ISSN 2616-5775 Vol. 6, Issue 4: 49-56, DOI: 10.25236/AJCIS.2023.060406

Component analysis and sub-classification of glass relics based on machine learning

Yeting Song[a,*], Yuzhu Mao[b], Yanrun Wang[c]

Northwest Normal University, Lanzhou, Gansu, 730070, China
[a]2103234620@qq.com, [b]3236721341@qq.com, [c]1469200339@qq.com
[*]Corresponding author

Abstract: *Ancient glass is easily affected by environmental weathering. In the process of weathering, the proportion of its chemical composition will change, thus affecting the accurate judgment of its category. Therefore, according to the information such as surface weathering degree, type, color and pattern, Spearman correlation coefficient is used to conduct correlation analysis on the chemical composition of weathered cultural relics. Therefore, according to the information such as surface weathering degree, type, color and pattern, Spearman correlation coefficient is used to conduct correlation analysis on the chemical composition of weathered cultural relics, K-means clustering algorithm and decision tree are used to explore the classification rules of glass cultural relics and subcategory division of glass cultural relics [1], which is of vital importance for the study of ancient glass cultural relics.*

Keywords: *Composition analysis of glass relics, Spearman correlation coefficient, Decision tree classification, K-means clustering algorithm*

1. Introduction

Glass products have a long history. The research on the composition and category of ancient glass relics is of great scientific significance to the cultural, economic and technological exchanges along the Silk Road and the technological origin and development of local glass crafts. At present, archaeological research mainly classifies glass products by making techniques. This paper starts with the chemical composition of glass products, adopts Spearman correlation analysis to explore the correlation among various components, and uses K-means clustering algorithm and CART decision tree model to establish subclass classification model from the perspective of chemical composition.

2. Questions to ask

In question 1, Analysis of the surface weathering information of glazed relics and its relationship with the type, pattern and color of glass; Combined with the type of glass, the statistical rule of whether there is weathering chemical composition content on the surface of cultural relics samples is analyzed, and the chemical composition content before weathering is predicted according to the detection data of weathering point.

In question 2, the classification rules of high potassium glass and lead barium glass were analyzed according to the known data. After selecting appropriate chemical components for each category, the specific classification method and classification results are given, and the rationality and sensitivity of the classification results are analyzed.

3. Questions analysis

3.1. Analysis of Question 1

This problem requires to give the relationship between the properties of glass relics and weathering, and according to the type of glass, classify and discuss whether the cultural relics surface weathering chemical content statistics, and predict the cultural relics before weathering chemical composition content. For this requirement, the analysis relationship is divided into correlation analysis and

ISSN 2616-5775 Vol. 6, Issue 4: 49-56, DOI: 10.25236/AJCIS.2023.060406

difference analysis. First, the four characteristic mean variables of cultural relics are pretreated, and then the correlation analysis is conducted by solving the Spearman correlation coefficient [2]. Secondly, the chemical composition content data of glass relics were preprocessed, and only the cumulative component ratio and the data between 85% and 105% were retained. Descriptive statistical analysis was conducted on the remaining data, so as to determine that silica was the main component. A linear regression model was established for the attribute data, and a prediction model formula was formed to solve the problem, and the chemical component content of cultural relics before weathering was calculated according to the formula.

3.2. Analysis of Question 2

The classification rules of glass types are required to be analyzed, and the data in Form 1 and Form 2 are required to be summarized. The decision tree model is trained with the sample, and the decision tree is constructed by Gini coefficient to complete the analysis of classification rules. It is required to subdivide the two categories, give the classification methods and results, and finally analyze the classification results. For this requirement, first of all, the missing value processing is carried out on the data, and the missing value of each component is filled as 0 according to the requirements of the topic. The categorical variable "type" was transformed into quantitative variable, with 1 representing the type of high potassium and 2 representing the type of lead barium. Secondly, the data are divided into lead barium glass and high potassium glass, and the elbow rule is adopted for the two types of data respectively to obtain a reasonable k value. The k value is substituted into the clustering algorithm to obtain the division of specific subclasses. Finally, the subclass data are taken as samples to analyze the classification through decision tree model. Aiming at the final classification results, the rationality of classification is analyzed from two aspects: k value of cluster analysis and P value of significance analysis. The sensitivity of the model is tested by slightly disturbing the original data and reclassifying it.

