Classifying the chemical content of cultural relics using the decision tree CART classification model

Zihan Wang *

Shandong University of Finance and Economic, Jinan, China

* Corresponding author: 15106330077@163.com

Abstract. In this investigation, the decision tree CART classification model was used to obtain the decisive relationship between various chemical components on glass type, and Matlab was used to calculate the classification criteria of the decisive chemical components, to make a decision tree diagram to determine the glass type, and to summarise the distribution law of barium oxide, silica and alumina content to determine the glass type. The most determining 8 chemical components were then selected based on the Fisher score expression in the feature selection model, which calculated lead oxide, potassium oxide and other components as classification criteria.

Keywords: Decision tree, CART classification model, Classification criteria.

1. Introduction

As a valuable testimony to trade along the Silk Road, glass is a historical crystallization of the economic exchanges between the East and West in ancient times, and is also an important source of material for modern research into ancient historical relics.

The ancient glass-making process in China drew on the technical methods spread by the ancient West, and was made locally from local materials, particularly in the refining of lead ore, a common local material, as a flux to make our unique lead and barium glass, while in the south of the country and in regions such as Southeast Asia and India, local materials such as grass ash, which has a high potassium content, were also chosen as fluxes to make potassium glass.

Ancient glass objects were susceptible to the effects of weathering in the environment in which they were buried. During the weathering process, the internal elements undergo a significant exchange with elements in the air and soil, resulting in changes in the proportions of the glass composition, which can affect the archaeologist's ability to make an accurate determination of its category. The classification patterns of high potassium glass and lead-barium glass are analysed on the basis of the attached data.

2. Materials and Methods

2.1. Data

To ensure the rigour and reliability of the data and to facilitate the collation and extraction of the data, the data in this study refer to the data of the second question in Question C of the 2022 National Student Mathematical Modelling Competition (mcm.edu.cn), and the data are collated and washed.

2.2. Introduction to the method

Firstly, the distribution pattern of chemical constituents under two glass types, lead-barium glass and high-potassium glass, was mainly explored. In order to obtain the decisive relationship between various chemical constituents on glass types, we applied the decision tree CART classification algorithm model, used the data in Form 2 as training data, centrally summarized a set of classification rules, and used Matlab to calculate the classification criteria of decisive chemical constituents to make decision tree diagram to determine the type of glass. The rules can also be supplemented by a summary of the rules in terms of the nature characteristics of the artefacts.

3. Model building and solving

3.1. Decision tree CART classification algorithm model

As it involves the analysis and calculation of the percentage and priority decisions of many components, we choose the decision tree analysis model [1-4], use MATLAB programming to calculate the components that have a decisive role in the type of artefacts, and create a decision tree diagram to obtain the flow chart of glass type judgement.

3.2. Decision tree CART classification algorithm model building

First we visualised the data in Form 2 to obtain a rough expression of the law, divided into weathered and unweathered data, calculated the mean value of each component under each type, in order to obtain an intuitive representation of the law, and made a histogram based on this to study the classification law of high potassium type glass and lead barium type glass in weathered data, respectively, and the law of high potassium type glass and lead barium type glass in unweathered data to make a rough prediction of the classification of glass In order to obtain an approximate prediction of the classification pattern of the glass, the pattern of high potassium type glass and lead-barium type glass in the unweathered data were studied separately.

In order to obtain accurate classification criteria for the above picture findings, we applied the decision tree CART classification algorithm model of the data mining classification method to summarise the classification laws for lead-barium and high-potassium glasses.

In constructing the chemical composition decision tree, we used the Gini coefficient [5] [6] as a measure and applied the formula

$$Gini(p)=\sum_{k=1}^{k} p_k(1-p_k)=1-\sum_{k=1}^{k} p_k^2 \tag{1}$$

We use the deterministic constituents of the artefacts as branch nodes of the decision tree [7] [8], and the leaf nodes of the tree are the types of glass that flow to the corresponding type branches when the branch nodes of the artefacts face decisions where the deterministic constituents of the artefacts are greater or less than the numerical criteria of the nodes. Due to the small amount of data, we do not need to prune the decision tree to get very reasonable results.

