

Ancient glass classification based on random forest algorithm and decision tree model

Zikai Wang^{*}, Yun Lu, Tianze Li

Jiangsu University of Science and Technology, Suzhou, Jiangsu, 215699

^{*} Corresponding Author Email: 2500405797@qq.com

Abstract. The texture change of ancient glass depends on its chemical composition and preservation environment, and it is highly susceptible to weathering by the burial environment. In this paper, a classification model based on random forest algorithm is established to solve the problem of analyzing the classification laws of high potassium glass and lead-barium glass; a decision tree model based on decision tree is established to solve the problem of subclass differentiation of glass categories and analyzing their rationality and sensitivity. The reasonableness of the model and the accuracy of the data are tested by the plain Bayesian algorithm.

Keywords: Glass Weathering, Random Forest Algorithm, Decision Tree Model.

1. Introduction

Glass is an amorphous inorganic non-metallic materials [1-3], generally a variety of inorganic minerals as the main raw material, the main raw material of this question glass is quartz sand [4], in addition to adding a small amount of auxiliary raw materials made [5]. Its main chemical composition is silicon dioxide and other oxides. As the melting point of pure quartz sand is high, a co-solvent can be added to lower the melting temperature during refining. Limestone is also added as a stabilizer, which is converted to calcium oxide after calcination. The chemical composition of the added co-solvent is different [6-8]. Among them, lead barium glass has a high content of lead oxide and barium oxide, which is strong in protection and resistant to oxidation, and the glass of Chu culture in ancient times was mainly lead barium glass. Potassium glass is mainly composed of silica, calcium oxide and potassium oxide, with a high potassium content, and grass wood ash is used as a combustion aid [9].

The texture change of ancient glass depends on its chemical composition and preservation environment, and it is highly susceptible to weathering by the burial environment. During the weathering process, the internal elements are exchanged with the environmental elements so much that the composition ratio changes, which affects the correct judgment of its category [10-11]. Therefore, this paper classifies ancient glass with the help of random forest algorithm and decision tree model [12].

2. Establishment of classification model based on random forest algorithm

2.1. Theoretical background of Random Forest

Random forest uses CART as a meta-classifier, generates a random training sample set for each decision tree using the Bagging method, and randomly selects features in the training sample to determine the node splitting of the decision tree when constructing a single tree. The combination of the Bagging method and CART, together with the random selection of features for attribute splitting, makes RF more tolerant of noise and has better classification performance.

Random forest is composed of multiple decision trees $\{h(x, k), k = 1, 2, \dots, n\}$ A combined classifier consisting of $\{k\}$ is an independent identically distributed random vector, The output is generated by combining the results of all decision trees and voting on them. The final classification of the random forest depends on the voting of multiple trees (weak classifiers).

A random forest consists of N decision tree composition, Each decision tree (e.g., decision tree $T_1, T_2, \dots, T_N$) is a classifier where the decision result of the random forest is derived from the combined strategy of all decision tree classification results.

A decision tree is a tree structure in which each internal node represents a test on an attribute and each branch represents a test output. The decision tree algorithm used in this paper is CART.

$$\text{Gini}(T) = 1 - \sum_{k=1}^K p[k|T] \quad (1)$$

When $\text{Gini}(T)$ maximum, the samples in this node are equal for classification into which class, i.e., they obey a uniform distribution, and therefore, the information is minimum.

Bagging method, also known as self-help aggregation method, is a method to generate decision trees by randomly sampling some samples from the training set.

(1) Data pre-processing

The distribution patterns of high potassium glass and lead-barium glass were analyzed. First, the data in Form 2 were processed. The data in Form 2 indicated the chemical composition content of high potassium glass and lead-barium glass, but did not distinguish whether they were weathered or not. According to the conclusion of Question 1, weathering can cause changes in the chemical composition of high potassium glass and lead-barium glass, so the data in Form 2 were divided into two parts, one for the chemical composition without weathering and the other for the chemical composition after weathering. On this basis, the patterns of high potassium glass and lead-barium glass with and without weathering were analyzed, in which there were 28 sets of sample data for weathering and 26 sets of sample data for no weathering.