4. Basic assumptions of the model

Assume that the samples are independent of each other; assume that the undetected chemical composition content on the surface of the cultural relic sample is 0;

5. Model establishment and solution

5.1. Solution of question 1

5.1.1. Spearman correlation coefficient analysis

In order to explore the relationship between the surface weathering of cultural relics and the type, decoration and color of glass, normality test was used to conduct descriptive statistics on the median and average value of the samples of decoration, type, color and surface weathering.

The specific formula of Spearman correlation coefficient analysis is shown in (1):

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)} \tag{1}$$

In the above formula,r_s is the correlation coefficient，n is the number of samples，d represents the rank difference between the data x_i and y_i. x_i and y_i respectively represent the i-th attribute value in the two groups of sample data.

Spearman correlation coefficient is used to further discuss the relationship between the four, so as to obtain the relationship between the surface weathering of glass relics and its glass type, ornamentation and color, as shown in Table 1.

The surface weathering of glass relics is positively correlated with the type and decoration of glass, and negatively correlated with the color of glass relics, and the degree of glass surface weathering is more strongly correlated with the type of glass.

ISSN 2616-5775 Vol. 6, Issue 4: 49-56, DOI: 10.25236/AJCIS.2023.060406

Table 1: Weathering factors.

	ornamentation	**type**	**color**	**surface weathering**
ornamentation	1.000(0.000***)	-0.370(0.006***)	-0.481(0.000***)	0.128(0.358)
type	-0.370(0.006***)	1.000(0.000***)	0.541(0.000***)	0.316(0.020**)
color	-0.481(0.000***)	0.541(0.000***)	1.000(0.000***)	-0.112(0.421)
surface weathering	0.128(0.358)	0.316(0.020**)	-0.112(0.421)	1.000(0.000***)

Note: The significance level of ***, ** and * is 1%, 5% and 10% respectively

5.1.2. Linear regression model

According to the use requirements of the polynomial linear regression equation [3], the weathering of cultural relics is related to various factors. By fitting the undetermined parameters of the influencing factors, the coefficients $b_0,b_1,\ldots,b_n$. The formula of correlation linear equation as shown (2)

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \quad (2)$$

In order to analyze the statistical rule of weathering chemical composition content on the surface of cultural relics samples, linear regression equation was established to process the data in the summary table. The decoration, type, color and surface weathering were taken as fixed x values, and different components were taken as y values to construct multiple linear regressions for iteration.

In the existing sample data, the hypothesis test is carried out by constructing statistics. Combined with the types of glass relics on high potassium and lead barium two types of relics. Descriptive statistical analysis was conducted on the composition differences before and after differentiation. After hypothesis testing, the main chemical component silica of the high-potassium and lead-barium cultural relics changed significantly before and after weathering. Finally, aiming at the data set of weathered glass relics, the linear regression of silica composition is made, and the analytical formula is solved as shown (3).

When y is SiO_2 content:

$$\begin{aligned} y = {} & 27.013 + 1.974 \times F_A + 40.86 \times F_B + (-15.821) \times F_C + 6.788 \times PB + 20.225 \times K + \\ & 16.834 \times Co_1 + 5.173 \times Co_2 + 15.416 \times Co_3 + (-12.636) \times Co_4 + (-4.095) \times Co_5 + \\ & 8.405 \times Co_6 + 1.477 \times Co_7 + (-3.563) \times Co_8 + 22.625 \times \Omega_0 + 4.388 \times \Omega_1 \end{aligned} \quad (3)$$

FA is decorative feature A, FB is decorative feature B, FC is decorative feature C, PB is the lead barium type characteristic of glass cultural relics, K is the high potassium type characteristic of glass cultural relics, C_1-C_8 are the eight light to deep color characteristics of cultural relics, Ω_0 is the weathered and no weathered characteristics of cultural relics surface, Ω_1 is the weathered and weathered characteristics of cultural relics surface.

