Classification model of glass relics based on decision tree algorithm

Zile Xu *

Faculty of Mathematics and Statistics of Chongqing University, Chongqing, 401331

* Corresponding Author Email: lio1249586941@163.com

Abstract. In order to assist in the compositional analysis and identification of glass artefacts, to clarify the correlations and differences between the individual chemical components, and to achieve the aim of analyzing the patterns of classification of glass artefacts, a classification model was developed in this paper. The data obtained from the chemical components are first processed accordingly, and then a multiple linear regression model is established to investigate the relationship between each chemical component variable and the variable of weathering, and a few chemical components that are significantly affected by weathering are selected as the basis, and a decision tree model based on a particle swarm algorithm is constructed to investigate the classification pattern and subclasses. The final result is that glass products can be classified into high potassium glass types and lead-barium glass types according to whether the chemical composition PbO content is greater than 6.078. The K-Means classification model was also used to classify three subclasses of high-potassium glass products and five subclasses of lead-barium glass products according to the content of the relevant chemical components.

Keywords: Chi-square test, Multiple linear regression, K-means, Decision tree classification, Particle swarm algorithm.

1. Introduction

There is quite a lot of information on glass artefacts, which archaeologists have classified into two categories based on the chemical composition of these artefacts and other testing methods: high-potassium glass and lead-barium glass [1]. Ancient glass is susceptible to weathering by the burial environment, and during the weathering process, internal elements are exchanged in large quantities with environmental elements, leading to changes in the composition ratios of the glass artefacts found, thus affecting the archaeologist's correct judgement of their category [2].

During archaeological excavations, the large amount of new glass unearthed by archaeology requires analytical work, and by discussing its chemical composition, the effects of weathering on chemical analysis and colorants, new evidence can be provided to trace the technological development of ancient China [3]. On the basis of the chemical composition obtained, the effect of surface weathering on the chemical content is explored, providing new and valuable information for the analysis and identification of the composition of new glass [4].

Therefore, this paper identifies the classification patterns of high-potassium and lead-barium glass based on the corresponding data. The chemical composition data of the artefacts were first processed initially and then analyzed using a multiple regression linear model, which makes it easier to find the chemical composition factors that are more influenced by weathering. Finally, the factors significantly affected by weathering are used as the basis for sub-category classification in this paper [5], and the K-Means algorithm is used to determine the number of sub-categories for classification and combined with a particle swarm-based decision tree algorithm.

2. Materials and methods

2.1. Data

This paper obtains data relating to the chemical composition of glass artefacts from the National Student Mathematical Modelling Competition

(http://www.mcm.edu.cn/html_cn/node/5267fe3e6a512bec793d71f2b2061497.html), and processes the data obtained accordingly.

This paper takes into account that in the process of collecting data, it is easy to obtain abnormal data due to the means of detection or technical means. Therefore, in this paper, data with composition ratios accumulating from 85% to 105% are considered as valid data and invalid data are removed, and then the non-quantitative data are quantified, and the process of quantifying the data is shown in Table 1.

Table 1. Data quantification process

Color	Value	Weathering or not	Value	Ornaments	Value	Type	Value
Black	1	Weatherization	1	A	1	High Potassium	1
Blue-green	2	Weather-free	0	B	2	Lead Barium	2
Green	3			C	3		
Light blue	4						
Light green	5						
Dark Blue	6						
Dark green	7						
Purple	8						

2.2. Introduction to the method

2.2.1 Multiple linear regression models

The general form of the multiple linear regression model is:

$$\begin{cases} y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \beta_k x_k + \varepsilon \\ \varepsilon \square N(0, \sigma^2) \end{cases} \tag{1}$$

Where $x_i \left(1 \le i \le k\right)$ is the i-th chemical component [6].

