Research on glass products classification based on machine learning

Wenhao Luo*, Yao Lu

Department of Electrical Engineering and Automation, Lanzhou Jiaotong University, Lanzhou, China, 730070

* Corresponding Author Email: mageag6@gmail.com

Abstract. In order to analyze the classification rules of ancient Chinese high-potassium glass and lead-barium glass and select the appropriate chemical composition for each category to divide them into subcategories, this paper models the type and chemical composition of glass by random forest algorithm, divides the data by 3:7 and uses 3-fold cross-validation, and uses Accuracy, Recall, Precision and F1-score to evaluate the model after training, and the results are optimal. The strongly correlated chemical components under each glass category were obtained by using the K-Means clustering algorithm, and the clustering results were visualized by using unsupervised learning, and the subclass division results were better when k was taken 2 after visualizing different cluster data. Finally, the results of subclassification are visualized and the rationality and sensitivity analysis of the method are carried out, and it is found that the method is accurate.

Keywords: Random Forest Algorithm, Artifact Identification, K-Means Clustering algorithm.

1. Introduction

The Silk Road occupied a very important position in ancient times and was an important channel for trade and cultural exchange between China and foreign countries, of which glass was a valuable physical evidence of early trade. As early as the Warring States period, glass was used as decorative items, and later on, from Han to Sui and Tang, Song Dynasty, all periods have been found. Especially the glass from the Warring States period to the Western Han period, the number and variety of glass is a rare material to study the history of the development of ancient glass in early China. Although the appearance of our ancient glass and foreign glass products are very similar, but the chemical composition contained in it is different. The main raw material of glass is quartz sand, the main chemical composition is SiO_2 [1]. The melting point of pure quartz sand is high, so a flux is added to lower the melting temperature during refining. Limestone is also added as a stabilizer, which is converted to Cao after forging. The main chemical composition of the added fluxes differs from one another, which allows us to distinguish the type of glass and its origin. In the meantime, ancient glass is subject to weathering due to environmental influences, and its internal elements are heavily exchanged with those of the environment, resulting in changes in composition ratios, so the correlation between glass type, texture, color and surface weathering, and internal chemical composition needs to be analyzed comprehensively. The research results and data processing methods will help to make some new breakthroughs in exploring the cultural and technological exchange of ancient glass on the Silk Road, which is of great significance to the exploration and development of Chinese culture and Chinese civilization.

Since these two ancient glasses were not pure enough, there were subclasses, and the number and characteristics of their subclasses were still unclear, so the team determined the number of subclasses by modeling and analyzed them.

2. Model selection

2.1. Random forest algorithm

The random forest algorithm can be used for classification problems and regression problems, and the algorithm has high prediction accuracy and good generalization performance. The idea of using

random forest for feature extraction is relatively simple, i.e., after establishing a random forest look at how much contribution each feature makes on each tree in the random forest, and then take the average of the contributions made, conduct a comparison between different features i.e., clarify the importance score of each feature on each tree in the random forest and then take the average value to compare the contribution size between the features [2-3]. The algorithm has been widely certified for its efficiency and practicality in data mining and fault diagnosis, and the random forest model diagram is shown in the Figure 1.

