A study on the composition analysis and identification of ancient glass products

Shubo Chang *, Zhipeng Gong, Qi Zhou

Hainan University, haikou, 570228

* Corresponding Author Email: 1002768960@qq.com

Abstract. The study of ancient glass, as an important branch of scientific and technological archaeology, is an important physical material for exploring the economic, technological and cultural exchanges between China and foreign countries on the Silk Road. It is very important for studying the development of ancient society and cultural exchanges between China and foreign countries. In this paper, we focus on the composition analysis and identification of ancient glass products, based on data mining in the annexes and with the help of software analysis, we give the weathering pattern of glass surface, predict the chemical composition content before weathering, as well as cluster analysis and correlation analysis of different types of glass and their chemical compositions. In this paper, a sample of glass types and the correlation of their chemical compositions is established for analysis. Based on this, a decision tree classification model was built to classify the artifacts in the sample and determine the magnitude of the chemical composition of the classification criteria. The Elbow method (elbow rule) is used to determine the optimal number of aggregated classes k, and K-Means clustering analysis is performed, and then the decision tree classification model is used to quasi-determine and classify the subclass classification criteria for the clustered results. Finally, the XGBoost model is trained to achieve the prediction of the given sample types.

Keywords: Decision Tree Classification Model; K-Means Cluster Analysis; XGboost Model.

1. Introduction

Glass, the ancient name of the wall glass, glaze, quite Li, Ming and Qing dynasties, also known as "material". The recipe of ancient glass has distinctive regional and contemporary nature. Different regions, different times of ancient glass often have different raw material formulations. The basic raw materials (SiO2), fluxes PbO, K2O, Na2O) and colorants (CuO, Fe2O3, MnO, CoO, Au) are taken from different origins [1]. Due to the raw material formula and the corresponding equipment and technology do not make the ancient glass there are differences in performance, texture, color, luster and other aspects. To study the formulation of ancient glass, it is necessary to observe the ancient glass objects, documentary evidence, and understand the conditions of its excavation or collection, as well as to take a comprehensive analysis of the assay formula dispel. Among them, the most effective method of laboratory analysis of ancient glass fragments [2]. Using this method, we have clarified that the ancient glass in China from the Western Zhou Dynasty to the Han Dynasty was lead-barium glass, which is different from the soda glass in the West, and proved that the evolutionary series of ancient Chinese glass is lead-barium glass, high lead glass, potassium-chalcogenide glass, soda-chalcogenide glass, potassium-lead glass, and lead-sodium-chalcogenide glass [3]. In this evolutionary series, there are ancient glasses with local characteristics, such as Chinese western glass, Guangzhou ancient glass, Quanzhou ancient glass, Fuzhou ancient glass and Suzhou ancient glass, all of which have different raw material formulations. Therefore, it is important to understand the formula of ancient glass to determine the origin, age, performance and process of ancient glass [4].

This paper analyzes the classification rules of high potassium glass and lead-barium glass based on the data; for each category, we select the appropriate chemical composition to classify them into subcategories, give the specific classification methods and classification results, and analyze the rationality and sensitivity of the classification results. For unknown categories of glass artifacts, the chemical composition is analyzed to identify the types to which they belong, and the sensitivity of the classification results is analyzed.

2. Model assumptions and notation

2.1. Assumptions [5]

1. It is assumed that the chemical composition content of the surface sampling points of the artifacts before weathering is approximately the same.
2. It is assumed that color and ornamentation do not influence the results in the classification discussion.
3. Assume that all artifacts are newly excavated
4. Assume that the type of glass artifacts is determined only by their chemical composition
5. Assume that the unweathered points on weathered artifacts are not directly related to each other and can be analyzed as two individuals.

2.2. Notations

Important notations used in this paper are listed in Table 1.

Table 1. Notations.

Symbols	Meaning
H_i	The i-th chemical composition of the glass artifact
Δ	infinitesimal amount relative to the same data set
y_i	is the predicted result of a chemical composition
$Gain(\Theta)$	Information gain of chemical composition Θ
p_j	Frequency of each chemical component to all chemical components
SSE	Sum of squared errors of aggregation
$\eta_i(j)$	The correlation coefficient of the jth chemical component of the ith artifact

