The Classification and Identification of Ancient Glass Artifacts Based on Decision Tree and Cluster Analysis

Jia Wang [#, *], Yiwei Wei [#], Yuqian Liu [#]

School of Mathematics and Statistics, Zhengzhou University, Zhengzhou, China, 450001

* Corresponding Author Email: wangjiamath123@163.com

[#]These authors contributed equally.

Abstract. The analysis of ancient glass compositions is crucial for accurately identifying artifact categories. This study establishes a classification standard for high-potassium glass and lead-barium glass by constructing a decision tree model, with PbO content less than 5.46% being classified as high-potassium glass, and the opposite as lead-barium glass. Furthermore, cluster analysis is used to further subdivide high-potassium glass into high-silicon and low-silicon categories, and lead-barium glass into high-copper and low-copper categories. This method not only helps to accurately identify unknown types of glass, improving the precision of manual classification, but also provides a reliable scientific basis for historical research.

Keywords: Decision Tree, K-means++ Cluster Analysis, Ancient Glass, Archaeological Artifact Identification.

1. Introduction

The authentication of cultural relics is crucial for preserving cultural heritage, maintaining historical authenticity, and advancing academic research. Scholars have conducted extensive research on the chemical composition analysis of cultural relics and the classification of artifacts. With the development of society and the advancement of science and technology, the authentication of cultural relics urgently requires more rigorous and accurate research methods.

In recent years, Wang et al. proposed a method for determining the provenance of stone cultural relics using handheld X-ray fluorescence spectroscopy [1]. This method has successfully identified the sources of stone artifacts. However, its analysis of complex chemical compositions is relatively simplistic. As a result, it is difficult to reveal the internal diversity of materials. Hu T conducted qualitative and quantitative analyses of protein substrates to study the sources and functions of materials in cultural heritage [2]. Although this research excelled in functional analysis, its applicability to non-biological materials, particularly glass artifacts, is limited. It does not provide comprehensive information on chemical compositions. Mastelaro performed X-ray absorption fine structure (XAFS) studies on oxide glasses [3]. Nevertheless, the high equipment requirements and operational complexity of XAFS restrict its application in actual artifact research. Guo et al. effectively classified glass artifacts by constructing a BP neural network for predictive modeling [4]. Nevertheless, the black-box nature of this model reduces the interpretability of the results. Zou Y established a multivariate time series model to predict the concentrations of various chemical components in glass artifacts prior to weathering [5]. Although this model faces challenges regarding accuracy when processing samples that have undergone weathering. Additionally, Wen Y employed statistical methods to examine the distribution patterns of chemical compositions across different types of glass [6]. Yet, traditional statistical approaches struggle to capture deeper relationships in complex data. Ai X conducted studies on compositional analysis and species identification of glass artifacts based on a multiple linear regression model [7]. Nonetheless, the strict assumptions of this model may lead to unreliable results. Zhang et al. utilized a gray relational analysis model to assess the correlations between chemical compositions across different categories of glass [8]. Yet, the effectiveness of this method in handling high-dimensional data has been questioned. Lastly, Shao et al. performed an in-depth study of glass composition using the weighted average method and

correlation analysis [9]. However, this approach still falls short in identifying subtle differences between samples.

Compared to previous research, this study employs a decision tree model to analyze the classification patterns of high-potassium glass and lead-barium glass. The decision tree model is known for its clear structure and interpretability. It makes the classification results more intuitive, facilitating the optimization of production processes and supporting scientific decision-making. Its interpretability enables researchers to easily understand the basis of the classifications. This provides reliable data support for historical research. The model can assist archaeologists in systematically classifying unearthed artifacts. By analyzing characteristics such as material composition, production techniques, and physical features, it reveals the historical and cultural information behind these artifacts. To address the shortcomings of traditional glass classification methods, we subsequently introduced the K-means++ clustering algorithm for subclass categorization. This method improves clustering outcomes through the intelligent selection of initial cluster centers. It allows for a more detailed differentiation between subclasses. This approach not only uncovers potential cultural exchange and technological transmission patterns but also provides scientific support for the preservation and restoration of artifacts. By combining these two methods, this study offers a more comprehensive and in-depth perspective for archaeological research. It contributes to the preservation and inheritance of cultural heritage. This innovative combination enhances the accuracy of classification and provides a more extensive data foundation for future research. The key research questions of this study have been clarified based on the above research methods and ideas, as follows:

RQ1: What is the classification basis for high-potassium glass and lead-barium glass according to data analysis?