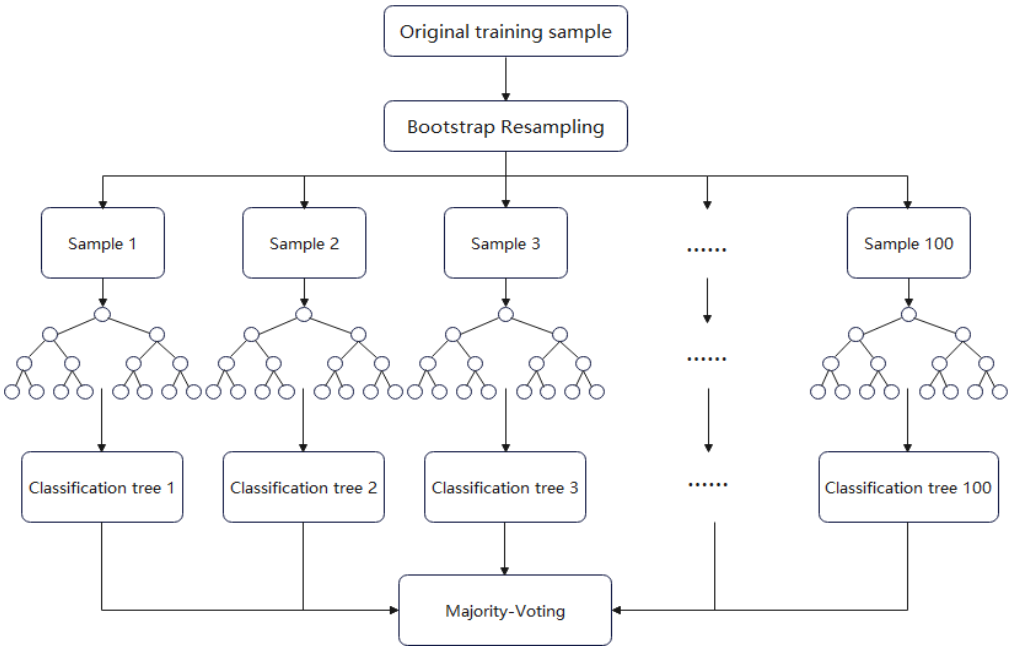

Figure 1. The random forest model with 100 number of classification trees in this paper

2.2. K-Means clustering algorithm

The K-Means Clustering algorithm is a cluster analysis algorithm, which is based on an iterative solution method to solve the problem, the steps of the algorithm are: first divide the given data into K groups and randomly select K objects from them as the initial clustering centers of the algorithm, then by calculating the distance between each element and all the clustering centers, then assign each element to the clustering center nearest to it according to the calculated result. The distance between each element and all the cluster centers is calculated, and then each element is assigned to the cluster center closest to it according to the result of the calculation. The cluster centers and all elements assigned to them represent one cluster [4-6].

For each sample assigned, the cluster centers are recalculated and reassigned based on the properties of all elements in that cluster. This algorithmic process will be repeated until some termination condition is met. The termination condition is that no (or a predetermined minimum number of) new elements are assigned to other clusters, or no (or a predetermined minimum number of) cluster centers are changed again, at which point the error sum of squares is locally minimized. The steps of the algorithm are shown in the Table 1.

Table 1. Steps of K-means algorithm

Input: sample set D=$\{x_1, x_2, \ldots, x_m\}$;
Cluster number k.
Process:
1: k samples are randomly selected from D as the initial mean vector $\{\mu_1, \mu_2, \ldots, \mu_k\}$
2: **repeat**
3: $C_i = \emptyset(1 \le i \le k)$
4: **for** j=1, 2, …, m **do**

5: Calculate the distance between sample x_j and each mean vector $\mu_i(1 \le i \le k)$: $d_{ji} = \|x_j - \mu_i\|_2$;
6: Determine the cluster marker of x_i based on the nearest mean vector: $\lambda_j = argmin_{i\in\{1,2,\ldots,k\}} d_{ji}$;
7: Divide sample x into the corresponding cluster: $C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$;
8: **end for**
9: **for** i=1, 2, …, k **do**
10: Compute the new mean vector:$\mu_i' = \frac{1}{|C_i|}\sum_{x\in C_i} x$;
11: **if** $\boldsymbol{\mu_i' \neq \mu_i}$ **then**
12: Update the current mean vector μ_i to μ_i'
13: **else**
14: Keep the current mean vector constant
15: **end if**
16: **end for**
17: **until** None of the current mean vectors have been updated

Output: Cluster division C=$\{C_1, C_2, \ldots, C_k\}$

3. Data pre-processing

3.1. Filling in missing values

By consulting the data that are derived from the 2022 Higher Education Cup National College Students Mathematical Contest in Modeling, this study obtained the basic information and related data of chemical composition of a batch of ancient Chinese glass products. Of the 69 pieces of chemical composition data collected, 14 kinds of chemical composition can be detected by archaeologists on the surface of glass relics. Due to technology, detection means that some chemical elements may not be detected, resulting in the position of the element being displayed as missing values (or null values). To facilitate the subsequent use of relevant models to model and analyze this data, the team uses Excel tools to process missing values based on model assumptions, assuming that undetected data does not affect the calculation of subsequent results. Therefore, replace these missing values (null values) with 0.

3.2. Max-Min normalization

The major chemical compositions of the glass artifacts include SiO_2, Na_2O, K_2O, AI_2O_3, and other 14 chemical elements, demonstrating significant differences in chemical composition at different artifact sampling sites, which are highly correlated with the glass types of glass artifacts.

After observation, it can be seen that there is a problem of excessive difference in the data range of each variable in this dataset, and in order to solve this problem, we performed Max-Min normalization on the collected data. The normalization formula is as follows [7].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Where x' denotes the normalized value by Max-Min, x is the original value, x_{min} is the minimum value of the current feature, and x_{max} is the maximum value of the current feature.

