Component Analysis and Classification Model of Ancient Glassware Based on K-means Clustering and Random Forest

Jian Yang #,1,*, Zelin Wang #,2, Song Gao #,2

[1] School of Electronic and Information Engineering, Changchun University of Science and Technology Changchun, China, 130022

[2] School of Artificial Intelligence, Changchun University of Science and Technology, Changchun China, 130022

* Corresponding Author Email: 2020001578@mails.cust.edu.cn

#These authors contributed equally.

Abstract. The analysis and identification of the composition of ancient glass objects has important application and practical significance to the world of archaeology. In this study, after classifying the artifacts according to their types, the corresponding classification models were analyzed and constructed. Using the analysis of classification laws, for each category to select the appropriate chemical components for subclassification, the K-means cluster analysis model was established, and then secondary components were added to the model to select the appropriate model to achieve clustering. Analysis of the clustering results to achieve subclasses, the glass artifacts into three categories: high lead and high barium artifacts, transitional artifacts, high silicon and high potassium artifacts. The classification results were analyzed for reasonableness, and sensitivity analysis was conducted using the OAT method, which showed that the sensitivity of the classification results was high and the chemical composition had good anti-disturbance ability within 10%. A random forest model was also established to determine the category to which the artifacts belonged based on the known chemical composition.

Keywords: Ancient Glassware, Classification Model, K-means, Random Forest.

1. Introduction

In the ancient times, through the Silk Road Chinese ancient glass became the physical evidence of trade between China and the West. The chemical composition of ancient Chinese glass is different from that of glass from other countries [1]. The main raw material of glass is quartz sand, and the main component is silicon dioxide. Ancient glassmaking used grass ash, natural soda ash, saltpeter and lead ore as fluxes, and added limestone as a stabilizer calcined into calcium oxide. It is because of the different fluxes added that the main chemical composition of ancient Chinese glass products differed. Most of the ancient glass artifacts excavated in China are divided into high potassium glass and lead-barium glass, while ancient glass is more prone to weathering due to the burial environment [2]. During the weathering process, the internal elements of the glass artifacts and the environmental elements undergo a large number of chemical and physical reactions, resulting in changes in their composition ratios [3,4]. This paper attempts to develop a mathematical model to find the relevant patterns of chemical composition in glass artifacts.

2. Materials and Methods

2.1. Data

The data in this paper were obtained from the Annex of Question C of the 2022 China Higher Education Society Cup National Student Mathematical Modeling Competition (http://www.mcm.edu.cn/html_cn/node/5267fe3e6a512bec793d71f2b2061497.html).

Pre-processing of the data. The classification information of these artifacts is given in Annex Form 1, and the corresponding proportions of the main components are given in Form 2. The data in Form

2 are compositional in nature, i.e., the sum of the proportions of the components should be 100%, but the sum of the proportions of the components may not be 100% due to testing methods and other reasons. In this paper, the data with the summation of component proportions between 85% and 105% are considered as valid data. Since the data in the Annex Form 1 may be missing, and the data in Form 2 may be invalid, some pre-processing of the given data is needed to prevent the existence of incorrect data from affecting the results of the calculation. For the data in Form 1, the colors of 19, 40, 48, and 58 artifacts were found to be missing; for the data in Form 2, the sum of 15 and 17 artifact components was found to be less than 85%, which was regarded as invalid data. Due to the performance of abnormal data is less, and the data of abnormal information is difficult to complete and correct, so this paper eliminates the attachment of cultural relics numbered 19, 40, 48, 58, 15, 17 of the cultural relics data.

2.2. Introduction to the method

2.2.1 K-means clustering

K-means clustering is a method of cluster analysis, which is a method of grouping sample data according to the similarity of the data itself without a given division into categories. K-means clustering algorithm idea is to cluster k points in space as the center, group the objects closest to them, and update the value of each cluster center one by one by iterative method until we get the best clustering result is [5].