RQ2: What specific subclasses can each category be further divided into?

RQ3: What significant differences exist in the chemical composition among the various categories?

Based on the RQs, the main contributions of this paper are:

1. The classification rules for high-potassium glass and lead-barium glass derived from the decision tree model provide clarity and interpretability, facilitating the optimization of production processes and supporting scientific decision-making.

2. A comprehensive evaluation of the classification model was conducted using confusion matrices, ROC curves, and other metrics, revealing that the model demonstrates exceptional accuracy and reliability in classification tasks, thereby substantiating its effectiveness and superiority in practical applications.

3. Further analysis of intra-subclass variations through Kmeans++ clustering enables the identification of underlying structures and subtle distinctions within each subclass, facilitating a more precise recognition of internal heterogeneity. This approach provides a robust foundation for optimization and tailored interventions, thereby enhancing both the accuracy and interpretability of the model in practical applications.

2. Methodology

2.1. Data Preprocessing

(1) Description of the Dataset

Our data is from the publicly available dataset of the National College Student Mathematics Competition, found at https://www.mcm.edu.cn/. It includes basic information and chemical composition ratios for both classified and unclassified glass artifacts. This dataset supports the analysis and classification of glass artifacts, providing valuable insights for research and applications in materials science.

(2) Outlier Processing

Data with a sum of component ratios between 85% and 105% are considered valid. Summing the component ratios of 69 samples revealed that samples 15 and 17 fall outside this range. Therefore, samples 15 and 17 are considered invalid and removed.

(3) Missing Value ProcessingIn Table

To address missing surface colors in artifacts 19, 40, 48, and 58, we match each with artifacts of similar chemical composition and use their colors as substitutes. For example, artifact 19's closest match (based on Euclidean distance) is the weathered sample from artifact 49, so we assign it the color "black."

(4) Centered Log-Ratio (CLR) Transformation

The centered log-ratio transformation is a common data transformation method, especially for compositional data. It transforms compositional data into relatively independent real-valued vectors by converting data through logarithmic and centralization operations, ensuring the transformed data sum to zero. This transformation eliminates the proportional dependency between components, highlighting relative changes between variables. The process of CLR transformation is as follows:

Step 1: Input data $A=(a_{ij})_{m\times n}$

Step 2: Calculate the logarithmic mean value:

$$\bar{a}_i=\frac{\sum_{j=1}^{n}\log(a_{ij})}{n}(i=1,2\ldots m) \tag{1}$$

Step 3: Calculate the geometric mean:

$$a_i=e^{\bar{a}_i}(i=1,2\ldots m) \tag{2}$$

Step 4: Take the natural logarithm of each component:

$$a_{ij}'=\log(a_{ij}) \tag{3}$$

Step 5: Subtract the logarithm of the geometric mean from each component:

$$\tilde{a}_{ij}=a_{ij}'-\log(a_i) \tag{4}$$

After the CLR transformation, the compositional data are converted into log-ratio values.

2.2. Classification Patterns of Glass Based on Decision Tree

The decision tree model is a widely used machine learning and statistical tool [10]. It systematically divides data into different branches through a series of conditional judgments, ultimately forming a tree-like structure where each leaf node represents the final prediction result. Analyzing the weathering classification of high-potassium glass and lead-barium glass using the decision tree model offers numerous advantages. This approach is not only intuitive and easy to interpret but also effectively identifies key chemical components and addresses complex nonlinear relationships.

2.3. Subclassification of Various Types of Glass

(1) Introduction to the K-means++ Clustering Algorithm

K-means++ improves the K-means clustering algorithm by selecting dispersed initial cluster centers, enhancing both convergence speed and clustering quality.

(2) Elbow Method for Determining the Optimal Number of Clusters.