3. Model construction and solving

3.1. Glass type classification law model

(1) For the establishment of the standard model of glass type classification

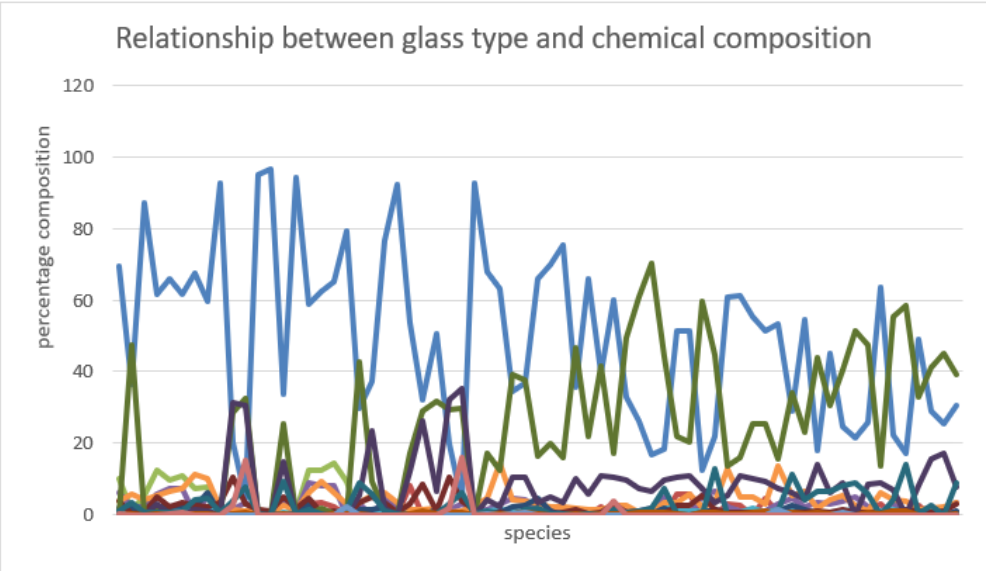

Figure 1. Glass type and chemical composition relationship distribution chart.

There are many chemical components contained in glass, and the relationship between the classification law of glass type and chemical composition cannot be obtained directly from the processed data, and the distribution of the relationship between glass type and chemical composition is shown in Figure 1.

So in order to analyze the classification laws of high potassium glass and lead-barium glass, it is first necessary to determine what chemical composition needs to be examined in the relationship between glass type and chemical composition, i.e., to establish a classification criterion.

A criterion called gain criterion is used to make a judgment, which is based on the concept of entropy in information theory [6]. So, in order to find this gain criterion, we first need to find the expected information required to classify a given sample, and let Y be the set of r "type-chemical composition" data samples after processing. Assuming that a given chemical composition has or is clustered into different values of m attributes, define m different classes $C_i(i=1\ldots m)$. Let r_i be the number of samples in class C_i, then the expected information is expressed as follows.

$$I(r_1, r_2, \ldots, r_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

The entropy or information expectation divided into subsets by Θ is given by the following equation

$$E(\Theta) = \sum_{j=1} \left(\frac{r_{1j} + r_{2j} + \cdots + r_{mj}}{r} \right) * I\left(r_{1j} + r_{2j} + \cdots + r_{mj}\right) \tag{2}$$

Then the information gain (gain criterion) of the chemical composition Θ

$$Gain(\Theta) = I(r_1, r_2, \ldots, r_3) - E(\theta) \tag{3}$$

Since the above expression for the selection criterion is too cumbersome, the gini indicator is used.

$$gini(\mathrm{r}) = 1 - \sum p_j * p_j \tag{4}$$

(2) Building a decision tree classification model

Once the selection criteria are constructed, the decision tree classification model is then used to begin the analysis of the relationship between glass type and chemical composition to determine the classification law.

Each internal node in the decision tree is a splitting problem: it specifies a test for an attribute of the instance, it splits the samples arriving at that node according to a particular attribute, and each subsequent branch of that node corresponds to a possible value of that attribute.

In this case it is the value of the judgment criterion gini that is expected to yield the classification effect by analyzing the information on the frequencies accounted for by various chemical components in a certain high potassium or lead-barium type of glass.

To calculate the frequency p_j of each column, i.e., of each chemical constituent in relation to all chemical constituents [7].

$$p_j = \frac{\sum_{i=1}^{m} p_{ij}}{\sum_{j=1}^{r} \sum_{i=1}^{m} p_{ij}} \tag{5}$$

From equation (4) and equation (5), the gini value of each chemical composition can be obtained, and the chemical composition corresponding to the minimum value of gini is taken as the classification criterion. Chemical composition p_j and its gini correspondence table is shown in

Table 2. Chemical composition pj and its gini correspondence.

p_j	$gini$
PbO	0.393
BaO	0.00021
Al_2O_3	Δ
MgO	Δ
Na_2O	Δ
…	…

From the calculation results in the above table, it is easy to see that for the two types of glasses, high potassium and lead-barium, since the content of other chemical components has little effect on their differentiation, it will be used as a criterion for their classification by PbO.

