Research on the composition of glass relics based on CART model

Mingyu An [1, a], Daolong Yuan [1, b], Suyang Liu [2, c], Jialei Ding [3, d, *]

[1]Shandong University of Science and Technology Jinan, China

[2] Ocean University of China Shandong, China

[3] Shandong University of Science and Technology Jinan, China

[a] 2762148027@qq.com,[b] 1051657207@qq.com,[c] 184465475@qq.com,[d] xmbrianding@foxmail.com

Abstract. Based on each index of glass samples, the indexes of all kinds of glass sample cluster analysis, and the class divide, it is to predict the chemical composition of glass samples which before weathering, the unknown glass samples for testing, and a correlation analysis of chemical composition in glass samples was carried out through the appropriate method. As for the relationship between the weathering degree of cultural relics surface and the type, ornamentation and color of glass samples, the weathering degree and the type, ornamentation and color of glass samples were corresponding to unilateral factors, and the relationship was preliminarily explored by bar charts. It was concluded that the weathering degree had a certain multiple relationship with the three indexes. In order to further study the relationship quantitatively, we used the correlation thermogram to analyze. It can be concluded that the type of cultural relics has the strongest correlation with surface weathering, the correlation degree reaches 0.4, while the correlation of color and ornamentation are 0.013 and 0.12, which are relatively low.

Keywords: CART; correlation analysis; components of glass cultural relics.

1. Introduction

First of all, it is necessary to analyze the relationship between surface weathering of glass relics and glass types, ornamentation and color. In this paper, the corresponding bar chart is drawn for qualitative exploration, and its internal connection is explored from the chart. In addition, the correlation thermal map is drawn to quantitatively analyze its internal relationship. Then it is necessary to analyze the rule of the content of chemical components on the surface of cultural relics samples. In this paper, by drawing the histogram of kernel density corresponding to each component and drawing the distribution curve, the rule of its existence can be found from the figure. Finally, it is necessary to predict the chemical composition content before weathering. Because the distribution characteristics of data are not discrete, it cannot be processed by classification. Therefore, we uses CART to predict the data, and then obtains the prediction results. Then, it is necessary to classify and process all data, make one-to-one correspondence between chemical components, and make one-to-one composition comparison table. Then, the correlation of data is preliminarily explored by drawing. After data integration and processing, the variance, mean and covariance of each group of data were calculated respectively. After integration and processing of the obtained data, Pearson's coefficient was calculated. The magnitude of Pearson's coefficient was calculated by calculating the quotient of mean and variance, and the correlation of chemical components was judged by observing the value of Pearson's coefficient. A grid diagram was made to show the correlation between components.

2. Model establishment and solution

First of all, the glass cultural relics were classified and the relationship between the quantities of lead-barium glass and high-potassium glass with weathering and without weathering was made. According to Figure 1, for the group without weathering on the glass surface, the amount of high-potassium glass is the same as that of lead-barium glass. For weathering group, the amount of lead barium glass is much higher than that of high potassium glass. From the perspective of the same type of glass, it can be seen that the unweathered quantity of high-potassium glass is twice that of weathered quantity, while the weathered quantity of lead-barium glass is 2.3 times that of unweathered quantity.

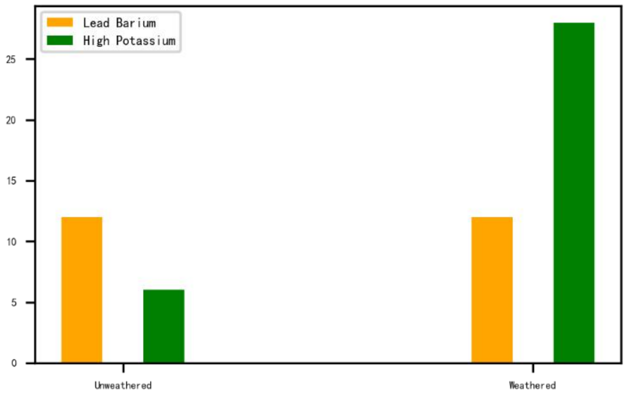

Figure 1. The bar chart depicts surface weathering and glass type.

