Classification of ancient glassware based on Random Forest

Jingdong Wang #,*, Ying Zhang #, Shuo Feng #, Xinyan Han #, Tianchen Zhou #

Shanghai Business School, Shanghai, China

* Corresponding Author Email: wjd13223633307@163.com

#These authors contributed equally to this work

Abstract. As a precious item in the foreign trade along the Silk Road, it is of great significance to analyze the composition and identify the type of glass relics. In this paper, through the training of random forest, the characteristic importance of each index is obtained, and the classification rule is obtained. It is determined that lead oxide (PbO) is the most important index in the classification of glass types. Then, the k value was determined by the elbow principle, and k=5 was selected as the cluster type. After that, k-means clustering was carried out on the high potassium and lead barium glass respectively to get the classification results, which were divided into 5 subclasses. Due to the small amount of data, in the random forest model, the robustness of the model is constantly enhanced through multiple hyperparameter tuning. In the trained random forest model, it is found that the results do not change, which proves the model is reasonable. At the same time, the correlation between the chemical components of the glass cultural relics samples of different categories was analyzed, and the difference of the correlation between the chemical components of different categories was compared.

Keywords: Random Forest, k-means clustering model, elbow principle, grey correlation.

1. Introduction

Glass relics, as precious objects traded along the Silk Road, are of great significance for the study of history and culture and the enhancement of cultural confidence. Glass is affected by various factors that make its internal chemical elements react with the external elements, resulting in the change of the proportion of its chemical components. Therefore, it is very important to accurately identify the unknown category of glass cultural relics [1].

Based on the basic information of a batch of glass cultural relics, in order to explore the classification law, this paper firstly selects the chemical composition of each glass cultural relic as the characteristic value, the type as the label value, establishes the random forest model for binary classification, obtains the importance of each characteristic value, and then obtains the classification law. Then, k-means clustering model was adopted to find out the most appropriate number of clustering through the elbow principle and drawing. k-means clustering was carried out on the high potassium and lead barium glass respectively to get the partitioning method and the partitioning results, which were divided into four subclasses. In addition, in order to predict the types of unknown glass relics, this paper determined the verification set according to the forecast requirements, and divided it into high-potassium glass and lead-barium glass by running the random forest model. Secondly, on this basis, the k-means clustering model of subclass division was used again to get its specific subclass division. The grey correlation analysis was carried out on the chemical composition of different types of glass cultural relics samples, and the grey correlation degree between each chemical component was obtained, so as to analyze the correlation between their chemical components. On this basis, the difference analysis is carried out.

2. Classification law of high potassium and lead barium glass

Before establishing the subclass division model, appropriate chemical components should be selected, so part of the index data should be screened. As the main component of glass is silica, the difference of silica is small, so it is reasonable to remove. The standard deviation of each component was calculated by classification, as shown in Figure 1.

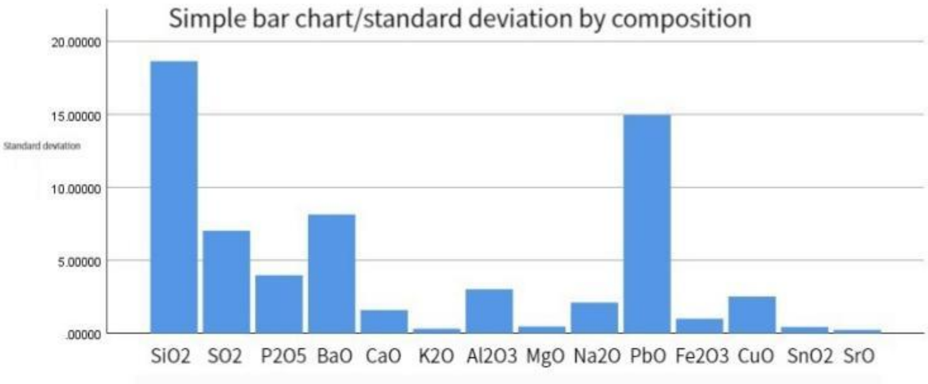

Figure 1. The standard deviation of each component

The larger the standard deviation is, the greater the difference between the data and the greater the information carried; while the smaller the standard deviation is, the larger the similarity between the data and the smaller the information is, the lower the contribution rate to the subclass division model and the effect of division will be affected to a certain extent. Therefore, the threshold is set as 1 and the index data with standard deviation less than 1 will be removed. After the above data screening, for high-potassium glass, 7 components including sodium oxide, potassium oxide, calcium oxide, alumina, ferric oxide, copper oxide and phosphorus pentoxide are selected to be retained, while for lead-barium glass, 8 components including sodium oxide, calcium oxide, alumina, barium oxide, phosphorus pentoxide and sulfur dioxide are selected to be retained.