(2) Establishing dataset

The dataset is pre-processed to form a dataset that can be recognized by the random forest algorithm, and then divided into a training set and a test set in a certain way.

(3) Constructing the random forest classifier

Step 1: Using Bagging method to form n tree training sets, the number of samples in the original training set is assumed to be N , from which N samples are randomly selected with put-back to form a new training set. The training set is different for each tree, and the n tree training sets are formed by repeating n tree times.

Step 2: For each training set, a CART classification tree without pruning is generated using the following procedure.

Let the feature dimension of each sample be M , given a constant m_{try} , satisfying $m_{try} \ll M$.

At each internal node, a subset of m_{try} features is randomly selected from the M features as the marquee feature subset for that split node. In the process of generating the entire forest, m_{try} remains unchanged.

The optimal splitting attribute for that node is selected from the m_{try} candidate feature subset using the Gini index.

For each tree, it is made to grow sufficiently to generate the maximum tree T_{max} such that each leaf node in the tree is very small, or is a pure node, or no longer has attributes that can be used as branching attributes. A small leaf node can be expressed in the algorithm as the number of samples contained inside the node is less than a given threshold. A pure node indicates that the samples inside the node belong to the same class. No pruning is performed for the maximum tree T_{max} , which makes the random forest less prone to overfitting and has good noise immunity.

Step 3: Repeat step 1 to step 2 until n tree classification trees are generated.

(4) Classification using classifier

When classifying samples of unknown class, the output class of the random forest is determined by the simple majority voting method.

$$c = \text{argmax}_c \left(\frac{1}{n_{\text{tree}}} \sum_{k=1}^{n_{\text{tree}}} I(h(x, k) = c) \right) \quad (2)$$

(5) Out-of-bag error rate

The random forest classification effect is related to the correlation of any two trees in the forest and the classification ability of each tree. Decreasing the number of feature selection mtry, the relevance and classification ability of the trees will decrease accordingly, and increasing mtry, both increase.

2.2. Solving the classification model based on random forest algorithm

The glass types high potassium glass and lead barium glass were used as labels and 14 chemical composition contents were brought into the model operation as features. The random forest algorithm was used to derive the feature importance magnitude of the 14 chemical composition contents in the unweathered condition of the surface of the artifact samples. The feature importance magnitude can reflect the correlation magnitude of each chemical composition content with the artifact sample type in the unweathered condition of the artifact samples.

Table 1. Table of importance of weathering-free features

Feature Name	Feature Importance
SiO ₂	3.10%
Na ₂ O	0.00%
K ₂ O	20.20%
CaO	8.00%
MgO	1.90%
Al ₂ O ₃	4.60%
Fe ₂ O ₃	3.30%
CuO	9.20%
PbO	25.10%
BaO	15.70%
P ₂ O ₅	6.10%
SrO	0.80%
SnO ₂	0.90%
SO ₂	1.10%

From Table 1, we can see the correlation between the types of artifact samples in the weathering-free state and 14 chemical components, among which, K₂O, PbO, and BaO have a greater degree of influence on the artifact samples. It shows that the type of artifact samples in the weathering-free state is closely related to its own material-related chemical composition, which is in line with the actual situation.

