Composition analysis and identification of ancient glass products based on population forest-KNN algorithm

Ziyuan Chen *

School of Information Engineering, Guangdong University of Technology, Guangzhou, Guangdong, 510006

* Corresponding Author Email: 3120002555@mail2.gdut.edu.cn

Abstract. This paper is based on the KNN and Random Forest algorithms in machine learning to analyze the differences and correlations in the compositions of ancient glass objects, determine their classification and identify the unknown categories of artifacts. The classification rules of different types of glass are analyzed; for each category, appropriate sub-classification processes are performed according to their chemical composition, specific classification methods and processing results are given, and sensitivity and rationality analyses are performed on the results. The chemical composition of glass artifacts of unknown categories is analyzed, and then their categories are identified and sensitivity analysis is performed on their classification results. Samples of glass artifacts from different categories and analyze the relationships between their chemical compositions, and compare the differences that exist in the relationships between the chemical compositions of different categories.

Keywords: Random Forest; sensitivity analysis; Euclidean distance; KNN nearest neighbor algorithm.

1. Introduction

Usually the main material for manufacturing glass is quartz sand, chemically known as silicon dioxide (SiO2) [1~3]. In ancient times, grass ash, saltpeter, lead ore and natural bubbling soda were commonly used as accelerants to lower the melting temperature of pure quartz sand, and limestone was added whose combustion produced calcium oxide as a stabilizer [4,5]. We in the role of different combustion aids, the main chemical composition of glass products are different, to our own invention of lead barium glass, for example, lead ore as a combustion aid fired lead barium glass in lead oxide (PbO) and barium oxide (BaO) content is relatively high, Chu culture of the glass to lead barium glass [6~10]. Another example is the potassium glass made of grass ash with high potassium content as a combustion aid is mainly popular in Lingnan, China, as well as India and Southeast Asia and other regions.

It is important to analyze the classification laws of different types of glass, give specific classification methods and treatment results, and compare the differences in the relationship of chemical composition between different categories to study the weathering of ancient glass.

2. Methods

2.1. Fisher's linear discriminant

According to the question, we were asked to investigate the classification law of lead-barium and high-potassium glass, and we presumed that their classification law was based on the content of chemical components. The Fisher linear discriminant method is used to draw a line in the plot, and each point is projected onto this line so that as many data of the same type fall together as possible.

We first combine each two variables and draw their scatter plots and then find the projection on the W line according to the two variables X1 and X2 to obtain a formula

$$Y = W^T X \tag{1}$$

In order to better obtain the W-line we first have to make sure that the projections are sufficiently spaced from each other and that the projections of the same kind are sufficiently dense to determine the mean value Di of the samples of class I of the two values and the mean value on the projection as Di*

$$D_i = \frac{1}{n_i} \sum_{X \in x_i} X \tag{2}$$

$$D_i = \frac{1}{n_i} \sum_{Y \in y_i} Y = \frac{1}{n_i} \sum_{X \in x_i} W^T X = W^T D_i \tag{3}$$

The mean value of inter-projection distance is

$$\| m_1^* - m_2^* \| = \| W^T (D_1 - D_2) \| \tag{4}$$

The degree of dispersion, the effect of intra-class aggregation, is indicated by the dispersion variance.

$$F_i^{*2} = \sum_{y \in Y_i} (y - D_i^*)^2 \tag{5}$$

The total dispersion F is the sum of their variances.

$$F_B^* = (D_1^* - D_2^*)(D_1^* - D_2^*)^T \tag{6}$$

The criterion function is

$$H(X) = \frac{\| D_1^* - D_2^* \|^2}{F} \tag{7}$$

After deducing the integrated Lagrange multiplier method, to find the extreme value W needs to satisfy

$$H(X) = F_W^{-1} (D_1 - D_2) \tag{8}$$

We determine the threshold δ after converting n dimensions into one dimension

$$\delta = -\frac{W^T (m_1 + m_2)}{2} \tag{9}$$

The final discriminant formula is obtained as:

$$\begin{cases} if W^T X + \delta > 0, X \in x_1 \\ if W^T X + \delta < 0, X \in x_2 \end{cases} \tag{10}$$

2.2. KNN nearest neighbor algorithm

We use the KNN nearest neighbor algorithm to solve the problem when the size of the data set is not very large. This method is used to train and learn the content of each chemical component of the known class of elements and then compare the features of the unknown elements to find the closest points to this feature. The main thing is to find the closest point to the target classified element, so the Euclidean distance can be used to determine their closest features.

