Solving the problem of composition and identification of ancient glass products based on classification model

Jingya Zou [1,*], Kangyu Zhu [2]

[1]School of Materials Science and Engineering, Xiangtan University, Xiangtan, Hunan, 411105 China

[2]School of Information Technology, Nanchang Vocational University, Nanchang, Jiangxi, 330007 China

*Corresponding author: 1871257580@qq.com

Abstract: Glass has been used as ornaments or containers in ancient times, and in recent years, with the development of the world, the existence of glass has been paid more and more attention. Examining the composition and identification of ancient glass artifacts reflects the rise and fall of glass production in each dynasty, thus reflecting the face of each dynasty, reflecting the historical understanding and importance of glass materials, which is quite helpful for archaeological research. The problem of predicting unknown classes of glass artifacts and performing sensitivity analysis is solved by a classification model based on random forest. In this paper, we first define lead-barium glass as a positive class, and build classification models for weathered and unweathered glass based on the random forest algorithm and the traditional classification algorithm, and then train them separately. ROC curve Verify the prediction effect of random forest.

Keywords: Random forest; AUC; ROC curves.

1. Introduction

The Silk Road is an ancient communication channel between China and the West, in which glass is a precious evidence of the initial trade. Early glass was often introduced into China as jewelry from the West Asian and Egyptian regions. However, as ancient Chinese glass absorbed this technicality and used local raw materials to make, the appearance is similar to foreign glass products, but the chemical composition is different[1].

The main raw material of glass is quartz sand, the main component is silicon dioxide (SiO2). Because of the high melting point of pure quartz sand, in order to reduce the melting environment temperature, it is necessary to add a melting agent when purification[2]. The common melting agents in ancient times were grass ash, natural bubble soda, saltpeter and lead ore, and the addition of limestone as a stabilizer, which was transformed into calcium oxide (CaO) by forging and firing. Depending on the combustion aid used, as its main components are not the same. For example, lead barium glass is added to lead barium glass as a flux in the calcination link, and its high content of lead oxide (PbO) and barium oxide (BaO) is generally considered to be a unique type of glass in China, and the glass of Chu culture is dominated by lead barium glass. Potassium glass is forged and fired by a substance with high potassium content, such as grass wood ash, and is prevalent in areas such as Lingnan, Guangdong, Southeast Asia and India[3].

Ancient glass is very susceptible to weathering due to the burial environment. In the weathering link, many internal structural elements and natural environmental elements produce interchangeable, resulting in changes in the percentage of its composition, which affects and judgment on the analysis of the classification[4]. The surface is not weathered, with significant color and grain, but does not exclude by the local shallow weathering of cultural relics; on the weathered surface of cultural relics, the surface of a large gray-yellow area for the weathering layer, a significant weathering area, the purple part is generally weathered surface. In some weathered artifacts, there are also areas with unweathered surfacesp[5].

Now there is a group of materials about ancient Chinese glass products. Archaeologists based on the composition of the different ways to test these artifacts into high potassium glass and lead-barium glass. The chemical composition of ancient glass artifacts of unknown types and analyzed[6].

2. Model assumptions and notation

2.1 Assumptions[7]

1. it is assume that the weathering of ancient glass is only chemical weathering.
2. assuming that the blanks in Form 2 all indicate that the constituent was not detected i.e. the content of the constituent is zero and that no missing values exist
3. assumes that the composition of high potassium glass and lead-barium glass is similar in all epochs
4. it is assumed that only geography and the duration of weathering are considered as influencing the causes of weathering.

2.2 Notations

Important notations used in this paper are listed in Table 1.

Table 1 Notations

Symbol	Definition
X^2	Number of statistics
J	Clustering coefficient
Sensitivity	Sensitivity
R_s	Spielman correlation coefficient
p	Spielman irrelevant probability

3. Model construction and solving

3.1 Idea analysis

In this paper, for predicting unknown categories of glass artifacts and performing sensitivity analysis ideas are mainly divided into two parts, the first part is based on the classification law of high potassium glass and lead barium glass, the second part is to add data perturbation to judge its sensitivity, the first part is based on the random forest model to predict the type of glass artifacts, the second part is appropriate language modeling to judge its sensitivity[8].