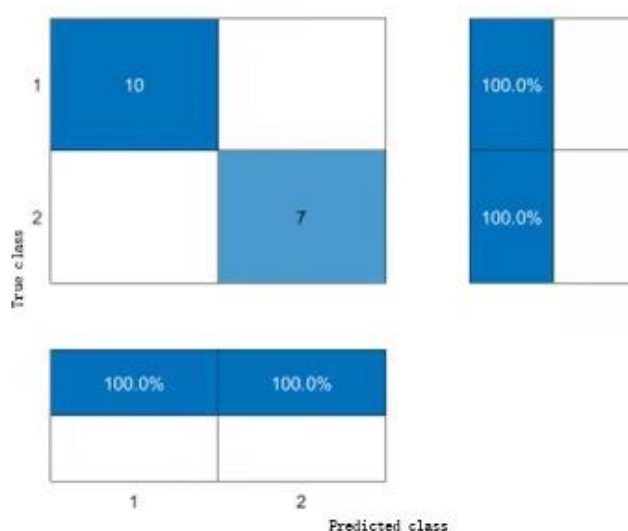


Figure 1. Map of the weathering-free training set

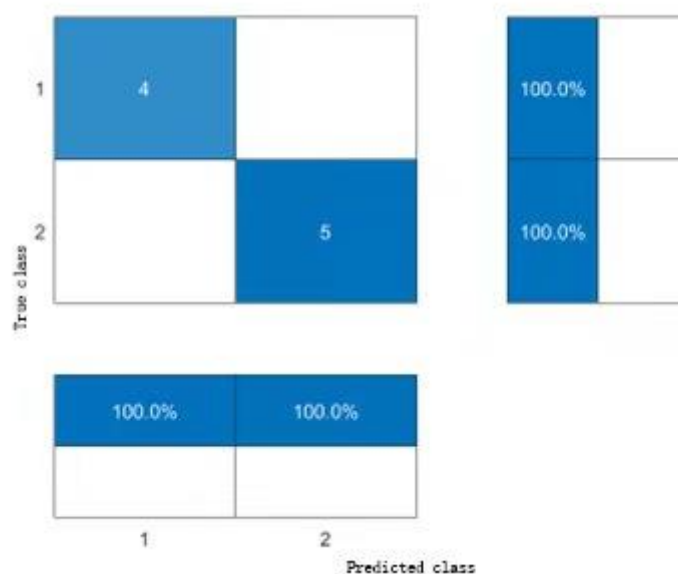


Figure 2. Weathering-free test set diagram

The number of training and test sets and their accuracy of training and testing can be seen in Figures 1 and 2. The unweathered sample data had 17 sets as training sets and 9 sets as test sets. The accuracy of both their training and test sets is 100%, which indicates the accuracy of the data and the reasonableness of the classification.

The correlation between the type of artifact samples in the weathered state and 14 chemical components, in which SiO_2 , PbO , BaO have a greater degree of influence on artifact samples, indicating that the correlation between the main components of glass and lead-barium glass after weathering is large, because lead-barium glass is harder to weather than high potassium glass, so the content of PbO , BaO is large after weathering, which has a greater influence on the type.

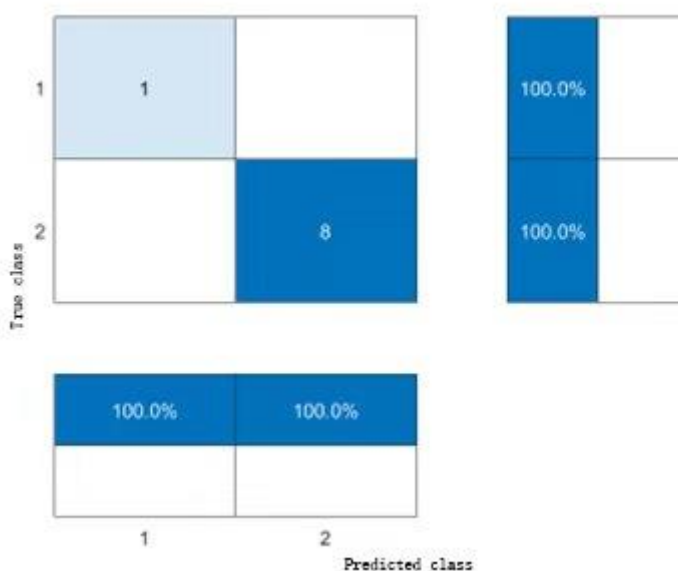


Figure 3. Weathering test set diagram

From Figure 3, it can be seen that the weathered sample data have 19 sets as training set and 9 sets as test set. The accuracy of their training and test sets is also 100%, which indicates the accuracy and reasonableness of the data.

The Form II data were first filtered and classified into two categories: high potassium glass and lead-barium glass, and then the two types of glass were subclassified separately. Here the subclass classification is solved by using the decision tree algorithm in the random forest algorithm in the

previous subquestion. Here the 14 chemical elements content is used as a feature and the presence or absence of weathering on the surface is used as a marker to apply the decision tree algorithm.

The decision tree algorithm was used to find out the characteristic importance of the 14 chemical elements of high potassium glass and lead-barium glass, and the specific data are shown in Table 2.

