Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.071113
ISSN 2371-8838 Vol. 7 Num. 11

Research Methods for Classification and Identification of Ancient Glass Types

Yang Chen[1,#], Yating Yang[1,#], Xinru Zhang[2], Xuan Zhu[1]

[1]*School of Physics, Changchun University of Science and Technology, Changchun, 130022, China*
[2]*School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun, 130022, China*
[#]*These authors contributed equally*

Keywords: CART, k-means, perceptual machine

Abstract: Ancient glass is susceptible to the influence of the environment of the burial site and then produce weathering, weathering will lead to changes in the proportion of its color and chemical composition, this paper analyzes the data of high-potassium glass and lead-barium glass, research on the weathering law of the glass artifacts, and classify and identify the type of glass. In order to classify the types of glass, this paper determines the best ccp_alpha of CART algorithm is located at [0,0.39296057] by cost pruning method, reduces the impurity of the classified tree to 0, and finds that the main difference between the classification of high-potassium glass and lead-barium glass lies in the content of PbO. The chemical compositions of different glasses are subclassified by K-means, and the number of nests of subclassified high-potassium glass and lead-barium glass is determined to be 4 and3 respectively with the help of SSE coefficients and profile coefficients, and the detailed subclassification is realized by CART algorithm. On the basis of the above, the prediction accuracy of Al-A8 glass types was accomplished by the perceptual machine model with 100% accuracy, and the results showed that the model stability and accuracy were high.

1. Introduction

Glass is a precious artifact from the Silk Road trade exchange [1]. In the process of making glass, different co-solvents need to be added in order to lower the melting point, and at the same time, different glasses are obtained [2]. The ancient glass is susceptible to the effects of the environment in which it was buried. Ancient glass is susceptible to weathering by the buried environment, and different weathering degree will have different weathering characteristics, which makes the analysis and identification of ancient glass products difficult. The analysis and identification of glass is primarily conducted through the utilization of chi-square testing and Q-clustering, as demonstrated by Huang Huiting, Li Chunming, Liu Siyu et al [3]. and Xu Hai, S. Hu, X et al., glass classification based on K-means clustering method [4], Zidong Z classifies glass by polynomial fitting mathematical expressions to determine the chemical composition of different glass. Ref [5]. Cao Jianyong, Xu Ting, Liu Yi et al used support vector machine algorithm to classify and predict glass types [6]. Jiang Shaoxuan, Chu Zhaoling, Li Jiaxiang et al found that PbO was the main difference between high potassium glass and lead barium glass through variability analysis [7]. In this paper,

with the help of 2022 mathematical modeling dataset and processing of missing values and outliers on the data, the difference division between high potassium glass and lead-barium glass is realized by CART algorithm, the subclassification of different glasses is realized by K-means algorithm, and the determination of the basis of the sub-non-classes is realized by the decision tree, and finally the prediction of A1-A8 glass types is realized by the perceptual machine algorithm.

2. Classification of glass types based on CART algorithm

2.1 CART algorithm

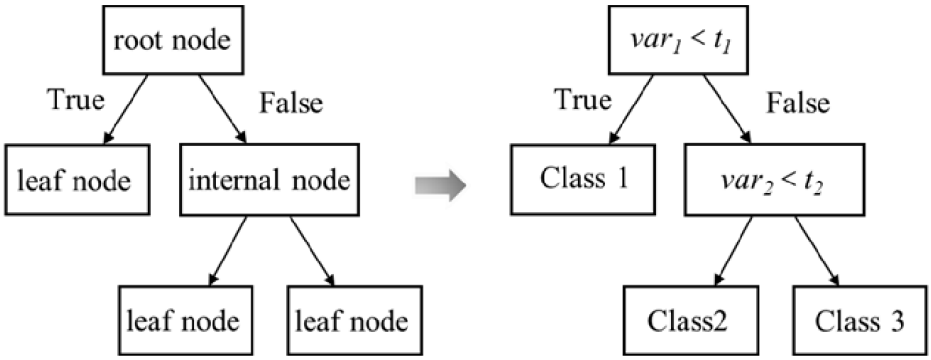


Figure 1: CART binary decision tree structure[8]