According to the analysis results of the model, P value is 0.00534<0.05, showing significance at the level, which can reject the original hypothesis that the regression coefficient is 0, so the model basically meets the requirements. The chemical composition before weathering can be obtained by comparing the actual composition of cultural relics with that of various weathering products.

5.2. Solution of question 2

5.2.1. Decision tree classification

Since the attribute types in the existing data sets are continuous, it is necessary to discretize the samples first, and then use Gini coefficient to construct the decision tree [4].The data discretization process is shown in Figure 1.

ISSN 2616-5775 Vol. 6, Issue 4: 49-56, DOI: 10.25236/AJCIS.2023.060406

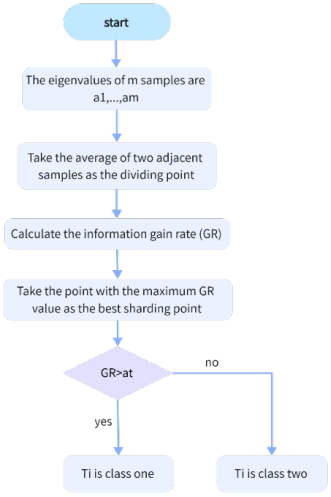

Figure 1: Flow chart of discretization.

The i th partition point T_i is expressed as $T_i = (a_i+a_{i+1})/2$. Because Gini coefficient can be interpreted as the probability that a sample randomly selected from the sample set is misclassified, in the construction process, the Gini coefficient of each attribute is calculated first, and the attribute with the lowest Gini coefficient is selected to divide the sample. The Gini coefficient formula is as shown (4):

$$G = 1 - \sum_{i=1}^{k} p_i^2 \tag{4}$$

In the above formula, k is the number of sample types in the data set; p_i is the proportion of the number of Class i samples in the total number of samples, G is the Gini coefficient

By calculating the Gini coefficient of each attribute, the minimum Gini coefficient of the lead oxide attribute is 0.412. According to the best division point of lead oxide content obtained by data discretization, all data sets are trained with the condition of whether the content of lead oxide is less than or equal to 5.46%. Since the two leaf nodes after partition are pure leaf nodes, the construction of the decision tree is complete. The structure of the decision tree is shown in Figure 2. The main parameters of the decision tree model are as follows: the minimum sample number of internal node splitting is 2; the minimum sample number of leaf nodes is 1. The maximum number of leaf nodes is 50; the maximum depth of the tree is 10.

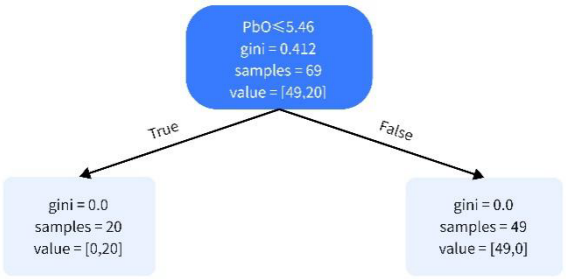

Figure 2: Decision tree structure diagram.

It can be concluded that glass relics may be classified according to the content of lead oxide. Pb oxide content > 5.46% is Pb barium type glass relics, ≤5.46% is high potassium type glass relics.

5.2.2. Subclass division

K-Means algorithm[5] is a clustering algorithm based on partition, and also an unsupervised learning algorithm. The core idea of K-Means algorithm is to use Euclidean distance as an index to measure the similarity between data objects for a given sample set. The similarity is inversely proportional to the distance between data objects. The greater the similarity, the smaller the distance.

The initial clustering number and each initial clustering center are specified in advance. According to the distance between samples, the sample set is divided into clusters. According to the similarity

ISSN 2616-5775 Vol. 6, Issue 4: 49-56, DOI: 10.25236/AJCIS.2023.060406

between data objects and clustering centers, the location of clustering centers is constantly updated to continuously reduce the sum of squares of errors of cluster (Sum of Squared Error，SSE），When SSE no longer changes or the objective function converges, the clustering ends and the final result is obtained.

