Classification and identification of glass products based on K-means clustering analysis and Fisher's linear discriminant method

Jiahuan Chen *, Yongan Li

School of Mechanical Engineering, Jiangsu University, Zhenjiang 212013, China

* Corresponding Author Email: 715781694@qq.com

Abstract. Ancient glass is very susceptible to weathering due to the buried environment, and the internal elements of the weathered glass will exchange with the elements in the environment thus leading to changes in the composition ratio. In order to classify and identify glass objects, this paper analyzes the classification laws of high potassium glass and lead-barium glass based on the data of samples of known types; selects the appropriate chemical composition for each category and then divides it into subcategories, gives specific division methods to establish a classification model, and uses this model for the chemical composition of glass artifacts of unknown categories to identify the types to which they belong

Keywords: Neural Network, Prediction Model, Big Data.

1. Introduction

Glass became an object of trade and cultural exchange between China and the West as early as the Silk Road period [1-3]. Early glass usually came from North Africa and West Asia, and was made into bead-shaped jewelry and introduced into China. Ancient glass is very susceptible to weathering due to the environment in which it is buried [4], and the internal elements of weathered glass are exchanged with elements in the environment, resulting in changes in the composition ratio [5-6].

According to the data of ancient glass products, archaeologists have classified these samples into two types of glass, high potassium and lead-barium, based on their chemical composition and other testing methods [7-9]. For this reason, this paper analyzes the classification pattern of high potassium glass and lead-barium glass based on the available data; selects the appropriate chemical composition to subclassify each category, gives the specific division method and results, and uses the established model to identify the types of unclassified glass products based on their chemical composition [10-12].

2. Modeling of glass artifact classification

The classification of glass artifacts relies heavily on the differences in the type and content of the chemical components contained in them; for example, high potassium glass is fired with a high content of potassium, while lead-barium glass contains a high content of lead oxide barium oxide. Therefore, we consider the difference in the chemical substances that make up the two types of glass to distinguish between the glass artifacts.

Considering that both glasses contain 14 different chemical components, we can use principal component analysis to remove some chemical substances that contribute less to the overall, leaving those that contribute more, and then perform binary logistic regression by using the values of the principal components of different chemical substances to obtain the classification laws of the two glasses

2.1. Chemical substance screening by principal component analysis

Principal component analysis is a dimensionality reduction method that uses fewer variables to explain most of the information in the original data by reducing the number of variables. With principal component analysis, fewer variables are obtained and these variables can contribute about eighty percent to the original data.First put 56 samples with 14 indicators into a matrix to get the matrix$X(i,j)(i = 1 \dots 56, j = 1 \dots 14)$，Re-standardization X，The processing formula is

$$\tilde{X}_{i,j} = \frac{x_{i,j}-\mu_j}{s_j}, \quad j = 1 \dots 14 \tag{1}$$

The obtained matrix is called the normalized matrix$\tilde{X}_{i,j}$.Then calculate the covariance matrix of the normalized matrix$R_{i,j}$， Then calculate the eigenvalues of the covariance matrix$\lambda_1 > \lambda_2 > \cdots > \lambda_j$with the eigenvector$u_1, u_2 \dots, u_j$，Among them$u_j = [u_{j1}, u_{j2}, \dots, u_{ji}]$，New indicator variables, i.e., principal components, are obtained from these eigenvectors as coefficients.

$$\begin{cases} y_1 = u_{11}\tilde{X}_{11} + u_{12}\tilde{X}_{12} + \cdots + u_{1i}\tilde{X}_{1i} \\ y_2 = u_{21}\tilde{X}_{21} + u_{22}\tilde{X}_{22} + \cdots + u_{2i}\tilde{X}_{2i} \\ \vdots \\ y_j = u_{j1}\tilde{X}_{j1} + u_{j2}\tilde{X}_{j2} + \cdots + u_{ji}\tilde{X}_{ji} \end{cases} \tag{2}$$

Then the contribution rate of each eigenvalue and the cumulative contribution rate are calculated,

$$b_j = \frac{\lambda_i}{\sum_{k=1}^{m}\lambda_k}, j = 1,2, \dots m \tag{3}$$

$$\alpha_p = \frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{m}\lambda_k} \tag{4}$$

The processed chemical composition data of high lead and lead-barium glass were imported into Matlab, and the corresponding eigenvalues and eigenvectors were calculated and obtained by using the program written in Matlab, and then the contribution rate of each eigenvalue was calculated, and the cumulative contribution rate was calculated to show that the contribution rate of the first six eigenvalues accumulated to 80%. Using the corresponding eigenvectors, six principal components were constructed.

