Classification and analysis of glass artifacts based on feature filtering

Hongkui Zhong [1, *], Xiuge Gu [2], Junyan Guo [2], Jianhong Wang [2]

[1] School of Information Science and Technology, Nantong University, Nantong, China

[2] School of Sciences, Nantong University, Nantong, China

* Corresponding Author Email: td_ljf@163.com

Abstract. Glass is an important physical evidence of the early trade between China and the West. In the early days, the glass introduced into China was mainly pearl shaped ornaments from Western Asia and Egypt. Later, China learned its glass manufacturing technology and used local materials to make glass. Therefore, although the appearance of local glass products is similar to that of foreign glass products, their chemical compositions are not the same. According to the data, the classification rules of high potassium glass and lead barium glass are analyzed; For each category, this paper choose appropriate chemical components to classify them into subclasses, give specific classification methods and results, and analyze the rationality and sensitivity of the classification results.

Keywords: Feature filtering; Mapping relationships; Sensitivity analysis; K-means clustering.

1. Introduction

Glass is an important physical evidence of the early trade between China and the West. The glass introduced into China in the early days was mainly bead-shaped jewelry in West Asia and Egypt, and later China learned its glass-making technology after the use of local materials to make glass, so although our local glass and foreign glass products look similar, but its chemical composition is not the same[1,2].

Glass is mainly made of quartz sand, and the main chemical composition is silicon dioxide (SiO2). Co-solvents are added during the refining process to lower the melting temperature of quartz sand, and the use of different catalysts can also lead to differences in its chemical composition. Limestone, commonly used as a stabilizer in the preparation of glass, is converted to calcium oxide (CaO) after calcination [3-5].

In addition, ancient glass is susceptible to weathering due to the burial environment. During weathering, the composition ratio of glass changes as elements inside the glass are exchanged with elements in the environment, making it difficult to correctly determine its category. In the weathered cultural relics, the degree of weathering varies from one cultural relic to another, and the degree of weathering varies from one part of the same cultural relic to another, and there are also unweathered areas on the surface of some weathered cultural relics [6].

In this paper, this paper analyze the classification rules of high potassium glass and lead-barium glass based on the data; for each category, this paper 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.

2. Model assumptions and notation

2.1. Assumptions [7]

1. It is assumed that the sample data are true and accurate and have reliability.

2. It is assumed that when two variables are analyzed for autocorrelation, the correlation coefficient is higher than 0.7 and the internal information is consistent.

3. Assume that there is a linear relationship between the logistic transformed values of the continuous independent variables and the dependent variables.
4. Assume that the initial clustering centers are far enough apart.
5. It is assumed that the randomly generated data are in good agreement with the true values.

2.2. Notations

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

Table 1. Notations.

Symbols	Meaning
X2	Chi-square test statistic
ε	Coverage
Gain(A)	Information gain rate
Loss xi, j	Loss function

3. Model construction and solving

3.1. Glass classification law

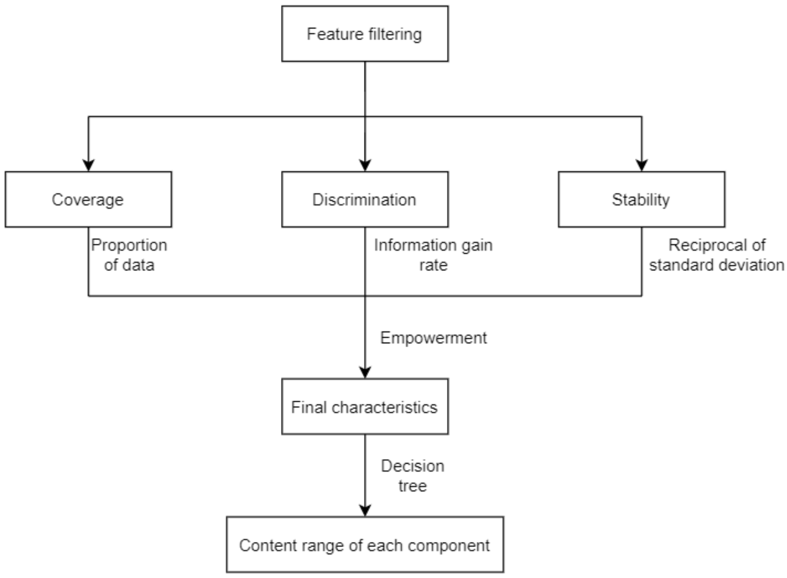


Figure 1. Flow chart of different categories of glass classification laws.