Thirdly, it explores the relationship between surface weathering of glass relics and glass ornamentation. It could be seen from Figure 2 that the surface of the glass relics with grain B is weathered, and the surface of the glass relics with grain A has the same amount of weathering or not, while the surface of the glass relics with grain C is 1.3 times of that without weathering.

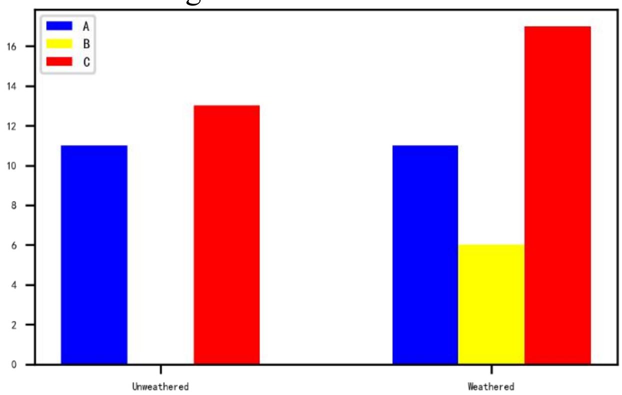

Figure 2. The bar chart depicts surface weathering of glass relics and ornamentation on glass.

Finally, it explores the relationship between surface weathering and color of glass cultural relics. According to Figure 3, the surface of cultural relics with dark blue and green colors is not weathered, while the surface of cultural relics with black color is weathered. However, the number of weathered and non-weathered cultural relics with light green, light blue, dark green, purple and blue-green surface has a small range of fluctuation.

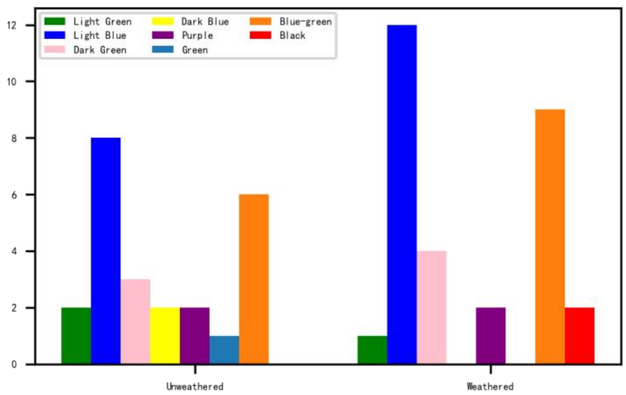

Figure 3. The bar chart depicts surface weathering and the color of glass relics.

In order to further show the relationship between the three factors and the surface weathering of cultural relics and their proportion, this paper chooses correlation thermography to explore.

Correlation thermogram is the correlation between two factors calculated by correlation analysis. The results are taken as a matrix and visualized as a thermal map for description, which can accurately and intuitively describe the correlation between various factors.

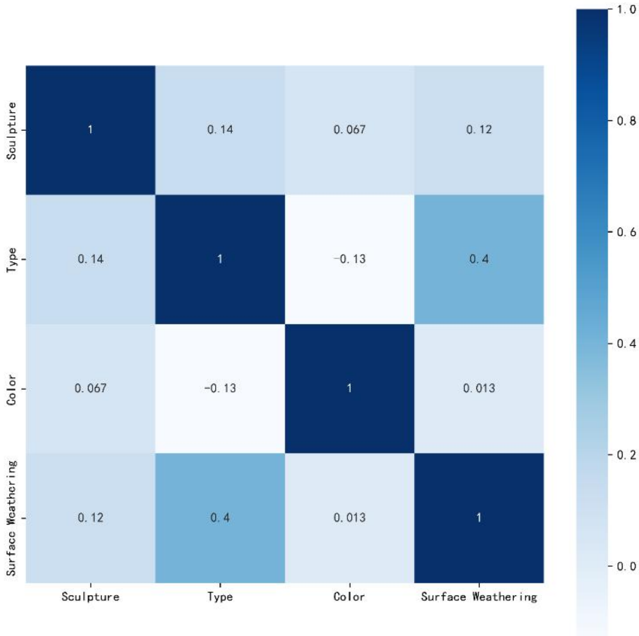

Figure 4. Thermal correlation diagram.