The CART algorithm is the most widely used decision tree learning method[9].It is suitable for handling discrete data with missing values by minimizing the Cini efficient Gini(p) criterion for feature selection. The CART tree is a bifurcated tree structure consisting of a root node, an intermediate node and a terminal node as shown in Figure 1 above. The CART algorithm splits each node into new child nodes based on its maximum features and the impurity of the split is measured by the error term of the loss function as in (2) below. Measurement.

$$Gini(p) = \sum_{i=0}^{i=k} p_i \overline{p}_i = 1 - \sum_{i=0}^{i=k} p_i^2 \tag{1}$$

$$Loss = \sum_{t=1}^{T} \frac{D_i}{D} L_i + \alpha T \tag{2}$$

Where D denotes the total number of samples, Di denotes the number of samples on the ith node, and Li denotes the loss function on the ith node.

In order to ensure that the impurity of the model is minimized, it is also necessary to prunethe CART algorithm, which is done in this paper using the cost complexity pruning(ccp) method.

The cost complexity pruning method is a top-down decision tree pruning method with a computational complexity $O(n^2)$ Compared with the error rate reduction pruning method, the complexity of pessimistic pruning method as well as the minimum

Error pruning method is $O(n)$ higher, but can get the optimal decision tree model.in order to improve the accuracy of the model, so this paper adopts the cost pruning method. After the pruning process to determine the appropriate value interval of ccp alpha, the impurity of the node is reduced to the minimum.

The main idea of the cost pruning method is that if clipping a node t in the decision tree reduces the complexity and impurity of the decision tree, the node will be clipped, otherwise the node is retained, the main setup of a metric α to realize the pruning method, and the node will be removed if

the value α after clipping the decision tree is less than the value α when the node is retained. The formula for calculating the value is as follows:

$$\alpha = \frac{R(T) - R(T_t)}{L(T) - 1} \tag{3}$$

Where $R(T)$ denotes the learning rate of the decision tree, $R(T_t)$ denotes the decision tree child, and $L(T)$ denotes the number of leaf nodes of the decision tree T.

The costly complexity pruning method is mainly done through the steps in the costly pruning method Table 1:

Table 1: Cost complexity pruning method

Cost complexity pruning method:
1) Input Decision Tree Model T; 2) Finding the node with the lowest cost complexity parameter among all non leaf nodes; 3) Do Loop pruning of nodes with the lowest cost complexity parameter; 4) Number of prunings output $\{T_0, T_1, ...T_n\}$ 5) Determine appropriate ccp_ Alpha interval and minimum impurity; 6) return T;

2.2 Modeling

In this paper, CART decision tree building is mainly divided into decision tree generation and pruning.

Step 1: Decision Tree Generation

The CART decision trees are classified based on the Gini coefficient, which is denoted by the Gini coefficient of CART given the dataset D and feature A:

$$Gini(D \mid A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{4}$$

Where D1 and D2 denote the 2 datasets after the classification of feature A. Gini (D|A) denotes the uncertainty after segmentation, and the smaller the Gini(D|A) the higher the model accuracy.

The Gini(D|A) can be calculated for each feature as in Table 2:

Table 2: Gini(D|A) coefficient

Type of feature	PbO	Other features
Gini(D|A)	0.0	0.393

From Table 2 $Gini(D \mid A = PbO)$ is the smallest, so PbO is chosen as the optimal cut-off point. Assign the dataset inside the two sub-nodes according to the features in turn.

Step 2: Pruning of decision trees

In order to reduce the complexity and impurity of the decision tree, the decision tree needs to be processed by the pruning algorithm, and the output CCP path through model fitting is.

$$ccp\_alpha = [0.0 , 0.39296057] \tag{5}$$

$$\text{impurities}=[0.0 , 0.39296057] \tag{6}$$

ccp path When$0 \leq \alpha \leq 0.39296057$, the impurity of the decision tree is 0.39296057. Setting the

ccp_alpha parameter of the decision tree in the interval [0, 0.39296057] reduces the impurity of the decision tree in the computation interval from 0.39296 at the default ccp_alpha to zero.