Input: Training data set Eq.

$$T = (x_1, y_1), (x_2, y_2), ..., (x_N, y_N) \tag{11}$$

Output category formula:

$$y = \arg\max a_j \sum x \in N_k(x) M(y_i = a_j) \tag{12}$$

where xi∈X; yi∈{1, 0}(1=high potassium, 0=lead barium); i=1,2,...,N, M is the discriminant function, i.e., M is 1 when yi=aj, otherwise M is 0. The k points in the training set T that are nearest neighbors to x are found, and are denoted as Nk(x). The distance algorithm is Euclidean distance, which does not differ much from the above approach, and is not described in detail.

According to the above figure we can clearly see that the two types of glass, high potassium and lead-barium, are well distinguished between iron oxide and lead oxide, lead oxide and copper oxide, barium oxide and strontium oxide, barium oxide and phosphorus pentoxide, and phosphorus pentoxide and strontium oxide, so we believe that the content of the six chemical components present above is the main factor affecting the classification of glass types.

Based on preliminary observations we found that 14 variables had an impact on the classification results, and found that this number of variables was too large and confounding such as the presence of too many zeros in some variables or their irregular distribution. In order to more accurately and rationally subclassify the different types of glass artifacts, we needed to pre-process the data. We decided to reduce the dimensionality of a large number of variables, i.e., to find a few important indicators of categories among multiple variables, so we adopted the method of PCA (principal component analysis) to find several categories of variables that have a greater impact on the classification of glass artifacts as the main basis for the study, and the results of data processing are shown in the following table 1.

Table 1. Principal component analysis of each type of artifact

High Potassium	Silicon Dioxide	Oxidation Na	Potassium oxide	Calcium oxide	Magnesium oxide	Aluminum oxide	Iron oxide			
Lead Barium	Silicon Dioxide	Oxidation Na	Potassium oxide	Calcium oxide	Magnesium oxide	Aluminum oxide	Iron oxide	Copper oxide	Lead oxide	Barium oxide

By the above pre-processing we get 2 sets of principal component variables, in this problem we do not know the specific categories, so it is defined as a clustering problem, according to our above selected principal components and their respective influence factors we can use the systematic clustering method to calculate their classification.

For the systematic clustering method, the "distance" between each artifact is calculated to combine two by two, and finally aggregated into a system through several iterations to generate a clustering spectrum map. The algorithm for "distance" is obtained by Euclidean distance.

The final clustering genealogy chart (tree diagram) can be obtained by SPSS as Figure 1

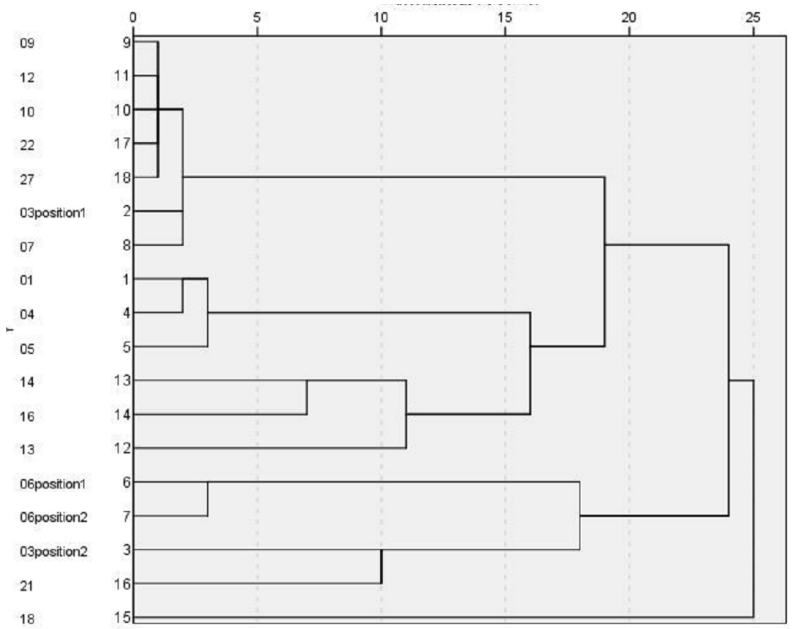


Figure 1. Genealogy chart

The same is true for lead barium glass.

In order to obtain the optimal number of clusters we examine the degree of distortion to obtain the following data, such as table 2.