The Euclidean distance between the data object and the clustering center in the space is calculated as shown (5):

$$d(X, C_i) = \sqrt{\sum_{j=1}^{m}(X_j - C_{ij})^2} \quad (5)$$

In the above formula, X is the data object; C_i is the ith clustering center; m is the dimension of the data object; X_j, C_{ij} are the J th attribute of X and C_i.

The calculation formula of SSE and error square of the whole data set is shown (6):

$$SSE = \sum_{i=1}^{k} \sum_{X \in C_i} |d(X, C_i)|^2 \quad (6)$$

SSE represents the quality of clustering results, and k is the number of clusters.

In the actual construction of the clustering model, k initial clustering centersC_i (i≤1≤k) need to be randomly selected from the data set, the Euclidean distance between the remaining data objects and the clustering center C_i is calculated, the clustering center C_i nearest to the target data object is found, and the data object is allocated to the cluster corresponding to the clustering center C_i. Then the average value of data objects in each cluster is calculated as the new clustering center, and the next iteration is carried out until the clustering center no longer changes or the maximum number of iterations is reached.

Therefore, using K-means clustering algorithm to study the subclass division of two categories, elbow rule [6] should first be adopted to determine the selection of K value of clustering quantity. The principle of this method is to minimize the distance from the point to the cluster center. The data of Pb barium glass type and high potassium glass type were analyzed respectively. The analysis results are shown in Figure 3 and Figure 4:

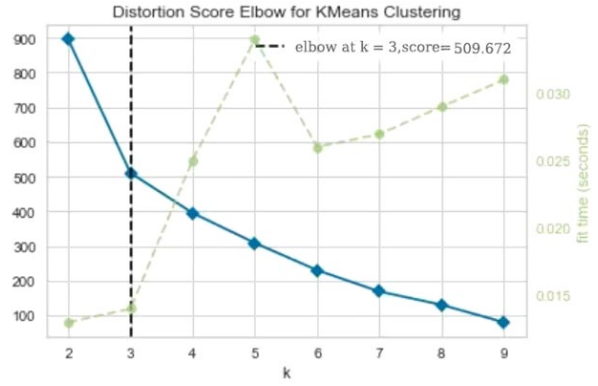

Figure 3: k value analysis 1.

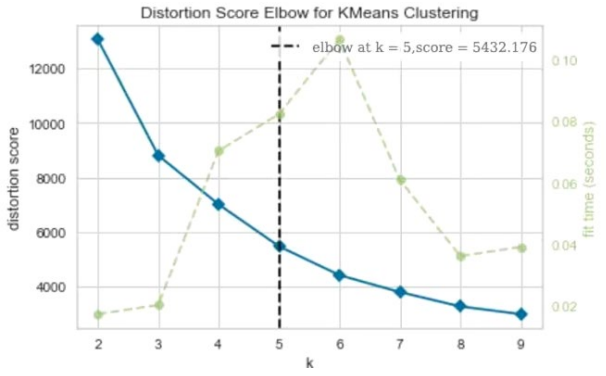

Figure 4: k value analysis 2.

According to the analysis in Figure 3, for high-potassium glass relics, k=3 is the k value at the

ISSN 2616-5775 Vol. 6, Issue 4: 49-56, DOI: 10.25236/AJCIS.2023.060406

elbow, that is, it is appropriate to divide high-potassium glass into three sub-categories. According to the analysis in Figure 4, for lead-barium glass relics, k=5 is the value of k at the elbow, that is, it is appropriate to divide lead-barium glass into five sub-categories.