When processing the data, we first divide this into a test set [9] and a training set, and then delineate the reference indicators and classification results. Immediately afterwards, we calculate the quantitative law of each chemical component for glass classification by creating a decision tree classifier in matlab [10] and applying the formulae in the model to solve for the final classification law.

We can also analyse the relationship between the nature of the artefact and the glass type of the artefact as a complement to the relationship between the artefact type and its chemical composition.

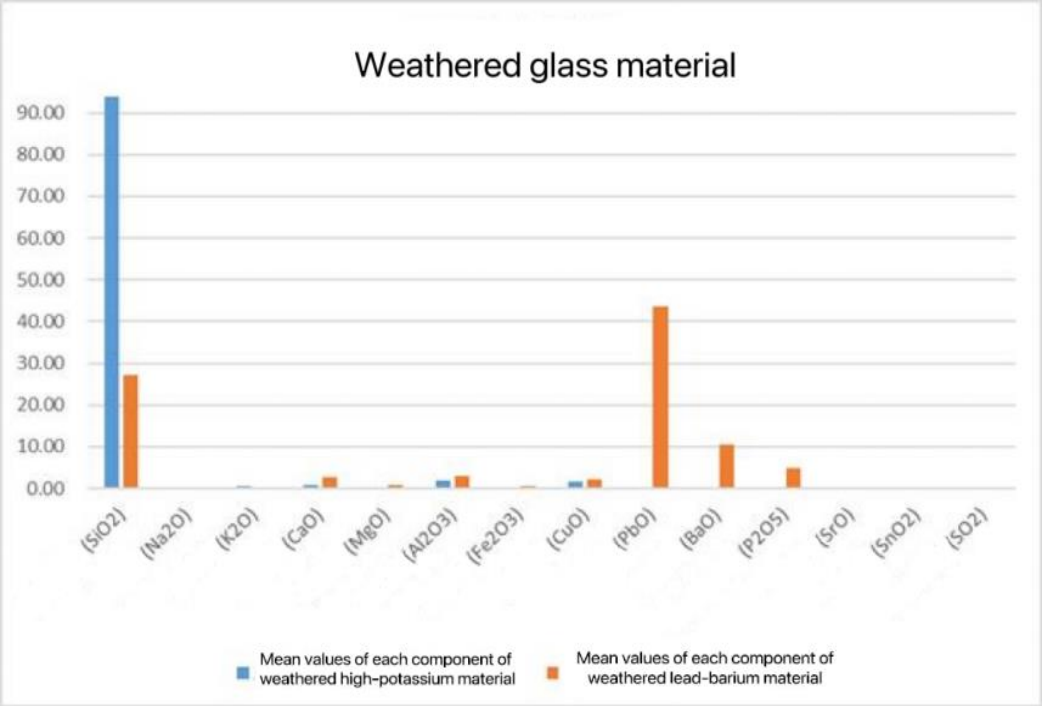


Figure 1. Distribution of weathered glass data

The results are shown in Figure 1, through Figure 1 we can see that in the weathered glass data, the chemical composition of the high potassium type has the largest proportion of silica components, aluminium oxide, copper oxide and calcium oxide are next, but the proportion is extremely low, and the proportion of the remaining chemical elements can be ignored: the chemical composition of the lead-barium type has the highest proportion of lead oxide, the second highest proportion of silica, barium oxide and phosphorus pentoxide have a smaller proportion, and the remaining The rest of the chemical composition is extremely low.

In addition to the high proportion of silica as the main component of the glass, it is tentatively determined that aluminium oxide, copper oxide and calcium oxide have a greater influence on weathering in the high potassium type, and that barium oxide and phosphorus pentoxide have a greater influence on weathering in the lead-barium type.