2.3 Model testing and results

The binary tree structure of this CART decision tree is shown in Figure 2:

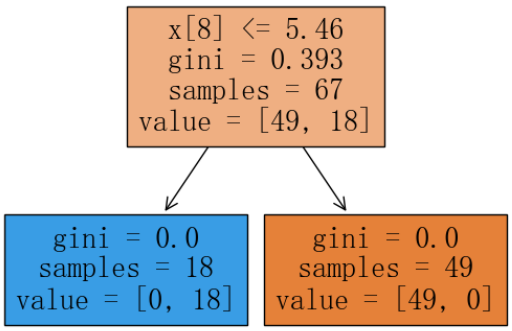
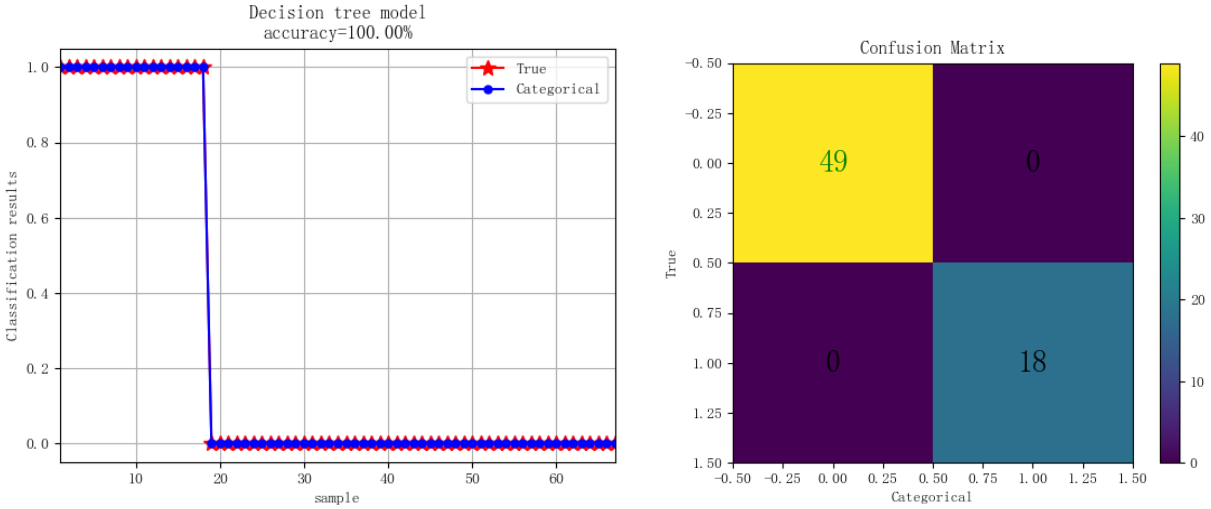


Figure 2: CART binary decision tree structure

The accuracy of the model and the confusion matrix for classifying the glass into high potassium glass ($PbO \leq 5.46$) and lead-barium glass ($PbO > 5.46$) by PbO content are shown below in Figure 3(a) and (b). As shown above the model is 100% accurate and the model is highly accurate.



(a)model accuracy (b)Confusion Matrix

Figure 3: The accuracy of the model

3. K-means based subclassification of chemically composed glasses

3.1 SSE coefficients and contour coefficients

The sum of squared errors (SSE) within a group is an important indicator in the clustering algorithm to determine whether the model is optimal or not, the smaller the SSE is, the better the model is under the same k-value clustering model. The formula for SSE is as follows:

