Composition analysis and identification of ancient glass objects based on AdaBoost and CART classification tree

Ruoying Yang [*, #], Guoxiang Tian [#], Junyang Luo [#]

School of Mathematics and Information, CHINA WEST NORMAL UNIVERSITY, Nanchong, China, 637009

* Corresponding Author Email: 18398116560@163.com

[#]These authors contributed equally.

Abstract. Ancient glass is highly susceptible to weathering by the burial environment, and a series of chemical reactions will occur in the process, which leads to changes in the chemical composition of glass artifacts. In order to identify and classify the composition types of glass artifacts, this paper uses high potassium glass and lead-barium glass as target training models to derive CART stump (CART tree with only 2 layers) combinations as a way to analyze the classification laws. Then, we analyzed the sub-classification results of weathering, color, and ornamentation, and analyzed the classification rules according to their chemical composition, and came up with the classification method based on CART stumps. In order to identify the type of unknown types of glass, this paper uses an integrated learning algorithm model based on CART classification tree and AdaBoost to train a prediction model using all the samples, with the objective of artifact type, and then performs type prediction on the data. This study is important for the correct classification of glass types.

Keywords: AdaBoost and CART Classification Trees; Ancient Glass; Classification and Identification.

1. Introduction

As a precious artifact of the early trade exchanges along the Silk Road (ancient glass) and foreign glass surface similarities, but the composition is different[1-2]. The melting point of quartz sand as its main raw material is high, and in order to lower its melting point, it is necessary to add different chemical composition of combustion agents when making glass, thus classifying glass into lead-barium glass and potassium glass. However, ancient glass is susceptible to weathering due to the burial environment, causing the ratio of its internal composition to change, thus affecting the judgment of its type [3]. In this paper, we will establish a model to classify and identify ancient glass based on the classification information and the proportion of the main components [4].

2. Glass artifact classification under the optimization algorithm based on the combination of Adaboost and CART

2.1. Subtype classification of glass artifacts based on AdaBoost and CART classification trees

Step1: CART tree algorithm

The CART algorithm is a kind of decision tree, except that its branches are always bifurcated. The CART binary tree is generated using the chemical composition of the glass artifact as the root node. When the CART tree is used as a weak classifier, it is specified to generate only a two-level stump, which also represents a chemical composition feature. The corresponding CART stump is shown in Figure 1 below.

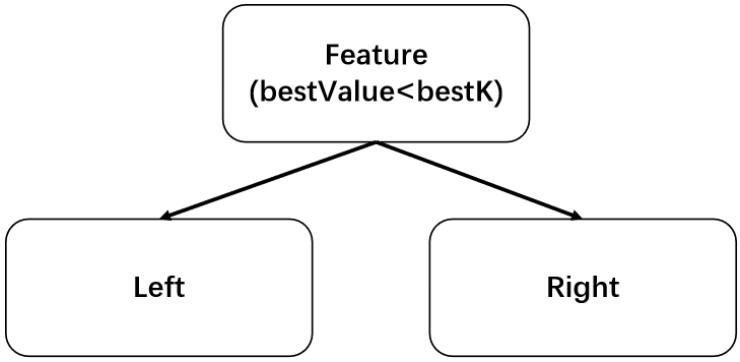

Figure 1. CART stump.

Step2: Apply the Gini index to generate the classification CART tree

First, calculate the Gini index of different chemical composition features of each glass artifact, if the smaller the Gini index of chemical composition features, the higher the purity of the sample, so the smallest Gini index of different chemical composition features of glass artifacts is selected as the optimal feature and the optimal cut point as the root node of the CART tree, the Gini index formula is as follows

$$\mathrm{Gini}(D) = \sum_{k=1}^{K} \sum_{j \neq k} p_k \, p_j = \sum_{k=1}^{k} p_k (1 - p_k) \tag{1}$$

((Assuming that there are K classes in a chemical composition dataset D, the probability that a sample belongs to class K is, one of the chemical composition dataset D is continuous irregular data, so it is converted into discrete data by using dichotomous discrete features, and then the minimum Gini index of the chemical composition feature data in the same class is calculated based on the discrete data using the above formula.

For a chemical component feature A, the set D is divided into D1 and D2, and the Gini index Gini (D, A) represents the minimum Gini index of the set after the division by A=a.

