Composition Analysis and Identification Scheme of Ancient Glass Products Based on K-Means Algorithm

Chunhui Wang[1, #, *], Jiajia Song[1, #], Mingrui Yin[2, #]

[1] School of Environment, Harbin Institute of Technology, Harbin, China, 150006

[2] School of Chemistry and Chemical Engineering, Harbin Institute of Technology, Harbin, China, 150006

* Corresponding Author Email: 120L040324@stu.hit.edu.cn

[#]These authors contributed equally

Abstract. Chinese ancient glass mainly includes high potassium glass and lead barium glass. In the process of weathering, its chemical composition is easy to change, affecting the judgment of its category. The data of this study are based on the chemical composition ratio of classified glass relics obtained by preprocessing. Firstly, a random forest model is constructed based on the Gini coefficient to obtain the contribution of lead oxide to the classification. The decision tree in the random forest is extracted, and the optimal classification rule is obtained by using the majority voting mechanism : when ω (PbO) is 10 %, it is lead barium glass. Secondly, the K-means clustering algorithm is divided into three sub-categories. The characteristic quantity with large standard deviation is used as the basis for classification, and the classification results are obtained : high potassium glass sub-categories include low calcium type, medium calcium aluminum type and high calcium aluminum type ; lead barium glass subclass includes low barium type, medium lead barium type, high lead barium type. For the identification of unknown types, the Euclidean distance between the unknown category and the sub-category centroid is calculated, and the minimum distance is taken to determine the sub-category it belongs to.

Keywords: Ancient Glass, Random Forest, K-means Clustering, Sensitivity Analysis.

1. Introduction

The Silk Road was a major artery linking Eastern and Western cultures in ancient times, and ancient glassworks and their production techniques were also introduced to China along the Silk Road and flourished [1]. The two main types of ancient glass in China include high potassium glass and lead barium glass. The analysis of the chemical composition of glass is an important method for the identification of ancient glass types [2]. As ancient glass is susceptible to weathering during storage due to the burial environment, its internal elements may exchange with environmental elements in large quantities, resulting in changes in its compositional content and affecting the accuracy of the identification results [3-7].

This study intends to address the following questions.

(1) Determine the classification rules for two types of glass based on the chemical composition ratios of classified glass artefacts (www.mcm.edu.cn).

(2) For different categories of glass, classify their subcategories based on their chemical composition and give specific classification methods and classification results.

(3) Combine the classification laws to establish a classification model to identify the corresponding types of glass artefacts by analyzing the chemical composition of unknown categories of glass artefacts (www.mcm.edu.cn).

2. Model building and solving

2.1. Classification rule based on random forest mode

Firstly, data processing is carried out to replace the information recorded in text in the original data with the corresponding numerical codes to facilitate the data mining and analysis work at a later stage. The specific replacement content is as follows:

(1) In “surface weathering”, assign “weathered” and “unweathered” as “1” and “0”.

(2) In “Type”, assign the values of “Lead-Barium” and “High Potassium” as “1” and “2”.

The next step is to build a random forest model. Random forest [8] is an integrated learning algorithm that classifies data by combining multiple decision trees, which has a wide range of application and is not easy to overfit, and has a wide range of applications.

The model construction process is as follows:

Step 1: Import the data into Python and check if there are any outliers

Step 2: Remove labels from the features and keep the feature names for subsequent use

Step 3: Use Skicit-learn to slice the data set

Step 4: divide the data into training set and test set and pass in parameters: features, labels, scales, random seeds

Step 5: Instantiate the model and train the data model

Step 6: Get the contribution of each feature and plot it

Step 7: Calculate the average absolute error and accuracy

Output: Histogram of feature contribution

The solution results are shown in Figure 1.