According to the obtained k value, the k-means algorithm is used in MATLAB for cluster analysis [7]. The visualization results are shown in Figure 5 and Figure 6:

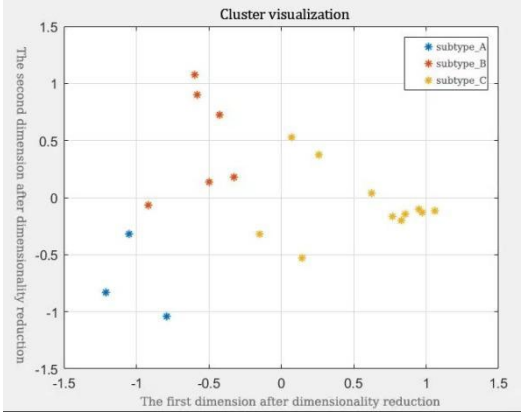

Figure 5: High potassium type.

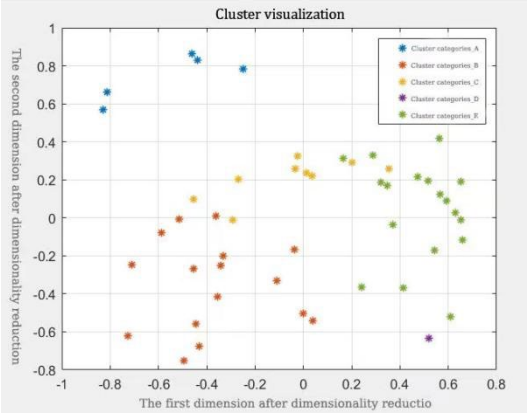

Figure 6: Lead-barium type.

5.2.3. Cluster model analysis

According to the above subclass classification results, the data were substituted into the CART decision tree model to specifically solve the critical content of classification. Thus, the structure of the high-potassium subclass decision tree is shown in Figure 7:

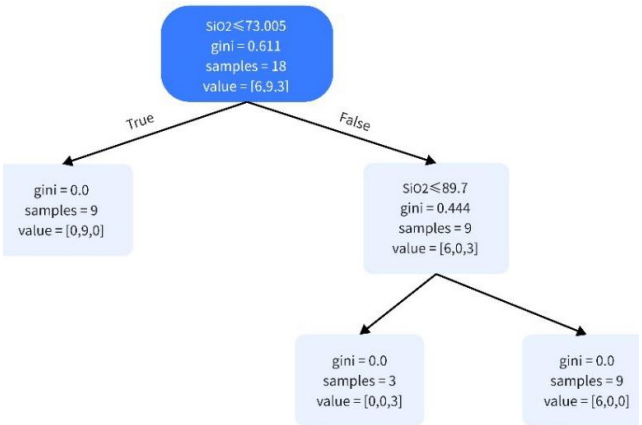

Figure 7: Decision tree structure of high potassium subclass.

According to Figure 7, high-potassium glass cultural relics can be divided into three sub-categories

ISSN 2616-5775 Vol. 6, Issue 4: 49-56, DOI: 10.25236/AJCIS.2023.060406

according to SiO2 content: SiO2 content less than or equal to 73.005%, SiO2 content between 73.005% and 89.7%, and SiO2 content greater than 89.7%. The structure of the decision tree of the lead barium subclass is shown in Figure 8:

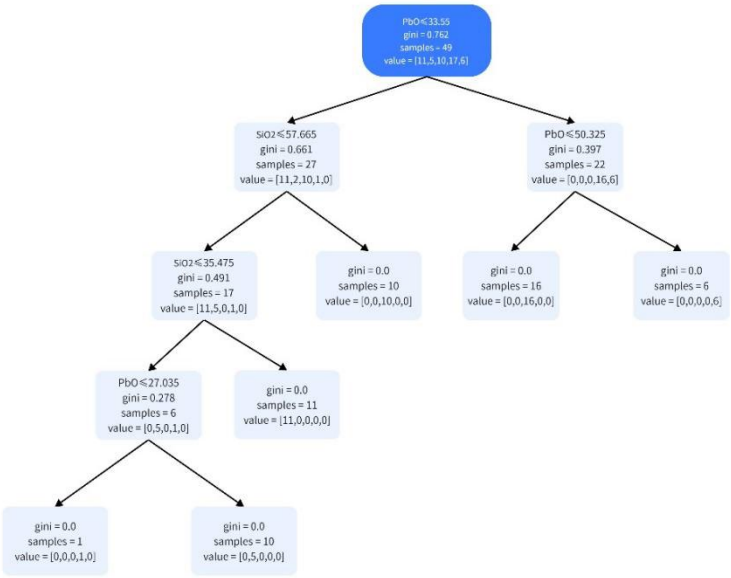

Figure 8: Decision tree structure of Pb barium subclass.