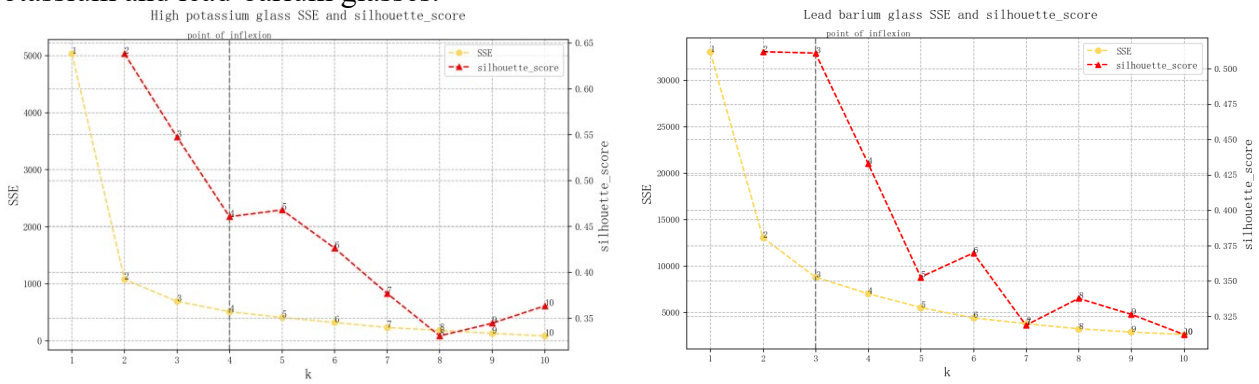
$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{m} w^{(i,j)} = \| x^{i} - \mu^{j} \|_{2}^{2} \tag{7}$$

The silhouette coefficient is an indicator of how good the clustering is, and it consists of the degree

of cohesion$a\ i$ by the degree of internal aggregation and the degree of separationbi. The formula is as follows:

$$S(i)=\frac{b_i-a_i}{\max\{a_i,b_i\}} \tag{8}$$

The optimum number of nests for classification can be determined by the relationship between the SSE coefficients and the profile coefficients, where the "elbow" of the SSE coefficients and the profile coefficients is the true number of nests. Figure 4 (a) and (b) below show the number of nests for high-potassium and lead-barium glasses.



(a) The number of clusters of high-potassium glass is K=4 (b) The number of clusters of lead barium glass is K=3

Figure 4: Count the result

3.2 K-means based subclassification

The main core of the K-means algorithm is the selection of the optimal division method D* by minimizing the loss function (9)[10], which uses Euclidean distance as the distance between samples $dist_{ij}$.

$$LOSS(D)=\sum_{l=1}^{k}\sum_{D(i)}\|x_i-x_l\|^2 \tag{9}$$

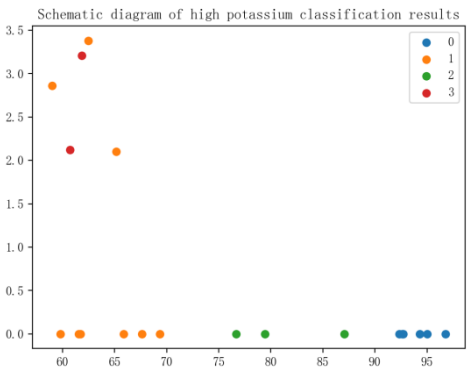
$$dist_{ij}=\|x_i-x_j\|^2 \tag{10}$$

In the above equation (9) x_l denotes the mean or center of the first l class.

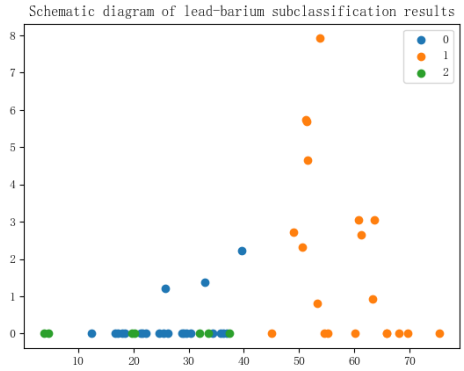
The K-means algorithm is an iterative process that first selects the centers of the k nests, assigns the samples to the nearest nests one by one, and then updates the expectation of each class as the new nest center, and repeats the above steps so as to solve the optimization problem to obtain the optimal division method D*:

$$D^*=\arg\min LOSS(D)=\arg\min\sum_{l=1}^{k}\sum_{D(i)=l}\|x_i-x_l\|^2 \tag{11}$$

The clustering diagrams of high potassium glass and lead barium glass obtained by solving the above model are shown in (a) and (b) of Figure 5 as follows.