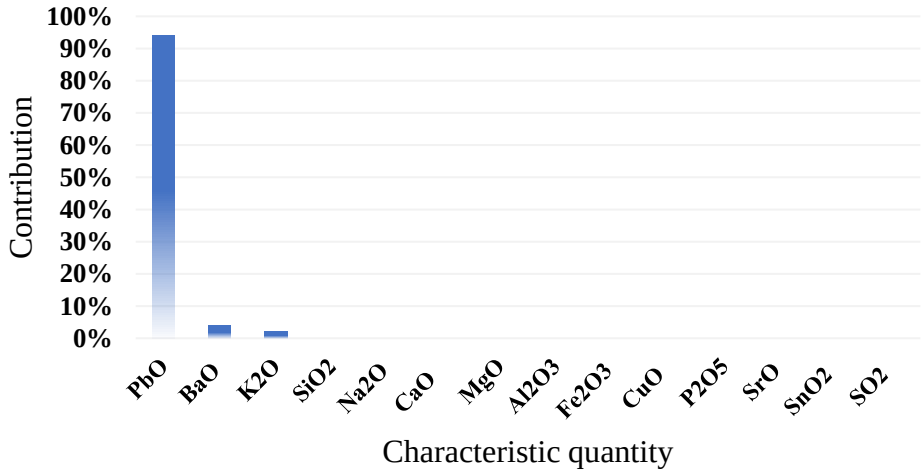


Figure 1. Feature volume contribution histogram

Based on the above solved results, the analysis can be concluded as follows: the contribution of PbO content to the classification reaches 0.94 and the contribution of BaO to the classification reaches 0.04; these two most important feature values are extracted, and the model is retrained to obtain the contribution of PbO to the classification reaches 100%, after that and the average absolute error and accuracy of the prediction results are calculated, and the results are obtained as 0.0 and 99.75%, from which it can be seen that the eigenvalue PbO has a high feasibility for category identification.

The decision trees in the random forest are extracted, and the most representative decision tree is shown in Figure 2 by taking the majority voting method

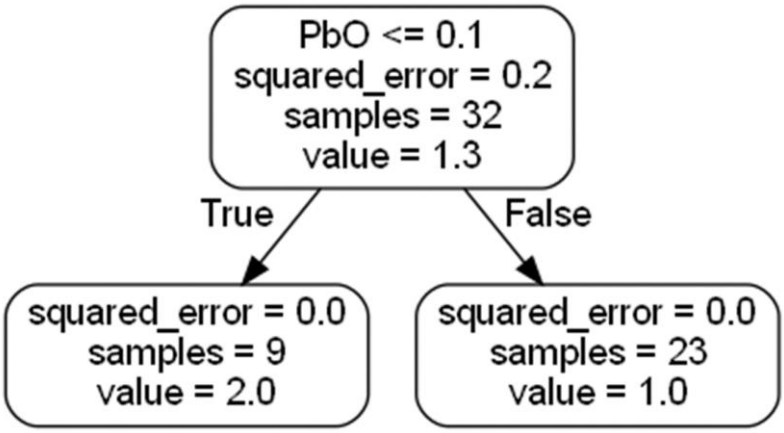


Figure 2. Decision trees in random forests

2.2. Subclassification model based on K-means algorithm

Based on the proportion of chemical composition of the classified glass artifacts obtained from the preprocessing, the K-means clustering algorithm was used to cluster the sample data and classify them into subclasses based on different chemical compositions.

The K-means clustering algorithm [9] is a simple selective generational clustering algorithm that clusters an n-dimensional vector data point set D, and finally divides the set D into K class clusters. The basis of clustering is mainly "tightness" or "similarity", the more similar the objects within a group and the larger the gap between groups, the better. The Euclidean distance is used as the similarity measure, and the sum of squared errors (SSE) is used as the objective function to measure the quality of clustering, and the data points are divided into K clusters according to the distance from the center of clustering by minimizing the objective function.

Definition: Euclidean distance between data points.

$$E(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}\ , \tag{1}$$

Definition: Euclidean distance formula between data points and clustering center c_i

$$E(x,c_i)=\sqrt{\sum_{j=1}^{n}(x_j-c_{ij})^2}\ , \tag{2}$$

Where x is the data object; c_i is the i-th cluster center; n is the dimension of the data object; x_j, c_{ij} is the j-th attribute values of x and c_i.

Definition: error squared and SSE calculation formula

$$SSE=\sum_{i=1}^{k}\sum_{x\in c_i}\left|E(x,c_i)^2\right|, \tag{3}$$

Where, the size of SSE indicates the good or bad clustering result; K is the number of clusters.