Gini (D, A) represents the uncertainty of the set D after the partitioning by A=a, and the formula is as follows:

$$\mathrm{Gini}(D, A) = \frac{| D_1 |}{| D |} \mathrm{Gini}(D_1) + \frac{| D_2 |}{| D |} \mathrm{Gini}(D_2) \tag{2}$$

Where |D|, |D1|, and |D2| denote the number of samples in datasets D, D1, and D2, respectively.

When the correct prediction rate of the test set of the algorithm reaches 100% or the root node of the CART tree continues to be selected without any change in the correct prediction rate of the test set, the generation of the CART tree is stopped.

Step3: Combining the Adaboost algorithm and CART tree

Adaboost is an iterative algorithm, the core idea is to train different classifiers for the same training set, and then combine these weak classifiers to form a stronger final classifier.

The basic idea of AdaBoost algorithm:

1. Multiple rounds of calls to the base learner;

2. In each round of calls, the weight of each sample in the loss function is adjusted for each sample in the sample set

3. Initially, all samples have equal weights, but after each round, the samples that are correctly classified are given smaller weights, and the samples that are not correctly classified have their weights increased.

4. Samples that are difficult to classify correctly will continue to receive high weights, allowing subsequent base learners to focus on and solve samples that are more difficult to classify.

Combining process:

First given the training sample set:

$$\mathrm{D} = \{(x_n, y_n)\}_{n=1}^{N}, x_n \in X, y_n \in \{-1, +1\} \tag{3}$$

Initializing the distribution:

$$D_1(n) = \frac{1}{N}, n = 1,2, \ldots, N \tag{4}$$

$$\varepsilon_t = P_{n\sim D_t}(h_t(x_n) \neq y_n) = \sum_{n=1}^{N} D_t(n)\, I(h_t(x_n) \neq y_n) \tag{5}$$

Take

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{6}$$

Perform the following update, for n = 1, 2, N

$$D_{t+1}(n) = \frac{D_t(n)}{Z_t} exp\left(-\alpha_t y_n h_t(x_n)\right) \tag{7}$$

$$\mathrm{H}(x) = sgn\left[\sum_{t=1}^{n} \alpha_t h_t(x)\right] \tag{8}$$

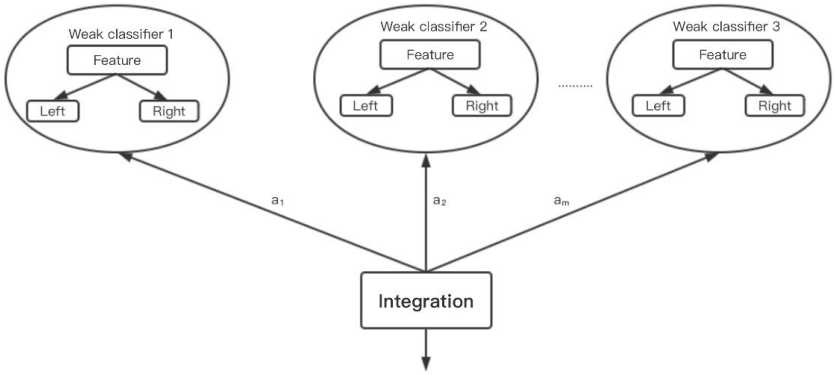

Figure 2. Integrated learner model.

The final integrated learner model is shown in Figure 2 above

Step4: The flow chart of the optimization algorithm based on the combination of Adaboost and the improved CART is as follows figure 3.