3.2 Development of a classification model for glass artifacts based on random forest algorithm

3.2.1 Introduction to the Random Forest Model

Random Forest Random Forest is a classical bagging model with a decision tree model as its weak learner, and all weak learners in the bagging algorithm are weighted equally. To ensure the generalization ability of the model, each decision tree built follows data randomization and feature randomization[9].

3.2.2 Random Forest Modeling

1. Training principle

One decision tree is trained using one training set at a time, and n decision trees are trained according to different training sets after n training sets are randomly selected with put-backs, and a sequence of classification models $\{h(X), h_2(X), \cdots, h_n(X)\}$ can be obtained after n rounds of training, and then they form a multi-classification model system, which has the final[10].

The final classification result is based on the majority voting method.

The final classification decision. $H(x) = \arg\max \sum I(h_i(x) = Y)$

$Y \quad i=1$

The equation illustrates the use of majority voting decisions to determine the final classification.

The training and test sets are determined using cross-validation to determine their generalization ability, as shown in Figure 1 below.

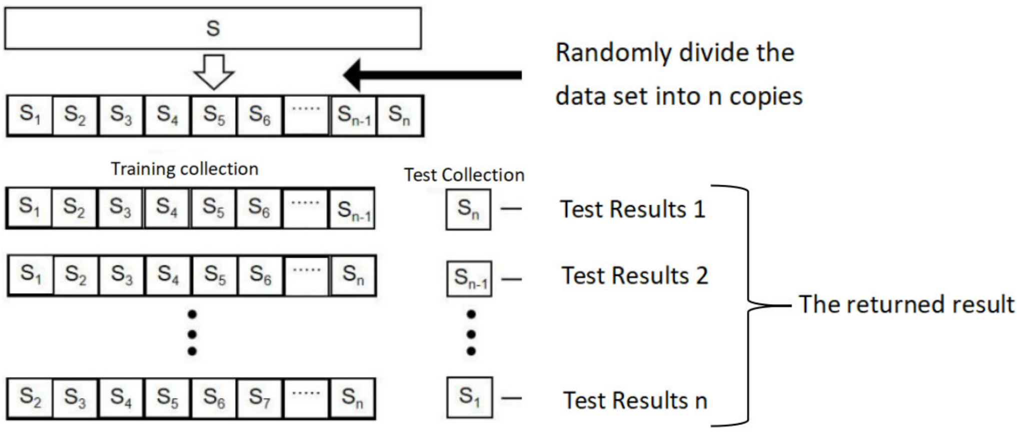

Figure 1 Cross-validation schematic

2. flowchart of random forest algorithm

The flow chart of the random forest algorithm is shown in Figure 2

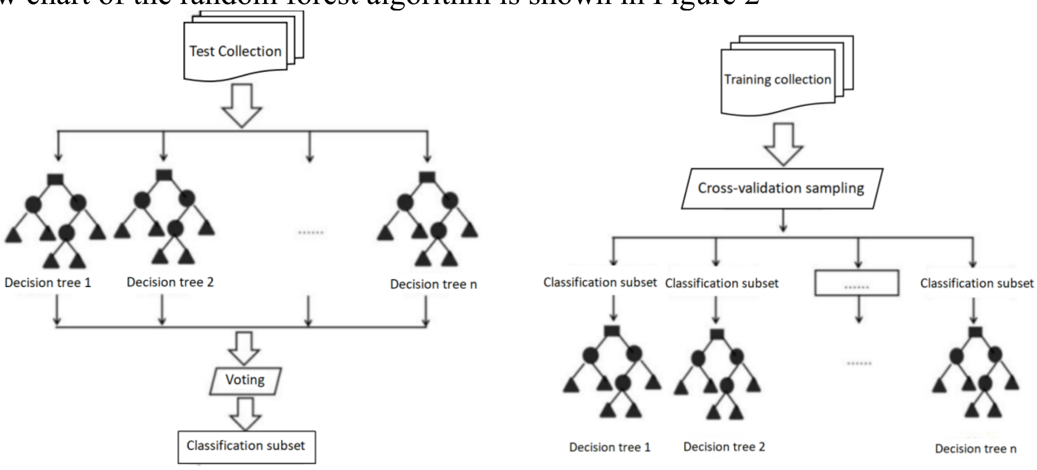

Figure 2 Random forest algorithm flow chart

3. Measuring the accuracy of the model

The F1 score is the summed average of the full rate and the accuracy rate, which can be simplified.