(a) Cluster plot of high-potassium glass



(b) Cluster plot of Lead-barium glass

Figure 5: Cluster plot

3.3 Determination of subcategorization components based on CART decision tree

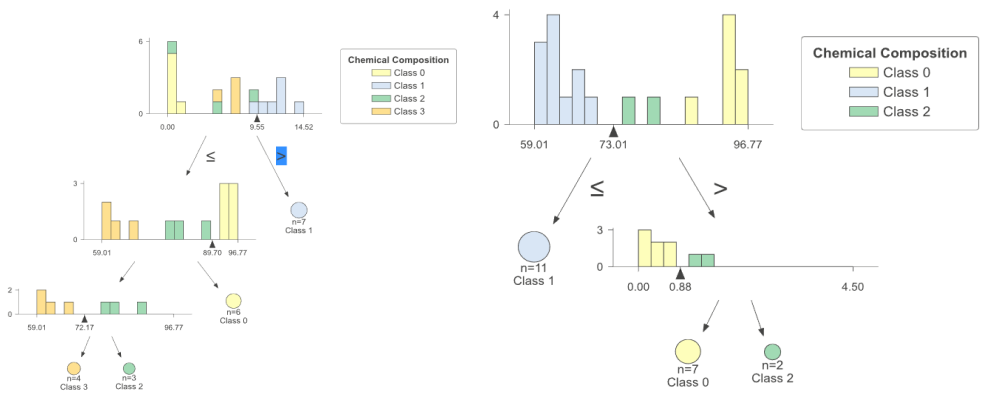
The classification of high potassium glass and lead-barium glass was determined by the k-means algorithm, but the specific differences in chemical composition between the subclassifications were not determined, and the subclassification chemical composition differences were determined by the CART decision tree. The steps of realization are shown in 2.2. The decision tree structure for high potassium glass and lead-barium glass is in Figure 6 , (a) and (b). The results of the classification are shown in Table 3 and Table 4.

Table 3: High-potassium glass subclassification results

0	1	2	3
High calcium-high silica glass	High calcium-low silica glass	High calcium-medium silica glass	Low calcium glass

Table 4: Lead-barium glass subclassification results

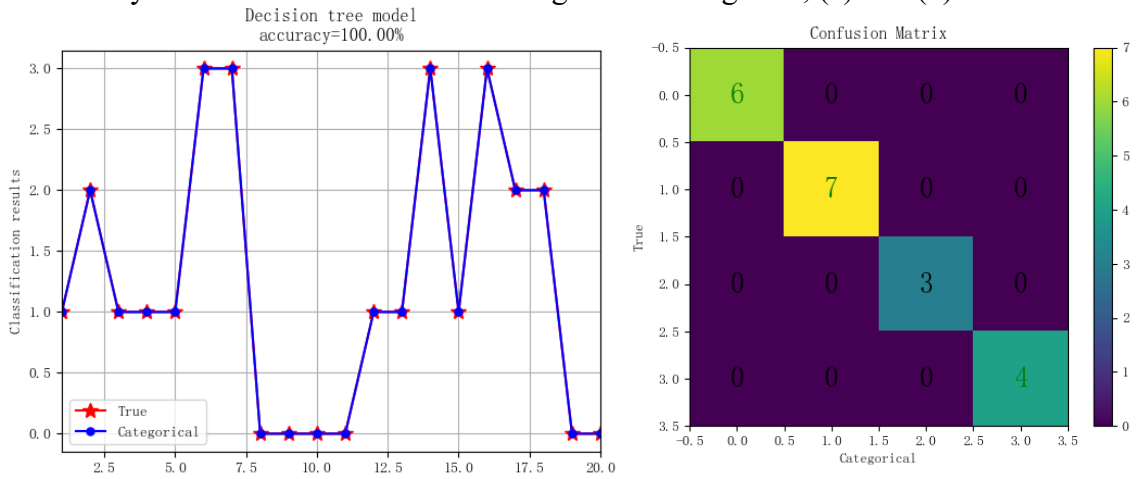
0	1	2
Low silicon-low phosphorus glass	High silicon-low phosphorus glass	High phosphorus glass