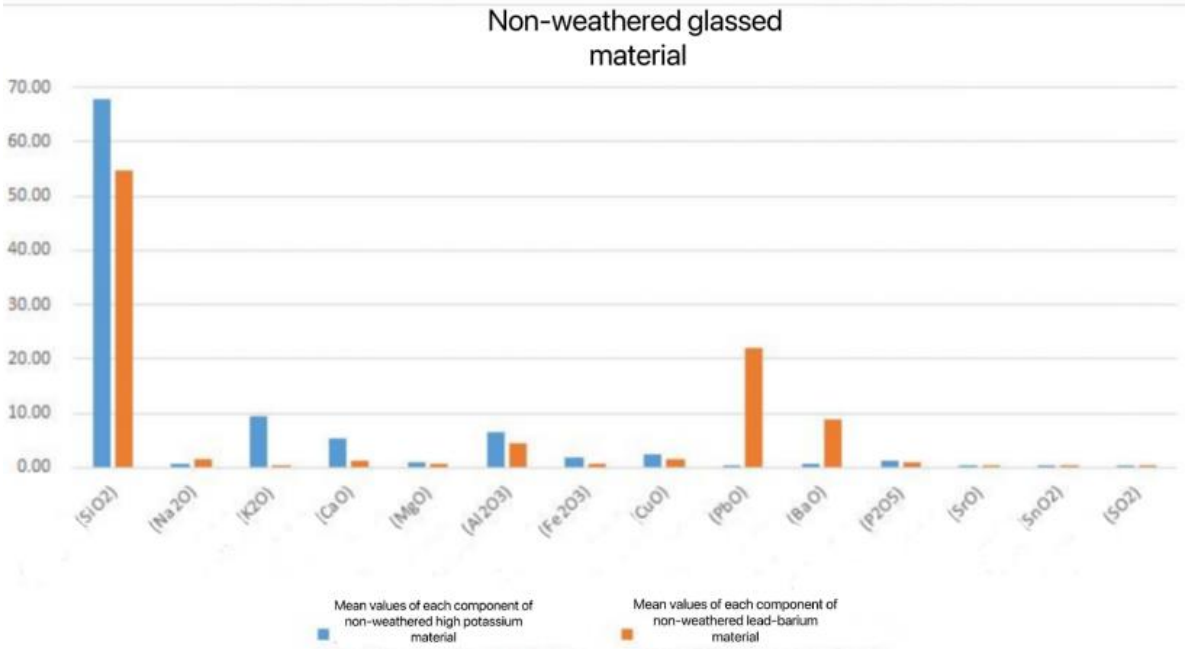


Figure 2. Distribution of data on unweathered glass materials

The results are shown in Figure 2, where we can see that silica is very high in the high potassium type of glass, followed by potassium oxide, calcium oxide and aluminium oxide, with the rest of the chemical composition being very low.

Again, in addition to the high proportion of silica as the main component of the glass, it is tentatively determined that the chemical compositions of potassium oxide, calcium oxide and aluminium oxide have a greater influence on unweathering for the high potassium type, and lead oxide, barium oxide and aluminium oxide have a greater influence on unweathering for the lead-barium type.

A comparison of weathered and unweathered data shows that barium oxide, silica oxide, calcium oxide and aluminium oxide have the greatest influence in determining the lead-barium type versus the high potassium type.

In order to obtain accurate data support, we will apply a decision tree model to the analysis.One point to note is that the high proportion of lead barium glass due to the intrinsic lead oxide component results in the other constituents not having a significant influence in the discrimination of lead barium glass from high potassium glass, so we ignore the influence of the lead oxide component and use MATLAB to program the data for the proportions of the remaining thirteen chemical constituents for analysis. , eventually achieving more reasonable results as shown in Figure 3.

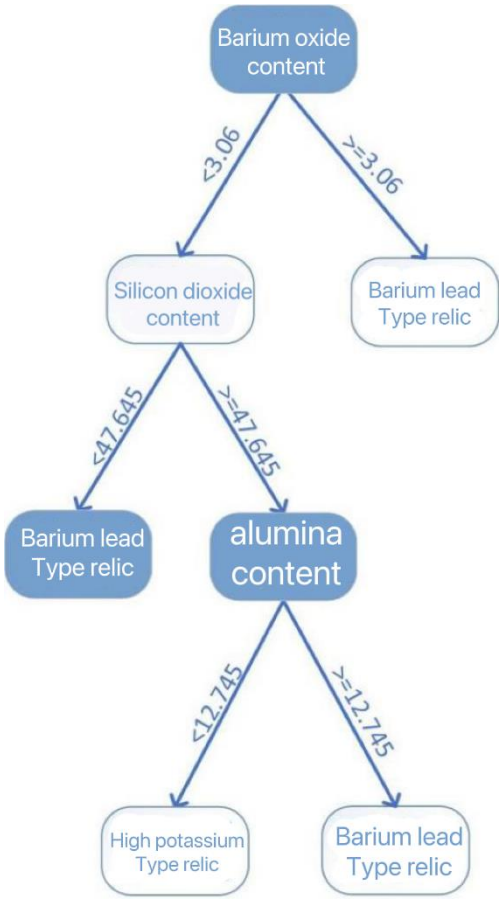


Figure 3. Decision tree diagram

Using the table filtering function, we have added to our analysis of the relationship between glass type and glass properties, i. e. colour, decoration and weathering state, the following conclusions have been roughly identified.