In the process of parameter estimation, an estimate of β, denoted as $\hat{\beta}$, is chosen so that the sum of squares of the random errors ε is minimized:

$$\min_{\beta} \varepsilon^T \cdot \varepsilon = \min_{\beta} (Y - X\beta)^T \cdot (Y - X\beta) = (Y - X\hat{\beta})^T \cdot (Y - X\hat{\beta}) \overset{def}{=} Q(\hat{\beta}) \tag{2}$$

From the requirements of the least squares method, the standard equation for the regression parameters can be solved from the necessary conditions for the multivariate function to obtain extreme values as follows:

$$\begin{cases} \frac{\partial Q}{\partial \beta_0}|_{\beta_0 = \beta_0} = 0 \\ \frac{\partial Q}{\partial \beta_i}|_{\beta_1 = \beta_i} = 0 (i = 1,2,3) \end{cases} \tag{3}$$

2.2.2K-Means Clustering models

The K-means clustering algorithm is an unsupervised learning algorithm distinguished from classification and is suitable for situations where the sample label is unknown. k in the K-means clustering algorithm stands for the number of clusters and means stands for taking the average of the values in each cluster as the center of that cluster, i.e. each cluster is described by the center of that class [7]. This clustering algorithm is easy to implement and the process of its implementation is as follows.

1) Determine the number of clusters k.

2) Calculate the distance from each sampling point to the center of the cluster, usually the Euclidean distance.

3) Update the "cluster centers" according to the newly divided clusters.

4) Repeat steps (2) and (3) until the "cluster center" no longer moves.

2.3. Decision tree model based on particle swarm optimization algorithm

2.3.1 Particle swarm algorithm

The particle swarm optimization algorithm is an algorithm proposed by Eberhart and Kennedy (1995) that was summarized and developed based on imitating a flock of birds foraging. Each individual in the population (called a particle) represents a possible solution to the problem and is adjusted using the following three factors: the particle's current knowledge (its problem-solving ability), its historical knowledge or previous experience (its memory), and the historical knowledge or previous experience of individuals located in its vicinity (social knowledge). In order to measure the performance of the particle as a problem solver, an adaptability function is used. This technique works by combining the three factors mentioned earlier, with each particle finding the optimal solution in the end by continuously searching generation by generation [8].

$$v_{id}^{k+1} = wv_{id}^{k} + c_1 r_1 \left(p_{id} - x_{id}^{k}\right) + c_2 r_2 \left(p_{gd} - x_{id}^{k}\right) \tag{4}$$

$$x_i d^{(k+1)} = x_i d^{k} + v_i d^{(k+1)} \tag{5}$$

where: the subscript "d" denotes the d-th dimension of the particle; "i" denotes particle i; "k" denotes the kth generation; c_1, c_2 is the learning factor, also called the acceleration constant; r_1, r_2 is a random number between $[0,1]$; w is the inertia weight function, $w = w_{\max} - \frac{k(w_{\max} - w_{\min})}{iter_{\max}}$; $w_{\max}$ is the initial weight; w_{min} is the final weight; $iter_{\max}$ is the maximum number of iterations.

2.3.2 Decision tree classification model based on particle swarm algorithm

Modeling process of the optimal decision tree algorithm:

The decision tree algorithm produces visual analysis results and the basic idea is to use the attribute selection metric to select the best attribute to split the record, producing a smaller subset of the dataset after that attribute becomes the next decision point, which can be used to calculate feature importance based on the decision tree. The use of a particle swarm algorithm optimized decision tree classification model, rather than using a separate decision tree classification model, allows for optimization of the decision tree classification results [9]. Our process of exploring classification laws for glass products and performing subclass classification is a multi-classification problem involving rule generation, where PSO techniques are used to determine the weights of conditional attributes and decision trees are used to generate rules [10]. The modelling process of the optimal decision tree algorithm is shown in Figure 1.