ROC curve: A curve based on a series of different classification thresholds, with the true class rate as the vertical coordinate and the false positive class rate as the horizontal coordinate, the closer the curve is to the upper left corner, the better the model classifies.

3.3 Model solving

3.3.1 Classification of weathered glass artifacts

First define the lead barium glass unpositive class event, the indicator is defined as the following Table 2

Table 2 Positive class event table

Indicators	Meaning
TP	The number of positive classes predicted to be positive classes
FN	Number of positive classes predicted to be negative classes
FP	Number of negative classes predicted as positive classes
TN	Number of negative classes predicted to be negative classes

Then the confusion matrix is obtained based on the training as shown in Figure 3 below.

The data were then fitted and tuned to the 4 base categories, and the better parameters were selected to derive the performance of the 4 base categories on the modified data set, as shown in Table 3 below.

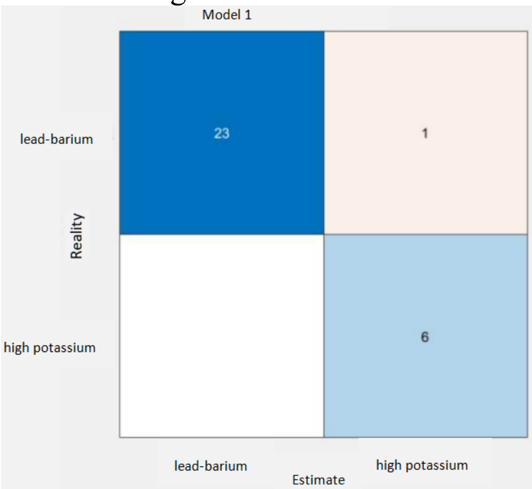

Figure 3 Confusion matrix for artifacts that have weathered glass

Table 3 Base classifier classification effect

Classifiers	Accuracy rate	AUC
Logistic Regression	93.3	0.99
Linear Discriminant Analysis	None	None
Decision Trees	None	None
Random Forest	96.7	0.99

It can be seen that Random Forest has the highest accuracy, so its ROC graph is drawn for further verification, as shown in Figure 4 below.

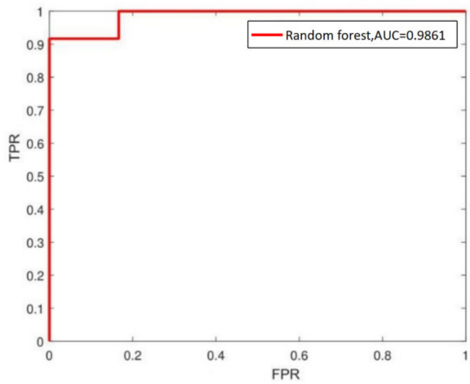

Figure 4 ROC Curve

3.3.2 Classification of artifacts that have been weathered glass

First define the lead barium glass unorthodox class of events with the indicators defined in Table 4 below.

Table 4 Positive Class Event Form

Indicators	Meaning
TP	The number of positive classes predicted to be positive classes
FN	Number of positive classes predicted to be negative classes
FP	Number of negative classes predicted as positive classes
TN	Number of negative classes predicted to be negative classes

The confusion matrix is then obtained based on the training as shown in Figure 5 below.