In order to analyze the classification rules of high potassium glass and lead barium glass, a random forest was established for classification [2], the random forest was used to conduct supervised training on the data amount, with 70% of the data amount as the training set and 30% of the data amount as the test set. After many times of training, the model has a high accuracy, and the characteristic importance of each indicator is obtained, as shown in Figure 2 below:

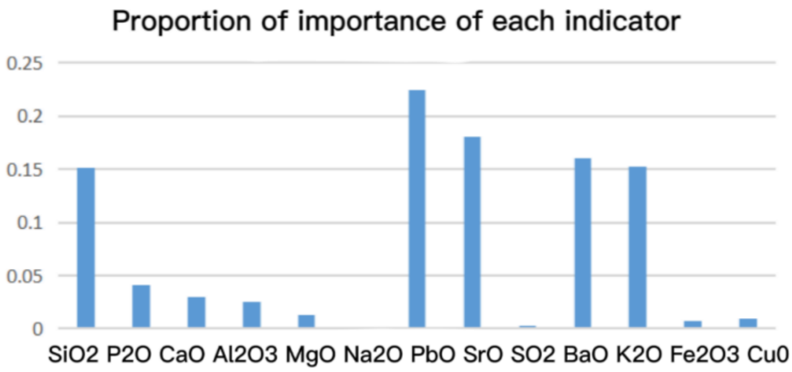

Figure 2. Proportion of importance of each indicator

Feature importance [3] refers to the importance of a feature in the whole model, which is measured by Gini coefficient in this paper. The Gini coefficient can be used to express the explanatory power of all variables, that is, to remove the effect on the accuracy of the model of a variable. Gini coefficient is the sum of the contribution values of different features in each decision tree to obtain the contribution value of the feature. Finally, all the importance scores are normalized. So the Gini coefficient can be approximated as the correlation coefficient of each feature and label, that is, the type. Therefore, according to the importance of features in the figure above, it can be found that silica, lead oxide, strontium oxide, potassium oxide and iron oxide are strongly correlated with the type, among which lead oxide is the most correlated with the type. It can be concluded that the content of lead oxide is the most important basis for the classification of glass types.

3. k-means subclass division model

3.1. K-means clustering

For subclass division, the k-means clustering model [4] is chosen to be built. However, for the given data, there are large differences in the order of magnitude between the data, which will greatly interfere with the k-means clustering model. Therefore, minMax standardization is adopted to process the data:

$$y_i = \frac{x_i - \min_{1 \le j \le n}\{x_i\}}{\max_{1 \le j \le n}\{x_j\} - \min_{1 \le j \le n}\{x_j\}} \tag{1}$$

means initializes k cluster centers randomly, and then calculates the Euclidean distance from each glass cultural relic to the cluster center, as shown in the formula: Compare the distance from each glass cultural relic to each cluster center in turn, and assign it to the cluster closest to the cluster center to obtain k cluster. The cluster center formula is as follows:

$$\mathrm{dis}(\mathbf{X}_i, \mathbf{C}_j) = \sqrt{\sum_{t=1}^{m} (X_{it} - C_{jt})^2} \tag{2}$$

$$C_t = \frac{\sum_{X_i \in S_i} X_i}{|S_1|} \tag{3}$$

K-means algorithm needs to specify k value, and the setting of k value will have a great impact on the model, so the selection of k value is crucial. For the selection of k value in this paper, the elbow rule is used to select k value. The elbow rule calculates the average deviation after the completion of each cluster iteratively, finds the inflection point through observation, and determines the value of k. The K-means subclass division model of high potassium glass and lead barium glass is set to k=5, and the model effect is the best.As shown in Figure 3 and Figure 4 below.