Through the normalization operation, all the data variables in the data set affect a new space, which can improve the predictive performance of the model, we randomly selected some data for display to facilitate the observation of the difference between the data before and after normalization, Table 2 is the normalized data.

Table 2. Normalized data

Sampling Point	SiO_2	K_2O	CaO	MgO
1	0.02	0.71	0.69	0.32
2	0.00	0.35	0.07	0.43
3	0.00	0.90	0.36	0.00
4	0.00	0.62	0.85	0.41
5	0.02	0.67	0.67	0.57
6	0.03	0.62	0.75	0.65
7	0.00	0.69	0.51	0.73
8	0.00	0.60	0.53	0.63
9	0.00	0.96	0.00	0.00
10	0.16	0.18	0.00	0.00

Since the random forest can use the out-of-bag error to measure the performance of the model, in order to reduce the time cost and improve the model performance, this paper does not divide the training set and the test set of the given data, but uses the out-of-bag error in the random forest to measure.

4. Model establishment and solution

For the random forest algorithm, the more important parameter is the number of decision trees, generally speaking, the more the number of decision trees, the better the prediction performance of the model, but the time cost required will also increase [8-9]. In this paper, the number of decision trees is set to 100 in order to obtain more stable prediction results while taking into account the time cost. commonly used methods include reduction of average impurity and reduction of average accuracy [10]. In this paper, the contribution of features is measured by the Gini Index using the average impurity method. The importance score of the features is expressed as VIM and the Gini is expressed as GI.

The total number of features involved in this paper is 14, so we use $c_1,\ c_2,\ c_3, \dots,\ c_{14}$ to denote all the features, so the index score of Gini for each feature $VIM_j^{(Gini)}$, which score also represents the average change in the degree of node non-division of the number j feature across all decision trees in the random forest. The Gini index is calculated as:

$$GI_m = \sum_{k=1}^{|K|} \sum_{k` \neq k} P_{mk} P_{mk`} = 1 - \sum_{k=1}^{|K|} P_{mk}^2 \tag{2}$$

Among them, K indicates that there are K categories, and $P_{mk`}$ indicates the proportion of category k in node m, which intuitively means the probability that two samples are randomly selected from node m, and their category labels are inconsistent. The importance of features X_j in node m, that is, the change of Gini index before and after branching of node m is:

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r \tag{3}$$

Where GI_l and GI_r represent the Gini indices of the two new nodes after branching, respectively.

If the node of feature X_j in the decision tree is in set M, then the importance of feature X_j in the first tree is:

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{jm}^{(Gini)} \tag{4}$$

A total of 100 trees are set in this paper, then:

$$VIM_{ij}^{(Gini)} = \sum_{i=1}^{100} VIM_{jm}^{(Gini)} \tag{5}$$

Finally, all importance scores are normalized:

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \tag{6}$$

5. Analysis and testing of results

The random forest feature importance algorithm calculates the proportion of influence of 14 chemical elements on the glass species of glass cultural relics when they are second classified, and the visualization results are shown in the Figure 2.

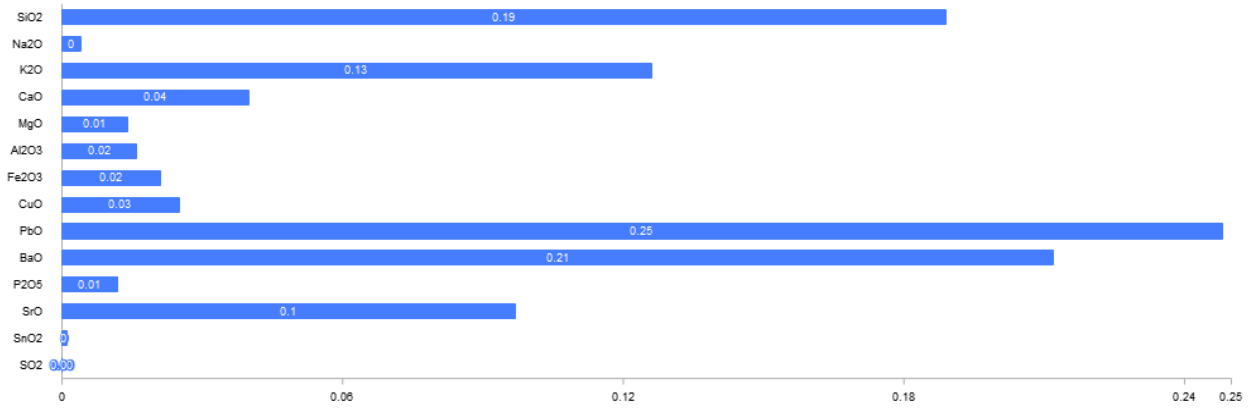

Figure 2. The proportion of the second classification of each chemical composition on glass type

The characteristic importance of each chemical composition is ranked from the largest to the smallest according to the numerical value, and the results are shown in the Table 3.