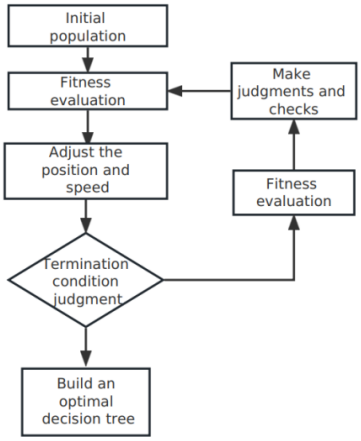

Figure 1. Modeling process of the optimal decision tree algorithm

3. Model building and solving

3.1. Correlation test

As the quantitative assignment of each definite category of indicators has been made, the corresponding discrete variables have been obtained, i.e. the correlation between the definite category of indicators needs to be analyzed, so Spearman correlation coefficients are constructed and correlation tests are carried out on the relevant indicators.

$$d_2 = rank_1 - rank_2 \tag{6}$$

$$\rho = 1 - \frac{n\sum d_2^{\ 2}}{n(n^2-1)} \tag{7}$$

The variables were first tested for statistical significance based on the results of the analysis, and as shown in Table 2 it can be analyzed that the p-values obtained between most of the indicators were significant, i.e. statistically significant.

Table 2. Results of statistical significance testing of indicators

	Ornaments	Type	Weathering of surfaces	Color
Ornaments	1.000(0.000***)	0.119(0.374)	0.037(0.781)	0.443(0.000***)
Type	0.119(0.374)	1.000(0.000***)	0.344(0.008***)	0.447(0.000***)
Weathering of surfaces	0.037(0.781)	0.344(0.008***)	1.000(0.000***)	-0.134(0.317)
Color	0.443(0.000***)	0.447(0.000***)	-0.134(0.317)	1.000(0.000***)

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

According to the size of the correlation coefficient between each indicator, a Spearman's correlation visualization image is drawn, the darker the color represents the stronger the correlation between two indicators. Through the visual image of correlation coefficient of each indicator, the size of correlation between each indicator can be analyzed visually, and the correlation of each indicator is shown in Figure 2.

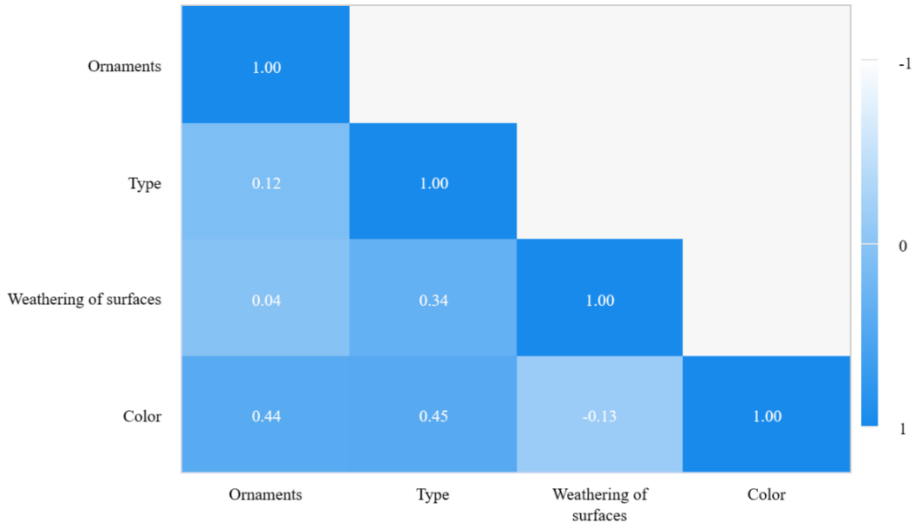

Figure 2. spearman correlation visualisation

3.2. Variance analysis

The four indicator variables of type, color, decoration and whether weathered were subjected to chi-square analysis, and the remaining three variables were tested for differences in the variable of whether weathered by the remaining three variables in turn, and the results of the chi-square test are shown in Table 3.

The quantitative analysis of effects and tests of significance led to the following conclusions:

1. Significant difference between surface weathering and category.
2. There is a more significant difference between surface weathering and decoration.
3. There is an intermediate difference between surface weathering and color, which is less than the difference between surface weathering and category and decoration.