It can be seen from Figure 4 that the type of cultural relic has the strongest correlation with surface weathering, with a correlation degree of 0.4, while the correlation of color and ornamentality is 0.013 and 0.12, which are relatively low. Silica, potassium oxide, phosphorus pentoxide, sulfur dioxide and lead oxide account for a large proportion in influencing whether the surface has weathering, as shown in Table 1. The influence of other factors on surface weathering have less effect, so they are not considered as the main factors to determine the statistical rule.

TABLE 1. Weathering chemical composition specific gravity table.

	SiO_2	K_2O	P_2O_5	SO_2	PbO
Unweathering	higher	higher	low	low	lower
weathering	low	low	high	high	Higher

It is necessary to integrate multiple chemical composition indexes and ensure that parameter Settings have little influence on the results. Therefore, use CART is the best way to solve this kind of problem, which has simple process, convenient parameter setting and it is reusable.

CART is a tree structure branch established according to the decision to select branches. In the decision tree, each node represents an object, each fork represents a possible attribute value or corresponding decision, and every node from the root to the leaf corresponds to every decision test sequence. In this problem, each node is a screening of glass features. When one of the satisfactory conditions is obtained, the next step will be taken[1].

Firstly, the feature vector matrix *X* and *X* are established to store the accessory data, and the space *X* is divided into several non-overlapping domains $R_1, R_2, \ldots, R_j$ and each space divided corresponds to a label value y to bear the prediction result. Label values are obtained by averaging the eigenvalues of glass samples in the region: (x, y is sample data)

$$y_{R_j} = \frac{1}{n}\sum_{j\in R_j} y_j \qquad (1)$$

All values of all features are divided one by one each time, and the optimal glass sample index is selected as the cut-off point according to the square error minimization criterion, and the following equation is solved: (is the variance)

$$\min_{j,s}\left[\min_{c1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c2}\sum_{x_i\in R_2(j,s)}(y_i-c_1)^2\right](2)$$

Solve the matrix of problem data with the selected values and determine the corresponding output value: (is the total number of samples)

$$c_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i(x\in R_m,m=1,2) \quad (3)$$

Among them$R_1(j,s)=\{x|x^{(j)}\le s\}$, $R_2(j,s)=\{x|x^{(j)}>s\}$. After completing the above steps, a decision tree is generated :(M is the total number of samples)

$$f(x)=\sum_{m=1}^{M}c_m I(x\in R_m) \quad (4)$$

Where is the indicator function:

$$I\begin{cases}1 & \text{if}(x\in R_m\\ 0 & \text{if}(x\notin R_m\end{cases} \quad (5)$$

In order to make the results more reliable, a loss function is needed to evaluate and analyze the regression effect. The square residuals and RSS are used to evaluate and analyze the sample data set:

$$RSS=\sum_{j-1}^{J}\sum_{i\in Rj}(yi-\tilde{y}_{Rj})^2 \quad (6)$$

The inner $\sum$ is the sum of the squared difference between the predicted index value and the true value of all the glass samples in the region, and the outer $\sum$ is the region divided by the matrix of all the glass samples[2].

In order to simplify the calculation, the recursive bisection method is used to simplify the results and verification process. Each split of the tree is split in the form of binary tree. After two child nodes were initially divided according to the characteristics and the best partition point, the relevant glass samples in this space were further divided into two parts from the current position. The partitioning method is top-down, constantly dividing the glass sample from the current position into two branches, with the appropriate greedy algorithm: each partition only considers the best of the current partition, and does not consider the previous partition.

