ISSN: 2688-8645

Glass Composition Analysis and Identification based on Improved Decision Tree Algorithm

Junyan Tian, Zhijing Wu, Xin Qu

School Of Electronic and Optical Engineering, Nanjing University of Science and Technology Nanjing, China, 210094, China

Abstract

Glass is a valuable material evidence of our early trade exchanges, and our glass is mainly divided into high potassium glass and lead-barium glass. The specific types can be distinguished by examining the content of their internal compounds. This paper delineates the broad types of each unknown artefact and then predicts the composition before weathering. Combining the results of existing predicted subclasses as known data, 80% of the data is selected as the training set and 20% as the test set. Suitable node splitting evaluation criteria and feature division point selection criteria were determined, and a glass component analysis and identification model based on an improved decision tree algorithm was established to predict the major unknown artefact types separately and obtain their specific subclasses. The sensitivity analysis of the established identification model is carried out by adjusting the proportion of training data as well as the principal component weights.

Keywords

Glass Artefacts; Decision Tree Model; Sensitivity Analysis.

1. Introduction

First produced in China, silk has a long history and the Silk Road was the artery that facilitated technological, economic, and cultural communication between Europe and Asia, with China as its central point in Asia [1]. In terms of the exchange of technology and culture between China and foreign countries in ancient times, the Silk Road was of great value and significance, and had a profound impact on the spread and exchange of ancient Chinese glass products and technology [2]. After the introduction of glass to China from the West Asian and Egyptian regions, the working people, after absorbing the production process and taking local materials, produced the unique ancient glass invented by China itself. Although the appearance of these glasses is similar, but due to the addition of different fluxes to make the chemical composition of the difference, there are two more popular varieties of glass in ancient China: the Chu culture of lead barium glass and popular in the Lingnan area of potassium glass. Under the influence of being buried for long periods of time, ancient glass is extremely susceptible to weathering, during which the internal elements are exchanged in large quantities with other elements in the environment resulting in changes in the composition and proportions contained in the glass [3-4]. This paper analyses the data for the high potassium glass and the lead-barium glass in the forms separately to identify classification patterns. For each category of glass, the chemical composition is analyzed and a refined subcategory is selected based on the most appropriate classification. The method and results of the classification are presented and analyzed for reasonableness and sensitivity [5]. The chemical composition of the glass artefacts in the data sample of unknown categories needs to be analyzed, and the data analyzed is used to identify the types, while the results are analyzed for sensitivity and accuracy [6].

ISSN: 2688-8645

2. The Fundamental of Cluster Analysis

Cluster analysis, also known as cluster analysis, is a multivariate statistical analysis method for quantitatively classifying multiple samples (or indicators). This question requires us to classify the sample detection points given in the Annex, and therefore Q-type cluster analysis is used. Similarity measures for samples [7].

If for the sample to be classified P variables are needed to describe it, then each sample point can be seen asR^p a point in the space. Distances can therefore be used to measure the degree of similarity between sample points. For quantitative variables, the most used is the Minkowski distance

$$d_q(x,y) = \left[\sum_{k=1}^{p} |x_k - y_k|^q\right]^{\frac{1}{q}} \tag{1}$$

When q=1,2 or q tends to infinity, then the absolute value distance is obtained

$$d_1(x,y) = \sum_{k=1}^{p} |x_k - y_k| \tag{2}$$

For each samplex_i ,assign it to the nearest centre

$$x_i^t < -argmin_k \left||x_k - y_k|\right|^2 \tag{3}$$

For each class centre k, recalculate the centre of that class

$$\mu_k^{(t+1)} < -argmin_\mu \sum_{i:c_i^t=k}^{b} \left||x_k - y_k|\right|^2 \tag{4}$$

The clustering results are shown in Figures 1 and 2 below.

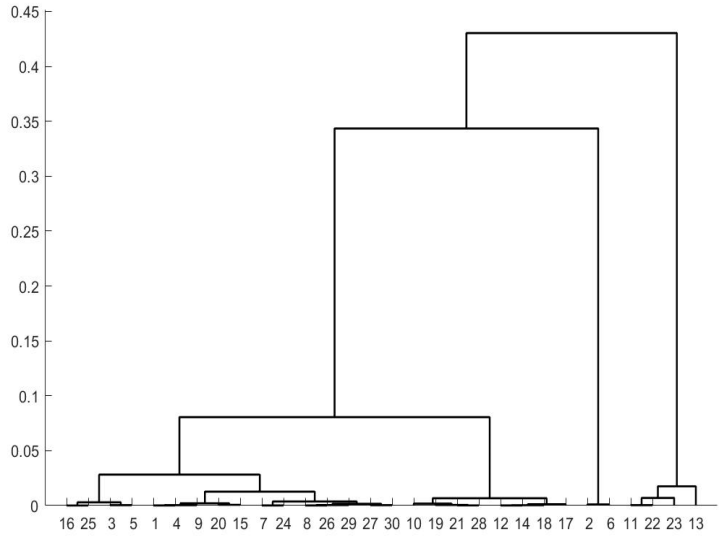

Figure 1. Barium and calcium cluster analysis

Artifacts belonging to the first category are: 2, 6, 34. The following objects belong to the second category: 1, 3, 4, 5, 7, 8, 9, 10, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 38, 39, 40, 41, 42, 43, 44, 45, 46. The following objects belong to the third category: 11, 13, 36, 37