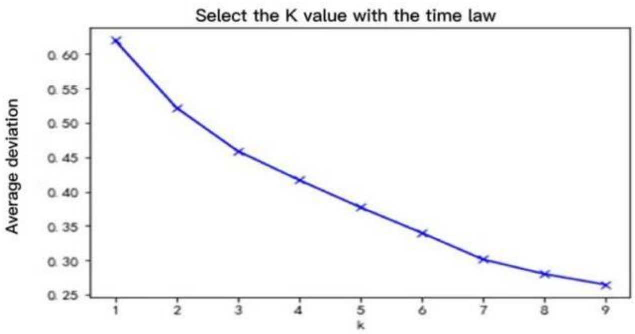

Figure 3. High potassium glass k-means model k value determination diagram

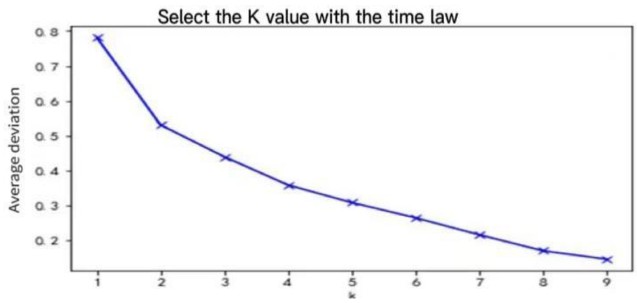

Figure 4. Lead-barium glass k-means model k value determination diagram

3.2. Rationality and Sensitivity analysis

To analyze the rationality of the classification results, the elbow rule is used in this paper to determine the k value. It can be seen from the following figure that when k=5, the degree of distortion is greatly improved, and then the significance tends to be gentle. k=5 is a more reasonable choice for k value. As shown in Figure 5 below.

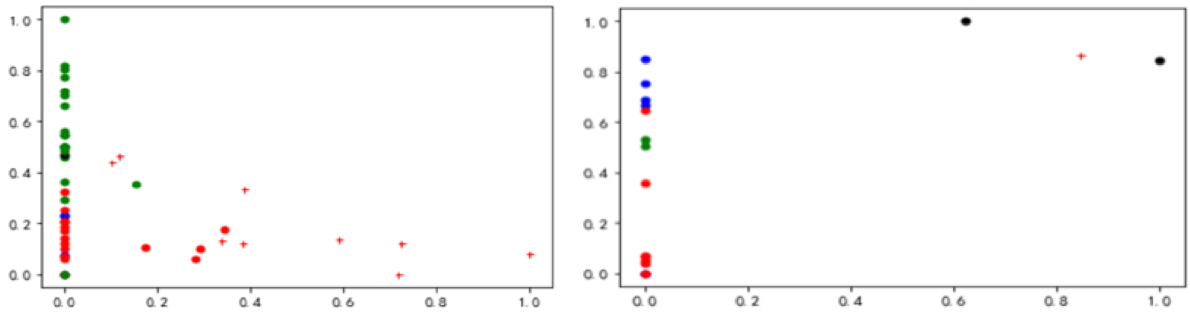

Figure 5. Cluster diagram of lead barium glass and high potassium glass

As for sensitivity, add data disturbance to the model, and test the change of accuracy of the model through the comparison of classification results before and after, so as to judge the sensitivity of classification.

In this data disturbance, within a certain error range, by increasing the value of 1%-5% to the original data, it was found that the classification results did not change, indicating that the sensitivity of the classification results was good. As shown in Table 1 below.

Table 1. Table of sensitivity analysis results

Heritaeg Number	1% perturbation	2% perturbation	3% perturbation	4% perturbation
A1	accurate	accurate	accurate	accurate
A2	accurate	accurate	accurate	accurate
A3	accurate	accurate	accurate	accurate
A4	accurate	accurate	accurate	accurate
A5	accurate	accurate	accurate	accurate
A6	accurate	accurate	accurate	accurate

4. Classification prediction model

4.1. Establishment of model

Through the random forest binary classification model, the glass relics were predicted by binary classification, and the preliminary binary classification results were obtained, that is, A1, A6, A7 are high-potassium glass, and the rest are lead-barium glass. Secondly, the two types of glass were divided into k-means subclasses respectively, and the final division results were obtained. Among them, high-potassium glass was again divided into 2 categories and lead-barium glass into 3 categories. The division results are shown in Table 2 below, where different numbers in different types represent different subclasses.