The solution process is shown below:

K-means clustering algorithm solving process [10].

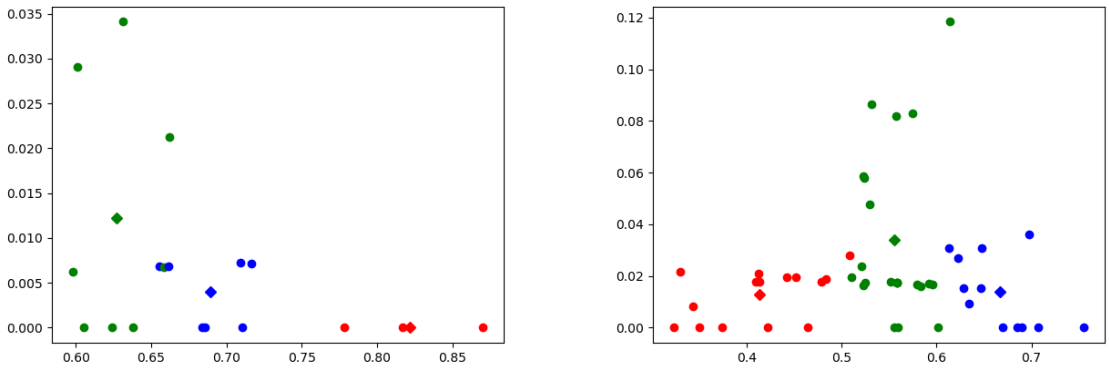
Step 1: Import the data into Python and obtain the number of samples and eigenvalues

Step 2: Construct a set of k random centers of mass for the given data set

Step 3: Iterate through all the centers of mass for each point and calculate the distance from the point to each center of mass, if the distance is smaller than the minimum distance and the error sum of squares SSE meets the requirement, then update the index of the minimum distance and the minimum center of mass.

Step 4: returns the value of the minimum number of centers of mass and the coordinate points of

each center of mass.
Output: Graphical presentation of clustering results in two random dimensions.
The solution result is shown in Figure 3.



(a) (b)

Figure 3. Clustering results of high potassium glass subclass (a) and lead-barium glass

Define the standard deviation calculation formula as follows.

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}, \tag{4}$$

The standard deviations of the corresponding characteristic quantities for different types of glass were obtained by calculation as shown in Figures 4 and 5.

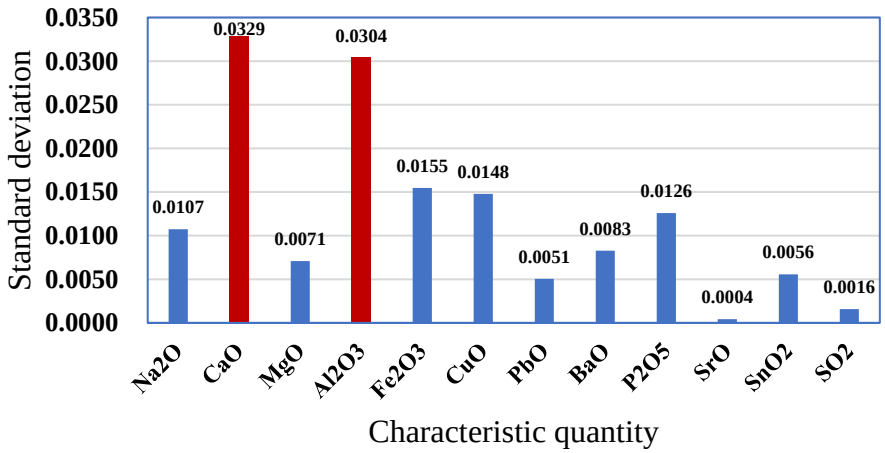


Figure 4. Standard deviation of characteristic quantities of high potassium glass

