Classification and identification of ancient glass objects based on decision tree model

Jiaqi Zhang [1, *], Ming Tang[1], Zhiyuan Gao[2]

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, SiChuan (zip code:611731), China

[2]Glasgow College, University of Electronic Science and Technology of China, Chengdu, SiChuan (zip code:611731), China

*Corresponding author: 2020010912016@std.uestc.edu.cn

Abstract: Glass products were one of the important goods on the Silk Road. In this paper, a decision tree model is adopted to determine the law of classification of high potassium and lead-barium glass and to obtain the criteria for classification thus illustrating the law of classification. In order to subclassify the two categories of high potassium and lead-barium, k-means clustering is used for unsupervised classification. Once a classification model was established, the unknown classes of glass artifacts were then classified using this model.

Keywords: Decision trees; k-means clustering; glass products.

1. Introduction

Glass products were one of the important goods along the Silk Road and were introduced to China through the regions of West Asia and Egypt[1-2]. The chemical composition of ancient glass made by Chinese learning methods varies depending on the raw material[3]. For glass products, the most important component is silicon dioxide (SiO2). During calcination, a flux is added to lower the melting point and limestone (with calcium content) is added as a stabilizer[4-6]. Depending on the type of flux, the glass may contain different chemical elements, such as lead oxide in lead-barium glass, barium oxide and potassium oxide in high-potassium glass [7-8]. However, weathering due to storage conditions can alter the chemical composition of the glass and thus affect the determination of the type[9]. To classify and identify these glass artifacts, this paper uses a decision tree model to develop a classification model to accomplish the above task[10].

2. Build a glass classification model

2.1 Decision tree analysis of glass classification

The question preemptively classifies glass artifacts into two categories: lead-barium glass and high-potassium glass. The basis for this classification is analyzed in this question below.

From the initial data analysis, it is clear that there are significant differences in the elemental content of silica, potassium, barium and lead between lead-barium glass and high-potassium glass; also the elemental content varies before and after weathering. We therefore considered using the data predicted in the first question to ensure that all data were in the unweathered condition for analysis.

Delineating the types of glass is a classification problem, and decision trees have the advantage of having fewer training samples, no data cleaning such as normalization, for example, and visualization. The data volume of this problem is relatively small, so the decision tree model is suitable for the analysis of the classification law of high potassium and lead-barium glass. The specific modeling and solving process is as follows。

(1) Decision tree model

A classification decision tree model is a tree structure that describes the classification of instances. A decision tree consists of nodes and directed edges, and there are two types of nodes: internal nodes, which represent a feature or attribute, and leaf nodes, which represent a class. Each path from the root

node to the leaf node of the decision tree constructs a rule, the features of the internal nodes of the path correspond to the conditions of the rule, and the classes of the leaf nodes correspond to the conclusions of the rule. The purpose of constructing a decision tree is to find out the relationship between attributes and classes and use it to predict the classes of records of unknown classes in the future.

The essence of constructing a decision tree is the process of continuously grouping the training sample set. The branches on the decision tree are gradually grown during the process of grouping the sample set, and the core technique is the test attribute selection problem. In order to adapt to the need of dealing with large-scale data sets, an improved algorithm was later proposed to use the Gini index as an attribute selection criterion.

The Gini coefficient is related to the information gain, and the smaller the Gini value, the greater the information gain, which means the better the classification effect. If a decision tree classifies a set of data well, it is often difficult to avoid the problem of overfitting, and to avoid overfitting, it is often necessary to prune the decision tree. Pruning is usually divided into pre-pruning and post-pruning. Pre-pruning is a top-down pruning, in which each node is estimated in the process of decision tree generation, and if the division of the current node cannot improve the generalization performance of the decision tree, then the division is stopped and the current node is marked as a leaf node; post-pruning is a bottom-up pruning, in which a complete decision tree is trained first, and then the non-leaf nodes are examined bottom-up, and if replacing the non-leaf nodes with leaf nodes can bring If replacing a non-leaf node with a leaf node can bring improvement in generalization performance, the node will be set as a leaf node.

(2) Solving the decision tree model

The above model is solved by SPSS, and the structure of the decision tree is shown in Figure 1.