(a) High-potassium glass decision tree structure (b) Lead-barium glass decision tree structure

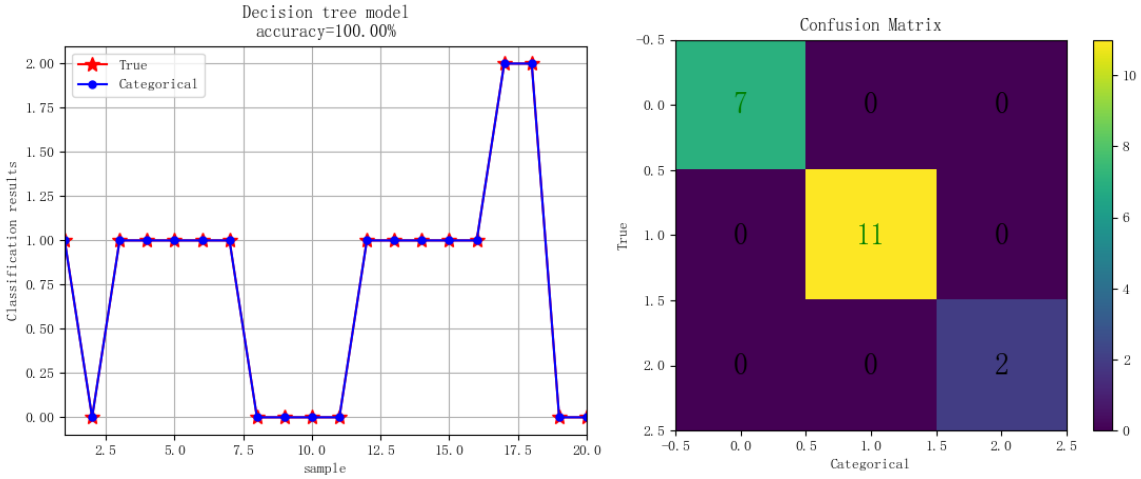
Figure 6: Decision tree structure

The accuracy of the model is as follows in Figure 7 and Figure 8, (a) and (b).



(a) (b)

Figure 7: High potassium decision tree accuracy



(a) (b)

Figure 8: Lead-barium decision tree accuracy

4. Glass prediction based on perceptual machine modeling

4.1 Perceptual machine model

Perceptron is a binary classification algorithm that inputs the corresponding feature space and outputs a categorical 1 or -1 expression for the feature space as in (12).

$$\begin{cases} f(x) = sign(w) \\ w = \omega \cdot x + b \end{cases} \tag{12}$$

Among them:

$$sign(x) = \begin{cases} 1, x \in D_1 \\ -1, x \notin D_0 \end{cases} \tag{13}$$

The input feature space is divided into 2 parts i.e. division plane by w when $x \in D1$, outputs 1,

i.e., high potassium glass, and when $x \notin D0$ when, output -1 i.e. lead barium glass.

In order to ensure that the loss function (14) is the smallest and the classification effect is the best delineation, this paper adopts the stochastic gradient descent method, through continuous iteration until there is no error classification point, and then optimize the location of the delineation plane, the specific method is as in (15):

$$L(w,b) = -\sum_{x_{i\in D}} (w \cdot x_i + b) \tag{14}$$

$$\begin{cases} \nabla_D L(\omega,b) = -\sum_{x_i \in D} y_i x_i \\ \nabla_D L(\omega,b) = -\sum_{x_i \in D} y_i \end{cases} \tag{15}$$

Step 1: Input feature space, due to the small dataset of unknown glass, this paper adopts the dataset of 2.2 as the training data of the perceptron.