The Elbow Method is a technique used to determine the optimal number of clusters in the K-means clustering algorithm. It analyzes the clustering performance at different values of K to identify a balance point, thereby selecting a suitable number of clusters. This paper presents two-line graphs of the glass clustering coefficients based on the known data, as shown in the figure 1 below:

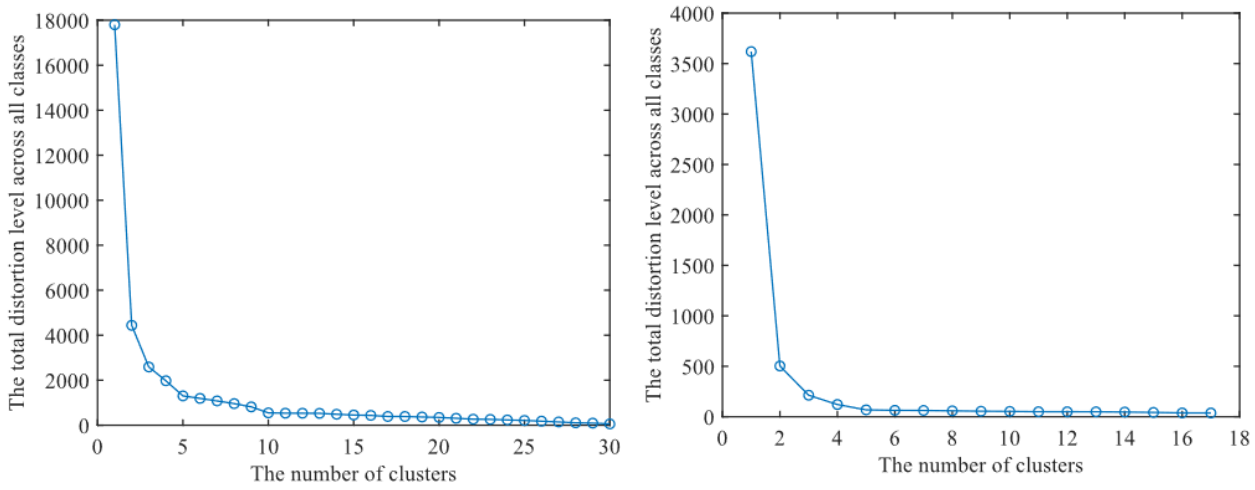

Figure 1. Polymerization Coefficient of Lead-Barium and High-Potassium Glass

The clustering coefficient line graphs show that for both lead-barium and high-potassium glass, the decrease in trend slows when the number of clusters is 2. Therefore, 2 clusters are chosen for each type.

(3) K-means++ Clustering Algorithm

Step 1: Randomly select a sample point from the dataset X as the first initial cluster center c_i;

Step 2: Select the remaining cluster centers:

Calculate the distance between all sample points and the selected cluster centers, and identify the shortest distance, denoted as di. Then, calculate the probability of each sample point being selected as the next cluster center based on this distance. Finally, select the sample point corresponding to the maximum probability value as the next cluster center:

$$P(\mathrm{x}) = \frac{d_i(x)^2}{\sum_{x \in X} d_i(x)^2} \tag{5}$$

Step 3: Repeat the above process until k cluster centers are determined.

Step 4: Calculate the Euclidean distance between each sample and each cluster center D_{ij}:

$$D_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ij} - K_{jk})^2} \tag{6}$$

Let the number of features be m and the number of samples be n. By comparing the distances, assign each sample to the cluster of the cluster center with the minimum distance.

Step 5: After one iteration, update the cluster centers, and the new cluster centers become $K_j^{'}$:

$$K_j^{'} = \frac{1}{n^{'}} \sum_{x_i \in K_j} x_i \tag{7}$$

Step 6: Repeat Step 4 and Step 5 until the cluster centers no longer change.

3. Result analysis and discussion

3.1. Results of the Decision Tree

To improve the generalization of the classification model, this study used 10-fold cross-validation and a holdout method, designating 61 samples as the training set and 6 as the test set. This approach yielded the classification criteria for high-potassium and lead-barium glass, with specific results shown in the figure 2 below:

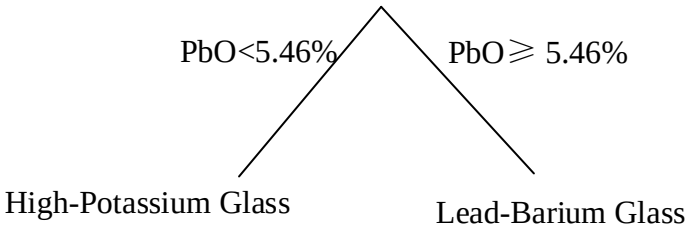

Figure 2. Classification Criteria for Glass Types

The results indicate that when the lead oxide content is below 5.46%, the glass sample is classified as high-potassium glass; when the lead oxide content is above 5.46%, the glass sample is classified as lead-barium glass.