Table 2. Characteristic importance of chemical composition in high potassium glass 14

Chemical composition	High Potassium
CaO	0.2974
P2O5	0.2974
MgO	0.2874
K2O	0.2381
Al2O3	0.1172
PbO	0.1172
SO2	0.0810
Fe2O3	0.0701
Na2O	0.0513
SrO	0.0505
BaO	0.0459
SnO2	0.0240
SiO2	0
CuO	0

From Table 3, it can be seen that the characteristic importance of the four chemical components, CaO, P2O5, MgO, K2O, in the high potassium glass is greater, which are 0.2974, 0.2974, 0.2874, and 0.2381, respectively, and have a greater influence on the presence or absence of weathering on the surface of the artifact samples, so the change of the content of these four chemical components can easily cause the presence or absence of weathering, so the CaO, P2O5, MgO, K2O, these four chemical components are divided into high potassium glass easy weathering part. The remaining ten chemical components are naturally divided into the non-weathering part of high potassium glass.

Similarly, in lead-barium glass, the characteristic importance of three chemical components, PbO, P2O5, MgO, are 0.3271, 0.2880, 0.2176, which have a greater influence on the weathering of the surface of the artifact samples. Therefore, the three chemical components of PbO, P2O5 and MgO are divided into the easy weathering part of lead-barium glass, and the remaining eleven chemical components are divided into the less weathering part of lead-barium glass. The results are shown in the table 3-4.

Table 3. Confusion matrix for high potassium glass

	Accuracy	Recall	Accuracy
Training set	1	1	1
Test set	1	1	1

Table 4. Confusion matrix of lead-barium glass

	Accuracy	Recall	Accuracy
Training set	1	1	1
Test set	95%	95%	95%

Meanwhile, the confusion matrix yields 100% accuracy for both the training and test sets of high potassium glass and 95% accuracy for both the training and test sets of lead-barium glass. It indicates that the classification results of both glass types are reasonable.

In the high potassium glass type, except for CaO, P2O5, MgO, K2O, which are four chemical components with greater importance, the importance of the other features is small, and changing their magnitudes has less influence on the presence or absence of differentiation, so the sensitivity of the

classification results is poor. Similarly, it can be concluded that the sensitivity of classification results in lead-barium glass is also poor.

3. Glass classification results based on random forest algorithm

The chemical compositions of K₂O, PbO, and BaO were more correlated with the types of artifact samples in the unweathered state, and the correlations of Na₂O, SrO, and SnO₂ were smaller. In the weathered state, the chemical composition of SiO₂, PbO, and BaO correlated more with the type of artifact samples, and the correlation of SnO₂, So₂, and Na₂O was smaller.

Among the high-potassium glasses, CaO, P₂O₅, MgO, K₂O were divided into the easy weathering part of high-potassium glasses. The remaining ten chemical components are naturally divided into the non-weathering part of high potassium glass. In lead-barium glass, PbO, P₂O₅, MgO are classified as the easy weathering part of lead-barium glass, and the remaining eleven chemical components are classified as the non-weathering part of lead-barium glass. The reasonableness of the classification results is better and the sensitivity is worse.

3.1. Test Analysis

Set input space $\mathcal{X} \subseteq R^n$ is the set of n-dimensional vectors, The output space is the set of class tags $y = \{c_1, c_2, \dots, c_k\}$, The input is the feature vector $x \in \mathcal{X}$, Output as a class tag $y \in \mathcal{Y}$. X is defined in the input space $\mathcal{X}$, Y is defined in the output space $\mathcal{Y}$. $P(X, Y)$ is the joint probability distribution of X and Y . The training data set

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \quad (3)$$

Powered by $P(X, Y)$

The prior probability distribution is as follows:

$$P(Y = c_k) = \frac{\sum_{i=1}^N I(y_i = c_k)}{N} \quad (4)$$

The posterior probabilities are as follows:

$$P(Y = c_k | X = x) = \frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)} \quad (5)$$

The basic formula of the plain Bayesian method of classification is obtained by combining the prior probability and the posterior probability as follows,

$$P(Y = c_k | X = x) = \frac{P(y=c_k) \prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k) \prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}, k = 1, 2, \dots, k \quad (6)$$

The plain Bayesian classifier can be expressed as

$$y = f(x) = \underset{c_k}{\operatorname{argmax}} \frac{P(y=c_k) \prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k) \prod_j P(X^{(j)}=x^{(j)}|Y=c_k)} \quad (7)$$

The training on the presence or absence of weathering and chemical content of high potassium glass using the plain Bayesian algorithm yielded an accuracy of 92.3% and a feature importance score.

