 2616-5880 Vol. 4, Issue 3: 39-46, DOI: 10.25236/AJMC.2023.040306

glass and lead-barium glass, which were 67 data, of which two pairs of classification errors occurred, and the correct rate reached 97%, which shows that the classification criteria of k-means cluster analysis can be used as classification criteria for high potassium glass and lead-barium glass.

For the classification criteria of high potassium glass and lead-barium glass, we first identified several chemical components that significantly affect the classification type of glass, and identified the mean and standard deviation corresponding to these chemical components, i.e., a certain chemical component between the mean ± standard deviation can be considered as a distance of 0. A sample to a certain type with the smallest sum of all squared distances can be considered to change the sample to belong to this type of glass.

2.2 Principal component analysis method

(1) Establishment of the method

From the above, it can be seen that the heritage samples can be initially divided into four categories: weathered high potassium glass, weathered lead-barium glass, unweathered high potassium glass, and unweathered lead-barium glass, for which we will conduct a second division, i.e., subcategory division. Since there are various chemical components detected in each category, it is more difficult and complicated to perform subcategory division, so we choose to use the principal component analysis method to retain as much information as possible from the original variables through the dimensionality reduction technique with fewer new variables, and these new variables we call the principal components.

(2) Detailed explanation of the steps

Step 1: Acquisition of data.

Assuming that there are n samples with p chemical components, a sample matrix x of size n × p can be formed.

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p) \tag{1}$$

Step 2: Data standardization.

The zero-mean method (z-score) is used to process the data, and the data obeying the standard normal distribution with mean 0 and standard deviation 1 can be obtained, i.e., the standardized data $X_{\mathrm{ij}} = \frac{x_{\mathrm{ij}} - \overline{X_{\mathrm{j}}}}{s_j}$, and the original sample matrix becomes after standardization.

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = (X_1, X_2, \cdots, X_p) \tag{2}$$

Step 3: Calculate the covariance matrix.

Obtained by $r_{ij} = \frac{1}{n-1} \sum_{k=1}^{50} (X_{ki} - \overline{X_i}) \left(X_{ki} - \overline{X_j}\right)$

$$R = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1p} \\ R_{21} & R_{22} & \cdots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n1} & R_{n2} & \cdots & R_{np} \end{bmatrix} \tag{3}$$

Step 4: Calculate the eigenvalues and eigenvectors of the covariance matrix.

Eigenvalues.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq 0 \tag{4}$$

Eigenvectors.

ISSN 2616-5880 Vol. 4, Issue 3: 39-46, DOI: 10.25236/AJMC.2023.040306

$$\alpha_1 = \begin{bmatrix} \alpha_{11} \\ \alpha_{21} \\ \vdots \\ \alpha_{31} \end{bmatrix}, \quad \alpha_2 = \begin{bmatrix} \alpha_{12} \\ \alpha_{22} \\ \vdots \\ \alpha_{32} \end{bmatrix}, \quad \alpha_3 = \begin{bmatrix} \alpha_{13} \\ \alpha_{23} \\ \vdots \\ \alpha_{33} \end{bmatrix} \tag{5}$$

Step 5: Calculate the contribution of principal components as well as the cumulative contribution.

$$S_i = \frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k} (i = 1,2,\cdots,p) \tag{6}$$

$$S_n = \frac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{p} \lambda_k} (i = 1,2,\cdots,p) \tag{7}$$

Where Si represents the contribution rate and Sn represents the cumulative contribution rate.

By the above method, the eigenvalues and contribution values of each glass category were finally obtained, where the eigenvalues and contribution rates of the unweathered high potassium glass category were calculated as shown in Table 2.

Table 2: Calculated eigenvalues and contribution rates for the unweathered high potassium glass category

Principal Component Number	Eigenvalue	Contribution rate	Cumulative contribution rate
1	4.1121	0.5140	0.5140
2	3.0068	0.3758	0.8899
3	0.6363	0.0795	0.9694
4	0.2120	0.0265	0.9959
5	0.0328	0.0041	1.0000
6	0.0000	0.0000	1.0000
7	0.0000	0.0000	1.0000
8	0.0000	0.0000	1.0000
9	0.0000	0.0000	1.0000
10	0.0000	0.0000	1.0000
11	0.0000	0.0000	1.0000
12	0.0000	0.0000	1.0000
13	0.0000	0.0000	1.0000
14	0.0000	0.0000	1.0000

2.3 Calculating principal components

(1) Principal component calculation

The first, second, and t-th (t≤p) principal components corresponding to the eigenvalues whose cumulative contribution exceeds 80% are taken.