2.2.2 Random Forest Classification

Random Forest is a supervised machine learning method built by integrating decision tree based learners, further introducing randomness in the training process of decision trees to make it have excellent resistance to overfitting and noise resistance [6].

3. Model building and solving

In this paper, correlation analysis between surface weathering and glass type, decoration and color of glass artifacts was conducted among discrete variables. K-means cluster analysis is used to re-establish a suitable classification system, and a reasonableness and sensitivity analysis of the classification system is conducted. For the classification prediction of unknown types of artifacts, this paper chooses to use the random forest model and conducts a sensitivity analysis on the model.

3.1. Analysis of variability

Since the variables in question 1 are discrete qualitative variables, the association between variables cannot be expressed by conventional methods such as person correlation coefficient, so this paper uses the chi-square test to analyze the variance of the data, which in turn reflects the correlation relationship between the variables [7].

3.1.1 Solving Process

Step1: Based on the original data matrix x_{ij} , construct the theoretical data y_{ij} [8]

Step2: Calculate the cardinality value

$$\sum_{\substack{i=1 \\ j=1}}^{\substack{i=2 \\ j=n}} \frac{(x_{ij}-y_{ij})^2}{y_{ij}} \tag{1}$$

$i=2$ indicates both weathered and unweathered conditions; the value of n is the type of decoration, type, and color.

Step3: Calculate the degrees of freedom

$$(i-1)*(n-1)=n-1 \tag{2}$$

Step4: Comparison with the critical value table of the chi-square test, whether the original hypothesis is rejected, and analysis to show the association between variables

3.1.2 Analysis of results

Since the hypotheses established in this paper are all that there is no correlation between variables, the rejection of the original hypothesis is considered as correlation between variables, as follows in Table 1.

Table 1. Cardinality test results

	u	X^2	P
Type	1	10.98	0.005
Ornamentation	2	7.8	0.03
Color	7	6.35	0.5

Comparing the results in the above table with the table of critical values of the chi-square test, we can get the following results: the degree of freedom of glass type is 1, and the chi-square test value is 10.98, which can reject the original hypothesis in 99.5% of cases and prove that type is correlated with surface weathering; the chi-square test value of glass decoration is 7.8, which can reject the original hypothesis in 97% of cases and prove that decoration is correlated with surface weathering; the chi-square test value of glass color is 6.35, which can reject the original hypothesis in 50% of cases and cannot prove that color is correlated with surface weathering. The chi-square test value for glass color is 6.35, which can reject the original hypothesis in 50% of cases and cannot prove that color is correlated with surface weathering.

3.2. K-means based clustering model

Since the category of cultural relics consists of four indicators: ornamentation, type, color, and surface weathering, this paper analyzes the correlation of 14 chemical components to each of the four indicators, and selects the chemical components with higher correlation to cluster the cultural relics.

In this paper, we choose to use Spearman's correlation coefficient for correlation analysis [9], n is the sample size of the sample.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})}} \tag{3}$$

The four qualitative variables of ornamentation, type, color, and surface weathering were quantified, and since there were no variables expressing trends among the four qualitative variables, the quantified values only represented attributes independent of degree, and the processing results are shown in Table 2 below.

Table 2. Table of quantified variables

	Variable Name	Quantification results		Variable Name	Quantification results
Ornamentation	A	1	Color	Pale Blue	1
	B	2		light green	2
	C	3		Deep Blue	3
Type	High Potassium	1		Dark Green	4
	Lead Barium	2		Blue-Green	5
				Green	6
				Purple	7
				Black	8

The results of the correlation analysis of each chemical composition with ornamentation, type, color, and surface weathering are shown in Figure 1 below.