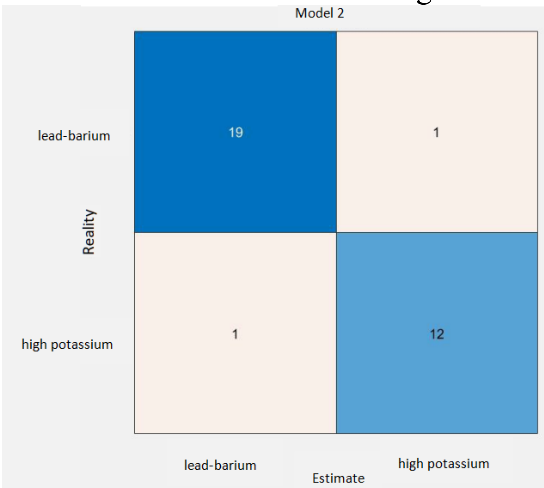

Figure 5 Confusion matrix for unweathered glass artifacts

The data were then fitted and tuned to the 4 base categories to select the better parameters, resulting in the performance of the 4 base categories on the modified dataset, as shown in Table 5 below.

Table 5 Base classifier classification effect

Classifiers	Accuracy	AUC
Logistic Regression	87.9	0.85
Linear Discriminant Analysis	87.9	0.83
Decision Trees	97	0.96
Random Forest	94	0.98

It can be seen that Random Forest has the highest accuracy, so its ROC graph is drawn for further verification, as shown in Figure 6 below.

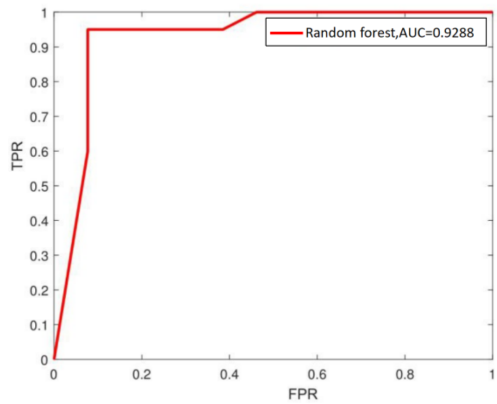

Figure 6 ROC Curve

3.4 Analysis of the model

3.4.1 Classification of weathered glass artifacts

The highest overall accuracy was found in the decision tree, the random forest was slightly higher, and the rest of the classifiers had varying degrees of overfitting or poor accuracy. On the whole, the linear discriminant analysis is obviously slightly inferior, with the lowest accuracy and poor fitting.

The AUC value is the area covered by the ROC curve, and the ROC curve is shown in Fig. 6, and its meaning can be considered as a combination of recall, precision and accuracy. In general, an AUC value of 0.8 or more is acceptable for classification models, and from this perspective, random forest meets the requirements.

Details of the classification results using Random Forest can be found in Supporting Material 4, where

Artifact A2, which has been weathered, has a 0.8 probability of being predicted as lead-barium glass and a 0.2 probability of being predicted as high-potassium glass, so its predicted result is lead-barium glass.

Weathered artifact A5 has a 0.8667 probability of being predicted as lead-barium glass and a 0.1333 probability of being predicted as high-potassium glass, so its predicted result is lead-barium glass.

Weathered artifact A6 has a 0.0333 probability of being predicted as lead-barium glass and a 0.9667 probability of being predicted as high-potassium glass, so it is predicted to be glass.

Weathered artifact A7 has a 0.1 probability of being predicted to be lead-barium glass and a 0.9 probability of being predicted to be high-potassium glass, so it is predicted to be high-potassium glass.

3.4.2 Classification of artifacts that have been weathered glass

As can be seen from Table 5, the highest overall accuracy is random forest, so the closest effect between the training and test sets is random forest, and the remaining classifiers have varying degrees of inability to fit or poor accuracy. On the whole, the logistic regression is obviously slightly inferior, and the middle two classifiers cannot be fitted.

The AUC value is the area covered by the ROC curve, and the ROC curve is shown in Fig. 20, and its meaning can be considered as a combination of recall, precision and accuracy. In general, an AUC value of 0.8 or more is acceptable for the classification model. From this perspective, random forest meets the requirements, so random forest is chosen for classification.