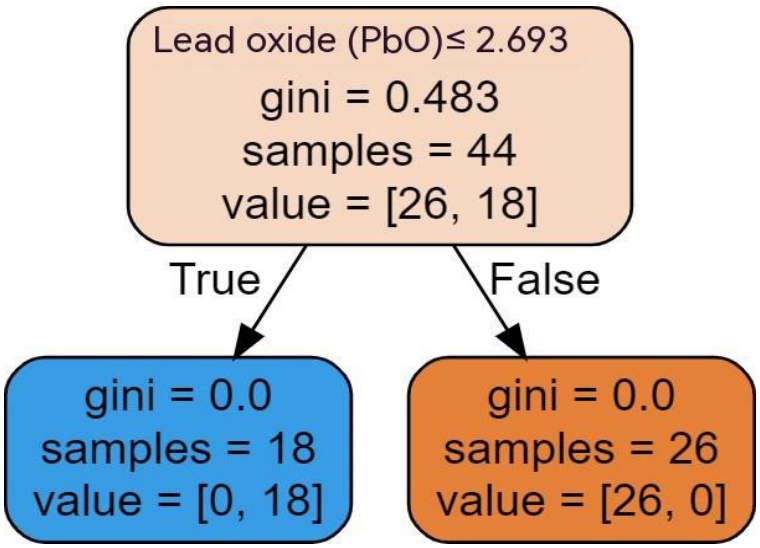


Figure 1 Decision tree structure diagram

Based on the results, it is clear that the only decisive variable for classification is lead oxide. For samples with PbO content less than 2.693, we consider them as high potassium glass, otherwise they are lead-barium glass. The accuracy of the model validation results is around 95%, which indicates that the decision tree model can accurately judge the glass category and the extracted judgment criteria are reasonable.

2.2 K-mean clustering algorithm for constructing subclasses

Since the relationship between the compounds is unknown, set for the subclass classification of glass artifacts, the overall consideration of the chemical composition of high potassium and lead-

barium glass separately for cluster analysis, to get the respective category labels that can be assigned, and then will hit on the category

(1) Steps of K-mean clustering algorithm

Input: training data set D, conditions for stopping the calculation; Output: CART decision tree.

According to the training data set, the following operations are performed recursively for each node, starting from the root node, to construct a bifurcated decision tree:

S1: Set the training data set of the node as D, and calculate the Gini index of the existing features for this data set. In this case, for each feature A, for each possible value a, D is divided into two parts D1 and D2 according to the "yes" or "no" test of the sample points for A=a, and the Gini index for A=α is calculated using equation below:

$$Gini(p) = 1 - \sum_{k=1}^{K} \frac{|C_k|^2}{|D|}$$

C_k is the subset of samples in D that belong to the kth class, and k is the number of classes.

S2: Among all possible features A and their possible cut points α, select the feature with the smallest Gini index and its corresponding cut point as the optimal feature and the optimal cut point. Based on the optimal feature and the optimal cut point, two sub-nodes are generated from the present node, and the training data set is assigned to the two sub-nodes according to the features.

S3: Call S1and S2 recursively for the two sub-nodes until the stopping condition is satisfied.

S4: Generate a CART decision tree. The algorithm stops when the number of samples in the node is less than a predetermined value, or the Gini index of the sample set is less than a predetermined value (the samples are basically in the same class), or there are no more features.

A decision tree analysis is performed on the labeled data to obtain a decision tree structure with good generalization of the importance of the indicators for classification. The clustering is done using the kmeans clustering method, and the elbow method is used to determine the number of clusters k。

(2) Clustering results and description

After determining the number of clusters, we programmed the solution to obtain the cluster centers and the classes to which each sample belongs. For sample, it is represented by the element content vector ci. The results are shown below.

(1) High potassium glass

For high potassium glass, it is divided into 4 subclasses.

(2) Lead-barium glass. For lead-barium glass, it is divided into 6 subclasses.

The results are shown in the figure 2 and 3.