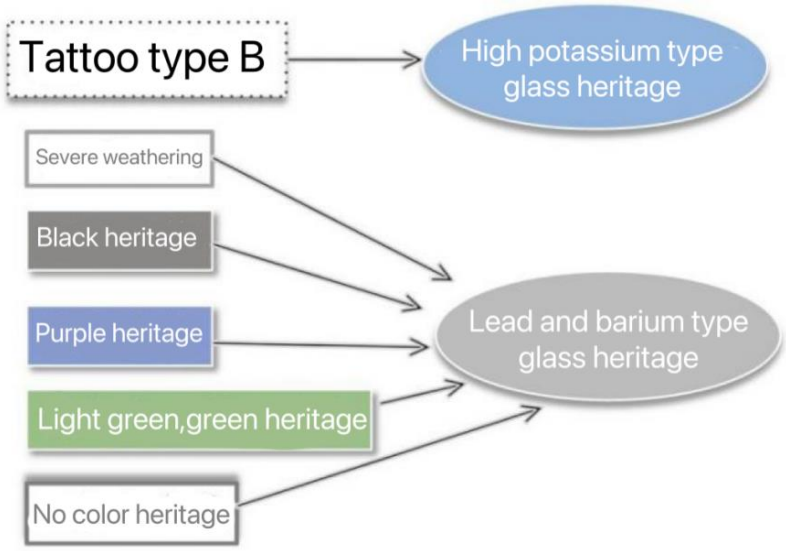


Figure 4. Partial relationship between nature and type of heritage Results

From Figure 4, we can determine the type of artefact in three ways: the type of decoration, if the type of decoration is B, then the type of glass can be judged as high potassium material type; the degree of weathering, if the artefact has severe weathering parts of the point, then the type of glass can be judged as lead barium material type; the colour of the artefact, if the colour of the artefact is black, purple, green, light green or no colour, then the type of glass can be judged as lead barium material type. green, light green or colourless, then the glass type of the artefact can be judged as a lead-barium material type.

The chemical composition of the artefacts and the nature of the artefacts complement each other and form the basis for the classification of lead-barium glass and high-potassium glass.

3.3. Influence value extraction

In order to select the appropriate chemical components among high potassium and lead-barium glasses and to classify subclasses within different glass categories, we propose a feature selection model, i.e. the 14 chemical components of the two glass categories using Fisher scores, so as to select the most decisive chemical components.

(1) Model development

① Feature selection model

For the appropriate chemical components in the two glass categories, the ideal feature should be that the values taken in the same category are more approximate, and the values taken between different categories are more different, for this reason we express the importance of the first chemical component by the Fisher score, and its expression is as follows.

$$S_i = \frac{\sum_{j=1}^{K} n_j(\mu_{ij}-\mu_i)^2}{\sum_{j=1}^{K} n_j \rho_{ij}} \tag{2}$$

where $n_j$is the number of samples in the different glass categories, j=1,2; $\mu_i$is the mean of the ith chemical component in all glass artefacts, i=1,2,...,14; $\mu_{ij}$and $\rho_{ij}$is the mean and variance of the ith chemical component in glass category j, respectively.

② Systematic clustering model

The systematic clustering model combines the two closest data points by calculating the distance between them, and iterates this process until all data points are combined into one class and a clustering spectrum graph is generated. Using the systematic clustering model to cluster the chemical components selected above, the most appropriate cluster centres can be selected among the different

glass classes, the optimal number of clusters can be judged using the elbow rule, and subclasses can be achieved by observing the clustering genealogy graph.