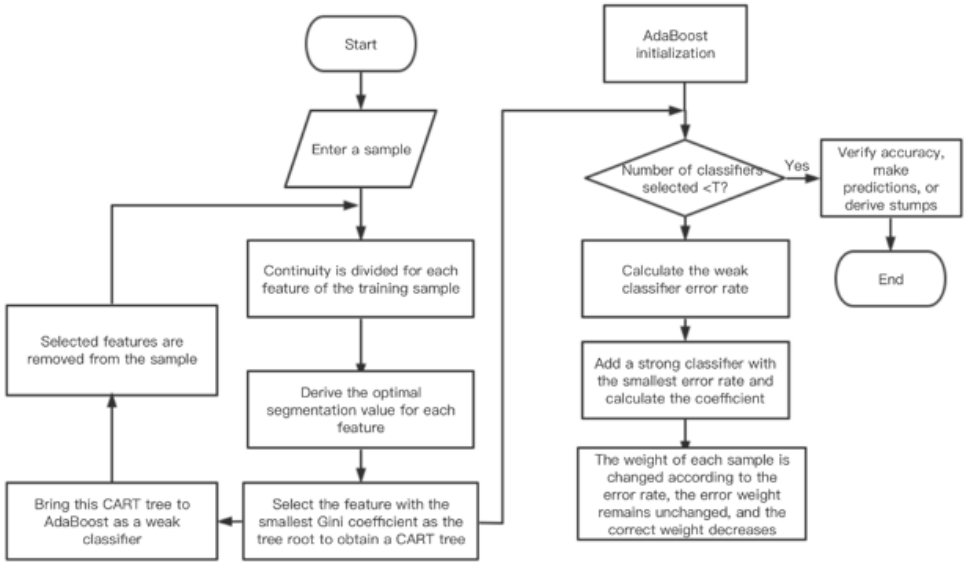

Figure 3. Flow chart of the optimization algorithm combining Adaboost and improved CART.

Features of the model algorithm:

1. Compared with the traditional AdaBoost and CART tree combination algorithm, the improvement of our algorithm is that the already selected features are removed from the sample database, and the already selected stumps are not selected when the next CART stump selection is performed.

2. AdaBoost has many advantages: AdaBoost makes good use of weak classifiers for cascading, and the weights of each classifier are fully considered to have high accuracy.

3, training using cross-validation, taking turns to select the test set and training set, traversing the entire data set, calculating the prediction accuracy, and finally the average prediction accuracy to evaluate the merits of the model.

Step5: By python programming, the following data were obtained as shown in the following table 1.

Table 1. Classification law of high potassium glass and lead-barium glass.

Priority	Best segmentation feature	Best segmentation value
1	Lead oxide	7.960204082
2	Barium oxide	3.047959184
3	Strontium oxide	0.13244898
4	Silicon dioxide	58.60673469

From the above table, it can be concluded that lead oxide dominates the classification of high potassium glass and lead-barium glass, followed by barium oxide, strontium oxide, and silica, which each classify the characteristic data with the best partition value, with decreasing effect from top to bottom. This shows the classification pattern of high potassium glass, lead barium glass [5-6].

Table 2. Subclassification results.

Main type	Subcategory	Priority	Select composition	Optimal split value	Accuracy
High Potassium	Weathering or not	1	Silicon dioxide	92.12102041	0.8
High Potassium	weathering or not	2	Aluminum oxide	3.878571429	0.8
High Potassium	weathering or not	3	Calcium oxide	1.957142857	0.8
High Potassium	Weathering or not	4	Potassium oxide	5.088163265	0.8
High Potassium	Color	1	Calcium oxide	7.7	0.8
High Potassium	Color	2	Sodium oxide	0	0.8
High Potassium	Color	3	Potassium oxide	12.49795918	0.8
High Potassium	Color	4	Silicon dioxide	65.84836735	0.8
High Potassium	Decoration	1	Aluminum oxide	3.85244898	0.15
High Potassium	Decoration	2	Potassium oxide	5.088163265	0.15
High Potassium	Decoration	3	Silicon dioxide	92.12102041	0.15
High Potassium	Decoration	4	Magnesium Oxide	0.98	0.15
Lead Barium	Weathering or not	1	Silicon dioxide	44.6	0.813
lead barium	Color	1	Barium Oxide	25.5755102	0.06
lead barium	Color	2	Aluminum oxide	1.692653061	0.06
Lead barium	Decoration	1	Aluminum oxide	3.633877551	0.836

The results of the subclass division are shown in Table 2 above. Classification method: When classifying into a certain category, the classification was made by the substance that plays a dominant role in the classification and the best division value. For example, under the subcategory of "weathered or not" of high potassium glass, the first division is based on silica (SiO2), which is greater than 92.121 and is classified as "weathered" and less than 92.121. "Then use alumina (Al2O3) as the secondary basis for similar classification [7-8].

Step6 Sensitivity analysis of the model:

Table 3. Sensitivity analysis.