A shard dimension and a shard point are selected for each shard to minimize the RSS results of the regression tree after shard:

$$R_1(j,s)=\{x|x_j<s\} \quad (7)$$

$$R_2(j,s)=\{x|x_j\ge s\} \quad (8)$$

$$RSS=\sum_{x_i\in R_1(j,s)}(y_i-\tilde{y}_{R1})^2+\sum_{x_i\in R_2(j,s)}(y_i-\tilde{y}_{R_2})^2(9)$$

In the process of optimizing the model by reducing the loss function, the model is easy to fall into the state of "overfitting". In order to avoid this kind of situation, regularization term and leaf node need to be introduced:

$$\sum_{m=1}^{|T|}\sum_{x_i\in Rm}(y_i-\tilde{y}_{R_2})^2+\alpha|T| \quad (10)$$

T represents the number of nodes of tree *T*. When the hyperparameter $\alpha>1$ is used, the more leaf nodes of the tree, the more complex the model is. The tree with more terminal leaf nodes will pay the price for its complexity, so the subtree with the minimum value of the above equation will become smaller.

3. Conclusion

This paper determined the existence rule of chemical composition on the surface of cultural relic samples by drawing the histogram of nuclear density and making a horizontal comparison. It can be concluded that silica, potassium oxide, phosphorus pentoxide, sulfur dioxide, and lead oxide accounted for a large proportion of the influence on whether the surface has weathered. In the prediction step, the CART was used to predict the data, and the detailed data of chemical composition content prediction before weathering was obtained. It is necessary to explore the correlation between data, and this paper selects correlation analysis as the solution. First of all, it is necessary to classify and process all the data, make one-to-one correspondence between the chemical components, show them in the form of charts, and visualize the processed data. Secondly, the correlation is preliminarily explored by drawing, and the variance, mean, and covariance of each group of data are calculated. Thirdly, the Pearson coefficient was used to further analyze the correlation, and the results were expressed in the form of a network graph, where the data with different correlation degrees were marked with different colors. Finally, the differences in chemical composition among different classes were analyzed. The final conclusion: in addition to tin oxide and sulfur dioxide, two kinds of chemical composition differences are weak, and the remaining chemical composition differences are strong.

4. Discuss

In this paper, the correlation between different chemical components was analyzed. In order to analyze the correlation of multiple variables, correlation analysis was selected to solve the problem. Correlation analysis refers to the analysis of two or more elements, so as to measure the degree of correlation between two factors, and further analysis of the strength and direction of the correlation. In order to make the calculation process data, Pearson's coefficient was introduced to measure the correlation degree (linear correlation) between two variables, and the specific Pearson's coefficient was obtained by calculating the quotient of covariance and standard deviation between two variables[3].

In the high potassium glass, the correlation between silica and calcium oxide and alumina is relatively strong, and the correlation with other components is weak. The correlation between potassium oxide and alumina, magnesium oxide, iron oxide and sodium oxide is relatively strong, and the correlation with other components is weak. The correlation between calcium oxide and magnesium oxide, iron oxide, lead oxide, phosphorus pentoxide is strong, and the correlation with other components is weak. Copper oxide has strong correlation with barium oxide, and relatively strong correlation with phosphorus pentoxide, tin oxide and sulfur dioxide. The correlation between magnesium oxide and alumina, iron oxide and phosphorus pentoxide is relatively strong. The correlation between alumina and iron oxide and tin oxide is relatively strong, and the correlation with other components is weak.

The correlation of iron oxide with phosphorus pentoxide and tin oxide is relatively strong and the correlation with other components is weak. The correlation of lead oxide with phosphorus pentoxide and tin oxide is relatively strong, but the correlation with other components is weak. Barium oxide has a strong correlation with sulfur dioxide and a relatively strong correlation with tin oxide. The correlation between phosphorus pentoxide and tin oxide and sulfur dioxide is relatively strong, and the correlation with other components is weak. The correlation between strontium oxide and sulfur dioxide is relatively strong, and the correlation with other components is weak.