Flow chart of different categories of glass classification laws is shown in Figure 1. Since not all features are valid, invalid features tend to influence the formation of the final results, and also for the sake of model and computational simplicity, feature screening around these features is required. The features screened using the established features can be used to explore the classification law of high potassium glass and lead-barium glass, so a feature screening model is established.

When performing feature screening, the amount of information contained in the features, the correlation between the features and the labels, and the data fluctuation of the information need to be considered. The proportion of missing information values to the total number is usually used to reflect the information inclusion, and in this paper, in order to adjust the data toward consistency, the indicator of feature value coverage is chosen to represent, and the coverage, which is the proportion of non-empty data to the total data, can be expressed as

$$\varepsilon = \frac{\text{count(Non-null values)}}{\text{count(Non-null values)}} \tag{1}$$

In order to determine the feature and label relevance index, i.e., to further determine the usefulness of the information contained in the features, the differentiation degree expressed as the information gain rate is chosen as the corresponding index in this question, and its calculation formula is

$$\text{Gain-ratio}=\text{Gain}\ (\mathbf{A}) \tag{2}$$

Among them

$$\text{Gain}(S,\ A)\ =\ E(S)\ -\ E(S,\ A) \tag{3}$$

The standard deviation is a representative indicator of data fluctuation, the larger the standard deviation the greater the data fluctuation, in order to make the trend of each indicator uniform, the standard deviation to do the inverse of the treatment so stability Calculation method is

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

On the basis of the three identified indicators, the entropy weighting method is used to analyze the importance of each feature's contribution to the classification judgment, and the weights of the type of glass coverage, differentiation, and stability are set as α, β, and γ in order, then the final evaluation equation is

$$\chi=\alpha\times\frac{1}{s}+\beta\times\varepsilon+\gamma\times\text{Gain}(\mathbf{A}) \tag{5}$$

The overall sample data was evaluated by applying the feature screening model and its evaluation equation was obtained as

$$\chi=0.204\times\frac{1}{s}+\times 0.309\varepsilon+0.487\times\text{Gain}(\mathbf{A}) \tag{6}$$

The best overall performance was obtained for lead oxide, and the classification rules for high potassium glass and lead-barium glass were: when the lead oxide content was less than 5.46%, the sample glass was high potassium glass; when the lead oxide content was greater than 5.46%, the sample glass was lead-barium glass.

In this paper, a decision tree algorithm was used to investigate the classification rules of high potassium glass and lead-barium glass.

To construct the decision tree, the Gini coefficient is used to divide the branches of the tree, assuming that there are K classes and the probability that the sample point belongs to the kth class is pk, then the Gini index of the probability distribution is

$$\text{Gini}(p)=\sum_{k=1}^{K}p_k\left(1-p_k\right)=1-\sum_{k=1}^{K}p_k^2 \tag{7}$$

During the Cart tree run, the Gini coefficient of each feature will be calculated at the time of each node division and the feature with the smaller Gini coefficient will be selected.

3.2. Subcategorization

The chemical composition of high potassium glass and lead-barium glass differs significantly, and therefore should be discussed by glass type when sub classifying glasses according to chemical composition.

