Glass classification and recognition based on genetic simulated annealing algorithm and decision tree algorithm

Zhijie Sun [1,*], Dayu Guan [2], Zhiliang Zu [2], Jing Wang [2], Hengchao Wei [2], Shuo Chen [2]

[1] Associate Professor and Researcher of Research Team of Science and Technology Application Innovation on Public Security, Public Security Research Center, Shandong Police College

[2.]Shandong Police College, Jinan, Shandong, 250000

* Corresponding Author Email: szj_sdpc@126.com

Abstract. The chemical composition and ratio of ancient glasses are highly susceptible to change due to the complex burial environment. In order to determine the classification rules of lead-barium glass and potassium glass, and to subclassify lead-barium glass and potassium glass according to the appropriate chemical composition index, this paper mines the linear combination of high potassium glass and lead-barium glass by linear discriminant analysis, and calculates the clustering centers of lead-barium glass and high potassium glass subclassifications and the subclassifications of glass to be classified based on the FCM (Fuzzy-C Mean Clustering) algorithm of genetic simulated annealing algorithm. The sub-clusters of lead-barium glass and high-potassium glass, as well as the affiliation degrees of the glass samples to be classified were calculated based on the genetic simulated annealing algorithm. The model results were then examined by using the decision tree algorithm, and the results of glass artifact type identification were obtained.

Keywords: Decision tree algorithm, genetic simulated annealing algorithm, glass composition.

1. Introduction

Glass is a silicate non-metallic material formed by cooling quartz sand through a high temperature melting process, with silicon dioxide (SiO2) as the main chemical component. Due to the high melting point of pure quartz sand [1-3], the melting process requires the addition of a flux to alkali-containing calcium silicate glass. The chemical composition of glass products varies from region to region, including our unique glass varieties of lead barium glass to lead ore using fluxes, while potassium glass mostly with high potassium content of grass ash flux. Because of the chemical stability of glass, a large number of ancient glass products have survived to this day, and become an important basis for the study of trade exchanges and cultural exchanges between the early regions [4-6].

At present, most of the studies on ancient glass at home and abroad focus on the study of the composition, production process and origin of glass products, and on this basis, the study of the development process and material and cultural exchanges of different glass products in different regions has been extended [7-8]. Among them, chemical analysis of glass composition is an important component of the study of ancient glass composition system, which has an important guiding value for identifying the raw materials and origin of glass production [9-10].

2. Glass category classification

2.1. Classification of glass categories based on linear discriminant analysis

In this paper, linear discriminant analysis is applied to mine the linear combination of high potassium glass and lead-barium glass in order to clarify the classification laws of the two glass types. Linear discriminant analysis is a classical linear learning method, which finds the linear combination of features of two types of things through the combination of statistics, pattern recognition and machine learning methods, so as to characterize and distinguish them.

First, this paper codes the definite class data of whether the surface of glass artifacts is weathered, type, decoration and color, and the corresponding rules are shown in Table 1.

Table 1. Coding rules for class-specific data

Surface Weathering	Codes	Type	Codes	Ornament	Codes	Color	Codes
Weathered	2	Lead barium glass	2	A	2	Blue Green	1
						Light Blue	2
						Violet	3
				B	3	dark green	4
						Dark Blue	5
						Void	6
Unweathered	1	High potassium glass	1	C	1	light green	7
						black	8
						green	9

The first 90% of the glass artifacts data were selected as the training set to train the fitted discriminant analysis model, and the remaining 10% data were used as the test set to verify the validity of the model. The discriminant functions (mathematical relationship expressions) for various chemical compositions, surface weathering, glass artifacts ornamentation and color of lead-barium glass and high-potassium glass were obtained by machine learning, as shown below table 2:

Table 2. Discriminant function for each classification of lead-barium glass and high potassium glass

Item	Discriminant function	Item	Discriminant function
Intercept	1.563	Lead oxide(PbO)	-0.134
Silicon dioxide(SiO2)	-0.034	Barium oxide(BaO)	-0.183
Sodium oxide(Na2O)	-0.215	Phosphorus pentoxide(P2O5)	-0.178
Potassium oxide(K2O)	0.298	Strontium oxide(SrO)	1.249
Calcium oxide(CaO)	-0.032	Tin oxide(SnO2)	-0.294
Magnesium oxide(MgO)	0.004	Sulfur dioxide(SO2)	0.076
Aluminum oxide(Al2O3)	-0.213	Surface weathering	2.053
Iron oxide(Fe2O3)	0.055	Decoration	-0.275
Copper oxide(CuO)	0.132	Color	0.066

According to the results of the operation, the expression of the classification law C_i for high potassium glass and lead-barium glass is:

$$\begin{aligned} C_i = {} & 1.563 - 0.034x_1 - 0.215x_2 + 0.298x_3 - 0.032x_4 + 0.004x_5 - 0.213x_6 \\ & + 0.055x_7 + 0.132x_8 - 0.134x_9 - 0.183x_{10} - 0.178x_{11} \\ & + 1.249x_{12} - 0.294x_{13} + 0.076x_{14} + 2.053B_i - 0.275\,W_i \\ & + 0.066Y_i \end{aligned} \tag{1}$$