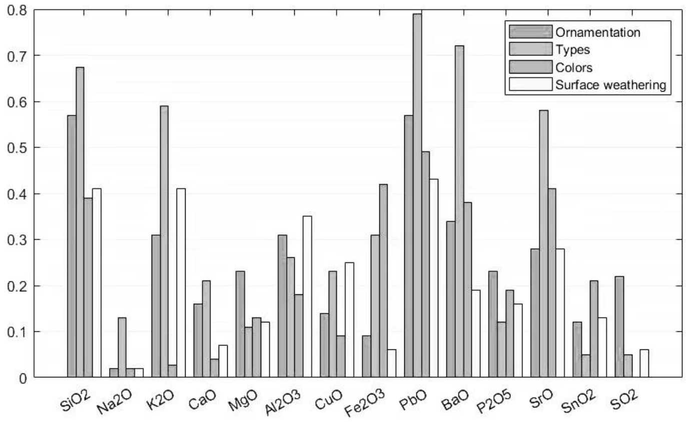

Figure 1. Correlation analysis

Analysis of the above graphs shows that the correlations between Na_2O and ornamentation, color, and surface weathering are all low, and the correlations between K_2O and SO_2 and color are zero. Based on the analysis of the histogram, this paper chose to consider the chemical components with correlations below 0.3 as minor factors and the others as major factors for clustering. It was summarized as SiO_2, K_2O, Al_2O_3, Fe_2O_3, PbO, BaO, SrO as the main chemical components for cluster analysis, and Na_2O, CaO, MgO, CuO, P_2O_5, SnO_2, SO_2 as the minor chemical components.

3.2.1 Modeling

Step1: Select a suitable data set and use the elbow method to determine the value of k, i.e. how many categories to divide into [10].

Step2: Construct a classification model using the major chemical components, fill in the model with the minor chemical components, and compare it with the previous model

Step3: Select the model with stronger explanation to get the final classification result

The core idea of the elbow method is that when the number of categories of classification is larger, the division of samples will be finer and the degree of aggregation of each class will be higher, then the error squared and SSE will naturally decrease, but the corresponding complexity of the model will also increase accordingly, and the elbow method chooses the inflection point out of the decreasing SSE to take as the k value of clustering.

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{4}$$

Figure 2 of the k-value inflection point derived from the elbow method is shown below.

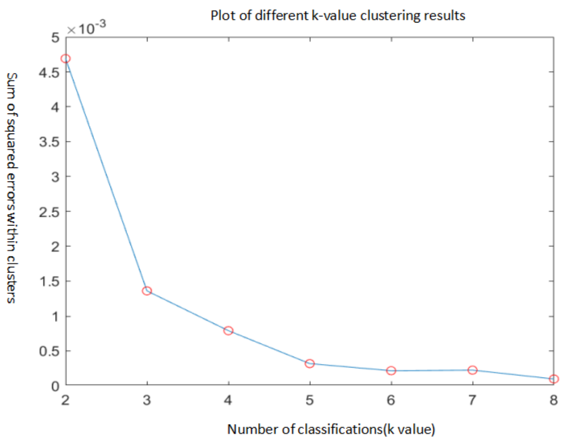

Figure 2. Elbow method inflection point diagram

The inflection point in the above figure is clearly seen when k is 3 out, so it follows that the clustering is relatively efficient when the samples are clustered into 3 classes.

3.2.2 Solving the model

The results of the cluster model variance analysis using the combination of major chemical components and minor chemical components are as follows: CaO: 0.002; MgO: 0.654; Al_2O_3: 0.01; Fe_2O_3: 0.147; CuO: 0.233; SnO_2: 0.534; SO_2: 0.196

The unlabeled chemical components in the above data are all 0. Compared to the clustering results of the model based on the main chemical components, the variables with insignificant differences in the model built using all chemical components appear as 5. The model classification requires the acceptance of some variables with more ambiguous classification.

Central values were taken for the chemical composition of artifacts in each category, as follows in Table 3.