Table 3. Cardinality analysis results

Cardinality analysis results						
Variables	Title	Weathering of surfaces(%)		Total	$\chi 2$	p
		0	1			
Ornaments	1	11(45.83)	11(32.35)	22(37.93)	4.957	0.084
	2	0(0.00)	6(17.65)	6(10.34)		
	3	13(54.17)	17(50.00)	30(51.72)		
Total		24	34	58		
Type	1	12(50.00)	6(17.65)	18(31.03)	6.88	0.009**
	2	12(50.00)	28(82.35)	40(68.97)		
Total		24	34	58		
Color	1	0(0.00)	2(5.88)	2(3.45)	7.234	0.405
	2	6(25.00)	9(26.47)	15(25.86)		
	3	1(4.17)	0(0.00)	1(1.72)		
	4	8(33.33)	16(47.06)	24(41.38)		
	5	2(8.33)	1(2.94)	3(5.17)		
	6	2(8.33)	0(0.00)	2(3.45)		
	7	3(12.50)	4(11.76)	7(12.07)		
	8	2(8.33)	2(5.88)	4(6.90)		
Total		24	34	58		
* p<0.05 ** p<0.01						

3.3. Equations based on multiple linear regression models

Considering the large variation in the content of individual chemical components and the fact that some of them were not detected, the contribution of each chemical component indicator to the relationship with whether it is weathered varies. We therefore developed a multiple linear regression model.

The model was used to predict the pre-weathering chemical content of the high potassium glass artefacts and the lead-barium glass artefacts by classifying them into two categories based on the corresponding weathering point data. Multiple linear regression analyses were carried out on the high potassium and lead-barium glass artefacts using weathering, color and decoration as independent variables, all chemical components as dependent variables and one chemical component indicator at a time as the dependent variable.

Multiple linear regression analyses were performed on each dependent variable, and the regression model was tested for the presence of multicollinearity and heteroskedasticity after each analysis. If the model did not have multicollinearity and heteroskedasticity and the fit was good, the model was considered reasonable and the parameters of the model were recorded; if the model had multicollinearity, forward stepwise regression was performed; if the model had heteroskedasticity, OLS+ was used to Robust standard error treatment was applied. The final multiple linear regression model was obtained and the significance of each core variable and the corresponding coefficients were obtained. Larger coefficients represent the greater the degree to which the chemical component is affected by weathering. The final corresponding coefficients obtained are shown in Table 4.

Table 4. Multiple linear regression model variable coefficients

chemical composition	coefficient	chemical composition	coefficient
SiO_2	28.44	MgO	-0.83
K_2O	-7.81	Na_2O	-0.71
CaO	-4.1	BaO	-0.63
Al_2O3	-3.74	SnO_2	0.28
Fe_2O3	-2.01	PbO	-0.24
CuO	-1.49	SO_2	-0.15
P_2O_5	-1.03	SrO	-0.03

3.4. Solution of decision tree model based on particle swarm optimization algorithm

3.4.1 An exploration of the laws of classification of glass artefacts

This paper uses a decision tree model based on a particle swarm algorithm to solve the classification laws for high potassium glass type artefacts and lead-barium glass type artefacts.

Considering that a larger inertia value is helpful for global search, the inertia weight was set to 0.9, the individual learning factor $c_1 = 1$, the social learning factor $c_2 = 2$, the initial number of particles was 100 and the maximum number of iterations was 100. The particle swarm algorithm was optimized for decision tree classification, and a graph of decision tree classification results was obtained. As shown in Table 5, the model achieved 100% accuracy on the training set, with good classification results.

The decision tree structure of the analysis results in a significant effect of lead oxide (PbO) for artefacts of the high potassium glass type and the lead-barium glass type. The type of artefact can be determined based on the percentage of lead oxide contained in the artefact. As shown in Figure 3, when the percentage of PbO in the artefact is higher than 6.078, the glass artefact can be classified as Barium lead glass type, while when the percentage of PbO in the artefact is less than or equal to 6.078, the glass artefact can be classified as a high potassium glass type.