Table 3. Ranking of the second classification impact of chemical compositions on glass types

Rank	Feature name	Feature importance
1	PbO	24.80%
2	BaO	21.20%
3	SiO_2	18.90%
4	K_2O	12.60%
5	SrO	9.70%
6	CaO	3.50%
7	CuO	2.50%
8	Fe_2O_3	2.10%
9	AI_2O_3	1.60%
10	MgO	1.40%
11	P_2O_5	1.20%
12	Na_2O	0.40%
13	SnO_2	0.10%
14	SO_2	0.00%

Table 3 shows that the four components of PbO, BaO, SiO_2 and K_2O have greater importance in the classification of glass cultural relics, that is, among the 14 chemical components detected on the cultural relics detection point, the proportion of these four chemical components is the most related to the glass type of glass cultural relics, and their characteristic importance is greater than 10%, which further verifies the strong correlation of the four chemical components. Based on this, four chemical components with strong correlation with glass categories were obtained, and then subclass analysis of the two glass classes was carried out based on these four chemical components.

By K-Means algorithm, k (k=2,3,4,5) objects were randomly selected as the initial clustering centers, and the clustering analysis was performed by Python for different k values of different glass

types, and the clustering results of lead-barium glass and high-potassium glass with different k values are shown in the Figure 3. When k=2, the clustering results of both barium glass and high potassium glass are better, so the subclasses are classified as 2. The final results of the subclasses of high potassium glass and lead-barium glass are shown in the Figure 4.