Table 3. Table of central values

Clustering type	SiO_2	K_2O	Al_2O_3	Fe_2O_3	PbO	BaO	SrO
1	24.83	0.17	2.63	0.67	42.57	13.07	0.45
2	56.76	0.18	4.75	0.64	20.30	8.24	0.23
3	76.58	6.07	4.91	1.30	1.11	0.56	0.03

3.2.3 Summary of model results

In this paper, cluster 1 is named as high lead and high barium artifacts, cluster 2 is named as transitional artifacts, and cluster 3 is named as high silicon and high potassium artifacts.

Since there are too many types of chemical components to visualize the higher dimensional data, two chemical components, silica and lead oxide, were selected in this paper to make the effect after cluster analysis, as shown in Figure 3 below.

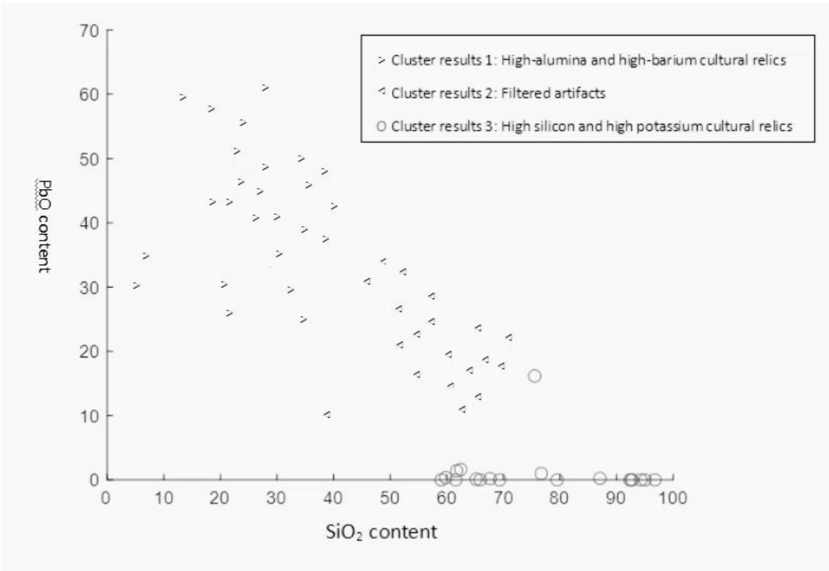

Figure 3. Clustering effect

3.2.4 Rationality and sensitivity analysis

In this paper, based on the clustering results of the reformulation of the subclass division, and the original 19 categories of glass artifacts re-divided according to the content of chemical composition as high lead and high barium artifacts, transitional artifacts, high silicon and high potassium artifacts. The principle of choosing such a division is similar to the method in the question, based on the content of each chemical component of the clustered artifacts to be divided, after a simple processing of the data, it is easy to find that in the clustering result 1, that is, high aluminum and high barium class artifacts, the content of alumina, barium oxide is very high, while the content of other chemical components or high or low, not outstanding; and in the clustering result 3, that is, high silicon and high potassium class artifacts, silica, potassium oxide content is too high; in the clustering results 2 that is, the transition class artifacts, where the chemical composition of all three categories in the middle value of the position, so in order to facilitate the analysis, this paper will be used as a transitional nature of the artifacts, so that researchers can more easily deal with the results.

The sensitivity analysis in this paper uses the OAT method to process the classification results and analyze whether the clustering effect changes by changing the content of some chemical components. If the degree of change is in a small range, the clustering can produce changes obviously, it indicates that the sensitivity of the classification results is high; if the clustering results still do not produce changes in the case of a larger change, it indicates that the sensitivity of the classification results is low.

In this paper, two chemical components, silica and lead oxide, were selected for visual sensitivity analysis. When changing their silica content for clustering effect 1, the clustering effect when raising their silica content by 20%, as follows Figure 4.