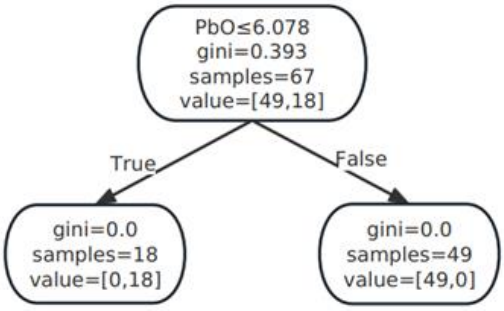

Figure 3. PSO-based decision tree classification results

Table 5. Model solution results

Model results	Accuracy	Recall rate
	100%	100%

3.4.2 Subclass classification based on K-Means and decision tree classification models

In order to further investigate the laws of classification, the paper then goes on to classify high potassium glass and lead-barium glass into subclasses based on their chemical composition, exploring the laws of subclass classification based on a decision tree model based on a particle swarm algorithm.

The results of the K-Means clustering algorithm were used to obtain the number of subclasses for high potassium glass and lead-barium glass, and the number of clusters was determined to be 3 for high potassium and 5 for lead-barium based on the elbow diagram of the cluster analysis, that is, unlabeled data is transformed into labeled data, so a decision tree was used to analyze the factors that were significant in distinguishing the subclasses.

Setting the ratio of training set to validation set as 7:3, inertia weight as 0.9, taking individual learning factor = 1, social learning factor as 2, initial number of particles as 100 and maximum number of iterations as 100, the corresponding decision tree classification result graph was obtained, as shown in Table 6, the model achieved 100% accuracy for the training set with good classification results.

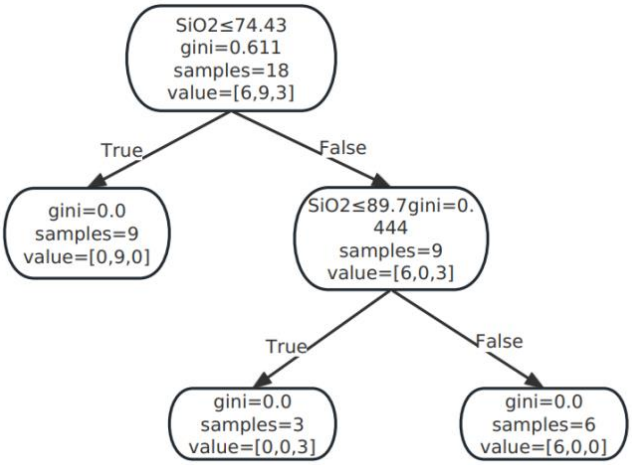

Figure 4. High potassium type decision tree classification results

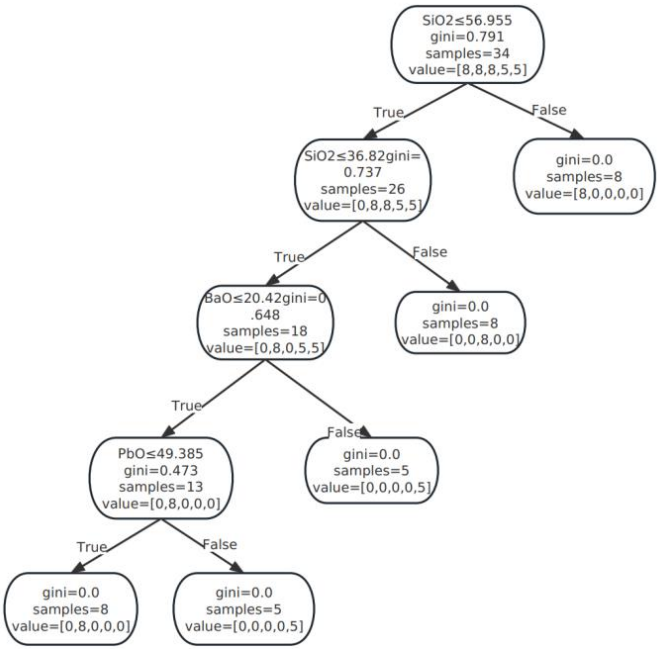

Figure 5. Lead-barium type decision tree classification results

Table 6. Model solution results

	Accuracy	Recall rate
Training set	100%	100%
Test set	100%	100%

1. As shown in Figure 4, According to the clustering number and decision tree structure diagram obtained based on K-Means clustering, high-potassium glass can be divided into the following three categories according to the chemical composition:

1)Type I

Silicon dioxide content is more than 89.7%.