According to the above feature screening method, the weights of high potassium glass coverage, differentiation, and stability were determined to be in the order of The weights of coverage, differentiation, and stability for high potassium glass were determined to be 0.272, 0.433, and 0.295, respectively; and the weights of coverage, differentiation, and stability for lead-barium glass were 0.266, 0.497, and 0.237, respectively. The final evaluation equations for high potassium glass and lead-barium glass were then derived as follows.

$$\chi_2 = 0.272 \times \frac{1}{s} + 0.295 \times \varepsilon + 0.433 \times \mathrm{Gain}(\mathbf{A}) \tag{8}$$

$$\chi_2 = 0.266 \times \frac{1}{s} + 0.237 \times \varepsilon + 0.497 \times \mathrm{Gain}(\mathbf{A}) \tag{9}$$

The performance of the 14 components on the three indexes was summarized and analyzed, and for high potassium glass, the best overall performance indexes were obtained: calcium oxide and aluminum oxide; for high potassium glass, the best overall performance indexes were obtained: calcium oxide, aluminum oxide and strontium oxide.

According to the composition characteristics of high potassium glass, it is divided into three categories according to the content of calcium oxide and alumina. According to the different contents of alumina and calcium oxide, the high potassium glass can be divided into three subcategories, and the value intervals of each subcategory are shown in Table 2.

Table 2. High potassium glass subclass classification table.

Subcategory	Calcium oxide content area	Aluminum oxide content interval	Clustering Center
l-Al-Ca	(0.21,2.01)	(0.81,4.06)	(1.03,2.23)
m-Al-Ca	(4.71,8.70)	(3.93,6.44)	(6.96,5.62)
h-Al-Ca	(5.41,8.23)	(9.23,11.15)	(6.92,10.14)

The subclasses of lead-barium glass were divided in the same way as above, and the valid features screened were calcium oxide, aluminum oxide, and strontium oxide.

The chemical composition content of lead-barium glass was analyzed, and due to the large dimension of lead-barium glass data features, an elbow diagram (Figure 2) was made to assist in determining the number of classifications.

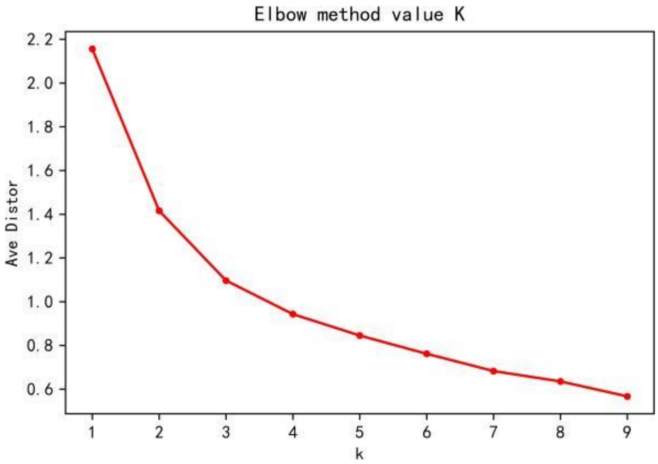


Figure 2. Lead barium glass subclass classification elbow diagram.

The elbow diagram decreases rapidly before the horizontal coordinate 3, and the data after the horizontal coordinate 3 tends to level off, which is considered as the inflection point of the image, and in order to take into account the conciseness of the classification category names and readability, the glass is classified according to other literature

The classification is divided into three categories: l-Ca-l-Al-h-Sr, h-Ca-m-Al-h-Sr and m-Ca-h-Al-l-Sr, where l(low) denotes low content, m(medium) denotes medium content, and h(high) denotes high content, for example, l-Ca-l-Al-h-Sr denotes low calcium content, low aluminum content, high strontium content. High strontium content category, and so on. High potassium and lead barium glass clustering results graph is shown in Figure 3.