Main type	Subcategory	Input Perturbation 1	Output change 1	Input Perturbation 1	Output change 2
High Potassium	Whether weathering	10%	2.21%	50%	24.20%
High Potassium	Color	10%	1.42%	50%	35.62%
high potassium	Decoration	10%	5.02%	50%	7.36%
Lead barium	Weathering or not	10%	0.41%	50%	45.52%
Lead barium	Color	10%	3.23%	50%	6.25%
Lead barium	Decoration	10%	0.31%	50%	57.65%

The sensitivity analysis is shown in Table 3.Sensitivity analysis is performed by calculating the rate of change of the output under 10% and 50% input perturbations. Taking 10% as an example, for a single value k of each set of input data, a perturbation value r is first generated randomly in the interval [-0.1, 0.1], and after multiplying k by (1+r), then k is randomly increased by the original r*100%, and the model is trained with the perturbed data, and the prediction is performed after training to detect the rate of change of the correct output value.

For a good model, the model should not change much with smaller input perturbations. After the input perturbation increases to a certain level, the sensitivity of the model should change more so that the model is discriminative to the input data.

The results show that under a 10% perturbation, most of the model outputs change less. At a 50% perturbation, most of the model outputs change more. The models with poor subclassification also perform poorly on sensitivity analysis, and conversely perform better and are generally consistent with the expected results. It indicates that under good subclassification, the model's input values do not cause large changes in the output values when they vary in a small range. In contrast, when the input values vary widely, the output values can change to a corresponding degree and the model is distinguishable. The model performs well in terms of sensitivity [9-10].

3. Type identification of artifacts to be tested by an optimized classification algorithm based on the combination of AdaBoost and CART

A combination of AdaBoost and CART is used for training, and the average correct rate is derived using cross-validation, with one model trained for each set selection method. After the dataset is traversed, the model with the highest correct rate is selected for prediction. For a set of chemical composition tuples, the prediction yields label values 1 and -1, and is reverse mapped back to high potassium and lead barium. Sensitivity analysis was performed by calculating the rate of change of the output at 10% and 50% input perturbations and obtaining the average rate of change after traversing the data set. The results showed that the output changed by 7.21% for 10% input perturbation and 49.32% for 50% input perturbation. With smaller input fluctuations, the output values do not change much. Under larger input fluctuations, the output values can change reasonably well. The model performs well in terms of sensitivity.

Table 4. CART Single-Story Stump List.

Select order	Best splitting characteristics	Optimal split value	Left division size	Right division size	Alpha factor
1	Lead oxide	7.960204082	18	36	6.907754779
2	Barium oxide	3.047959184	20	34	5.088455845
3	Strontium oxide	0.13244898	24	30	6.646787941
4	Silicon dioxide	58.60673469	27	27	6.907088006

Table 4 above shows the intermediate product CART stump at the corresponding training in this paper. Table 5 shows the predicted results.

Table 5. Prediction results.

SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO	BaO	P_2O_5	SrO	SnO_2	SO_2	Type
78.4	0	0	6.0	1.86	7.23	2.15	2.11	0	0	1.06	0	0	0.51	High Potassium
37.7	0	0	7.6	0	2.33	0	0	34.3	0	14.2	0	0	0	Lead Barium
31.9	0	1.36	7.1	0.81	2.93	7.06	0.21	39.5	4.69	2.68	0.5	0	0	lead barium
35.4	0	0.79	2.8	1.05	7.07	6.45	0.96	24.2	8.31	8.45	0.2	0	0	lead barium
64.2	1.2	0.37	1.6	2.34	12.75	0.81	0.94	12.2	2.16	0.19	0.2	0.49	0	Lead barium
93.1	0	1.35	0.6	0.21	1.52	0.27	1.73	0	0	0.21	0	0	0	High Potassium
90.8	0	0.98	1.1	0	5.06	0.24	1.17	0	0	0.13	0	0	0.11	high potassium
51.1	0	0.23	0.8	0	2.12	0	9.01	21.2	11.3	1.46	0.3	0	2.26	lead barium

4. Conclusions

Ancient Chinese glass artifacts are long-established, and at the same time have their own characteristic varieties of glass from our ancient people. The ancient glass is highly susceptible to weathering by the burial environment, and a series of chemical reactions will occur in the process, which will lead to changes in the chemical composition of glass relics. Therefore, exploring and analyzing the changes in the chemical composition of glass relics before and after weathering, and then identifying and classifying the relics, will help to gain a deeper understanding of the ancient glass-making technology in China, and thus fully demonstrate the national cultural confidence.