According to Figure 8, it can be concluded that lead-barium glass cultural relics can be divided into 5 sub-categories, which are: SiO2 content is less than or equal to 35.475% and PbO content is less than 27.035% and PbO content is between 33.55% and 50.325%, SiO2 content is less than 35.475% and PbO content is between 27.035% and 33.55%, PbO content is less than 33.55% and SiO2 content is 35. Between 475% and 57.665%, the content of PbO is less than 33.55%, and the content of SiO2 is more than 57.665% and the content of PbO is more than 50.325%.

5.2.4. Rationality and sensitivity analysis

For the rationality analysis of the model, this paper starts with K value of cluster analysis and P value cluster analysis of significance analysis. In the cluster analysis, the elbow algorithm is used in this paper, and the optimal K value can be intuitively seen, so the classification is reasonable. In addition, the significance P value can represent the relevance of the data. If the data presents significance at the level, then the classification is reasonable. About 85% of the content of the comprehensive content of the classification model in this paper presents significance, so the classification in this paper is reasonable.

For sensitivity analysis, we analyzed the original data by small disturbance processing. Reasonable reduction of the data between 0.9 and 1.3 times, and comparison between the predicted results of the processed data and the original data showed that the predicted results after reasonable reduction were almost consistent with the original data, with an accuracy of 94.658%, indicating that the model had good robustness and high accuracy.

6. Conclusions

The prediction model established and used in this paper is widely applicable to the spearman correlation analysis model, and the analysis and prediction results obtained are relatively accurate, which has certain universality for predicting the chemical composition content of weathered glass cultural relics before weathering and predicting the unknown category of glass cultural relics. It has a certain reference value for studying the chemical composition content of ancient glass relics before weathering and the unknown category of glass relics. It is suitable to be applied to the research of cultural relics. Cultural relics can be restored reasonably to the maximum extent by analyzing the chemical composition content of cultural relics before weathering, and the unknown category of cultural relics can be identified by analyzing the existing chemical composition, which has great reference significance for cultural relics research.

ISSN 2616-5775 Vol. 6, Issue 4: 49-56, DOI: 10.25236/AJCIS.2023.060406

References

[1] Gan Fuxi. The Origin and development of glass in ancient China [M]. Chinese Journal of Nature. 2006. 28(004):187-193.
[2] Jia Xiaofen, Guo Yongsheng, Huang Yourui. Salt and pepper noise detection algorithm for color images based on Spearman rank correlation [J]. Journal of University of Science and Technology of China. 2019. 49(001):63-70.
[3] Zhou Nan. Case Study of Prediction Analysis Based on Multiple Linear Regression Model [J]. Enterprise Review. 2013 48(009):122-154.
[4] Si Dajun, Hu Wenyue, Deng Zilin, Xu Yanhui. Fair hierarchical clustering of substations based on Gini coefficient [J].Global Energy Interconnection. 2021.04(006):12-16.
[5] Dong Shirong. Application of linear regression method of K-mean cluster analysis in correlation analysis [J]. Journal of Changchun Normal University. 2018.25 (008): 62-66.
[6] Yang Zhenzhen, Li Hongyao, Yang Keyi, Liu Lin. Clustering Analysis of Residents' Consumption Level Based on Systematic Clustering Algorithm and Elbow Criterion [J]. China Science and Technology Information. 2021. 16(012):121-122.
[7] Li Mingmei. Research on Fusion clustering method based on data feature selection [D]. Hangzhou Dianzi University. 2022.