Table 2. Lead-barium and high potassium aberrations

Lead-barium distortion	65.448	21.938	11.706	7.832	6.409	4.359	2.829	1.482	
	42.922	20.005	10.304	7.46	5.713	4.275	2.521	1.46	
	33.31	16.542	9.748	7.33	5.239	3.808	2.376	1.092	
	32.661	14.374	8.304	7.33	5.139	3.495	2.005	0.614	
	31.604	14.343	8.186	7.234	4.796	3.262	1.774	0.349	
	24.635	13.137	7.857	6.491	4.584	3.1	1.76	0.178	
High Potassium Aberration	39.013	36.232	28.265	26.588	24.045	16.852	14.671	10.335	3.921
	3.608	2.34	1.992	1.594	1.013	0.457	0.254	0.024	

Make their scatter plots separately as figure 2 and 3:

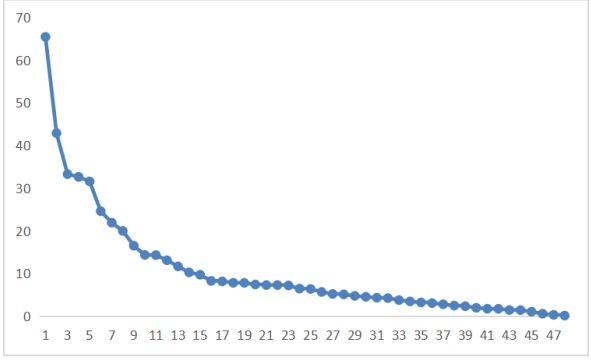


Figure 2. Lead-barium high-potassium aberration map

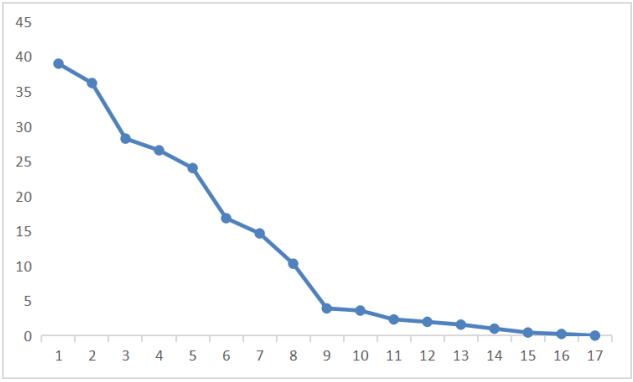


Figure 3. High potassium distortion map

We can find their optimal number of clusters K=3 according to the elbow principle, and we can synthesize the above clustering spectra to get the clustering table as table 3.

Table 3. Subclassification of different types of glass clusters

Lead Barium				High Potassium	
Sample Points	Group	Sample Points	Group	Sample Points	Group
2	1	44 Unweathered spots	1	1	A
11	1	45	1	03 part 1	A
19	1	46	1	4	A
20	1	47	1	5	A
23 Unweathered spots	1	49	1	7	A
25 Unweathered spots	1	49 Unweathered spots	1	9	A
28 Unweathered spots	1	50	1	10	A
29 Unweathered spots	1	50 Unweathered spots	1	12	A
30 part 1	1	51 part 1	1	13	A
30 part 2	1	51 part 2	1	14	A
31	1	52	1	16	A
32	1	53 Unweathered spots	1	22	A
33	1	54	1	27	A
34	1	54 Severe weathering	1	03 part 2	B
35	1	55	1	06 part 1	B
36	1	56	1	06 part 2	B
37	1	57	1	21	B
38	1	58	1	18	C
39	1	8	2		
40	1	08 Severe weathering	2		
41	1	24	2		
42 Unweathered spots 1	1	26	2		
42 Unweathered spots 2	1	26 Severe weathering	2		
43 part 1	1	48	3		
43 part 2	1				

The above division results are divided into three groups, 1, 2 and 3 in the lead-barium group, which correspond to the previous artifact labeled sample points in the high potassium group as A, B and C, which correspond to the previous artifact labeled sample points.

1 represents all cases except low barium oxide, high oxidation and high barium oxide, low alumina cases 2 represents high barium oxide, low alumina, 3 represents low barium oxide, high alumina; A represents all cases except B and C, B high alumina, high potassium oxide, low silica, C low alumina, low potassium oxide, high silica.