Step 2: Select the initial, bring in (13)-(15), start the iterative process until the loss is minimized and there is no error point, in this paper, through 56 iterations, the loss function of the model is reduced to $loss = 0.00339607$ close to 0, the model works well.

Step 3: Output the classification result as Table 5:

Table 5: Predict the outcome

A1	A2	A2	A3	A4	A5	A6	A7	A8
1	-1	-1	-1	-1	-1	1	1	-1

5. Conclusion

In this paper, through the in-depth study of glass data, CART classification model, the impurity of classification is reduced to 0 by cost complexity pruning method, and PbO is determined as the main difference between high potassium glass and lead-barium glass. The K-means-CART classification model was also established, i.e., the class of each glass was determined by the K-means algorithm, and then the difference between the subclasses was determined by the decision tree algorithm, which categorized the high potassium glass into four classes: high calcium-high silica glass, high calcium-low silica glass, high calcium-medium silica glass, high calcium-medium silica glass, high calcium-medium silica glass, high calcium-low silica glass, and high potassium glass. The high potassium glass is divided into four categories: High calcium-high silica glass, High calcium-low silica glass, High calcium-medium silica glass, Low calcium glass; the lead-barium glass is divided into three categories: Low silicon low phosphorus glass, The lead-barium glass is classified into three categories: Low silicon low phosphorus glass, High silicon low phosphorus glass, High phosphorus glass, and finally, through the perceptual machine model, it is predicted that A1, A6, and A8 are high potassium glass, and A2-A5, and A7 are lead-barium glass, with an accuracy rate of 100%.

This paper predicts and divides the types of ancient glass by their chemical composition content. However, for the glass artifacts unearthed in the future, it is difficult to know their specific chemical composition, and out of the principle of protection of cultural relics, it is also difficult to directly measure the chemical composition content of cultural relics. The problem can be well solved by establishing a known detection model through the classification algorithm, i.e., extracting the color, chemical composition and other characteristic data through the information of the known cultural relics of glass, establishing a database of glass relics, training the relics model, and providing a

solution for the classification and prediction of cultural relics of glass.

References

[1] Liu Shuna. Study on Glassware of Nomadic People in Ancient Northern China [D]. Inner Mongolia Normal University,2022.
[2] Li Mo. Preparation of faience products and research on lead-barium glass during the Warring States, Qin and Han Dynasties [D]. Beijing University of Chemical Technology, 2015.
[3] Huang Huiting, Li Chunming, Liu Siyu et al. Compositional analysis and identification of ancient glass products based on compositional data[J]. Mathematical Modeling and its Applications, 2023, 12(02):52-62+124.
[4] H. Xu,S. Hu,X. Yao et al. Research on the composition of glass artifacts based on K-Means clustering and gray correlation analysis[J]. Journal of Xinjiang Normal University (Natural Science Edition), 2023, 42(03):66-73+96.
[5] Zidong Z. Mathematical model for composition analysis and identification of ancient glassware[J]. Modern Information Technology, 2023, 7(14):88-93+98.
[6] Cao JY, Xu TY, Liu Y, et al. Composition prediction and classification of ancient glassware based on RBF and SVM[J]. Science and Technology Innovation and Application, 2023, 13(18):41-43+48.
[7] Jiang Shaoxuan, Chu Zhaoling, Li Jiaxiang et al. Correlation analysis of chemical elements in ancient glass artifacts[J]. Chemical Engineering and Equipment, 2023(07):23-25.
[8] Zhou, Meiqin. Research on unit cost gain sensitive decision tree classification algorithm and its pruning algorithm [D]. Guangxi Normal University, 2016.
[9] Li H. Machine learning methods [M]. Beijing: Tsinghua University Press, 2022.
[10] Zhou Z. Machine learning [M]. Beijing: Tsinghua University Press, 2016.