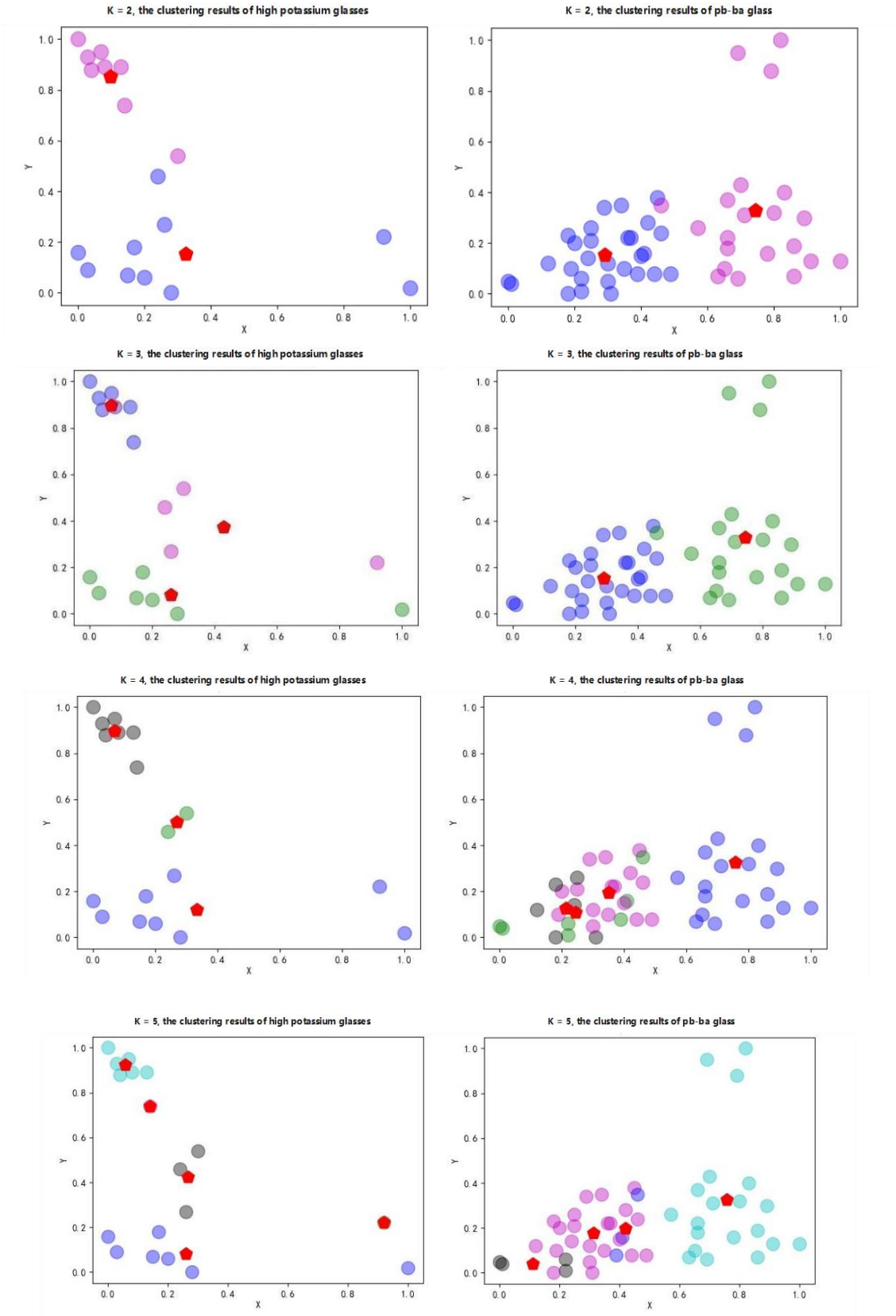

Figure 3. The proportion of the second classification of each chemical composition on glass type