3.2. Comparison of test results

Comparing them with the eigenvalues obtained through the random forest algorithm, the following line graph was drawn using MATLAB software, see Figure 4

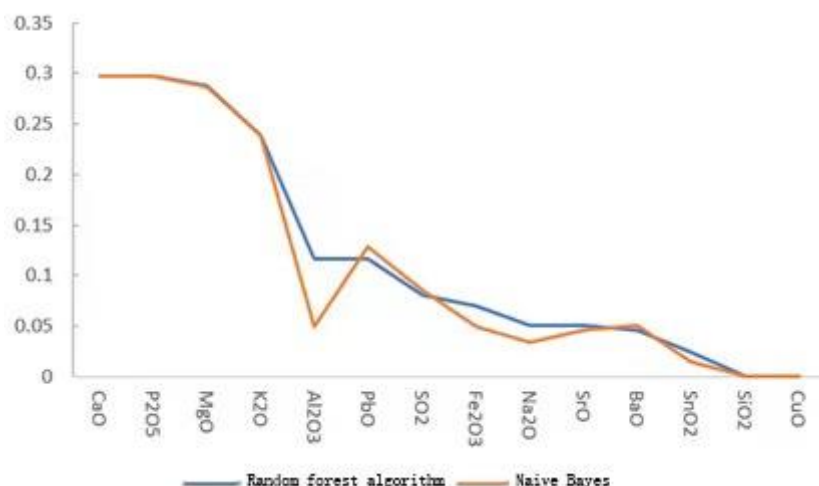


Figure 4. Line graph of feature importance scores of high potassium glass under two algorithms

As can be seen from Figure 4, the characteristic importance scores of chemical composition under both algorithms are nearly equal under the allowable error, thus, the data accuracy is high.

Similarly, the plain Bayesian algorithm was used to train the presence or absence of weathering and chemical composition content of lead-barium glass to obtain the accuracy of 75% as well as the feature importance scores. This was compared with the eigenvalues obtained through the random forest algorithm, and the following line graph was plotted using MATLAB software, see Figure 5.

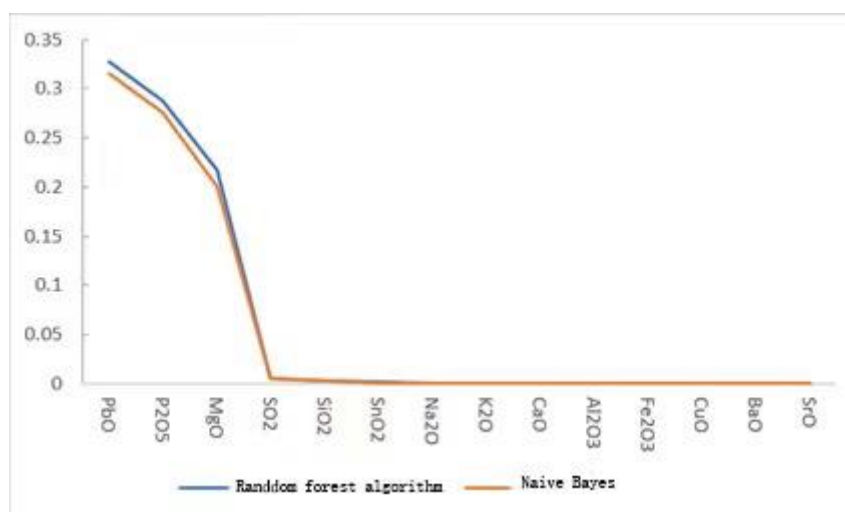


Figure 5. Line graph of characteristic importance scores of lead-barium glass under two algorithms

As can be seen from Figure 5, the characteristic importance scores of the chemical components are nearly equal under the allowed error for both algorithms, and therefore, the data accuracy is high.