We observe the classification after the change by changing the content of different chemical components so as to determine whether it is reasonable or not to obtain the table 1 as follows

Table 4. High potassium and lead-barium sensitivity test

Change amount (Lead barium)	Change potassium oxide			Change of barium oxide			Changing lead oxide		
	total	Number of errors	Accuracy	Total	Number of errors	Accuracy	Total	Number of errors	Accuracy
Original content *0.8	48	0	100%	48	0	100.00%	48	0	100.00%
Original content *0.9	48	0	100%	48	0	100.00%	48	0	100.00%
Original content *1	48	0	100%	48	0	100%	48	0	100%
Original content *1.1	48	0	100%	48	0	100.00%	48	0	100.00%
Original content *1.2	48	3	93.75%	48	0	100.00%	48	0	100.00%
Original content *1.3	48	3	93.75%	48	0	100.00%	48	0	100.00%

Change potassium oxide (high potassium)	total	Number of errors	Accuracy	Total	Number of errors	Accuracy	Total
Original content *0.7	18	2	88.89%	Original content *1	18	0	100%
Original content *0.8	18	2	88.89%	Original content *1.1	18	0	100.00%
Original content *0.9	18	0	100.00%	Original content *1.2	18	0	100.00%

According to the above table, for the sensitivity analysis of high potassium, based on the seven principal components identified by the principal component analysis, by changing the content of potassium oxide, barium oxide and lead oxide, with the other components unchanged, it was found that under 20% disturbance fluctuation, two numbers were different from the previous clustering results without changing the content of potassium oxide by clustering, and the overall accuracy was 88.89%.

For the sensitivity analysis of lead and barium, based on the ten principal components identified by the principal component analysis, by changing the content of potassium oxide, lead oxide and barium oxide one by one, with the other components remaining unchanged, it was found that under the 20% fluctuation of disturbance, the results of three numbers by clustering differed from the previous clustering results without changing the content of potassium oxide, with an overall accuracy of 93.75%.

For the classification of the unknown artifacts in Form III, what we know is their chemical composition, that is, to put them into the category that is most similar to their chemical composition. It can be judged that this is a classification model that

The final classification results were obtained as shown in Table 2.

Table 5. lassification of unknown artifacts

Artifact Number	A1	A2	A3	A4	A5	A6	A7	A8
Type	High Potassium	Lead Barium	Lead Barium	Lead Barium	Lead Barium	High Potassium	High Potassium	Lead Barium

In addition to this we used binary logistic regression to predict the type of unknown artifacts as high potassium or lead-barium on top of question 2. Based on SPSS software, we obtained 100% accuracy of the processed classification data and concluded that the unknown artifact types were high potassium, high potassium, lead-barium, lead-barium, lead-barium, high potassium, lead-barium, and on the basis of these predicted types, we combine each chemical composition and its corresponding type in Annex II into a total data for the next step of subclassification. After predicting the types from the above, they were classified into subcategories under their respective types. We obtained the classification of unknown artifacts with the help of SPSS software for the high potassium type into category A and B respectively, and the unknown artifacts with the lead-barium type into category 1, 1, 2, 1 respectively. Sensitivity analysis: Based on the same principle of sensitivity analysis as in Problem 2, the predicted results were found to be more robust based on SPSS software, and the accuracy of the predicted results was still 100% after certain perturbations were generated.

Table 6. Classification of unknown artifacts

Artifact Number	A1	A2	A3	A4	A5	A6	A7	A8
Type	High Potassium	Lead Barium	Lead Barium	Lead Barium	Lead Barium	High Potassium	High Potassium	Lead Barium
Subcategories	1	A	A	A	C	1	1	A

2.3. Random Forest Prediction

Random forest is an algorithm that integrates multiple combinations of classifications to predict a thing (1) each time n samples are taken from the training set and put back to get a new training set; (2) with the new training set, multiple sub-models are trained; (3) for the classification problem, a simple prediction model that uses a voting method to give each category a voice. The main random forest completion is mainly divided into the construction of decision trees and the generation of results. According to the random forest algorithm we get the following feature correlation coefficient plots for the two types of glass as Figure 12 and Figure 13

We analyze the variability of the correlation between the chemical composition of different types of artifacts based on the tables of composition characteristic coefficients generated by the random forest for high potassium and lead-barium glasses. By observing the above component characteristic correlation coefficients, the negative correlation between magnesium oxide and silica in high potassium glass is extremely strong at -0.68, the positive correlation between iron oxide and phosphorus pentoxide is extremely strong at 0.71, and the correlation between potassium oxide and tin oxide is very weak at 0.0072; observe the negative correlation between lead barium silica and barium oxide is extremely Figure 4 and 5.