The i-th principal component: FI=c1iX1+c2iX2+...+cpiXp (i=1, 2, ..., t).

Table 3: Eigenvectors of the first four principal components of the unweathered high potassium glass category on the original data

Chemical composition	Principal Components			
	F1	F2	F3	F4
SiO2	0.1536	**-0.4759**	-0.1009	-0.1854
Na2O	0.2605	0.3147	-0.2287	0.3426
K2O	0.1546	0.3002	0.1487	0.3804
CaO	0.1667	**0.4755**	0.0884	-0.0799
MgO	**-0.3792**	-0.1082	0.2254	0.094
Al2O3	**-0.3111**	0.2292	-0.1071	0.2875
Fe2O3	**-0.3855**	0.2152	0.1197	-0.071
CuO	-0.1575	0.2768	0.0593	**-0.468**
PbO	-0.0228	0.1388	**-0.5661**	-0.1012
BaO	**-0.3084**	0.0139	-0.392	-0.3134
P2O5	**-0.4287**	-0.0702	0.0546	0.1883
SrO	**-0.3979**	-0.051	-0.029	0.2823
SnO2	0.064	-0.3683	0.0644	0.3101
SO2	0.0323	0.0921	**0.5877**	-0.2487

Note: The data in bold in the table indicate the contribution weight of the principal components reflecting the relevant chemical components.

For example, if the cumulative contribution of the eigenvalues of the category of unweathered high potassium glass exceeds 80%, the corresponding first, second, third and fourth principal components' eigenvectors on the original data are shown in Table 3.

ISSN 2616-5880 Vol. 4, Issue 3: 39-46, DOI: 10.25236/AJMC.2023.040306

(2) Analysis of the significance of principal component representatives based on coefficients

For a certain principal component, the larger the coefficient in front of the indicator, the greater the influence of the representative indicator on that principal component.

For example, the category of unweathered high potassium glass.

1) First principal component F1 mainly reflects the contribution of MgO, Al2O3, Fe2O3, BaO, P2O5 and SrO, which are mainly related.

2) Second principal component F2 mainly reflects the contribution of SiO2, CaO, and K2O:Na2O <1. According to the online information query, it can be regarded as Na2O(K2O)-CaO-SiO2 glass system.

3) The third main component F3 mainly reflects the contribution of SO2, PbO, and PbO, K2O as the main co-solvent in our glass system, so it is judged to be a co-solvent component.

4) Fourth main component F4 is mainly the contribution weight of CuO, which is a coloring oxide and can be classified as a colorant component [4].