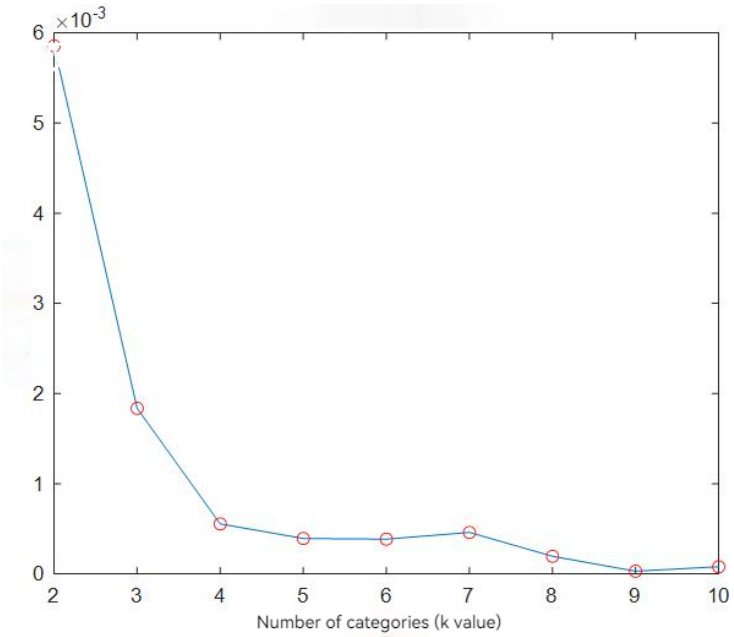


Figure 2 Variation of SSE of high potassium glass with the number of clusters

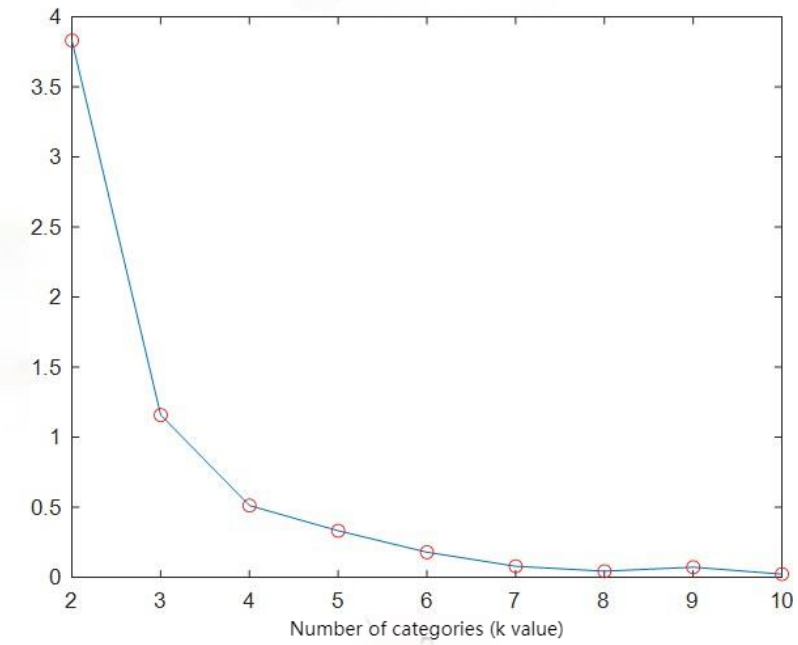


Figure 3 SSE of lead barium glass

2.3 Decision tree to determine the division basis

The results obtained from clustering are only the classified samples, and the basis of sample division is not clear. In the first part of Problem 2, we used the decision tree model to make a quantitative analysis of the classification of the classified data, so this question uses the same method to clarify the chemical composition of the classification. Using SPSS to solve the model, the structure of the decision tree was obtained as shown in Figure 4.