Table 2. Results of subcl

Heritaeg Number	Type	The class
A1	High in potassium	2
A2	Lead barium	2
A3	Lead barium	2
A4	Lead barium	2
A5	Lead barium	4
A6	High in potassium	0
A7	High in potassium	0
A8	Lead barium	3

4.2. Analysis of sensitivity

Since it is a machine learning algorithm, it needs to constantly adjust the overparameters for tuning in the training process. In this sensitivity test, the n_estimators parameter is constantly adjusted, which means the number of base estimators in the forest, namely the number of trees. The influence of this parameter on the accuracy of the stochastic forest model is monotonous. The appropriate n_estimators parameter is selected through constant adjustment. When the parameter n_estimators is about 50, the accuracy reaches the maximum value and tends to be stable, with the accuracy being 1. Through continuous tuning, the robustness of the model is greatly enhanced, that is, the model sensitivity is better. As shown in Figure 6 below.

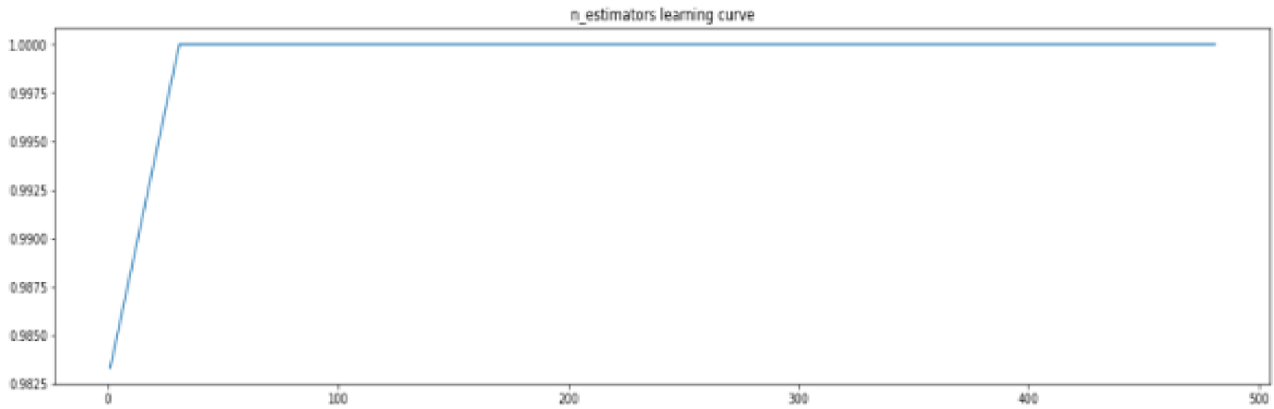

Figure 6. n_estimators Parameter diagram

4.3. Grey correlation analysis

For the glass cultural relics samples of different categories, [5] the data are first divided into two categories, namely high-potassium glass and lead-barium glass. Then, each chemical component is set as the parent sequence, and the other chemical components are set as the sub-sequence. Through the gray correlation analysis, the gray correlation degree between each component is obtained, that is, the correlation relationship. The correlation degree between silica and alumina is the highest, and that between silica and sodium oxide is the lowest. For lead-barium glass, the correlation between silica and alumina is the highest, and that between sulfur dioxide is the lowest.As shown in Table 3 and Table 4 below.

Table 3. High potassium glass silica correlation results ranking table

	Result of correlation degree	
Item of evaluation	Degree of correlation	ranking
Al_2O_3	0.932	1
CuO	0.931	2
P_2O_5	0.917	3
MgO	0.911	4
K2O	0.907	5
CaO	0.903	6
Fe_2O_3	0.897	7
PbO	0.873	8
SnO_2	0.866	9
SrO	0.864	10
BaO	0.855	11
SO_2	0.85	12
Na^2O	0.849	13

Table 4. Lead barium glass silica correlation results ranking table

Result of correlation degree		
Item of evaluation	Degree of correlation	ranking
Al_2O_3	0.953	1
BaO	0.936	2
PbO	0.934	3
MgO	0.931	4
SrO	0.924	5
CaO	0.921	6
K_2O	0.92	7
CuO	0.915	8
Fe_2O_3	0.911	9
Na_2O	0.9	10
P_2O_5	0.896	11
SnO_2	0.884	12
SO_2	0.882	13