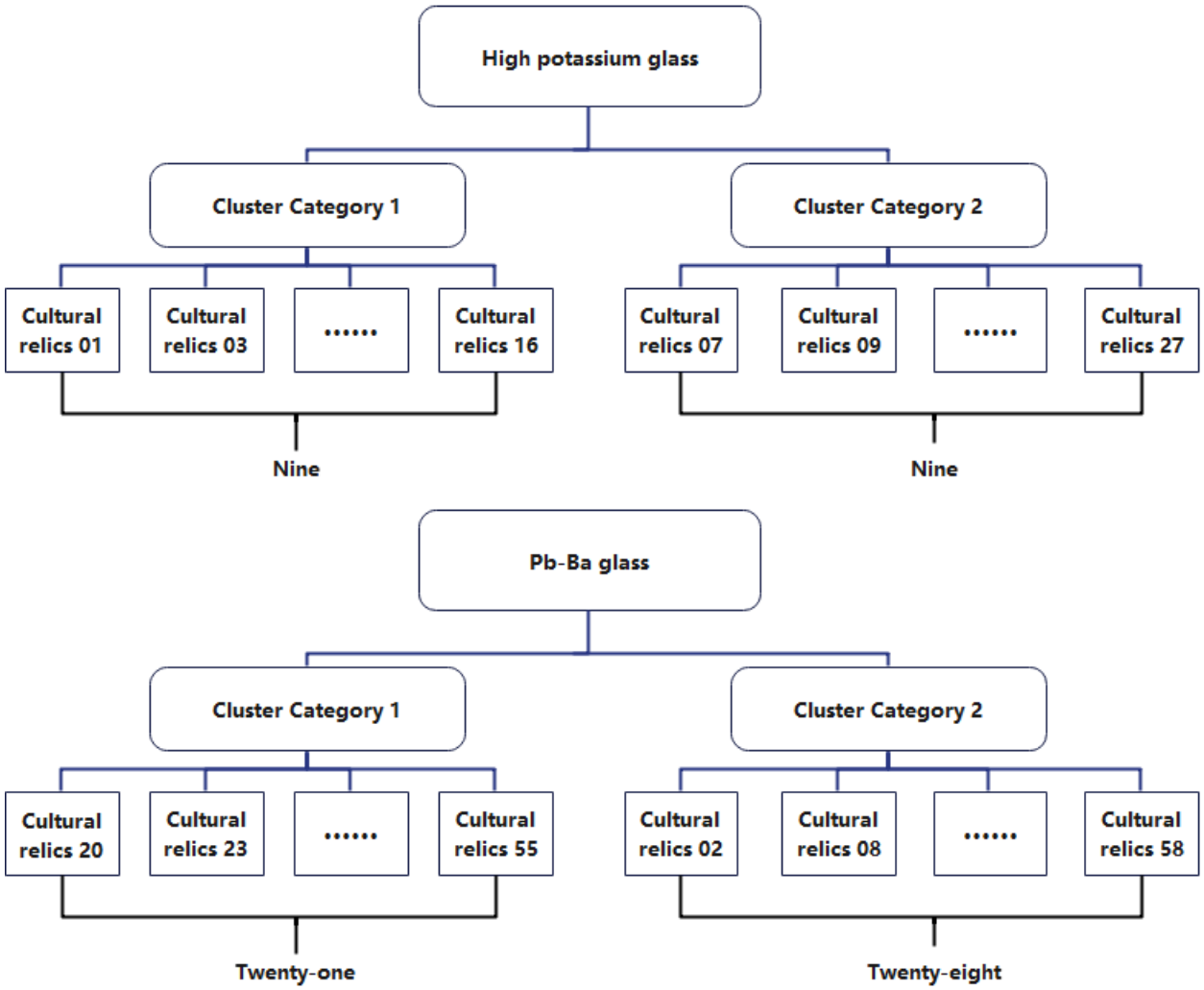

Figure 4. The proportion of the second classification of each chemical composition on glass type

After obtaining the random forest model we evaluated the prediction performance of the random forest model using four evaluation metrics: accuracy, recall, precision, and F1-score, and in the training process of the model, we set the cut ratio of the training set and the test set to 7:3. We also conducted a five-fold cross-validation, and the four metrics were calculated to reach 1, which proved that the model has good classification effect.

6. Conclusions

In this study, the types and chemical compositions of glasses were modeled by random forest algorithm, and the data were divided by 3:7 and cross-validated using 3-fold, and the models were evaluated using Accuracy, Recall, Precision and F1-score after training and the results were optimal. Means clustering algorithm was used to obtain the strongly correlated chemical components under each glass class, and unsupervised learning was used to visualize the clustering results, resulting in: two subclasses for high potassium glass and lead-barium glass. The subclasses may be caused by impure refined materials. The results of the study and the data processing methods will help to make some new breakthroughs in exploring the cultural and technological exchange of ancient glass on the Silk Road, which is important for the exploration and development of Chinese culture and Chinese civilization.

Among them, four chemical components PbO, BaO, SiO_2 and K_2O are of great importance to the classification of glass cultural relics. The research results and data processing methods will help to make some new breakthroughs in the cultural and scientific exchanges of ancient glass on the Silk

Road, which is of great significance to the exploration and development of Chinese culture and civilization.

References

[1] Li Qinghui, Gan Fuxi, Gu Donghong. Several issues on the study of ancient Chinese glass[J]. Studies in the History of Natural Sciences,2007,No.102(02):234-247.

[2] Zhang Jianrong, Zhang Wei, Xue Nannan, Zhao Tingsheng. Prediction and causation analysis of tower crane safety accidents based on random forest algorithm[J]. Safety and Environmental Engineering,28(05):36-42,2021.

[3] Guo Guangshan, Guo Jianhong, Sun Lichun, Liu Lifang, Tian Yongjing. 3D fine modeling of coal seam gas content based on random forest algorithm [J] . China offshore oil and gas, 2022,34(04) : 156 -163.

[4] Wang Sen, Liu Chen, Xing Shuaijie. Overview of K-means clustering algorithm [J]. Journal of East China Jiaotong University, 2022,39 (05): 119-126.

[5] Ma Yufeng, Dai Shaowu, Wang Rui, Dai hongde, Zheng Baodong. A nonlinear spatial K-means clustering algorithm [J/OL] for zero-speed interval detection in inertial pedestrian navigation. Journal of Beijing University of Aeronautics and Astronautics: 1-15[2022-09-18].

[6] Dong Wenjing. Overview of K-means algorithm [J]. Information and Computer (Theoretical Edition), 2021,33(11):76-78.

[7] Yang Won-Chol,Choe Chol-Min,Kim Jin-Sim,Om Myong-Song,Kim Un-Ha. Materials selection method using improved TOPSIS without rank reversal based on linear max-min normalization with absolute maximum and minimum values[J]. Materials Research Express,2022,9(6)

[8] Chen Yunying, Wu Jiqin, Xu Kejia, “Attribute splitting method based on Gini index in decision tree” Financial Theory and Practice, Papres38(5),66-68 (2004).

[9] Lai, Chun-Ting. Research on decision tree classification algorithm [J]. Information and Computer(Theory Edition),2020,32(14):59-62.

[10] Gong Xu, Lv Jia, Pi Jiatian. Cooperative training algorithm combining information gain rate and K-means clustering [J]. Journal of Chongqing Normal University (Natural Science Edition), 2020,37(02):112-119.