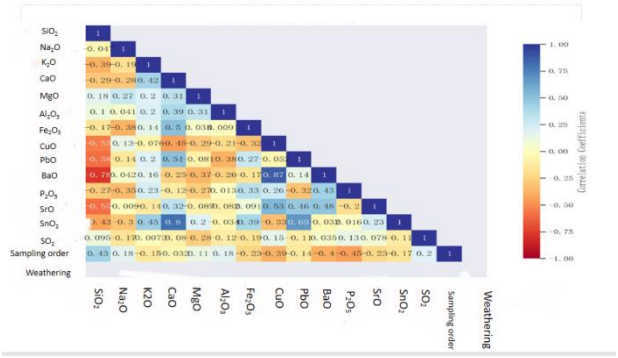


Figure 4. Correlation coefficient of lead-barium composition characteristics

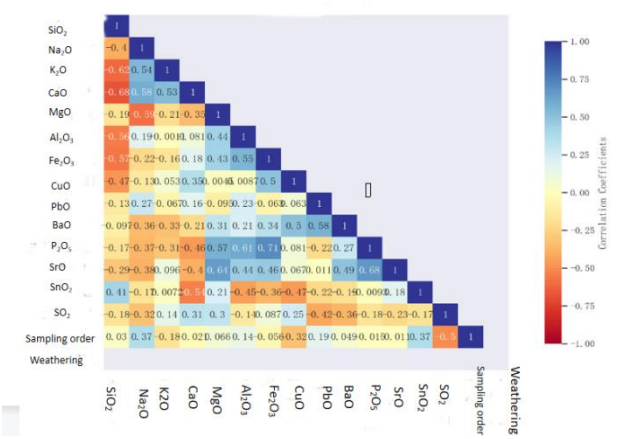


Figure 5. Correlation coefficients for the characteristics of high potassium components

The positive correlation between copper oxide and barium oxide is 0.87, and the correlation between potassium oxide and sulfur dioxide is the weakest, only 0.0073. We compare the two types of charts, we can see that the correlation between magnesium oxide and silica is reduced to 0.18 compared to high potassium glass but with a positive correlation trend, and the positive correlation coefficient between iron oxide and phosphorus pentoxide is reduced to 0.33. Reduced to 0.33; high potassium glass compared to lead barium glass silica and barium oxide negative correlation coefficient reduced from a strong negative correlation to -0.097, meaning almost no correlation, copper oxide and barium oxide only reduced to a positive correlation coefficient of 0.5, relatively little change in the correlation of these two groups of chemical composition.

3. Conclusion.

In this paper, we analyze the classification laws of different types of glass, perform suitable subclassification processing, give classification methods and results, and perform sensitivity and rationality analysis on their results.

Sensitivity analysis of the classification results for unknown categories of glass artifacts.

The differences in chemical composition relationships between different categories are compared for samples of glass artifacts from different categories.

References

[1] YANG Yanzhao, WU Wenjie. A method for rapid preparation of silica aerogel composites by single solvent exchange from water glass:, CN112592149A [P]. 2021.

[2] LUO Dan 1, LONG Lijuan 2, QIN Shuhao 1,2*, HE Min 1. Preparation and properties of glass fiber and carbon fiber reinforced silica aerogel composites[J]. New chemical materials, 2022, 50(4): 5.

[3] Xie EJ, Lin JP, Gan GC, et al. Experimental study on the purification of low-iron quartz sand from kaolin tailings for photovoltaic glass [J]. Building Materials World, 2021, 42(3): 4.

[4] Chen K. Geological characteristics of metallurgical fused limestone deposits in Zhoujiazhai, Shuicheng County, Guizhou [J]. Western Resources, 2022(1): 3.

[5] Sun Lijuan, Yang Maolin, Du Yinggang, et al. Distribution and migration characteristics of Ca^(2+), Mg^(2+) and Si^(4+) in thermal power plant desulfurization systems [J]. Power Generation Equipment, 2022, 36(5): 6.

[6] Yang, B.D.. An overview of homemade glass from the Western Zhou to the North and South [J]. Journal of the National Palace Museum, 2003(5): 6.

[7] Li Ruming. The effect of different combustion aids on the combustion traces of pine wood [J]. Fire Protection: Electronic Edition, 2021.

[8] Yin Zhenyu, Sheng Minrong, Sun Lihua, et al. A glass fiber reinforced rigid PVC plastic cable conduit and its production method:, CN112175328A [P]. 2021.

[9] Zhou Jing. The pearl connection and unique wisdom of Chu glass art and foreign craft [J]. Journal of Suzhou College of Arts and Crafts, 2021, 000(001): P.6-11.

[10] Zhou Jing. The general view of Chinese lead and barium glass art obsessions [J]. Journal of Fine Arts, 2021, 000(003): 184-186.