ISSN: 2688-8645

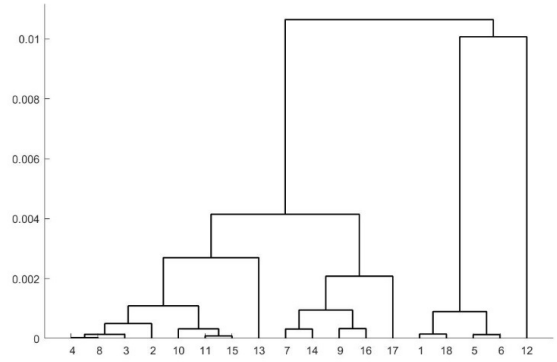

Figure 2. High Potassium Cluster Analysis

The artefacts belonging to the first category are: 12. The artefacts belonging to the second category are: 1, 5, 6, 18. The artefacts belonging to the third category are: 2, 3, 4, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17. High potassium glass is mainly classified based on calcium and aluminium content, and in conducting the sensitivity analysis, both were chosen to be weighted more heavily and re-substituted into the Q-test for analysis, resulting in a classification The results were significantly different from the original classification results, therefore the calcium and aluminium content had a greater impact on the model analysis, the results of which are shown in Figures 3 and 4 below.

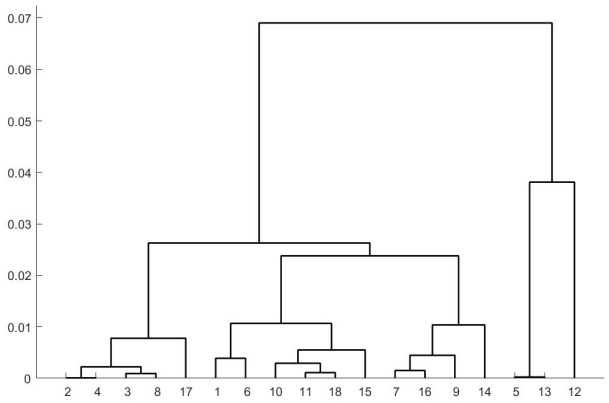

Figure 3. High potassium sensitivity analysis

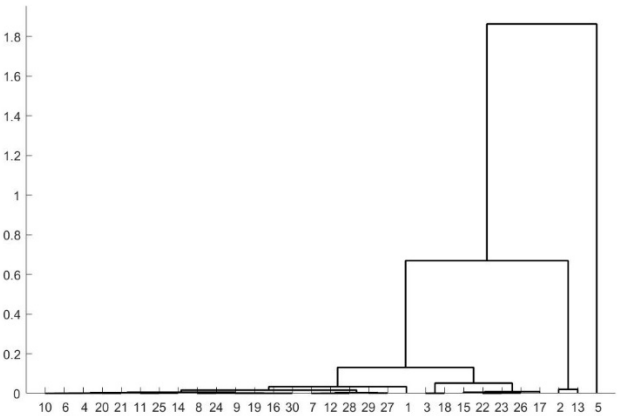

Figure 4. Lead and barium sensitivity analysis

ISSN: 2688-8645

Similarly, the main classification of lead-barium glass is based on the lead-barium compound content, and when sensitivity analysis is carried out, the weights of both are increased and substituted into the Q-type test for analysis, yielding classification results that are also significantly different from the original classification results, so the lead-barium compound content has a greater impact on the model analysis.

3. Results

3.1. The Establishment of Simulation Model

The specific algorithmic steps of the random forest decision model are as follows.

(1) Training set drawn from the original sample set. In each round, n training samples are drawn from the original sample set using the Bootstraping method (with put-back sampling). A total of k rounds are performed to obtain k training sets. (The k training sets are independent of each other)

(2) One model is obtained using one training set at a time, and a total of k models are obtained from k training sets.

(3) For the classification problem: the k models obtained in the previous step are used to obtain the classification results by voting; for the regression problem, the mean value of the above models is calculated as the final result.

The percentage of each chemical component is an indication of the importance of that chemical to the determination of its category, with larger percentages indicating greater importance to the decision.

In this paper, we supplement the unknown artefact data given in the title by performing a zero-filling operation on all vacant compounds. We have given the species categories: weathered high potassium, unweathered high potassium, severely weathered high potassium, weathered lead barium, unweathered lead barium, and the average of the values for each compound, so we scored them using the following formula.

$$Q = \sum_{i=1}^{14} |\omega_i - W_i| \times \varepsilon_i\% \tag{5}$$

whereω_i denotes the content of element i of the artefact given in the appendix to the title, andW_i denotes the average of the content of element i calculated in 5.3.1, andε_i% Q is the score of the artefact, which is the sum of the weighted distances from each element of the artefact to the mean.