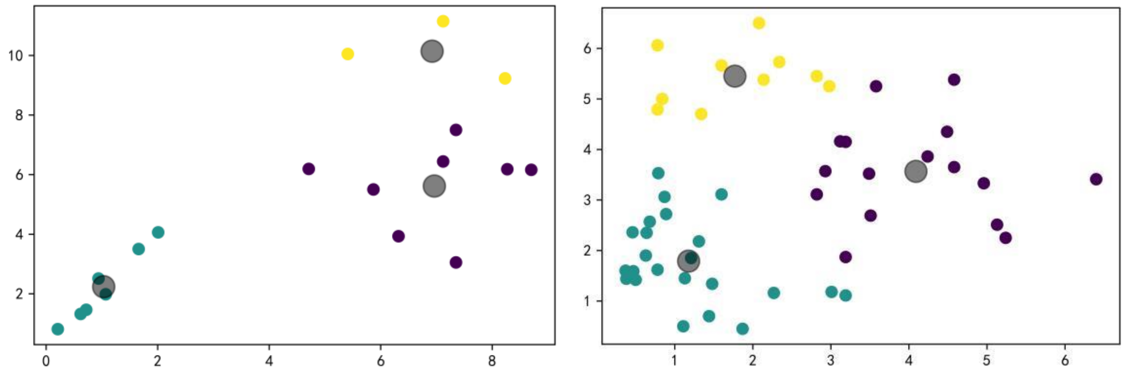


Figure 3. High potassium and lead barium glass clustering results graph.

3.3. Optimization model

Initially, the coverage in the subclass classification table using K-means clustering is not wide enough and the loss is large. In order to make the loss function as small as possible, an optimization model is established with the minimum loss function as the optimization objective.

Firstly, the matrices X_1, X_2 with dimensions of 3 × 2 and 3 × 3 were established for the classification information of high potassium glass and lead-barium glass, respectively, where $X_{i,j}$ denote the jth feature in the i-th class. Then, the optimization model is established with the 3σ upper and lower limits of each index with the minimum loss function as the optimization objective. The genetic algorithm is used to find the optimal classification model, based on which more accurate range results are obtained through training.

The optimization objectives established for high potassium glass are as follows [8].

$$\min_K \mathrm{Loss}(x_i) \tag{10}$$

The constraint of the content of each component is the optimum interval of the chemical composition, then the optimum interval is specifically expressed as

$$\begin{aligned} &\min_K \quad \mathrm{Loss}(x_i) \\ &\text{s.t.} \begin{cases} 0.21 \le x_{1,1} \le 2.75 \\ 0.81 \le x_{1,2} \le 5.55 \\ 4.71 \le x_{2,1} \le 10.60 \\ 3.05 \le x_{2,2} \le 9.67 \\ 5.41 \le x_{3,1} \le 10.40 \\ 9.23 \le x_{3,2} \le 12.50 \end{cases} \end{aligned} \tag{11}$$

Similarly, a loss function optimization model is developed for lead barium glass as follows.

$$\begin{aligned} &\min_{PbBa} \quad \mathrm{Loss}(x_i) \\ &\text{s.t.} \begin{cases} 0.37 \le x_{1,1} \le 3.48 \\ 0.45 \le x_{1,2} \le 4.23 \\ 0 \le x_{1,3} \le 1.15 \\ 2.82 \le x_{2,1} \le 7.16 \\ 1.87 \le x_{2,2} \le 6.06 \\ 0 \le x_{2,3} \le 1.28 \\ 0.78 \le x_{3,1} \le 4.67 \\ 4.7 \le x_{3,2} \le 6.97 \\ 0 \le x_{3,3} \le 0.62 \end{cases} \end{aligned} \tag{12}$$

3.4. Analysis of results

The classification of high potassium glasses into low, medium and high alumina is based mainly on the calcium oxide and alumina content of the sample. In a review of the literature on the components of ancient Chinese potassium glass, the classification is also "based on the Al2O3 and CaO content, these potassium glasses can be divided into three subclasses" [9].