Further analysis of the classification criteria, at what value of PbO can be taken to distinguish the two types of glass, the frequency p_j of each chemical composition of all chemical compositions can be obtained from equation (5), and then the value of PbO should be taken as:

$$x = \frac{p_j}{\sum_{i=1}^{m} p_{ij}} \tag{6}$$

Where m is the chemical composition has m different values of properties, and x is the value of the requested division.

It is found that x = 5.46, so the following distinction can be made for a particular piece of glass.

$$\begin{cases} Y \in High\ potassium\ glass & PbO \leqslant 5.46 \\ Y \in Lead\ barium\ glass & PbO > 5.46 \end{cases} \tag{7}$$

3.2. Sub-category classification model for different categories

(1) Analysis of optimal clustering k values using Elbow method

The subclassification of the different classes can be considered as a further analysis and division of their properties, i.e., chemical compositions, on each of the two classes after already separating the high potassium class of glass from the lead-barium class of glass. Therefore, the first thing to do is to perform a cluster analysis of these chemical compositions by dividing the samples into K classes so that the sum of the distances between each sample and the center or mean of the class to which it belongs is minimized for further analysis [8].

Here we choose k-mean method for clustering.

The k-means is to minimize the squared error between the sample and the prime as the objective function, and the sum of the squared distance errors between the prime of each cluster and the sample points within the cluster is called distortions (distortions), then, for a cluster, the lower its distortions, the tighter the members of the cluster, and the higher the distortions, the looser the structure of the cluster. The degree of distortion decreases as the category increases, but for data with a certain degree of differentiation, the degree of distortion improves greatly when a certain critical point is reached, and decreases slowly afterwards, and this critical point can be considered as the point with better clustering performance.

So, the Elbow method (elbow rule) is used to determine the best k value, and the steps are as follows.

(1), for the chemical composition of n points, let k iteratively calculate the sum of squared errors (SSE) from to n. Calculate the sum of squares of the distances from each point to the cluster center to which it belongs after each clustering is completed.

(2) The sum of squares is asymptotically smaller until the k-sum of squares is 0, because at this point each point is the cluster center itself where it is located.

(3) In this sum of squares change process, there will be an inflection point that we need to require the "elbow" point, the rate of decline suddenly slowed down that is considered the best k value.

The error sum of squares formula for aggregation is.

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{8}$$

The data of high potassium-chemical composition and lead-barium-chemical composition were brought into the equation taking k∈ [2, 8] respectively and iterated to obtain the results as shown in Figure 2.

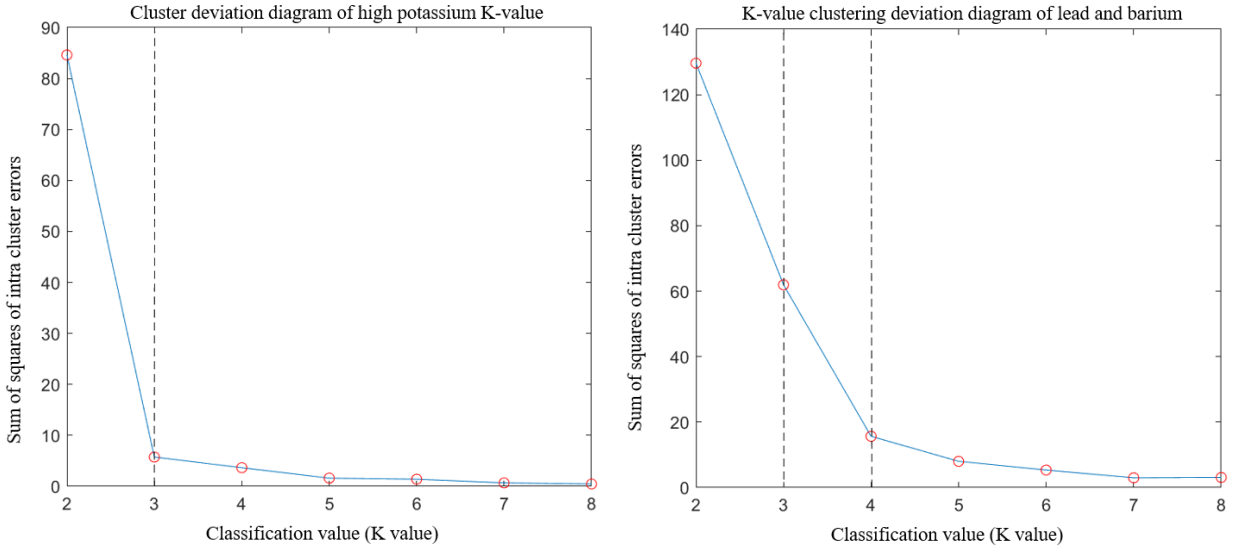

Figure 2. Cluster deviation diagram for lead-barium type and high potassium type glass.