The weight distance and Q value were calculated separately for each artefact and the five categories, with a smaller Q value indicating that the artefact is closer to the category, so it is classified as being in that category.

3.2. Analysis of Experimental Results

Decision tree construction can be carried out in two steps.

The first step, decision tree generation: the process of generating a decision tree from a training sample set. In general, the training sample dataset is the dataset used for data analysis and processing according to the actual needs, and we choose to use the values of each category after the second question prediction as the dataset to build the decision tree.

The second step, pruning of the decision tree: pruning of the decision tree is the process of checking, correcting, and trimming down the decision tree generated in the previous stage, mainly by using the data in the new sample dataset (called the test dataset) to check the initial

ISSN: 2688-8645

rules generated during the generation of the decision tree and to cut out those branches that affect the accuracy of the rebalance.

A sensitivity test of the subclass decision tree classification model by varying the weight test of the main classified compounds in the sample to be tested yielded that the selected main compounds have a greater influence on the model classification, thus proving that the model is suitable for compound-based glass classification [8-9].

4. Conclusion

This paper describes the process of classifying ancient glassware with a degree of accuracy and ingenuity, and can make scientifically sound decisions through a decision tree model. The random forest is extremely accurate, works effectively on large data sets, introduces randomness, and is not prone to overfitting. Random forests are very resistant to noise, but can overfit when the data is relatively noisy. Can handle very high dimensional data without dimensionality reduction. Can handle not only discrete data but also continuous data and does not require normalization of the data set. Decision trees are easy to understand and interpret, and require less data for training than other machine learning models that typically require data normalization, such as constructing dummy variables and removing missing values. The number of data points used to train a decision tree result in an exponential distribution of the overhead of using decision trees (the time complexity of a training tree model is the logarithm of the number of data points participating in the training). Although the random forest algorithm is fast enough, when faced with many decision trees in a random forest, the space and time required for training can be significant, resulting in a slower model. Therefore, in practice, it is better to choose other algorithms if real-time requirements are very high. The decision tree model tends to produce an overly complex model, which has a poor generalization performance to the data. This is known as overfitting and some strategies like pruning, setting the minimum number of samples needed for a leaf node or setting the maximum depth of the number are the most effective ways to avoid this problem. The analysis of this problem can also be applied to the problem of siting and dispatching of other transport exchanges, e.g., e-bikes, where the parameters can be adjusted to solve the problem with reference to this model. In addition, the model is also useful for optimal path planning and selection.

References

[1] Tian Hongpeng, Wei Tian. Modular decision forest for blockchain transaction fraud detection model [J]. Computer Engineering and Applications:1-13.

[2] Wang Fuyue, Ren Yi, Zhao Tan, Cui Fuxiang. Optimization of steel plate flaw detection model based on decision tree algorithm[J]. Anshan Steel Technology,2022, (06):33-38.

[3] LI Wei, CHEN Jianhua, WU Shaowei, XIA Rujun, CHEN Xiaoquan. Identification of production anomalies in offshore oil and gas wells based on SPC control charts and weighted decision trees[J]. Frontiers in Marine Geology,2022,38(12):84-91.

[4] Wei, Dongni, Che, Bin, Zhang, Zelong, Tang, Mengyuan, Qi, Caijuan. Information processing technology for think tank talents based on concomitant data collection and decision tree algorithm [J]. Electronic Design Engineering,2022,30(23):56-60.

[5] Lei Chao, Cui Xin, Wang Zhifei, Xie Yanming. Pharmacoeconomic evaluation of epilepsy treatment with epilepsy capsules based on decision tree model[J]. China Hospital Drug Evaluation and Analysis, 2022, 22(11):1365-1367+1374.

[6] ZHU Jinglong, YU Leiyan, WU Guiwei, SUN Dezhi, XU Yongkang, XU Kai. A virtual simulation experiment platform for driverless vehicle lane change decision[J]. Experimental Technology and Management,2022,39(11):99-104+110.

ISSN: 2688-8645

[7] Yan Shaowei, Li Zhuo, Wan Kai, Zhang Ming, Lin Shonan. Decision tree-based algorithm for optimal reconfiguration of multi-voltage grid fault recovery[J]. Information Technology, 2022, 46 (11):183-188.

[8] WANG Lijun, DU Jianhua, LIU Jichao, WANG Shuangshang, XIE Hansheng, ZHAO Bing. Design of meteorological big data cloud platform based on decision tree mining algorithm[J]. Computer Measurement and Control,2022,30(11):140-146.

[9] Wang Jiao. The application of decision tree algorithm in employment prediction of college graduates--a case study of Pu'er College[J]. Digital Technology and Applications,2022,40(11):85-87.

[10] Serious, He Suyu, Huang Ping, Jiang Huamei, Gong Changhao, Zhu Xiaofeng. A study of DRGs inpatient cost criteria grouping for cirrhotic patients based on decision tree model[J]. China Medical Case, 2022, 23(11):55-59.