This paper selected the following metrics to evaluate the classification performance:

(1) Confusion Matrix

A confusion matrix is a tool for evaluating classification model performance, presenting a comparison of predicted versus actual results in a matrix format. In addition to showing overall accuracy, the confusion matrix helps identify categories where the model performs well or poorly.

This paper utilizes a confusion matrix to visualize the prediction results, as shown in the figure 3 below:

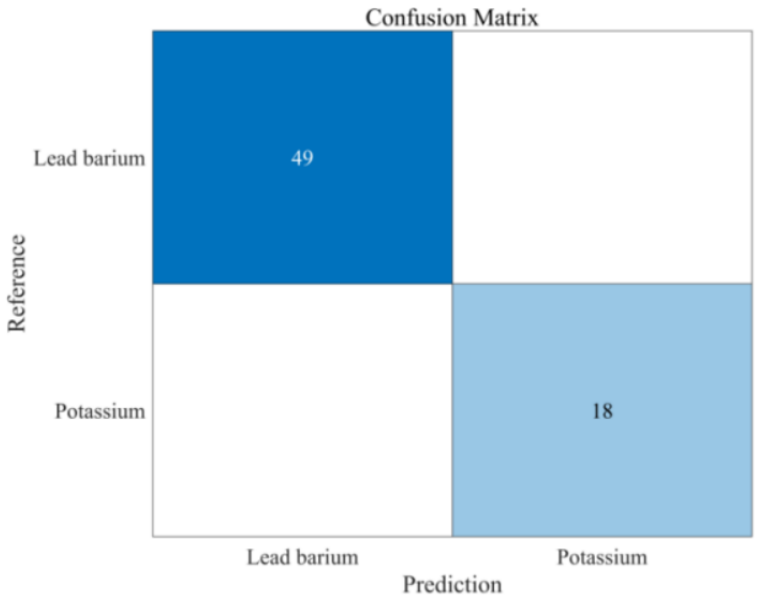

Figure 3. Confusion Matrix of Classification Results

(2) Values of metrics

Values of metrics is shown in Table 1 below:

Table 1. Values of Evaluation Metrics

Accuracy	*R*	*P*
1	1	1

The high accuracy of this model reflects strong overall performance in effectively handling the classification tasks of the current dataset. The high recall indicates the model's capability to correctly identify most positive samples, while the high precision signifies accuracy in predicting positive classes. In summary, this model demonstrates good classification efficiency.

(3) F_1 score

The F_1 score is the harmonic mean of recall and precision and is also a metric used to assess the performance of a classification model.

The calculation yields an F_1 score of 1, indicating that the classification performance is excellent.

(4) ROC Curve

The ROC curve can be used to compare the performance of multiple models, and the closer the curve is to the top left corner, the higher the true positive rate and the lower the false positive rate at various thresholds, indicating better model performance. The ROC curve for this classification model is shown in Figure 4 below:

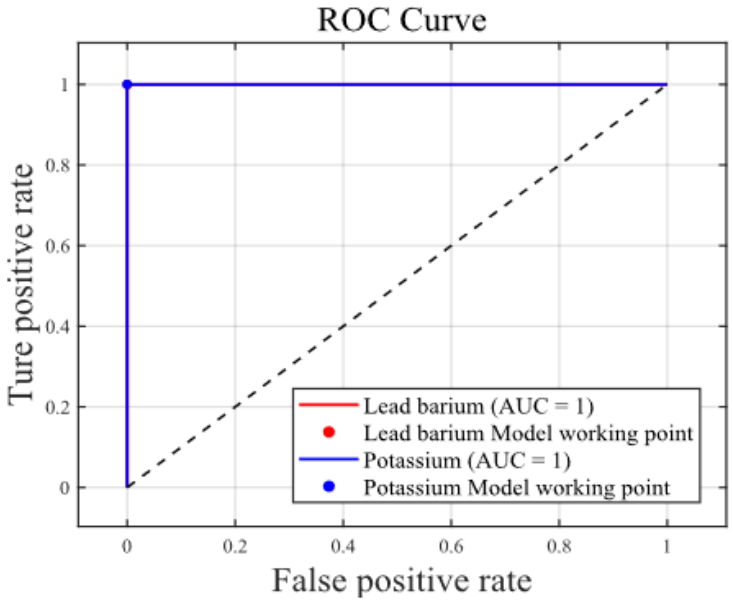

Figure 4. ROC Curve

The area under the ROC curve is 1, indicating that the classification performance of the model is good.

3.2. Results of Kmeans++ Clustering Algorithm

(1) Classification Results

The classification results of high-potassium artifact samples are shown in Figure 5 below:

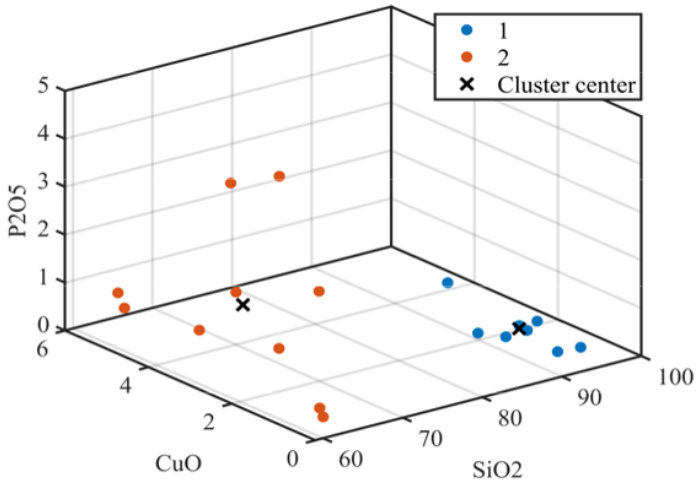

Figure 5. 3D Scatter Plot of High-Potassium Glass

From the graph, it can be observed that Category 1 is mainly concentrated in the region with lower silicon dioxide content, while Category 2 is primarily concentrated in the area with higher silicon dioxide content.

The 3D scatter plot of lead-barium is shown in Figure 6 below:

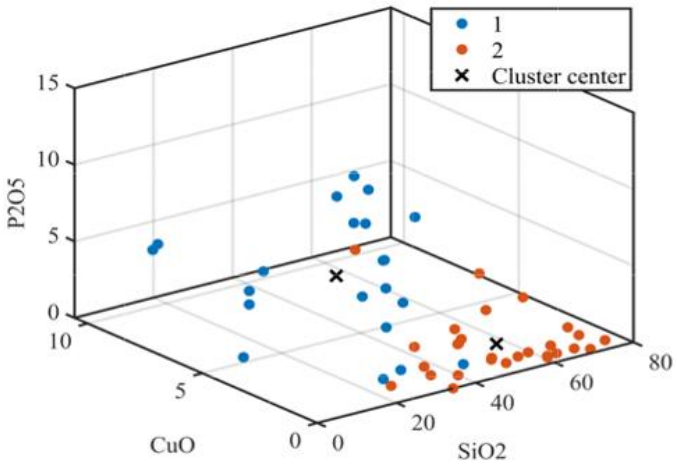

Figure 6. 3D Scatter Plot of Lead-Barium Glass

Analysis of the above figure shows that the data points in Category 1 are mainly concentrated in the region with higher copper oxide content, and Category 1 is relatively dispersed. In contrast, the data points in Category 2 are primarily concentrated in the region with lower copper oxide content, and the data are more clustered.

(2) Analysis of the Validity of Classification Results

The significant differences in key variables among different clusters indicate that the clustering results are valid for these variables. Therefore, in this study, the validity of the clustering results can be analyzed through the significant differences in chemical composition content across different classifications. This paper evaluates the classification results using lead-barium glass as an example. The evaluation steps are as follows:

Step 1: Shapiro-Wilk Test

The Shapiro-Wilk test is a statistical method used to determine whether sample data come from a normal distribution. It is particularly suitable for testing the normality of small sample sizes. The Shapiro-Wilk test was conducted to assess the normality of the various categories of lead-barium glass, and the results are presented in the following table 2:

Table 2. Normality Test Results for Different Categories of Lead-Barium Glass

Variable Name	Sample Size	S-W Test
SiO2	18	0.852(0.009***)
CuO	18	0.949(0.408)
P2O5	18	0.692(0.000***)

The p-values for both silicon dioxide and phosphorus pentoxide are less than 0.05, indicating significance. Therefore, we can reject the null hypothesis, suggesting that the data do not follow a normal distribution. Consequently, only non-parametric methods can be used to verify their differences.