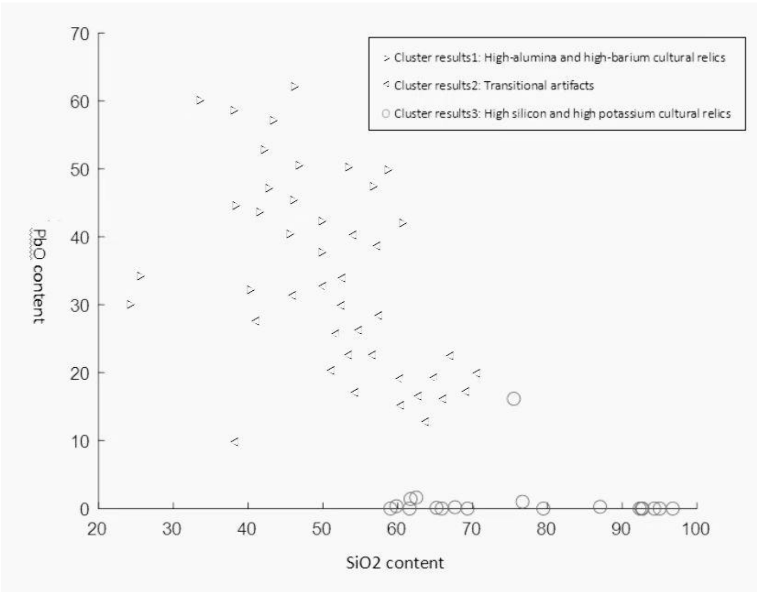

Figure 4. Clustering effect

Comparing the above graph with the original clustering graph it is obvious that there are significantly more points in the clustering result 2. This is because after increasing the silica content of the samples in cluster 1, when the re-clustering is carried out, some points will be divided into cluster 2, which proves that the classification result is more sensitive to the change of variables. At the same time, during the adjustment of silica content, we found that the clustering effect changed only when the adjustment range was above 10%, thus indicating that the classification result has a good anti-disturbance effect in the range of 10%.

The classification results are sensitive to changes in chemical composition, and the classification results are well resistant to perturbation in the 10% range.

3.3. Random Forest Classification Model

3.3.1 Modeling

Step1: The training set is extracted from the attached original dataset. In each round, n training samples are extracted from the original data using the Bootstraping method. A total of m rounds of extraction are performed to obtain m mutually independent training sets.

Step2: One decision tree model is obtained using one training set at a time. m training sets are used to obtain a total of m decision tree models. m decision tree models together form a random forest model.

Step3: For the classification problem of artifacts: the data to be predicted are input to the m decision tree model obtained in the previous step using voting to obtain the classification results.

3.3.2 Solving the model

For the importance ratio of each chemical component, 0.09 for SiO_2, 0.00 for Na_2O, 0.10 for K_2O, 0.05 for CaO, 0.01 for MgO, 0.03 for Al_2O_3, 0.02 for Fe_2O_3, 0.02 for CuO, 0.3 for PbO, 0.23 for BaO, 0.02 for P_2O_5, 0.13 for SrO, and 0.00 for SnO_2.

The data were predicted using the established random forest model and the results were collated as shown in Table 4 below.

Table 4. Prediction results of random forest model

Predicted results	Predicted outcome probability (lead barium)	Predicted outcome probability (high potassium)	Artifact Number	Surface weathering
High Potassium	0.01	0.99	A1	No weathering
Lead Barium	0.56	0.44	A3	No weathering
Lead Barium	0.82	0.18	A4	No weathering
Lead Barium	0.91	0.09	A8	No weathering
Lead Barium	0.86	0.14	A2	Weathering
Lead Barium	0.00	1.00	A5	Weathering
High Potassium	0.01	0.99	A6	Weathering
High Potassium	1.00	0.00	A7	Weathering

3.3.3 Sensitivity Analysis

For the prediction model, a total of eight groups of artifact data were obtained for the test, which were classified into high potassium glass and lead-barium glass artifacts. In this paper, A1 was chosen as an example to do a sensitivity analysis study of the classification results. The classification results after the mobilization of artifact number A1 for silica, lead oxide, and barium oxide are shown in Table 5 below.