The degree of distortion (y-value) of the high potassium class glass is improved substantially at k=3, and k=3 can be considered as the number of clusters Obviously, the elbow has the highest curvature for a k-value of 3, so for the clustering of this data set, the optimal number of clusters should be chosen 3; similarly, for the lead-barium class glass the optimal number of clusters should be chosen 4.

(2) Using K-means cluster analysis (K-Means)

Since there are more types of chemical components, the next step is to perform a cluster analysis on these chemical components into K classes, so that the sum of the distance between each sample and the center or mean of the class it belongs to is minimized for further analysis.

The K-Means algorithm generally uses the Euclidean distance as a measure of similarity between data objects. The similarity is inversely proportional to the distance between the data objects. The greater the similarity, the smaller the distance [9]. The algorithm needs to specify the initial number of clusters k and the initial cluster center k. According to the similarity between the data objects and the cluster center, the position of the cluster center is constantly updated and the SSE of the cluster is continuously reduced. When the SSE no longer changes or the objective function converges, the clustering ends and the final result is obtained [10, 11].

When the high potassium-chemical composition sample points and lead-barium-chemical composition sample points are put into the space separately, the Euclidean distance between them and the clustering center is calculated as:

$$d(H, C_i) = \sqrt{\sum_{i=i}^{m} (H_i - C_{ij})^2} \tag{9}$$

The clustered data were obtained, and the clustering summary results are shown in Table 3 and 4.

Table 3. Summary table of clustering of high potassium class glasses.

Clustering categories	Frequency	Percentage %
Clustering category_1	3	16.667
Clustering category_2	9	50.0
Clustering category_3	6	33.333
Total	18	100.0

Table 4. Cluster summary table of lead-barium class glass.

Clustering categories	Frequency	Percentage %
Clustering category_1	19	38.776
Clustering category_2	14	28.571
Clustering category_3	10	20.408
Total	6	12.245

(3) Subclassification using the decision tree classification model

The clustered data are brought into the decision tree classification model for solving.

(4) Rationalization and sensitivity analysis of classification results and subclassification results

The reasonableness of the classification results mainly depends on the accuracy, recall, precision and their interrelationship of the test set in our subclassification decision tree classification model.

The accuracy rate: the proportion of correct samples to the total samples, the larger the accuracy rate the better. Recall: The proportion of positive predicted samples out of the actual positive samples. Accuracy rate: the proportion of the results predicted to be positive samples that are actually positive samples, the larger the accuracy rate, the better.

Their mathematical expressions are described as follows.

$$\rho_c = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{m} r_m} \tag{10}$$

$$\rho_z = \frac{\sum_{i=1}^{n+} r_{i+}}{\sum_{i=1}^{n} r_n} \tag{11}$$

$$\rho'_z = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n+} r_{i+}} \tag{12}$$

The results of the test set of the decision tree classification model for the two subclasses analysis show that the subclass classification results have a considerable degree of confidence and reasonableness.

3.3. XGboost classification model building

XGboost is the abbreviation of "Extreme Gradient Boosting", and XGboost algorithm is a class of synthetic algorithms that combine basis functions and weights to form a good fit to the data.

For a data set of n chemical components containing m dimensions, the XGboost model can be expressed as follows.

$$\hat{y} = \sum_{k=1}^{m} f_k(x_i), \quad f_k \in F \quad (i = 1, 2, ...n) \tag{13}$$

When constructing the XGboost model, it is necessary to find the optimal parameters according to the principle of minimizing the objective function to build the optimal model.The objective function of the XGboost model can be divided into the error function term L and the model complexity function term. The objective function can be written as.

$$Obj = L + \Omega \tag{14}$$

$$L = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{15}$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {w_j}^2 \tag{16}$$

When training the model optimally using the training data, it is necessary to keep the original model unchanged and add a new function to the model so that the objective function is reduced as much as possible.