4.4. Analysis of difference

One-way analysis of variance (ANOVA) was conducted for chemical composition and type. Through ANOVA, the P-value of each analysis was judged to be significant. If P<0.05, it indicated that there was a significant difference between the two groups of data; otherwise, it indicated that there was no difference. The analysis results are shown in the Figure 7 below:

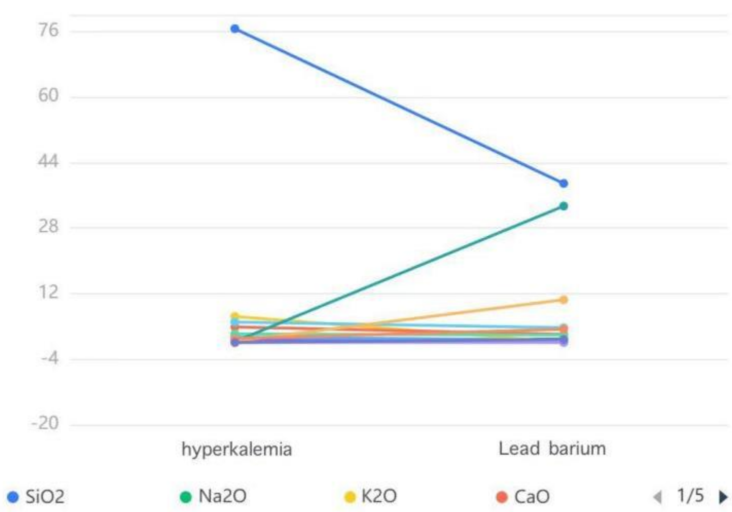

Figure 7. Difference analysis results

It can be seen that there is a significant difference between silica and sodium oxide types.

5. Conclusions

This paper adopts the algorithm of machine learning -- random forest for classification, because random forest is not easy to overfit and is good at processing high-dimensional data, which can be more convenient and efficient implementation. Random forest was used to conduct supervised training on the data amount, with 30% of the data amount as the test set and 70% of the data amount as the training set. After many tests, the accuracy of the model is 1 and the importance ratio of each index is obtained. Among them, the correlation with lead oxide (PbO) is the strongest, and the correlation with potassium oxide (K2O) and barium oxide (BaO) is strong. In addition, appropriate chemical components are selected for different categories to be divided into subclasses, and detailed subclass division, division methods and results are provided.

For the glass cultural relics of unknown category, this paper determined the verification set according to the predicted demand, and divided them into high-potassium glass and lead-barium glass by running the random forest model. Secondly, on this basis, the k-means clustering model of subclass division was used again to get its specific subclass division. The results show that A1, A6 and A7 are high potassium glass, and the rest are lead barium glass. Secondly, the two types of glass were divided into k-means subclasses respectively, and the final division results were obtained, among which the high-potassium glass was again refined into 2 categories and the lead-barium glass into 3 categories. Finally, in the process of training the random forest model, the robustness of the model is continuously enhanced through multiple hyperparameter tuning. In the trained random forest model, the sensitivity is better.

References

[1] Shao Hongfei, Huang Hui, Deng Weihua, Ji Kejian. Research and application of reference materials for glass composition analysis [J]. Chemical Analysis and Metrology,2011,20(02):9-12. (in Chinese) J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Cheng Shaochi, You Guangrong. Research on Reduction Method of Evaluation Index based on improved Random Forest Algorithm [J/OL]. Systems Engineering and Electronics :1-9[2022-09-18]. K. Elissa, "Title of paper if known," unpublished.

[3] Zhao Qingyuan, Ye Chunmao, Lu Yaobing. Importance evaluation of fretting features based on Random forest [J]. Modern Defense Technology,2022,50(4):124-131. (in Chinese) DOI: 10.3969/j.issn.1009-086x.2022.04.014. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[4] Du Ying, Wang Chao, Yang Jie, Wang Qian. Automatic Monitoring Method for Mechanical Properties of Transformer Switches Based on Improved K-means Clustering [J]. Manufacturing Automation,202,44(06):183-185+198. (in Chinese)

[5] Xin Changqing and Xu Jian. Research on energy hub configuration optimization problem based on analytic hierarchy and grey correlation analysis[J]. Journal of Physics: Conference Series, 2022, 2354(1)