The prediction accuracy of the training set is examined, where the correct rate is the proportion of glass artifact samples that actually belong to the category when the samples are judged to be lead-barium glass or high-potassium glass according to the classification law. The recall rate is the proportion of glass artifact samples correctly judged to be of that category when the artifact sample is actually lead-barium glass or high-potassium glass. F1-score value is a weighted composite index of the correct rate P and recall rate R, calculated as

$$F1 - score = \frac{2*P*R}{P+R} \tag{2}$$

Table 3. Prediction accuracy of training set

Prediction category	Sample size	Correct rate P	Recall rate R	$F1 - score$
Category 1 (lead barium)	44	97.67%	100.00%	98.82%
Category 2 (high potassium)	16	100.00%	94.44%	97.14%
Aggregate	60	98.37%	98.33%	98.35%

As shown in Table 3, the classification law expressions for high potassium glass and lead-barium glass solved by the model have high accuracy for the training set. In order to verify the validity of the model, the correctness of the classification law expressions was tested based on the test set data. From the test results, Table 9 shows that the above-mentioned classification expressions have high accuracy and can be used to classify high potassium glass and lead-barium glass. The results are shown in the table 4.

Table 4. Test set prediction accuracy

Prediction category	Sample size	Correct rate P	Recall rate R	$F1-score$
Category 1 (lead barium)	5	100.00%	100.00%	100.00%
Category 2 (high potassium)	2	100.00%	94.44%	97.14%
Aggregate	7	100.00%	100.00%	100.00%

2.2. Subclassification of glass categories based on genetic simulated annealing clustering analysis

The principal component analysis method is a mainstream dimensionality reduction method. In this paper, the chemical composition data of glass artifacts were dimensioned down to three principal components using principal component analysis at a cumulative variance explained rate of 71.6%, and the results are detailed in Appendix Table 23 and Table 24.

The FCM (fuzzy-C mean clustering) algorithm is a local search optimization algorithm based on identifying data points in Euclidean space, assigning the data to different clusters, and then determining the distance between clusters. In this paper, we set 58 glass artifact samples as $X=\{x_1,x_2,\dots,x_{58}\}, c\in[2,58]$ is the number of glass artifact types to be subclassified, $\{A_1,A_2,\dots,A_c\}$ denotes the corresponding c categories, U is its similarity classification matrix, and the clustering centers of each category are $\{v_1,v_2,\dots,v_c\}$,$\mu_k\ (x_i)$ is the affiliation of glass artifact sample x_i for class A_k, then the objective function

$$J_b(U,v)=\sum_{i=1}^{58}\sum_{k=1}^{c}(\mu_{ik})^2\,(d_{ik})^2 \tag{3}$$

where the Euclidean distance $d_{ik}=d(x_i-v_k)=\sqrt{\sum_{j=1}^{m}\left(x_{ij}-v_{kj}\right)^2}$ is used to measure the distance between the glass artifact sample xi and the centroid of class k; m is the characteristic number of the artifact sample;b is its weighting parameter ($1\le b\le\infty$). fcm requires that the affiliation value of a single sample to each cluster sum to 1, i.e., it satisfies.

$$\sum_{j=1}^{c}\mu_j(x_i)=1, i=1,2,\dots,58 \tag{4}$$

Equation (2)(3) calculates the affiliation μ_{ik} of glass artifact sample x_i for class A_k and c clustering centers $\{v_i\}$, respectively.

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}(\frac{d_{ik}}{d_k})^{\frac{2}{b-1}}} \tag{5}$$

Set up $I_k=\{i|2\le c\le 58;d_{ik}=0\}$,for all i classes,$i\in I_k,\mu_{ik}=0$.

$$v_{ij}=\frac{\sum_{k=1}^{58}(\mu_{ik})^b x_{kj}}{\sum_{k=1}^{58}(\mu_{ik})^b} \tag{6}$$

Using Eqs. (4)(5) to iteratively modify the clustering centers, data affiliation and perform the classification, when the algorithm converges, the clustering centers of lead-barium glass and high-potassium glass subcategories and the affiliation of glass artifact samples for each subcategory are theoretically obtained. Since the clustering center is based on local search optimization, it is easy to fall into local optimum. The simulated annealing algorithm is a stochastic optimization-seeking algorithm based on Monte-Carlo iterative solution strategy, and the genetic algorithm is a method to search for the optimal solution by simulating the natural evolutionary process. In this paper, we use a combination of simulated annealing algorithm and genetic algorithm for FCM analysis, which can effectively overcome the premature defects of traditional genetic algorithm, and at the same time, we

can set the function according to the specific situation of the clustering problem, so that the algorithm can reach the global optimal solution more effectively.