The obtained principal components then calculate the principal component scores of the 56 samples to obtain the principal component score table.

Table 1. Table of principal component scores

	y_1	y_2	y_3	y_4	y_5	y_6
x_1	-2.5256	1.0298	1.8764	-0.4710	0.5755	0.9190
x_2	-0.2062	1.1214	-1.5864	-0.4107	0.5104	-0.2670
x_3	-2.1589	0.0108	1.3609	-0.4175	0.3608	0.6244
……			……			
x_{54}	1.5191	-0.8613	-0.3103	-0.6096	0.4318	0.5365
x_{55}	1.5758	-1.0638	-0.1363	-0.3253	0.3676	0.5423
x_{56}	1.2292	1.6621	-0.3370	-0.7038	0.4776	0.1175

2.2. Logistic regression using principal component scores

Previously, we have used principal component analysis to reduce the number of chemical substance variables and to obtain principal component scores corresponding to 56 glass sample indicators. We now use the principal component scores and perform binary logistic regression using SPSS.

The significance of the test result is greater than 0.5, which indicates that its regression effect is good, and the significance of 1 in the table indicates that the logistic regression is significant. The regression classification table is as follows.

Table 2. Regression classification table

		Predictions		
		Lead Barium	High Potassium	Percent correct
Real test	Lead Barium	40	0	100
	High Potassium	0	16	100
	Overall percentage			100

The table shows that the correct rate of the regression reached 100%, which indicates that the data we obtained through the principal component analysis can explain the difference between the two types of glass very well.

In summary, for the classification law of high potassium glass and lead-barium glass, we can first use principal component analysis to reduce the number of variables so that a small number of variables can be used to explain most of the information, and then use the obtained principal component indicator values to replace the original indicator values for binary logistic regression, which can well distinguish the two types of glass.

2.3. Subclass classification of high potassium and lead-barium glass based on K-means clustering algorithm

The process of subclassifying high potassium glass and lead-barium glass is shown in the following diagram:

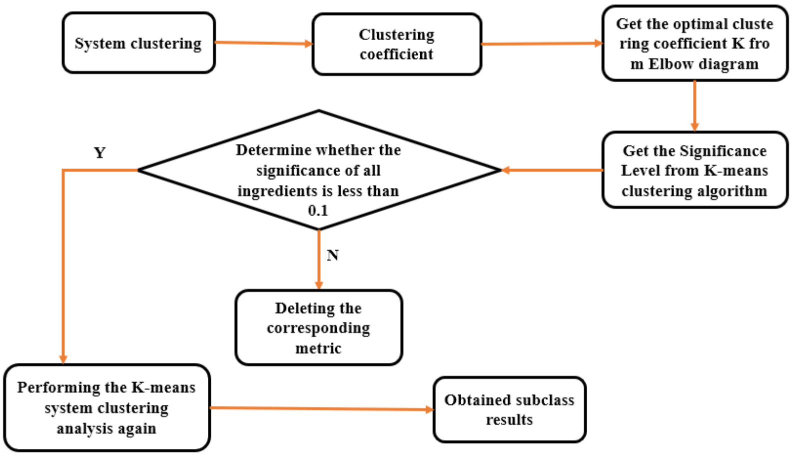

Figure 1. Flow chart of two glass subclasses division

2.3.1 K-means Clustering Algorithm

The basic principle of the K-means clustering algorithm is to first randomly select K sample points as the initial cluster centers, and then calculate the distances between other sample points and the cluster centers separately, assign each sample point to the cluster center with the shortest distance to it, and then recalculate the new cluster centers after each sample point is assigned to a cluster center, and repeat the process until all sample points have a cluster center. To calculate the distance from the sample points to the cluster centers, we use the squared Euclidean distance, i.e.

$$d_i = \sqrt{(X_m - x_n)^2} \tag{5}$$

where X_m is the mth clustering center and x_n is the nth sample point.