(2) Solving of the model

We first calculated the mean values of the 14 chemical components in all glass artefacts (excluding severely weathered artefact types), then classified them by high potassium glass and lead-barium glass, found the number of samples in different glass categories, and calculated the mean and variance of the 14 chemical components in high potassium glass and lead-barium glass respectively Substituted the data into the Fisher score expression, used matlab to solve, and The data were sorted in descending order by score and the results are shown in Table 1 below.

Table 1. Score ranking chart

i	Chemical composition	S_i
9	(PbO)	1.3367
3	(K2O)	0.9819
1	(SiO2)	0.9364
10	(BaO)	0.5064
12	(SrO)	0.4122
4	(CaO)	0.1298
11	(P2O5)	0.0728
7	(Fe2O3)	0.0683
6	(Al2O3)	0.0352
2	(Na2O)	0.0180
13	(SnO2)	0.0080
5	(MgO)	0.0069
14	(SO2)	0.0068
8	(CuO)	0.0042

We selected the first eight chemical components in Table x as suitable chemical components and eliminated the last six, i. e. the suitable chemical components are lead oxide (PbO), potassium oxide (K2O), silicon dioxide (SiO2), barium oxide (BaO), strontium oxide (SrO), calcium oxide (CaO), phosphorus pentoxide (P2O5), and Iron oxide (Fe2O3).

4. Conclusions

The decision tree CART classification model was used to determine the relationship between the various chemical components on the glass type, to make a decision tree diagram to determine the glass type, and to summarise the distribution pattern of barium oxide, silica and alumina content on the glass type.

In determining the type of artefact, the glass type of the artefact can be judged as a high potassium material type if the artefact has a heavily weathered spot; the glass type of the artefact can be judged as a lead-barium material type if the colour of the artefact is black, purple, green, light green or colourless. and the finally selected the most decisive 8 chemical components as a regular supplement.

The chemical composition of the artefacts and the nature of the artefacts complement each other and constitute the classification rules for lead-barium glass and high-potassium glass.

References

[1] Gorgan Mohammadi Faezeh, Rajaee Taher, Zounemat Kermani Mohammad. Decision tree models in predicting water quality parameters of dissolved oxygen and phosphorus in lake water [J]. Sustainable Water Resources Management, 2022, 9 (1).

[2] Shichao Zhang. Decision tree classifiers sensitive to heterogeneous costs [J]. The Journal of Systems & Software, 2012, 85 (4).

[3] Ranjit Panigrahi, Samarjeet Borah. Rank Allocation to J48 Group of Decision Tree Classifiers using Binary and Multiclass Intrusion Detection Datasets[J]. Procedia Computer Science, 2018, 132.

[4] Kitzler Florian, Wagentristl Helmut, Neugschwandtner Reinhard W., Gronauer Andreas, Motsch Viktoria. Influence of Selected Modeling Parameters on Plant Segmentation Quality Using Decision Tree Classifiers[J]. Agriculture, 2022, 12 (9).

[5] Mansour Toufik, Schork Matthias, Shattuck Mark, Wagner Stephan. Gini index on generalized r-partitions[J]. Mathematica Slovaca, 2022, 72 (5).

[6] Orovi/ Irena, Stankovi/ Srdjan, Beko Marko. On the Use of Gini Coefficient for Measuring Time-Frequency Distribution Concentration and Parameters Selection [J]. Mathematical Problems in Engineering, 2022, 2022.

[7] Gorgan Mohammadi Faezeh, Rajaee Taher, Zounemat Kermani Mohammad. Decision tree models in predicting water quality parameters of dissolved oxygen and phosphorus in lake water [J]. Sustainable Water Resources Management, 2022, 9 (1).

[8] Zhou Yichen, Hooker Giles. Decision tree boosted varying coefficient models [J]. Data Mining and Knowledge Discovery, 2022, 36 (6).

[9] Raghu N Kacker, D Richard Kuhn, Yu Lei, Dimitris E Simos. Measuring the Adequacy of a Test Suite With Respect to a Modeled Test Space[J]. IEEE Software, 2022, 39 (5).

[10] S. Durga, Esther Daniel, S. Deepa Kanmani, Jinsa Mary Philip. Cardiac arrhythmia classification using sequential feature selection and decision tree classifier method [J]. International Journal of Innovative Computing and Applications, 2021, 12 (4).