(1) High potassium glass

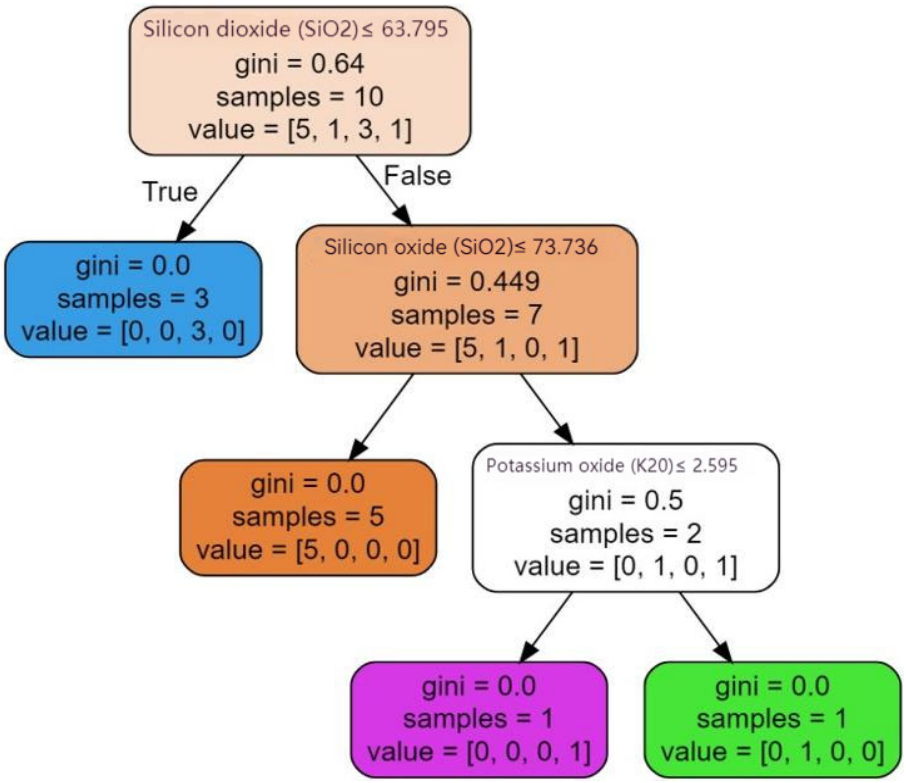


Figure 4 Decision tree structure of high potassium glass

As can be seen from Figure 4, the classification of subclasses for high potassium glasses is based on the content of silica and potassium oxide. By setting the thresholds for silica and potassium oxide, the high-potassium glasses are classified into four categories： w(SiO2 ⩽ 63.795); w(63.795 ⩽ SiO2 ⩽ 73.736);

w(SiO2 ⩾ 73.736andK2O ⩽ 2.595); w(SiO2 ⩾ 73.736andK2O ⩾ 2.595).

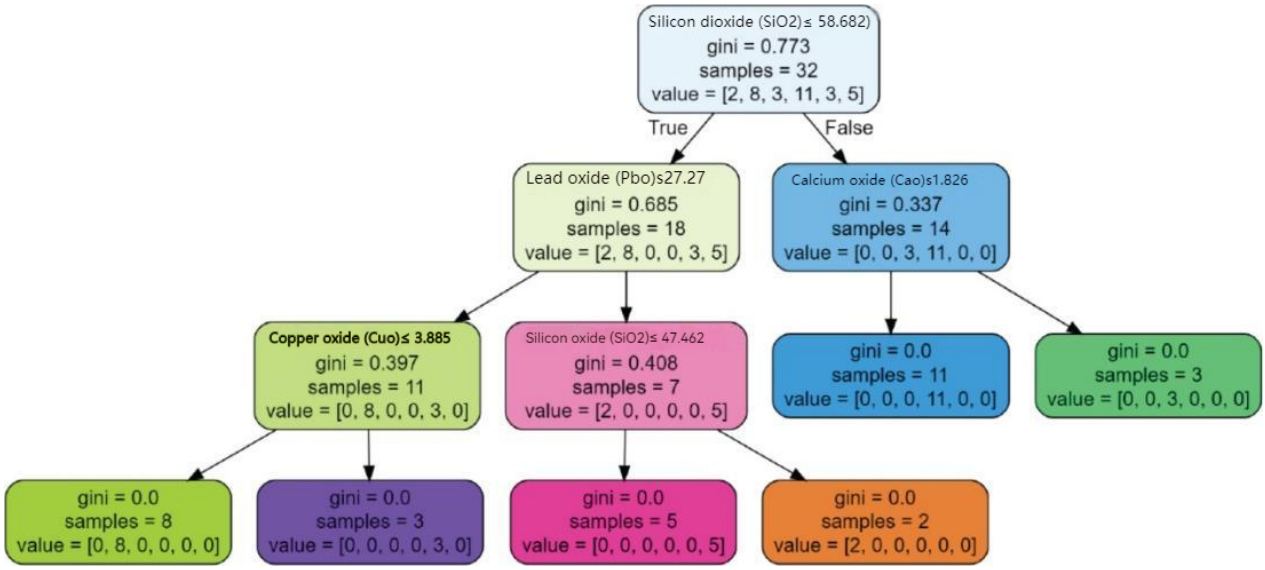


Figure 5 Decision tree structure of lead barium glass

As can be seen from Figure 5, the classification of subclasses for high potassium glasses is based on the content of silica, lead oxide, copper oxide and calcium oxide. By setting the thresholds for silica and potassium oxide, the high-potassium glasses were classified into six categories.

3. Classification of the glass to be tested

3.1 Classification and detection of weathering data

The weathered data were computationally transformed into unweathered data for classification. This question considered first reversing the process to determine the category of weathered glass in Form 3 by deriving the basis for classification of high potassium and lead-barium glass in the weathered case. For the unweathered data, the prediction model of Problem 1 was used to convert them into weathered data. For the weathered data, a decision tree was used to determine the basis for classifying high potassium and lead-barium glasses. The results of the decision tree are shown in Figure 6.