Before using K-means clustering, we also have to determine the initial number of cluster centers K. Here we consider using the systematic clustering method first to get the clustering coefficients J of the sample points, then using the clustering coefficients as the vertical axis and K as the horizontal axis, we draw the elbow diagram, find the point on the elbow diagram where the curve suddenly becomes flat, and use the value of K corresponding to this point as the initial number of cluster centers for K-means clustering.

2.3.2 Determining the number of clusters K values using elbow plots

Firstly, we need to determine the number of clusters K value of K-means algorithm, and we can get the clustering coefficients of high potassium glass and lead-barium glass according to the systematic clustering algorithm of SPSS. The above table will be drawn with EXCEL software elbow diagram, as shown in Fig. Observing the elbow diagram, we can find that the trend of the curve starts to become flat from K=3. According to the principle of elbow diagram to determine the K value, we directly choose the initial number of clustering centers K=3.

Similarly, the number of clusters for lead barium glass can be derived as 4

2.3.3 Subclassification of two types of glass using K-means clustering

Before proceeding with the K-means clustering algorithm, it is necessary to determine whether each chemical composition is suitable for use in the algorithm. We used the systematic clustering of SPSS software to obtain the significance levels of the chemical components of high potassium glass and lead-barium glass, as shown in the following table.

Table 3. Significance level of each chemical composition of high potassium and lead-barium glass

	High Potassium Glass	**Lead barium glass**
Chemical Indicators	Significance level	
Silicon dioxide	0	0
Sodium oxide	0.1138	0.016
Potassium oxide	0	0.59
calcium oxide	0	0.008
magnesium oxide	0.03003	0.246
Aluminum oxide	0.00034	0.129
Iron oxide	0.05429	0.634
copper oxide	0.46503	0
Lead oxide	0.19394	0
barium oxide	0.01216	0
phosphorus pentoxide	0.25929	0.001
Strontium oxide	0.18754	0.002
Tin oxide	0.10904	0.817
Sulfur dioxide	0.10596	0

Note: 0 indicates a very high level of significance

Since a significance greater than 0.1 indicates that the difference between this chemical composition and other chemical compositions is not significant, i.e., a better clustering result cannot be obtained by using this chemical composition for cluster analysis, we removed the high potassium glass and lead-barium glass with their respective chemical compositions having a significance greater than 0.1, and then re-clustered them by K-means.

The above retained chemical indicators were subjected to systematic cluster analysis using SPSS software, and the clustering results of high potassium glass and lead-barium glass were obtained separately as shown in Table.

Table 4. Cluster classification of high potassium glass

Clustering	Artifact Number
1	01,04,05,06,13,14,16
2	03,18,21
3	07,09,10,12,22,27

Table 5. Cluster classification of lead barium glass

Clustering	Artifact Number
1	39,40,43,54
2	08,20,24,26,
3	02,11,19,30,36,38,41,49,50,51,52,56,57,58
4	23,25,28,29,31,32,33,35,37,42,44,45,46,47,48,53,55

2.4. Rationalization and sensitivity analysis of classification results

2.4.1 Rational Analysis

According to the results of the K-means cluster analysis in the previous question, the significance of the F-test of each chemical indicator involved in the cluster analysis was obtained, and the significance of the chemical indicators were all less than 0.1, indicating that these components can better reflect the characteristics of each classification in the clustering process, and therefore the results obtained have good reasonableness.

2.4.2 Sensitivity Analysis

For sensitivity analysis, we changed the values of the chemical components with higher significance values in the tables and then analyzed their K-means clustering with SPSS software, and compared the results obtained with those in Tables 10 and 11. If there was a significant change in the color categories of each sample in the results obtained after the change, it indicated that our model was highly sensitive, and if the results did not change much, it indicated that the model was less sensitive.
(1) Sensitivity analysis of high potassium glass
According to the significance magnitude of each chemical component in the K-means clustering results, we adjusted the content of the substances with greater significance: magnesium oxide, iron oxide and barium oxide to 0.95 times of the original one, and then used SPSS software to perform K-means cluster analysis, and the following results were obtained.:

Table 6. Cluster classification of high potassium glass

Clustering	Artifact Number
1	01,04,05,06,13,14,16
2	03,18,21
3	07,09,10,12,22,27

Comparing the above table with the table in the second subquestion The clustering classification of high potassium glass1, we can clearly see that it changes a lot, so the following conclusion is drawn: the sensitivity of the subclassification results of high potassium glass is very high.
(1) Sensitivity analysis of lead-barium glass
Similarly, the chemical indicators of sodium oxide, calcium oxide, lead oxide and strontium oxide were adjusted to 0.95 times of the original ones, and then cluster analysis was performed using SPSS software, and the following results were obtained.:

Table 7. Cluster classification of lead barium glass

Clustering	Artifact Number
1	39,40,43,54
2	08,20,24,26,
3	02,11,19,30,36,38,41,49,50,51,52,56,57,58
4	23,25,28,29,31,32,33,35,37,42,44,45,46,47,48,53,55

This can be clearly seen that there is no change from the results before the data change.
To make the results more credible, we then adjusted the corresponding data to 0.9 times the original and then used SPSS software for cluster analysis, The results show that the results remain the same as before the data were changed, so we conclude that the sensitivity of the subclassification results for lead-barium glass is particularly low.

3. Classification of glass to be tested

After considering the improvement of the model that solved the above, the number of variables was first reduced using principal component analysis, and the obtained principal component values were subjected to Fisher's linear discriminant analysis to classify the glass to be tested according to the results of the analysis. The data representing silica, sodium oxide, ..., and sulfur dioxide were normalized, respectively. The original data were standardized and substituted into the principal component equation obtained above, and the four principal component values of the eight unknown glass artifacts were obtained. The principal component analysis process is more cross single and is omitted here.

The principle of Fisher linear discriminant analysis is to project the samples onto a one-dimensional line so that sample points of the same class are as close and dense as possible, and sample points of different classes are as far away as possible. Using Fisher linear discriminant analysis, two overall Fisher discriminant functions based on the idea of ANOVA can be obtained

$$C(Y) = C_1Y_1 + C_2Y_1 + \cdots + C_pY_p \tag{6}$$

where the coefficient $C_1, \dots, C_p$ The principle of determination is to make the maximum inter-group distance difference between two groups and the minimum intra-group divergence for each group. Once the discriminant function is obtained, we can bring the new index values into the two discriminant functions separately to find the discriminant value, and whichever group has a larger discriminant value indicates which group the sample belongs to.

First, the chemical composition content of high potassium glass and lead-barium glass were put into SPSS for discriminant analysis, and the discriminant function of high potassium glass was obtained

$$G(y_i) = -2.045y_1 - 0.121y_2 + 1.579y_3 - 1.189y_4 - 3.944 \tag{7}$$

and the discriminatory function of lead barium glass

$$Q(y_i) = 0.818y_1 - 0.048y_2 - 0.631y_3 + 0.475y_4 - 1.213 \tag{8}$$

in the formula: $y_i (i = 1,2,3,4)$ indicates the principal component value.

Then, we imported the principal component values of the eight samples obtained in the previous subquestion into SPSS and performed Fisher discriminant analysis with the known categories of glass together with these eight glasses to be distinguished, and obtained the following results:

Table 8. Results of Fisher's linear discriminant analysis for eight samples

	y1	**y2**	**y3**	**y4**	**Predicted group**	**in lead barium glass**	**in high potassium glass**
A1	-1.26	0.09	-0.49	-0.86	1	0.232	0.768
A2	1.59	-0.40	-2.02	-2.51	0	0.999	0.001
A3	2.58	-1.83	-0.19	1.74	0	1.000	0.000
A4	1.53	-1.38	0.30	0.60	0	0.999	0.001
A5	-3.32	-2.58	1.63	-0.36	1	0.000	1.000
A6	-1.21	1.97	-1.30	1.31	0	0.991	0.009
A7	-1.41	1.65	-1.14	0.76	0	0.941	0.059
A8	1.50	2.48	3.22	-0.68	1	0.310	0.690

Meanwhile, we use the Fisher linear discriminant function obtained earlier to calculate the discriminant values of the two discriminant functions for samples A1 to A8, respectively, and the following results are obtained.:

Table 9. Two discriminant function values for the eight types of glass

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
High potassium glass	-1.129	-7.352	-11.368	-7.146	6.159	-5.318	-3.964	-1.419
Lead barium glass	-2.347	0.189	1.932	0.200	-5.004	-0.855	-1.365	-2.460

According to the principle of discriminant function, we can see from the table that the discriminant value of the high potassium glass function of glass artifacts A1, A5 and A8 is greater than the discriminant value of the lead-barium glass discriminant function, and combined with the results of SPSS discriminant analysis, we can conclude that A1, A5 and A8 glass belong to high potassium glass, and the other five glass artifacts belong to lead-barium glass.