Step 2: Kruskal-Wallis Test

The Kruskal-Wallis test is a non-parametric statistical method used to compare the distributions of two or more independent groups for significant differences. The basic idea is to rank all sample values and then assess whether there are significant differences between groups by comparing the ranks. Using lead-barium glass as an example, this test can be employed to verify the significant differences in chemical composition content among different categories, thereby demonstrating the validity of the model. The test results are shown in Table 3:

Table 3. Results of the Kruskal-Wallis Test Analysis

Analysis Item	Grouping Variable	Sample Size	Statistic	p-value	*f*-value
SiO2	1	29	34.802	0.000***	0.188
	2	20			
	Total	49			
CuO	1	29	5.524	0.019**	0.064
	2	20			
	Total	49			
P2O5	1	29	12.295	0.000***	0.108
	2	20			
	Total	49			

The p-values for each analysis item are all less than 0.05. This further validates the reasonableness and effectiveness of the clustering method in revealing the inherent structure of the data.

4. Conclusion

By establishing a decision tree model and conducting clustering analysis on ancient glass compositions, this research enhances classification accuracy and reveals internal variations among similar glass types. This method facilitates accurate artifact identification, leading to more effective preservation strategies while providing reliable support for understanding the evolution of ancient culture, trade, and technology. Moreover, it can be applied to the analysis of other artifacts and materials, offering new tools for archaeological and historical studies.

Ensemble machine learning methods offer significant potential for artifact composition identification by combining various algorithms to enhance accuracy and robustness. They excel in processing high-dimensional data and effectively capture subtle differences in compositions. This improves the reliability of artifact identification and optimizes feature extraction. Investigating their application can strengthen the scientific rigor of artifact analysis.

References

[1] Wang Z, Zhang Z, Wang F, et al. A pXRF-based approach to identifying the material source of stone cultural relics: a case study [J]. Minerals, 2022, 12 (2): 199.

[2] Wu Q, Zhang B, Hu Y. Comparison and Research Progress of Protein Detection Technology for Cultural Relic Materials [J]. Coatings, 2023, 13 (8): 1319.

[3] Mastelaro V R, Zanotto E D. X-ray absorption fine structure (XAFS) studies of oxide glasses—a 45-year overview [J]. Materials, 2018, 11 (2): 204.

[4] Guo K, Qiao Y, Gao Z. Based on BP neural network glass cultural relics chemical category and composition prediction model construction [J]. Highlights in Science, Engineering and Technology, 2023, 42: 111 - 117.

[5] Zou Y. Molecular-composition analysis of glass chemical composition based on time-series and clustering methods [J]. Molecules, 2023, 28 (2): 853.

[6] Wen Y. Prediction of chemical composition of cultural relics glass based on moving average algorithm of least square method [J]. Highlights in Science, Engineering and Technology, 2023, 58: 179 - 187.

[7] Ai X. Study on Composition Analysis and Species Identification of Glass Relics Based on the Multiple Linear Regression Model [J]. Advances in Computer, Signals and Systems, 2023, 7 (4): 55 - 63.

[8] Chang S, Yang Y, Xu Y H, et al. Composition analysis and identification of ancient glass products based on gray correlation [J]. Highlights in Science, Engineering and Technology, 2023, 42: 188 - 196.

[9] Shao K, Du R, Xiong H. Composition Analysis of Glass Based on Weighted Mean and Correlation Analysis [J]. Highlights in Science, Engineering and Technology, 2023, 42: 55 - 62.

[10] Li X, Yi S, Cundy A B, et al. Sustainable decision-making for contaminated site risk management: A decision tree model using machine learning algorithms [J]. Journal of Cleaner Production, 2022, 371: 133612.