The color of a substance is determined by its constituents, so the factor that plays a major role in the classification of ancient glass is its constituents, and the known classification results seek to classify based on the CART classification tree and integrated learning based on the AdaBoost algorithm in the decision tree model in particular, followed by a second classification for both categories we continue to classify according to whether they are weathered or not, respectively, and then finally according to the characteristics of the texture In order to take into account the problem of small sample data for classification, the need for later model rationality and sensitivity testing, firstly, the samples were divided into training and testing sets, and cross-training was used in training the classification model. The reasonableness of the model is analyzed in two aspects: first, the accuracy of the test set results, the higher the accuracy, the stronger the reasonableness; second, the higher the final results of the two sub-classification models, i.e., decision tree and clustering, the higher the degree of proximity to the category to which the component content belongs, the stronger the reasonableness. The sensitivity analysis is divided into two aspects, one by fine-tuning the test set data if the category does not change or the category changes after a major adjustment of the data is considered as good sensitivity, otherwise it is considered as poor sensitivity; the other by substituting the non-important classification based on the component index data if the category does not change is considered as good sensitivity, otherwise the opposite.

In order to carry out type identification, this paper uses CART classification tree and integrated learning algorithm model based on AdaBoost algorithm to train a prediction model with artifact type as the target, and also takes cross-validation to test the accuracy of the model, after which the type prediction is carried out for the data in Form III, and the results are shown in Table. Sensitivity analysis was performed to observe the change of output data when the input data fluctuated. The results show that the model performs well in terms of sensitivity.

References

[1] Yin Weijie, Yang Lijun, Gu Yunlou, Mei Zhifan. Analysis of the composition of ancient glass based on cluster analysis method to identify the type to which it belongs [J]. Journal of Beijing Printing Institute, 2023, 31(06):64-67.

[2] Zhai Sixun. Composition analysis and identification of glass products based on decision tree [J]. Heilongjiang Science, 2023, 14(08):47-49.

[3] Xiong TW, Chu ZG, Lv FJ. Current status and progress of research on the natural derivation pattern of pulmonary ground glass nodules and CT differential diagnosis [J]. Chinese Journal of Lung Diseases (Electronic Version), 2023, 16(02):290-292.

[4] Zhang Xingliang, Xu Ke, Chen Junfang, Wu Hong. Quality of life classification tree analysis of heterosexual HIV-infected and AIDS patients in Hangzhou [J]. Chinese Journal of Preventive Medicine, 2023, 24(05):406-413.

[5] Tian Hao, Lu Bo, Yang Yandong, Bu Jianchong, Deng Jianxin, Li Dongchang. Analysis and prediction of substation construction safety accidents based on CART regression tree model [J/OL]. Journal of Xiangtan University (Natural Science Edition):1-8 [2023-06-30].

[6] Lv Fei,Fu Hangwei, Liu Chenglin. Composition analysis and identification of ancient glass products based on machine learning [J]. Information and Computer (Theory Edition), 2023, 35(04): 98-102.

[7] Kong, H.P., Ruan, David. Random forest-based traction motor fault feature selection method for rolling stock [J]. Railway Vehicles, 2023, 61(01):110-115.

[8] Y. J. Xu, J. H. Liang. Construction of student performance evaluation model based on K-means fusion decision tree classification algorithm [J]. Wireless Connected Technology, 2022, 19(22):134-137.

[9] Zhang X. F., Yusuf Jiang-Rusuli, Qiu Zongli, Yashar Esker, Abdulgehman Guzman. Research on remote sensing classification and accuracy evaluation of agricultural crops based on different machine learning--Fukang City, Xinjiang Uygur Autonomous Region as an example [J]. Journal of Xinjiang Normal University (Natural Science Edition), 2022, 41(03):17-28.

[10] Lin Shengcheng, Liu Haihua, Huang Yonglin, Chen Lin, Gong Xiang. An analysis of the identification method of sapphire glass for watches [J]. Science and technology innovation and application, 2017(02):88.