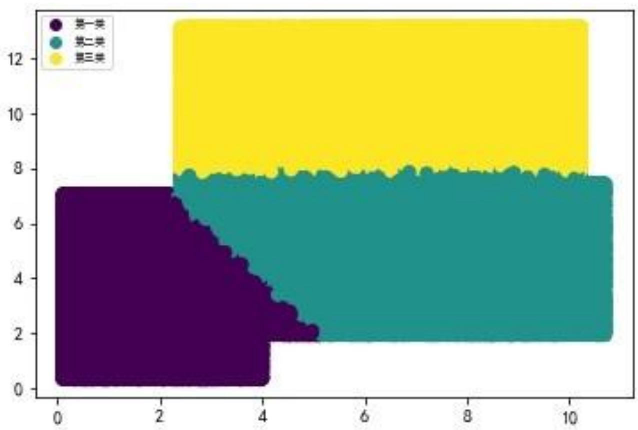
Since the glass artifacts chosen in the literature are different from those in the present question, the resulting classification intervals are slightly different. However, this literature mainly uses LA-ICP-AES method to analyze glass artifacts, which gives more complete data and more comprehensive information on trace elements of potassium glasses, and its classification is reasonable, so it is reasonable to subclassify high potassium glasses by calcium oxide content and alumina content in this paper.

In the sub classification of lead-barium glass, the content of strontium oxide is also considered in addition to the content of alumina and calcium oxide. In the literature on nondestructive analysis of Han dynasty glass from Guangzhou, the sub classification of glass is based on the content of CaO and Al2O3 and the proportion of trace elements Rb and Sr, and the sub classification of potassium silicate glass, which has the highest proportion [10]. Therefore, the classification of glass subclasses used in this paper is consistent with the classification of many literatures and has a high reasonableness.

3.5. Sensitivity analysis

The above subcategory classification model obtains the change rule of classification category caused by the change of model boundary under the condition of different proportion of characteristic component content. In order to directly reflect the influence of different parameters on subcategory classification, the sensitivity analysis results are obtained by randomly selecting parameter control variables around the boundary.

In the process of generating random data, the upper and lower bounds of the results are first used, and then the optimized classification model is used for prediction. In the simulation process, the boundary between each two categories is fixed to the other two categories for large sample simulation, and the data points of each time are superimposed on the same two-dimensional plan. For the lead barium glass category, through comparison, although three dimensions are considered in the clustering model, two-dimensional images can still be used to show the effect of the model. The sensitivity simulation diagram of two types of glass is as follows (Figure 4-7):



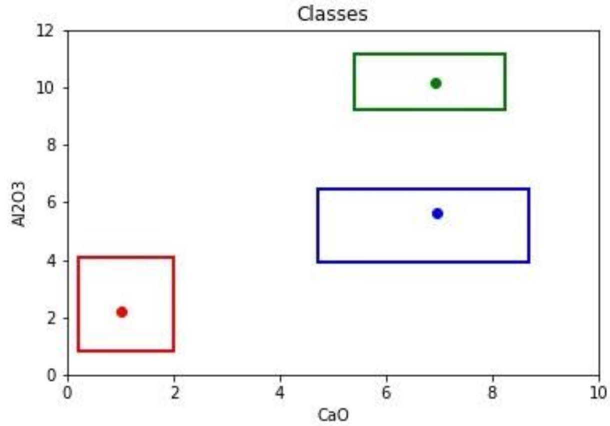


Figure 4. Clustering of high potassium glass (1). **Figure 5.** Clustering of high potassium glass (2).

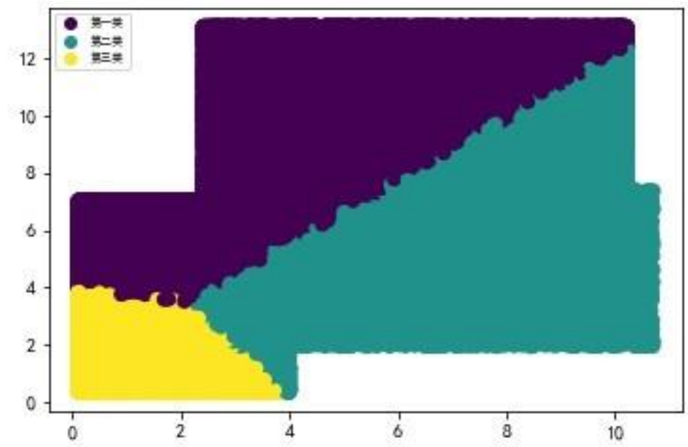


Figure 6. Clustering result diagram of lead barium glass (1).