4. Conclusions

In order to identify the glass artifacts, this paper first used principal component analysis to reduce the dimensionality to get six principal components and obtained the principal component values of two kinds of glass according to the calculation formula. The principal component values were imported into SPSS using binary logistic regression and the regression results were obtained. The clustering coefficients were obtained by using the SPSS software to analyze the clustering coefficients of high potassium glass and lead-barium glass, and the elbow diagrams were drawn by using the EXCEL software to determine the optimal number of clusters for high potassium glass and lead-barium glass as 3 and 4. K-means clustering analysis was then performed for high potassium glass and lead-barium glass, and the appropriate clusters were selected according to the significance values in the results. The K-means clustering analysis was then performed on the high potassium glass and the lead-barium glass, respectively, and the appropriate chemical components were screened out based on the significance values in the results. The significance values in the last K-means cluster analysis results were less than 0.1, indicating that the clustering results were reasonable.

After the model was built, the classification model was modified. The discriminant functions of high potassium and lead-barium glass were obtained separately by Fisher's linear discriminant analysis$G(y_i)$， $Q(y_i)$.The unknown glass samples were then analyzed using principal component analysis to obtain the four principal components and the corresponding principal component values. These principal component values were put into SPSS for Fisher's linear discriminant analysis to obtain the analysis results.

References

[1] Chemistry; New Findings from University of Sofia in Chemistry Provides New Insights (Multivariate Analysis for the Classification of Copper Lead and Copper Zinc Glasses) [J]. Journal of Technology & Science,2020.

[2] Chemistry; Reports from University of Sofia Advance Knowledge in Chemistry (Multivariate analysis for the classification of copper-lead and copper-zinc glasses) [J]. Chemicals & Chemistry,2020.

[3] Harshit Mathur,Aditya Surana. Glass Classification based on Machine Learn-ing Algorithms[J]. International Journal of Innovative Technology and Exploring Engineering (IJITEE),2020,9(11).

[4] Dimitar Dimitrov, Miroslava Nedyalkova, Sergio Madurga, Ludmila Naneva, Vasil Simeonov. Multivariate analysis for the classification of copper–lead and copper–zinc glasses[J]. Open Chemistry,2020,18(1).

[5] Vedrtnam Ajitanshu, Pawar S. J.. LAMINATED GLASS: CLASSIFICATION, CHARACTERIZATION AND FUTURE PERSPECTIVES[J]. JOURNAL OF MATERIALS EDUCATION,2020,42(1-2).

[6] G.A. Gogotsi. Classification of ceramics and glass (edge chipping and fracture toughness) [J]. Ceramics International,2014,40(4).

[7] Megan N. C. Grainger, Merilyn Manley-Harris, Sally Coulson. Classification and discrimination of automotive glass using LA-ICP-MS[J]. Journal of Analytical Atomic Spectrometry,2012,27(9).

[8] Daniel Ramos, Grzegorz Zadora. Information-theoretical feature selection using data obtained by Scanning Electron Microscopy coupled with and Energy Dispersive X-ray spectrometer for the classification of glass traces[J]. Analytica Chimica Acta,2011,705(1).

[9] Vandini Mariangela, Fiori Cesare, Cametti Rachele. Classification and technology of byzantine mosaic glass. [J]. Annali di chimica,2006,96(9-10).

[10] Schmidt T,Vogt T, Surmann J P, Kössler W. Glass classification by linear discriminant analysis of LA-ICP-MS data.[J]. Die Pharmazie,2002,57(6).

[11] K. Scholz, E. Henneberg, B. Thomas. Classification of electrode glasses including structural parameters[J]. Fresenius' Journal of Analytical Chemistry,1994,349(1-3).

[12] J Michael Marcouiller. A Revised Glass Annealing Method to Distinguish Glass Types[J]. Journal of Forensic Science,1990,35(3).