For the given chemical composition, it is possible to construct.

$$Obj = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{17}$$

$$Obj^{(t)} = \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_i(x_i)\right)\right)^2 + \Omega \tag{18}$$

The data were brought into XGboost for learning and using SPSSPRO software to learn the results its accuracy, recall, and precision can be higher, so the model can be used for prediction.

The following prediction results were obtained after removing whether the differentiated data were brought into the model as follows Table 5.

Table 5. Data in XGboost model training situation.

Artifact Number	Predicted results	Lead Barium	High Potassium
A1	High Potassium	0.009278297	0.9907217
A2	Lead Barium	0.9936732	0.006326808
A3	Lead Barium	0.9936732	0.006326808
A4	Lead Barium	0.9936732	0.006326808
A5	Lead barium	0.9936732	0.006326808
A6	High Potassium	0.009278297	0.9907217
A7	high potassium	0.009278297	0.9907217
A8	lead barium	0.9936732	0.006326808

The above predictions show that the probability of the predicted results is all greater than 99%, indicating that the predictions are better.

4. Conclusion

This paper analyzes the classification pattern of high-potassium glass and lead-barium glass based on the data; for each category, we select the appropriate chemical composition to classify them into subcategories, give specific classification methods and classification results, and analyze the reasonableness and sensitivity of the classification results. For unknown categories of glass artifacts, the chemical composition is analyzed to identify the types to which they belong, and the sensitivity of the classification results is analyzed.

The multiple linear regression model developed in this paper can predict the surface weathering points of high potassium glass artifacts and lead-barium glass artifacts to a certain extent, while being able to be accepted by most people due to the simplicity of the model. The correlation analysis model developed in this paper can also evaluate each chemical composition in different categories of glass

to a certain extent. The combination of Elbow method and K-Mean clustering is used to make better clustering results and guarantee the reasonableness of the subsequent results. Most of the models in this paper are based on assumptions and lack some generalizability. The model developed in this paper can be applied to some extent to the research of cultural relics such as the prediction of chemical composition of cultural relics and the classification of chemical composition of cultural relics, and the model can also be improved in future research based on the model developed in this paper.

References

[1] Gan Fuxi, Zhao Hongxia, Li Qinghui. Scientific and technological analysis and research of Warring States glassware excavated in Hubei Province [J]. Jianghan Archaeology, 2010(02):108-116+151+0.

[2] Wen Rui, Zhao Zhiqiang, Ma Jian. Composition analysis of glass beads excavated from the Shirenzigou site group in Balikun, Xinjiang [J]. Spectroscopy and spectral analysis, 2016, 36(09):2961-2965.

[3] Wang YB, Peng XY, Zhang SB. Experimental study on the recovery of lead from high lead glass [J]. Mineral Comprehensive Utilization, 2016(04):90-93.

[4] Cao M.X. Research on grey correlation analysis model and its application [D]. Nanjing University of Aeronautics and Astronautics, 2007.

[5] He Xiuli. Multiple linear models and ridge regression analysis [D]. Huazhong University of Science and Technology, 2005.

[6] X.B. Yang, J. Zhang. Decision tree algorithm and its core technology [J]. Computer Technology and Development, 2007(01):43-45.

[7] Zhang Zihao, Li Xiangcheng, Wu Haotian, et al. Screening nitrogen efficient wheat varieties based on principal component analysis and cluster analysis [J/OL]. Journal of Hunan Agricultural University (Natural Science Edition): 1-7 [2022-10-19]

[8] Huang Yupu, Wu Dazhang, Wang Sen, et al. Study on HPLC Fingerprint of Compound Huangqin Tablets and Determination of Six Components [J]. West China Journal of Pharmacy, 2022, 37 (05): 531-535. DOI: 10.13375/j.cnki.wcjps.2022.05.0013

[9] Fu Zhiyuan. Research on Digital Marketing Strategies in the 5G Era Based on K-means Cluster Analysis -- Taking Hangzhou Tourism Marketing as an Example [J]. China Business Review, 2022 (16): 75-80. DOI: 10.19699/j.cnki.issn2096-0298.2022.16.075

[10] Wu Shunchuan, Li Yujie, Zhang Huajin, et al. K-means cluster analysis method of dominant occurrence of rock mass discontinuity based on improved density peak value [J]. Mining Research and Development, 2022, 42 (07): 137-143. DOI: 10.13827/j.cnki.kyyk.2022.07.023

[11] Gao Hairong, Zhang Zhonglili, Yue Huanfang, et al. Estimation method of lettuce crop coefficient based on Bayesian XGBoost [J]. Shanxi Agricultural Science, 2022, 50 (10): 1482-1488.