Table 5. Classification results after mobilization

Artifact Number	Predicted results	Silicon Dioxide	Lead oxide	Barium oxide
A1	High Potassium	58.45	0	0
A1	High Potassium	38.45	0	0
A1	Lead Barium	38.45	20	20

Analysis of the above table shows that when the silica content is reduced by 20%, the predicted results do not change A1 is still predicted to be a high potassium glass type; when lead oxide and barium oxide are adjusted to increase by about 20% respectively, the predicted category of A1 is converted from high potassium to lead-barium glass type. When small changes in chemical composition were made, the classification results did not differ because of these changes, which indicates that the classification results are more difficult to differentiate in the 20% adjustment range and have some resistance to interference.

The model can play a very good classification effect, for the higher dimensional data for the classification of the problem, the model can have a more accurate judgment ability, for some classification of the more ambiguous individuals, will not be affected.

3.3.4 Sensitivity Analysis

Classification results: A1, A6, A7 for high potassium glass, A2, A3, A4, A5, A8 for lead barium glass.

4. Conclusions

In this paper, we established a component division based on K-means clustering, analyzed its rationality, and conducted a sensitivity analysis using the OAT method, which yielded an analysis result with high sensitivity. The classification prediction model of ancient glass products based on random forest was established, and the sensitivity analysis of the final classification results detected that the established model has strong anti-disturbance ability, which has very important application value and practical significance in archaeological career.

Acknowledgements

First of all, I would like to express my sincere thanks and gratitude to my mentor, Mr. Wu Guohui, for his guidance throughout the process. He always made me feel confident in my abilities. I would also like to thank my research partners, Zelin Wang and Song Gao, who have been extremely supportive and helpful during the course of my research project. Finally, I sincerely thank all the people who helped and cared for me during the research process.

References

[1] Dussubieux Laure, Fenn Thomas R., Abraham Shinu Anna, Kanungo Alok Kumar. Tracking ancient glass production in India: elemental and isotopic analysis of raw materials[J]. Archaeological and Anthropological Sciences,2022,14(12).

[2] Yu D, Xu H, Chen C L P, et al. Dynamic coverage control based on k-means[J]. IEEE Transactions on Industrial Electronics, 2021, 69(5): 5333-5341.

[3] Li T, Kou G, Peng Y, et al. An integrated cluster detection, optimization, and interpretation approach for financial data[J]. IEEE Transactions on Cybernetics, 2021.

[4] Gu Liangliang, Zhou Jing. Some problems in the study of ancient Chinese glass art [J]. Art and Design, 2021, 2(04): 112-114.

[5] Wang Sen,Liu Chen,Xing Shuaijie. A review of K-means clustering algorithm research[J]. Journal of East China Jiaotong University,2022,39(05):119-126.

[6] Wu, Wei-Jie. Application and optimization of random forest algorithm[D]. Jiangnan University, 2021.

[7] Zhang Yajin,Hu Dianshun,Yao Benwu. Cardinality test for the intellectual understanding of probability and statistics[J]. Mathematics Letters,2021, (14):1-4+28.

[8] Lu Yunqing. Issues to note when using Pearson's chi-square statistic for statistical testing[J]. Statistics and Decision Making, 2009 (15): 32-33.

[9] Chen Xinguang,Zou Yueyue,Zhu Yingyue,Ma Zhipeng,Zhao Rui,Jia Yong. A longitudinal protection method for DC distribution network using Spearman correlation coefficient with current integration ratio[J]. Journal of Chongqing University of Technology (Natural Sciences),2022,36(03):172-180.

[10] Saroj, Kavita. review: a study on simple k mean and modified K means clustering technique [J]. International Journal of Computer Science Engineering and Technology, 2016, 6(7): 279-281.