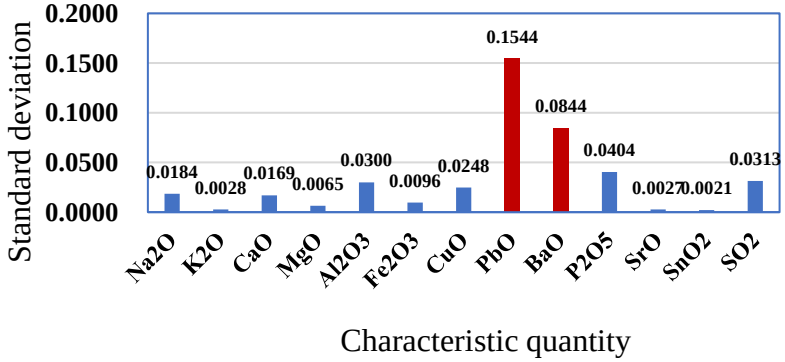


Figure 5. Standard deviation of characteristic quantities of lead-barium glass

The standard deviations of different components of different types of glass were calculated separately, and the standard deviations of alumina and calcium oxide were the largest in high potassium glass, which were 0.0304 and 0.0329, respectively, and had a significant influence on the classification of subclasses, so these two chemical components were selected as the basis for the classification of high potassium glass; in lead-barium glass, the standard deviations of lead oxide and barium oxide were the largest, which were 0.1544 and 0.0844, respectively, and had a significant influence on the classification of subclasses. The standard deviations of lead oxide and barium oxide were 0.1544 and 0.0844, which had significant effects on the classification of subclasses, so these two chemical compositions were selected as the basis for the classification of subclasses of lead-barium glass.

From the analysis of the results, it can be seen that

For high potassium glass, the obtained K value is 3 by K-means clustering algorithm analysis, that is, there are three subclasses of high potassium glass, which are low calcium type, medium calcium-aluminum type and high calcium-aluminum type, and the specific division results are shown in Table 1 below.

Table. 1. Subclasses of high potassium glass

Artefact number	**SiO_2**	**K_2O**	**CaO**	**Al_2O_3**	**…**
Low calcium type	84.38%	7.44%	1.01%	3.60%	…
Medium Calcium Aluminum type	71.41%	9.08%	3.99%	5.97%	…
High calcium aluminum type	63.11%	11.53%	7.29%	7.43%	…

For lead barium glass, the analysis was carried out by K-means clustering algorithm, and the obtained K value is 3, that is, there are 3 subclasses of lead barium glass, which contain low barium type, medium lead barium type and high lead barium type respectively. The specific division results are shown in Table 2 below.

Table. 2. Subclasses of lead barium glass

Artefact number	**SiO_2**	**PbO**	**BaO**	**MgO**	**…**
Low lead barium type	64.65%	16.82%	6.19%	0.60%	…
Medium lead barium type	48.34%	25.00%	12.54%	0.88%	…
High lead barium type	27.85%	32.58%	22.49%	0.78%	…

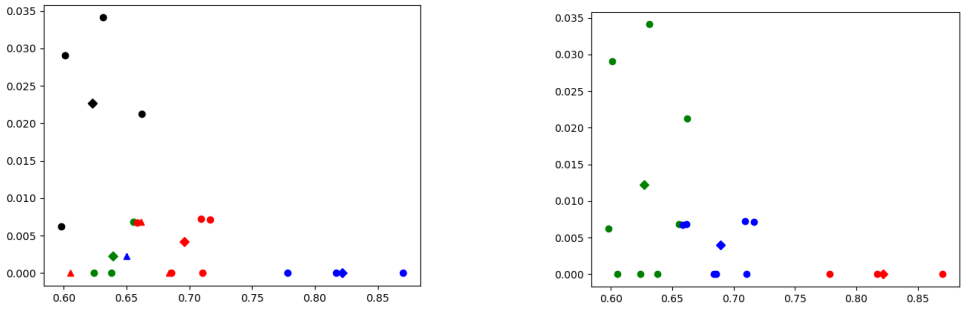
Next, the rationality and sensitivity of the classification results are analyzed.

Rationality analysis:

By comparing the analysis using the K-means clustering algorithm and the division methods in the literature, it can be seen that the number of subclasses we obtained for the division of high potassium glass and lead-barium glass, and the division method based on chemical composition are basically consistent with the existing classification methods and have good rationality.