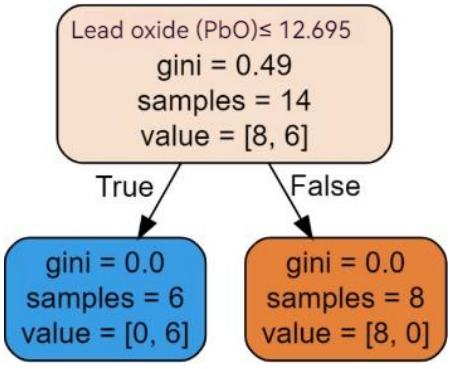


Figure 6 Decision tree structure of weathered glass

As shown in Figure 6, samples with lead oxide content below 12.695 were classified as high potassium glass, and those above 12.695 were lead-barium glass. The accuracy of the decision tree model was verified to be 100%. Applying the above analysis, four of the final weathering samples

were classified as high potassium glass and the other four as lead-barium glass. The specific results are shown in Table 1.

Table 1 Sample binary classification results

Artifact Number	Glass Types	Artifact Number	Glass Types
A1	High potassium glass	A5	High potassium glass
A2	Lead barium glass	A6	High Potassium Glass
A3	Lead barium glass	A7	High Potassium Glass
A4	Lead barium glass	A8	Lead barium glass

3.2 Classification of unlabeled data

For the data that have been dichotomized, we bring them into the decision tree model of Problem 2 separately to obtain the classification results of their subclasses. The model solution results are shown in Table 2.

Table 2 Classification results of sample subclasses

Artifact Number	Glass subcategories	Artifact Number	Glass subcategories
A1	High potassium glass 4	A5	High Potassium Glass 3
A2	Lead barium glass 3	A6	High Potassium Glass 3
A3	Lead barium glass 6	A7	High potassium glass 1
A4	Lead barium glass 2	A8	Lead barium glass 2

4. Conclusions

In order to establish the classification model of glass, this paper firstly adopts the decision tree model to determine the classification law of high potassium and lead-barium glass, and obtains the classification law: high potassium glass with lead oxide content less than 2.693%, otherwise it is lead-barium glass; secondly, the k-means clustering method is used to classify the two categories of high potassium and lead-barium into subclasses, and obtains the result that high potassium glass is divided into 4 classes and lead-barium glass is divided into 6 classes. After the classification model was established, this model was applied to classify the glass artifacts of unknown categories. First, by predicting the chemical composition before and after weathering and the classification law of high potassium and lead-barium, the classification law of high potassium and lead-barium under weathering was obtained, that is, the content of lead oxide is less than 12.695% for high potassium glass, otherwise it is lead-barium glass, and then substituted into the model for solution and sensitivity analysis, the category of artifacts can be obtained.

References

[1] Wu B., Yu D.. Deep convolutional neural network-based classification and detection of cell phone glass cover surface defects[J]. Software Engineering,2021,24(12):6-10.DOI:10.19644/j.cnki.issn2096-1472.2021.012.002.

[2] Huang JY. Research on the detection and classification method of cell phone panel glass defects based on digital image processing [D]. Xi'an University of Electronic Science and Technology, 2020. DOI:10.27389/d.cnki.gxadu.2020.000709.

[3] Wang Yumei. Glass fiber classification definition and code [C]//. Glass fiber products standard dissemination meeting and inorganic fiber technology for industrial use seminar proceedings. ,2019:9-20. doi:10.26914/c.cnkihy.2019.068339.

[4] Kong Yang. Study on the precrystallization behavior of fluoroaluminosilicate system glass[D]. Qilu University of Technology,2019.

[5] Lu Yue. Deep learning-based classification and detection of cell phone glass defects[D]. Zhengzhou University,2019.

[6] .Luo Yan,Cui Yan,Liu Qiang,Li Xiao. Research on classification and grading standards for waste glass[J]. China Resource Comprehensive Utilization,2018,36(08):80-83.

[7] The basic classification of glass curtain wall[J]. Building materials development orientation,2017,15(04):94.

[8] Jin Dabing,Feng Fuming,Xu Shiqing,Liu Shimin. Classification of glass homogeneity streak images and its algorithm[C]//.Proceedings of the 2015 National Conference on Glass Science and Technology. ,2015:256-259.

[9] Fu Xingliang. Preparation, characterization and adsorption properties of porous glass[D]. Tianjin University, 2010.

[10] Ai Jieyan. Research on color classification and defect detection of wall and floor tiles based on computer vision[D]. South China University of Technology, 2003.