Details of the results of the classification using random forest are shown in Supporting Material 4.

Unweathered artifact A2 has a probability of 0 being predicted as lead-barium glass and a probability of 1 being predicted as high-potassium glass, so it is predicted as high-potassium glass.

Unweathered artifact A2 has a 0.9333 probability of being predicted as lead-barium glass and a 0.0667 probability of being predicted as high-potassium glass, so it is predicted as lead-barium glass.

Unweathered artifact A2 has a 1 probability of being predicted to be lead-barium glass and a 0 probability of being predicted to be high-potassium glass, so it is predicted to be lead-barium glass.

Unweathered artifact A2 has a 1 probability of being predicted as lead-barium glass and a 0 probability of being predicted as high-potassium glass, so it is predicted as lead-barium glass.

3.4.3 Sensitivity Analysis

The sensitivity analysis of the classification model for the weathered glass artifacts using the formula was 95.84%, indicating that the classification model is very sensitive when there is a 5% fluctuation in the data.

The sensitivity analysis of the classification model for unweathered glass artifacts using the formula was 95%, indicating that the classification model is sensitive to 5% fluctuations in data.

4. Conclusion

This paper analyzes the chemical composition of unknown categories of glass artifacts, and then identifies whether they belong to the type of high potassium glass or lead-barium glass. Since the composition of weathered glass and unweathered glass differs greatly, the classification models for weathered glass and unweathered glass can be developed separately. Based on the data, a classification model based on the random forest algorithm and a traditional classification model for glass artifacts were established, and then the optimal classification model was selected based on the generalization ability of these models. Based on the optimal classification model, the glass artifacts of unknown categories are classified separately according to their weathering or non-weathering to obtain their glass types. The sensitivity of the classification results can be analyzed by calculating the recall, i.e., sensitivity, of the classification results of the two classification models, weathered and unweathered.

In addition, the model established by random forest can be tuned by grid search to improve the classification accuracy.

References

[1] Author: Wang Jintao Publication source: Shanghai Experimental Animal Science Published: 2000-12-25

[2] Li Xiongwei. Effect of rare earth doping on the chemical stability and weathering resistance of ZnO-B2O3-SiO2 glass [D]. Inner Mongolia University of Science and Technology,2014

[3] Wang Chengxian, Tao Ying, Gu Xiumei, eds. Glass surface decoration [M].2011

[4] Yang, Z., Li, H. Y., +1 author, Yang, K. Y., Liu, L. (supervised by). China Science and Technology Information. 2021, Vol. 12, No. 121-122, Total 2 pp.

[5] Zhang Liyan, Li Hong, Chen Shubin, Li Zhong Dy, Ruan Tiachi, Xue Tianfeng, Qian Min, Fan Sijun. Simulation methods for the composition and properties of glass[J]. Journal of Silicates,2022,50(08):2338-2350.DOI:10.14062/j.issn.0454- 5648.20220115.

[6] Zhang HW, Wang MW, Gan LX. Research on text classification model based on random forest. [J]. Journal of Shandong University (Science Edition).2006(03)

[7] Li Xiongwei. Effect of rare earth doping on the chemical stability and weathering resistance of ZnO-B_2O_3-SiO_2 glass[D]. Inner Mongolia University of Science and Technology,2014

[8] Liu Shaobo. Research on parameter estimation algorithm and sensitivity analysis algorithm based on ecohydrological model[D]. Lanzhou University,2012.

[9] Lin Feng,Cai Yanguang,Roy Cai,Wang Jiancheng. A random forest model based on improved fruit fly algorithm optimization[J]. Journal of Changshu University of Technology,2020,34(05):51-57.DOI:10.16101/j.cnki.cn32-1749/z.2020.05.013.

[10] Tang Donglai,Zhong Sheng,Li Qingyu,Chen Zeyu,Gong Yiyu,Nie Xiao. A multi-classifier-based approach for ultra-short-term rooftop PV power prediction[J]. Power Big Data,2022,25(03):1-8.DOI:10.19317/j.cnki.1008-083x.2022.03.008.