2)Type II

The silicon dioxide content is less than 74.73% and there is also about 11% potassium oxide and about 6% calcium oxide.

3)Type III

Silicon dioxide content between 74.73% and 89.7%, with around 4% alumina.

2. As shown in Figure 5, According to the clustering number and decision tree structure diagram obtained based on K-Means clustering, lead barium glass can be divided into the following five categories according to the chemical composition:

1) Type I

The silicon dioxide content is greater than 56.995%.

2) Type II

The silicon dioxide content lies between 36.82% and 56.995%.

3) Type III

Silicon dioxide less than 36.82%, barium oxide less than 20.42% and lead oxide less than 49.385%.

4) Type IV

The silicon dioxide content is less than 36.82% and the barium oxide content is greater than 20.42%.

5) Type V

Less than 36.82% silicon dioxide, less than 20.42% barium oxide and greater than 49.385% lead oxide.

4. Conclusions

Ancient glass is sensitive to the weathering of its burial environment, and during the weathering process, internal elements exchange heavily with those in the environment, altering the compositional proportions of the glass artefacts found and thus affecting the archaeologist's ability to correctly identify their category. In this paper, by optimizing the decision tree classification model using a particle swarm algorithm, the classification patterns of lead-barium and high-potassium glass artefacts were found, a deeper analysis of the two types of artefacts was carried out, and K-means analysis was combined to derive the number of categories into which the two types of artefacts could be divided into sub-categories, with five sub-categories for high-potassium artefacts and three sub-categories for lead-barium artefacts. The experimental results show that the classification laws derived in this paper have practical significance and provide a more accurate law and reference basis for the classification of artefacts.

References

[1] Zhang X, Jiang X, Jiang J, et al. Spectral–spatial and superpixelwise PCA for unsupervised feature extraction of hyperspectral imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-10.

[2] Yu D, Xu H, Chen C L P, et al. Dynamic coverage control based on k-means[J]. IEEE Transactions on Industrial Electronics, 2021, 69(5): 5333-5341.

[3] J. Q. Dong, Q. H. Li, S. Liu. The native development of ancient Chinese glassmaking: a case study on some early lead–barium–silicate glasses using a portable XRF spectrometer[J]. X-Ray Spectrometry,2015,44(6).

[4] Chang Su and Jingjing Wang. Research on composition analysis and type identification of ancient glass products based on data mining[J]. Automation and Machine Learning, 2022, 3(2)

[5] Yilei Wang, Wenxuan Liu, Ziqiang Lin. Component Analysis and Identification Model of Ancient Glass Products Based on Correlation Analysis[J]. Analytical Chemistry A Journal,2022,1(1).

[6] Patonai Zoltán and Kicsiny Richárd and Géczi Gábor. Multiple linear regression based model for the indoor temperature of mobile containers[J]. Heliyon, 2022, 8(12)

[7] Zhang X, Jiang X, Jiang J, et al. Spectral–spatial and superpixelwise PCA for unsupervised feature extraction of hyperspectral imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-10.

[8] Patricia Jimbo Santana and Laura Lanzarini and Aurelio F. Bariviera. Variations of Particle Swarm Optimization for Obtaining Classification Rules Applied to Credit Risk in Financial Institutions of Ecuador[J]. Risks, 2019, 8(1): 2-2.

[9] Ping-Feng Pai et al. A group decision classifier with particle swarm optimization and decision tree for analyzing achievements in mathematics and science. [J]. Neural Computing and Applications, 2014, 25(7-8): 2011-2023.

[10] Ping-Feng Pai, Chen-Tung Chen, Yu-Mei Hung, Wei-Zhan Hung, Ying-Chieh Chang. A group decision classifier with particle swarm optimization and decision tree for analyzing achievements in mathematics and science. [J]. Neural Computing and Applications,2014,25(7-8).