2.4 Subclassification of principal components by cluster analysis

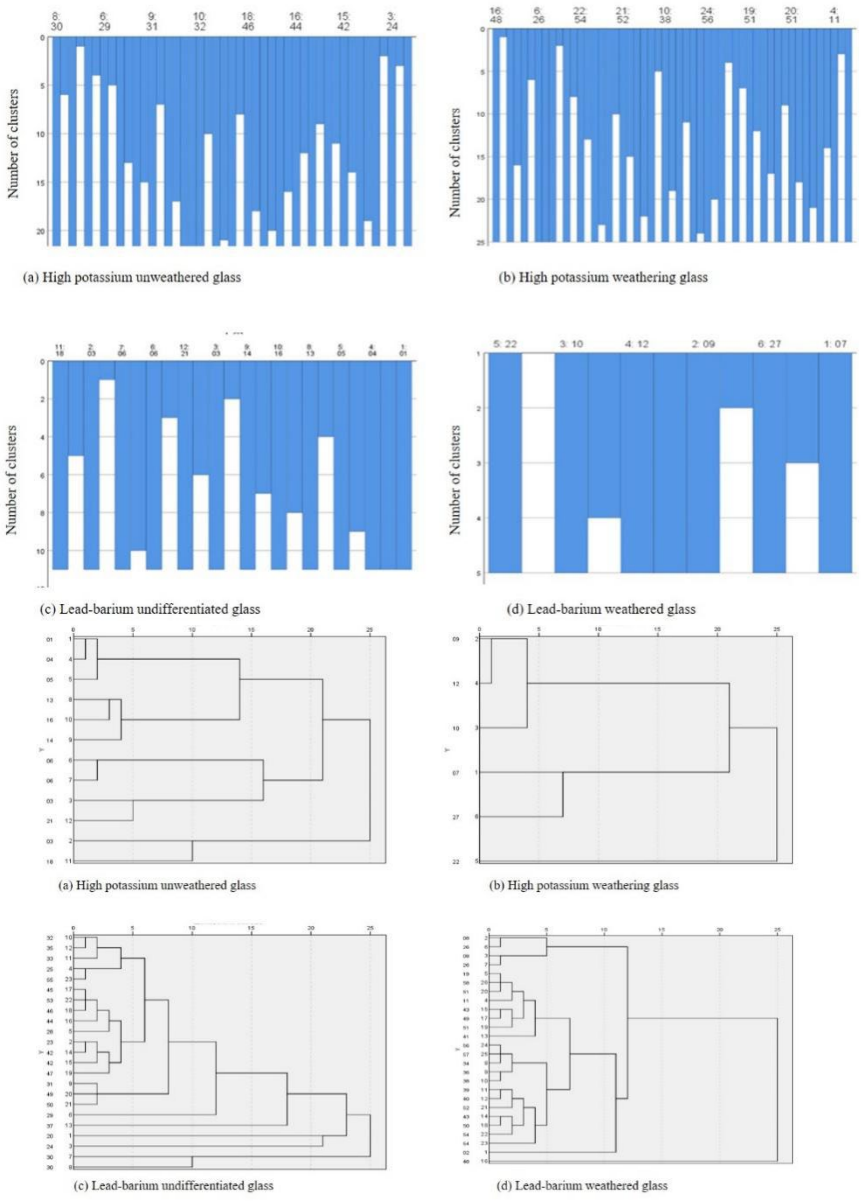

Figure 1: Spectrum diagram using average links (between groups)

ISSN 2616-5880 Vol. 4, Issue 3: 39-46, DOI: 10.25236/AJMC.2023.040306

From the above principal component analysis method we obtained the principal component indicators that can reflect the whole data set more comprehensively with a small number of composite variables. Then, by using SPSS software, we performed cluster analysis with four categories of principal components as variables and each sample as an event, respectively, to derive a cluster dendrogram, and finally we divided the samples in each category more carefully by observing the cluster dendrogram (Figure 1).

2.5 Rationality and sensitivity analysis of ROC curves on subclassification results

(1) Definition analysis

The RCO curve reveals the interrelationship between sensitivity and specificity by applying the constitutive method, and is a composite indicator of the continuous variables of response sensitivity and specificity. Its horizontal coordinate is the true rate FPR and the vertical coordinate is the false positive rate TPR.

This question needs to analyze and test the rationality of the subclassification results as well as the sensitivity, and the definition of ROC shows that this class method meets the requirements of this question.

(2) Application of SPSS to plot ROC curves (using unweathered high potassium glass as an example)

The ROC curve visualizes the relationship between the false positive rate (1-specificity) and the true positive rate (sensitivity) [5].

Step 1: AUC, the area under the ROC curve, has a value between 0 and 1. The closer the AUC is to 1, the better the diagnosis.

Step 2: AUC is judged by the following criteria: below 0.5 does not match the actual situation, 0.5 indicates no diagnostic value at all, between 0.5 and 0.7 has very low diagnostic value, 0.7 to 0.9 indicates some diagnostic value, and above 0.9 indicates high diagnostic value (Table 4).

Table 4: Summary of AUC of ROC results

Title	AUC	Standard error	*p*	95% CI
F1	0.778	0.157	0.076	0.471 ~ 1.085
F2	0.833	0.124	0.007**	0.590 ~ 1.077
F3	0.722	0.163	0.173	0.403 ~ 1.042
F4	0.556	0.178	0.755	0.206 ~ 0.905

* $p<0.05$ ** $p<0.01$

We constructed ROC curves with F1, F2, F3, and F4 as variables to determine their diagnostic value for type, as seen in the above table.

F1 corresponds to an AUC value of 0.778 (95% CI:47.09%~108.47%), implying that F1 has a higher diagnostic value for the type.

F2 corresponds to an AUC value of 0.833 (95% CI: 58.99% to 107.68%), implying that F2 has a higher diagnostic value for the type.

F3 corresponded to an AUC value of 0.722 (95% CI:40.27%~104.17%), implying that F3 has a higher diagnostic value for the type.

F4 corresponded to an AUC value of 0.556, implying that F4 has a lower diagnostic value for the type.

In conclusion, it can be seen that F1, F2 and F3 have a high diagnostic value for type.

Table 5: Results of ROC best bounds

Title	AUC	Optimum Boundary	Sensitivity	Specificity	Cut-off
F1	0.778	0.667	1.000	0.667	-0.652
F2	0.833	0.667	1.000	0.667	-0.971
F3	0.722	0.500	0.500	1.000	0.418
F4	0.556	0.333	0.833	0.500	-0.871

The optimal cut-off value of the ROC curve (the maximum value of the Jorden index) is the point of maximum diagnostic value.

Step 1: The optimal cut-off value is the critical point corresponding to the maximum diagnostic value.