Sensitivity analysis:

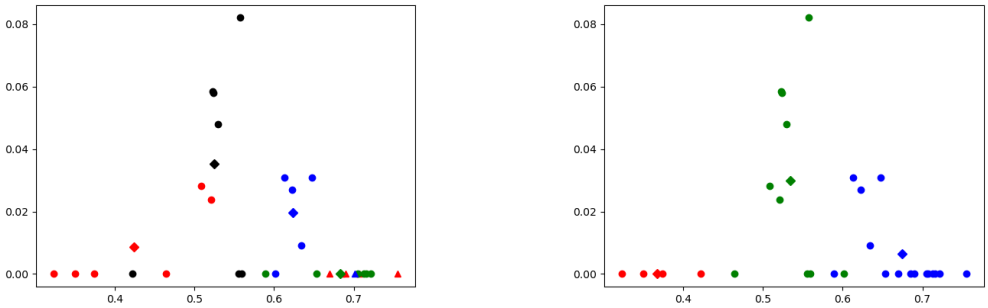
Sensitivity analysis is the study of how uncertainty in the output of a mathematical model or system (numerical or otherwise) is assigned to different sources of uncertainty in its input. In order to analyze the sensitivity of the classification results, we used a dimensionality reduction method to eliminate the alumina and calcium oxide components of the high potassium glass, respectively, and randomly eliminate the other chemical components, after which we used the K-means clustering algorithm to obtain subclassification results, which were compared with the original results to obtain the results shown in Figure 6.



(a) (b)

Figure 6. Removal of alumina and calcium oxide (a) and removal of other components (b)

Using the same method, the subclassification results of lead-barium glass after dimensionality reduction were obtained and compared with the original results, and the results shown in Figure 7 were obtained.



(a) (b)

Figure 7. Removal of lead oxide, barium oxide (a) and removal of other components (b)

Analyzing the above results, it can be seen that if the stability of the classification model is influenced by eliminating important feature quantities, the model has a higher sensitivity to important feature variables, which is manifested by the change of the optimal number of centers k and a poorer classification superiority; if the stability of the classification model is less influenced by eliminating non-important feature quantities, the model has a lower sensitivity to important feature variables and the classification results are similar to the original results.

The above analysis shows that the model has better stability with the retention of important feature quantities.

2.3. Classification model based on decision tree and Euclidean distance

Firstly, the chemical composition of the unknown category of glass artifacts was normalized, and then the type of glass artifacts belonged to the unknown category was identified according to the classification law of high potassium glass and lead-barium glass in question 2: when the content of PbO is less than 10%, the type of glass artifacts can be judged as high potassium glass; when the content of PbO is greater than 10%, the type of glass artifacts can be judged as lead-barium glass. The results of the classification of glass categories shown in the table below were obtained.

After that, the Euclidean distance between the chemical composition of the unknown category of glass artifacts and the chemical composition corresponding to the center of mass was calculated by combining the methods of subclassification of glass artifacts in Problem 2, and the results of subclassification of glass were obtained as shown in Table 3.

Table. 3. Type classification of unknown categories of glass artifacts

Artefact number	Type	Subclass
A1	High potassium content	Low calcium type
A2	Lead and barium	High lead barium type
A3	Lead and barium	High lead barium type
A4	Lead and barium	High lead barium type
A5	Lead and barium	Medium lead barium type
A6	High potassium content	Low calcium type
A7	High potassium content	Low calcium type
A8	Lead and barium	High lead barium type

Sensitivity analysis:

In order to analyze the sensitivity of the classification results, the data were first processed for the chemical composition content of the unknown category of glass artifacts as follows.

(1) the samples to be tested with predicted results of low/high X type were replaced with the lowest/highest X content values from Form II.

(2) For the samples to be tested with medium X-content, the average value of X-content in Form II is used for replacement.

The samples after replacement of chemical composition X were subclassified and compared with the subclasses obtained from the above table according to the method of predicting the subclasses of glass artifacts of unknown category in question three.

The subclasses of the samples after replacement of chemical composition X are shown in Table 4.