4. Conclusions

In this paper, a classification model based on random forest algorithm is developed to solve the classification law of high potassium glass and lead-barium glass. Firstly, the data are classified into non-weathered case and weathered case. Then the random forest algorithm is applied separately. The random forest algorithm-based classification model was constructed by first forming a training set using the Bagging method and then generating a CART classification tree for each training set. The random forest algorithm was used to calculate the characteristic importance of chemical content in 14 without weathering, and the comparison showed that the three chemical components, K₂O, PbO, and BaO, had a greater influence on the glass type. Similarly, it can be obtained that the three chemical components, SiO₂, PbO, and BaO, have a greater influence on the glass type in the weathered state.

Finally, the importance of the features derived by the plain Bayesian algorithm is basically consistent with the above model, and the reasonableness of the model is verified.

Subsequently, a classification model based on decision tree was established to solve the problem of how to select the chemical composition for subclassification of glass types. Firstly, the tabular data were divided into two categories, high potassium glass and lead-barium glass, and each of the two glass types was classified into two categories, easy to weather and not easy to weather. A decision tree model was used to subclassify the chemical composition under the two glass types using a classifier. By comparing and analyzing the feature importance, it was concluded that the chemical compositions CaO, P₂O₅, MgO, K₂O in high potassium glass and PbO, P₂O₅, MgO in lead-barium glass were classified as easily weathered, and the rest of the compositions were not easily weathered. The accuracy of the model is 100% and 95%, respectively, which indicates that the results are reasonable. The characteristics of chemical components that are not easily weathered are less important and less sensitive.

References

- [1] Ke Di, Lin Shiwei. More detailed! The "public transportation" of recyclables [N]. Xiamen Daily, 2022-07-22(A04). DOI: 10.28890/n.cnki.nxmrb.2022.001991.
- [2] Liu C. X., Deng P. F., Pan S. H., Luo W. D., Zhang S. L., Li T. M.. Classification and research status of fireproof glass[J]. Guangzhou Chemical Industry, 2021, 49(15):16-18.
- [3] Huang Fan. Introduction to the application of super white glass in architecture [J]. Glass, 2021, 48(06):54-56.
- [4] Huang JY. Research on defect detection and classification method of cell phone panel glass based on digital image processing[D]. Xi'an University of Electronic Science and Technology, 2020. DOI: 10.27389/d.cnki.gxadu.2020.000709.
- [5] Wang Hongcheng, Liu Qiming. Effect of high voltage electric field on the weathering properties of soda lime silicate glass[J]. Journal of Wuhan University of Technology, 2009, 31(22):63-65+70.
- [6] WANG Chengxian, QI Ji, HUANG Ming, PANG Shihong. Weathering of float glass[J]. Journal of Wuhan University of Technology, 2009, 31(22):98-102.
- [7] WANG Chengxian, QI Ji, HUANG Ming, PANG Shihong. Weathering of float glass [C]//. Proceedings of the 2009 Annual National Conference on Glass Science and Technology of the Glass Branch of the Chinese Silicate Society. [publisher unknown], 2009:44-51.
- [8] Zhang B.L., Hong H.J., Liu F., Pan X.D.. Transmission electron microscopy study of volcanic glass weathering layers[J]. Journal of Mineralogy, 2008(03): 251-254. DOI: 10.16461/j.cnki.1000-4734.2008.03.008.
- [9] QI Ji, GURISTAN, WANG Chengxuan, NING Guiling. Study of alkali precipitation on the surface of silicate glass[J]. Glass and Enamel, 2006(03):9-13.
- [10] Wang Chengxian, Chen Min. Study on the effect of nitrogen ion injection on weathering of silicate glass by XPS[J]. Journal of Shanghai Institute of Building Materials, 1992(04):312-316.
- [11] Liu, Yuan-Xin, Jiang, Tsuen. Selection of evaluation methods for weathering of flat glass [J]. China Building Materials Science and Technology, 1992(06):29-33.
- [12] Wang Chengxian, Tao Ying, Chen Min, Huang Ming. Weathering of soda-lime alumino-magnesium silicate glass and alkali-lead silicate glass[J]. Silicate Bulletin, 1989(06): 1-9. DOI: 10.16552/j.cnki.issn1001-1625.1989.06.001.