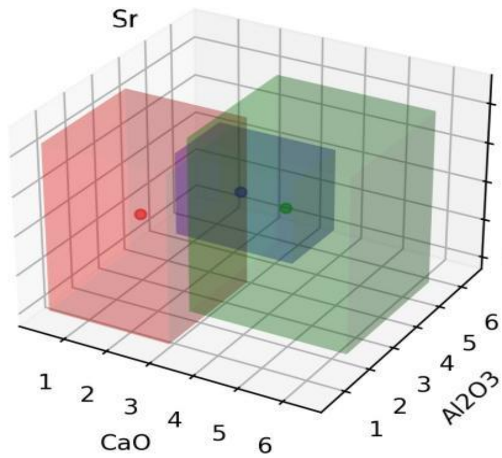


Figure 7. Clustering result diagram of lead barium glass (2).

The sawtooth boundary of the two-dimensional model can clearly reflect the randomness of the sensitivity test sample selection. Observation of the pictures shows that the optimized clustering model still gives good results in such a random learning sample. Through comparison, the boundary value is basically consistent with the diffusion range, indicating that the data disturbance has little impact on the model. It can be considered that the optimized classification model has strong stability and high reliability.

4. Conclusion

In this paper, the best indicator was PbO, and the specific classification law was obtained through the decision tree: if the PbO content was less than 5.46%, the sample was high potassium glass; otherwise, the sample was lead-barium glass. Then, the effective features of the two glasses were screened by the feature screening model, and three subclasses were obtained by the K-means clustering method. Then, for the resulting classification model, an optimization model was established with the initial condition of the range of historical data values combined with the classification range, and the optimization objective of minimizing the loss function, and the genetic algorithm was used to find the best. In order to analyze the rationality and sensitivity, random numbers were selected to simulate and analyze the boundary values, and the optimized classification model was obtained with high robustness and reliability.

References

[1] Zhao Zhiqiang. Research on the composition system and production process of glass beads excavated from the Balikun Shirenzigou site group in Xinjiang [D]. Northwestern University, 2016.

[2] Yuan Yingchun, Zhou Yi, Song Yufei, et al. A method for estimating chlorophyll content in wheat canopy based on information entropy feature selection [J]. Journal of Agricultural Machinery, 2022, 53(08):186-195.

[3] Zhu J-P. Applied multivariate statistical analysis [M]. Fourth edition. Science Press. Journal of Agricultural Machinery, 2021.

[4] Feng Guoshuang. Statistical Analysis of Medical Cases with SAS Applications [M]. Beijing University Medical Press, 2011.

[5] Yang Nan, Lin Songxiang, GAO Qiang, et al. A method for discovering potential WEB community features from Markov clusters [J]. Journal of Computer Science, 2007(07):1086-1093.

[6] Sichin Bilig, Li Qinghui, Gan Fuxi. Analysis of ancient Chinese potassium glass components by laser exfoliation-inductively coupled plasma-atomic emission spectrometry/mass spectrometry [J]. Analytical Chemistry, 2013, 41(09):1328-1333.

[7] Fu Qiang, Kwong Guirong, Lu Liangbo, et al. Nondestructive analysis of Han dynasty glass artifacts excavated in Guangzhou [J]. Journal of Silicates, 2013, 41(07):994-1003.

[8] Yi, J.; Yang, H.R.; Lee, C.H. Compositional Variation and Color Diversity of Glass Beads from the 4th Century Tomb Complex in Korea. Appl. Sci. 2021, 11, 5233. , 5233.

[9] Anastasia Cholakova, Thilo Rehren, Ian C. Freestone, Compositional identification of 6th c. AD glass from the Lower Danube, Journal of Archaeological Science: Reports. Volume 7. 2016. Pages 625-632.

[10] [10] Zhao Zhiqiang. Detection and Analysis of Glass Products Unearthed in Hunan Liyemai Tea Warring States Cemetery [J]. Hunan Archaeological Series, 2020 (00): 288-301.