Table. 4. Classification of the types of glass artifacts after changing the composition content

Artefact number	Type	Subclass	Subclass after substitution of chemical component X
A1	High potassium content	Low calcium type	Low calcium type
A2	Lead and barium	High lead barium type	High lead barium type
A3	Lead and barium	High lead barium type	High lead barium type
A4	Lead and barium	High lead barium type	High lead barium type
A5	Lead and barium	Medium lead barium type	Medium lead barium type
A6	High potassium content	Low calcium type	Low calcium type
A7	High potassium content	Low calcium type	Low calcium type
A8	Lead and barium	High lead barium type	High lead barium type

Sensitivity analysis of the classification results is performed using the Sensitivity metric of the classifier.

Definition:

True positives (TP): the number of instances that are correctly classified as positive cases, i.e., the number of instances that are actually positive and classified as such by the classifier.

Definition:

False negatives (FN): number of instances that are incorrectly classified as negative, i.e., the number of instances that are actually positive but classified as negative by the classifier.

Definition:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

Sensitivity indicates the proportion of all positive instances classified correctly and measures the classifier's ability to recognize positive instances.

By analyzing the subclasses in the above table, it is obtained that TP=8 and FN=0. According to the definition formula, Sensitivity=1.

That is, all positive cases are correctly classified by the classifier, indicating that the classification results obtained by this classification method have strong stability within the sample.

3. Conclusions

This paper mainly solves three problems by building three models: random forest, K-Means clustering and Euclidean distance, respectively: classification of glass artefacts, subclass classification of glass artefacts and identification of types of glass artefacts by chemical composition. These three problem-solving methods have great application in medicine and biology, in addition to the identification of ancient artefacts, such as identifying the chemical components that make the greatest contribution to the classification of drugs, simplifying the difficulty of classification; subclassifying substances such as proteins to find the most specific subclasses precisely; and classifying the types of herbs by identifying the components of herbs.

References

[1] J. Henderson,J. An,H. Ma. The Archaeometry and Archaeology of Ancient Chinese Glass: a Review [J]. Archaeometry,2018,60(1).

[2] Zhou Xueqi,Lv Hongshu,Cui Jianfeng,Dong Xinlin,Wang Ying. Fluorite used in ancient Chinese glassmaking during the 10th to 12th centuries: Evidence from glass products excavated in the capital city site of the Liao dynasty [J]. Archaeometry,2022,64(5).

[3] H. C. BECK, C. G. SELIGMAN. Barium in Ancient Glass [J]. Nature,1934,133(3374).

[4] J. Q. Dong,Q. H. Li,S. Liu. The native development of ancient Chinese glassmaking: a case study on some early lead–barium–silicate glasses using a portable XRF spectrometer [J]. X - Ray Spectrometry,2015,44(6).

[5] LI Qinglin, XU Chengtai, LING Xue, YAO Zhengzheng. Nondestructive Analysis of Some Ancient Glass in Jin and Yuan Dynasty,by EDXRF Probe [J].Spectroscopy and Spectral Analysis,2011,31(07):1960-1963.)

[6] Siqin Bilige, Li Qinghui, Gan Fuxi. Analysis of Ancient Chinese Potash Glass by Laser Ablation Inductively Coupled Plasma-Atomic Emission Spectrometry/Mass Spectrometry [J]. Chinese Journal of Analytical Chemistry,2013,41(09):1328-1333.)

[7] WANG Jie,LI Mo,MA Qinglin,ZHANG Zhiguo,ZHANG Meifang,WANG Julin. Weathering of an Octagonal PbO-BaO-SiO2 Glass Stick from the Warring States Period [J]. Glass and Enamel,2014,42(02):6-13.)

[8] Wang Guijin. Research on Optimization and Improvement of Random Forests Algorithm and Its Parallelization [D]. Nanchang University,2019.)

[9] WANG Sen,LIU Chen,XING Shuaijie. Review on K-means Clustering Algorithm [J]. Journal of East China Jiaotong University,2022,39(05):119-126.)

[10] JIA Ruiyu,SONG Jianlin. K-means Optimal Clustering Number Determination Method Based on Clustering Center Optimization [J]. Microelectronics and Computer,2016,33(05):62-66+71.