In lead-barium glass, silica has no correlation with other components, sodium oxide has a strong correlation with calcium oxide and potassium oxide, but has a weak correlation with alumina and lead oxide. The correlation between alumina and iron oxide, phosphorus pentoxide, strontium oxide is high, and the correlation with other chemical components is low. Potassium oxide is highly correlated with calcium oxide, alumina and strontium oxide, but weakly correlated with other chemical components. Calcium oxide has high correlation with alumina, iron oxide, copper oxide and lead oxide. The correlation between iron oxide and copper oxide, phosphorus pentoxide, strontium oxide is strong, and the correlation with other components is weak.

The correlation between copper oxide and barium oxide is strong, but the correlation between copper oxide and other components is weak. The correlation between lead oxide and barium oxide is strong, but the correlation with other components is weak. The correlation between barium oxide and strontium oxide and phosphorus pentoxide was relatively strong, and the correlation between barium oxide and other components was weak. Phosphorus pentoxide was strongly correlated with strontium oxide, but weakly correlated with other components. The correlation between strontium oxide and tin oxide is relatively strong. There is no correlation between tin oxide and sulfur dioxide except as described above.

In many practical applications, K-means subspace clustering algorithm is used to detect and classify unknown glass samples. Although simple, the existing K-means subspace clustering algorithms usually use eigenvalue decomposition to generate approximate solutions, which reduces the efficiency of the model. To solve this problem, Wang et al. proposed a fast adaptive K-means subspace clustering model[4]. The model designed an adaptive loss function and provided a flexible clustering index calculation mechanism, which was suitable for data collection of different unknown glass samples[5]. The relationship between the weathering degree of the cultural relic surface and the type[6], ornamentation and color of the glass sample is of great importance for the detection of cultural relic glass. Selvi et al. proposed a new supervised adaptive genetic neural network model for the examination of glass samples[7], so as to facilitate the detection of the weathering degree of cultural relics. In view of the reliability of composition detection and data processing efficiency of ancient glass samples, Chen et al. The artificial neural network learning algorithm is adopted to efficiently optimize the detection of glass samples[8]. It is proved that ML algorithm has important application value, and provides a new design idea for the direction of precise development of AM manufacturing technology. The antiquities dating process uses experimentally derived diffusivity predicted from glass composition to convert the amount of surface diffusive molecular water into a calendar age. Its internal structural water content has been identified as a highly influential variable controlling the rate of water diffusion at ambient temperature.

References

[1] Stevenson, Christopher Rogers, Alexander Rogers and Michael Glascock. Variability of water content in obsidian structures and its importance in the hydration dating of cultural relics. Acta Archeologica Sinica: Report, vol. 23, pp. 231-242, 2019.

[2] Cheng Qi, Zhang Xu, Guo Jun, et al. Application of computed tomography in the analysis of glass beads unearthed from the Shanpu Tomb (Yutian) in Xinjiang Uygur Autonomous Region. Archaeology and Anthropology, vol. 11, no. 3, pp. 937-945, 2019.

[3] Sun H, Liu M, Li L, et al. A new classification method of ancient Chinese ceramics based on machine learning and component analysis. Ceramics International, vol.46, no. 6, pp. 8104-8110, 2020.

[4] Lijian Zhang, Guangfu Liu, "Mathematical Modeling for Ceramic Shape 3D Image Based on Deep Learning Algorithm", *Advances in Mathematical Physics*, vol. 2021.

[5] S. Wang et al., "K-Means Clustering With Incomplete Data," in IEEE Access, vol. 7, pp. 69162-69171, 2019.

[6] X. -D. Wang, R. -C. Chen, F. Yan, Z. -Q. Zeng and C. -Q. Hong, "Fast Adaptive K-Means Subspace Clustering for High-Dimensional Data," in IEEE Access, vol. 7, pp. 42639-42651, 2019.

[7] Selvi C, Sivasankar E. An Adaptive Genetic Neural Network (AGNN) model for Recommendatory Systems based on improved K-means clustering method. Multimedia Tools and Applications, vol.78, no.11, pp.14303-14330, 2019.

[8] Chen J, Almajed R. Progress in Auxiliary Ceramic Art Design Under Big Data.The International Conference on Cyber Security Intelligence and Analytics. Springer, Champ, pp.138-144, 2022.