Step 2: The optimal cut-off value is the point corresponding to the best combination of true positive rate and true negative rate.

From the Table 5, we can see that F1 corresponds to an AUC value of 0.778, which means that F1 has

ISSN 2616-5880 Vol. 4, Issue 3: 39-46, DOI: 10.25236/AJMC.2023.040306

a higher diagnostic value for the type and corresponds to an optimal cut-off value of 0.667 (at this time, the sensitivity is 1.000 and the specificity is 0.667).

F2 corresponds to an AUC value of 0.833, which means that F2 has a higher diagnostic value for the type and corresponds to an optimal cut-off value of 0.667 (at this point the sensitivity is 1.000 and the specificity is 0.667).

F3 corresponds to an AUC value of 0.722, which means that F3 has a higher diagnostic value for the type and corresponds to an optimal cut-off value of 0.500 (at this point the sensitivity is 0.500 and the specificity is 1.000).

F4 corresponds to an AUC value of 0.556, which means that the diagnostic value of F4 for the type is relatively low.

In conclusion, it can be seen that F1, F2 and F3 will have a high diagnostic value for the type (Figure 2).

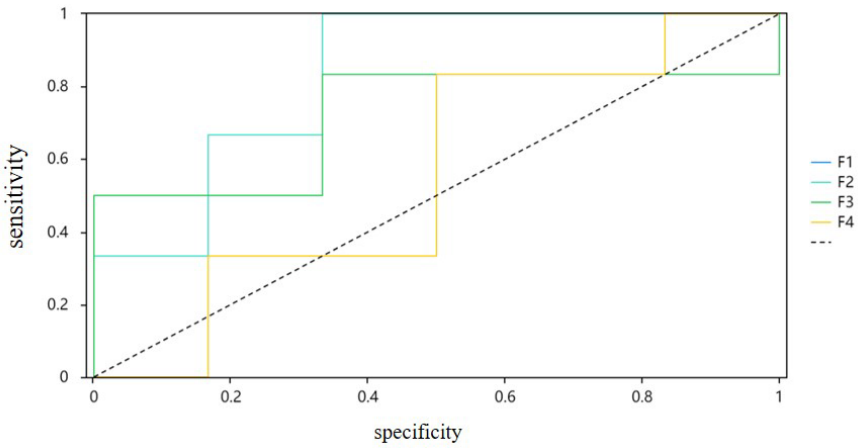

Figure 2: ROC curve

3. Conclusions

In order to study the classification laws of glass types, the data were first divided into two categories of weathered and unweathered points, and then k-means cluster analysis was used to subdivide each category of data into two categories. It was found that the artifacts in these two categories corresponded to high potassium glass and lead-barium glass, respectively, indicating that k-means cluster analysis could be used as a classification law for high potassium glass and lead-barium glass. Since there are 14 chemical components in each of the four categories, it is more difficult and complicated to use them as the basis for subcategory classification, so principal component analysis was applied to reduce the dimensionality, and the 14 chemical components were replaced by comprehensive indicators (principal components) filtered by the cumulative contribution of eigenvalues over 80%. Then, the sample glass was classified into 15 classes by applying SPSS software to classify the principal components of each class as variables, respectively, and the samples as one event for clustering. In order to verify whether the classification method established by this model is realistic, the results of the division of each category into classes were analyzed separately using ROC curves for reasonableness and sensitivity in this paper, and the final reasonableness and sensitivity were both good.

References

[1] Wang J-Tao, Zhou L-F, Gao E-S. Sixth lecture on chi-square test [J]. Experimental Animals and Comparative Medicine, 2000(4): 251-254.
[2] Dong JQ, Li QH, Gan FX, Hu YQ, Cheng YJ, Jiang HJ. Nondestructive analysis of a batch of glassware from Eastern Zhou to Song dynasties excavated in Henan [J]. China Materials Progress, 2012, 31(11): 9-1.
[3] Pu Huizhong. K-means clustering analysis algorithm in artificial intelligence + personalized learning system [J]. Intelligent Computers and Applications, 2022, 12(08): 152-156.
[4] Fu Xiufeng, Gan Fuxi. Study on the composition of a group of ancient glass from South and Southwest

ISSN 2616-5880 Vol. 4, Issue 3: 39-46, DOI: 10.25236/AJMC.2023.040306

China based on multivariate statistical analysis [J]. DOI: 10.16334/j.cnki.cn31-1652/k. 2006.04.002.
[5] Wang Jinghan. ROC curve in clinical medical diagnostic experiments [J]. Chinese Journal of Hypertension, 2008, 16(2): 175-177.