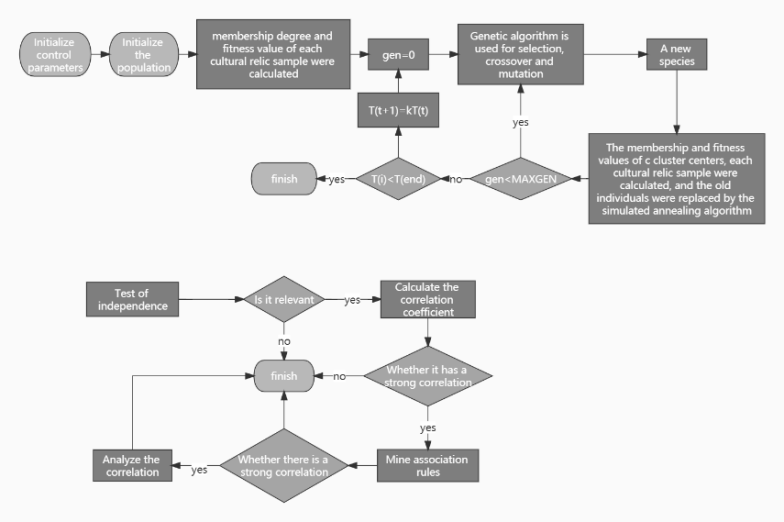

Fig. 1 Flow chart of FCM based on simulated annealing genetic algorithm

According to the calculation results such as the clustering diagram of lead-barium glass and high-potassium glass shown in fig 1, we can determine the results of the subclass division of lead-barium and high-potassium glass artifacts as shown in Table 5. The results are shown in the fig 2.

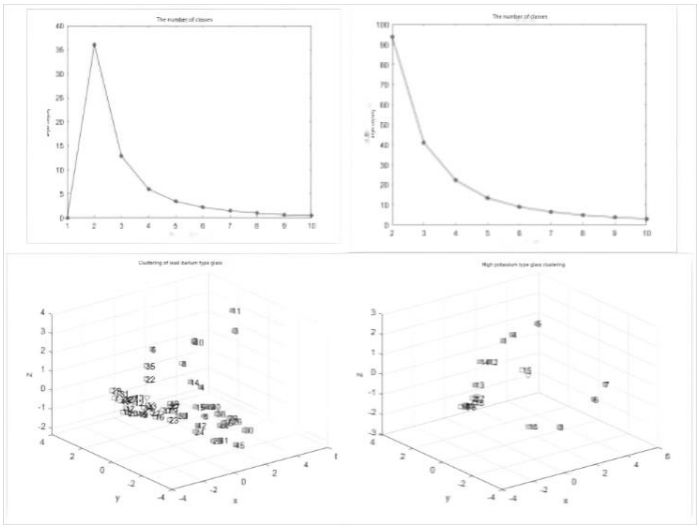

Fig 2. Cluster diagram of lead-barium glass and high-potassium glass

Table 5. Sub-classification results after clustering of lead barium and glass

Glass category subclass division	Lead-barium post-clustering category 1	Lead-barium post-clustering category 2	High potassium post-clustering type 1	High potassium post-clustering type 2
Artifact Number	02, 08, 08 severe weathering point, 11, 19, 26, 26 severe weathering point, 30 parts 1, 30 parts 2, 39, 40, 41, 43 parts 1, 43 parts 2, 49, 50, 50 unweathered point, 51 parts 1, 51 parts 2, 52, 54, 54 severe weathering point, 58	20, 23 unweathering point,24,25unweathering point,28unweathering point,29unweathering point,31,32,33,34,35,36,37,38,42unweathering point1,42unweathering point2,44unweathering point,45,46,47,48,49unweathering point,53unweathering point,55,56,57	01, 03part2,04,05,06part1,06part2,13,14,16,21	03part1,07,09,10,12,18,22,27

2.3. Reasonableness analysis and sensitivity analysis of glass category subclass classification results

(1) Rationality analysis

The JB value is the value based on the SEE value and the contour coefficient to measure the clustering effect. In this paper, the JB objective function is used as a judgment index to measure the reasonableness of the glass category subclass classification results. According to the test results, the JB value of the clustering items in this paper is the largest, indicating that the above glass category subclass classification results are reasonable.

(2) Sensitivity analysis

In this paper, perturbations are added to each component of the source data in the range of [90%, 110%]. Firstly, the perturbation coefficients are determined and the perturbation coefficients are perturbed by adjusting the mean value of normal distribution to achieve the perturbation, and then this paper weights the components of the sample values of the artifacts and fixes the linear weighting values determined by principal component analysis to obtain a principal component matrix consisting of three principal components, thus re-clustering them and obtaining the corresponding JB values. Subsequently, this paper measures the sensitivity of the clustering model by adjusting the perturbation coefficients and observing the changes of the corresponding JB values.

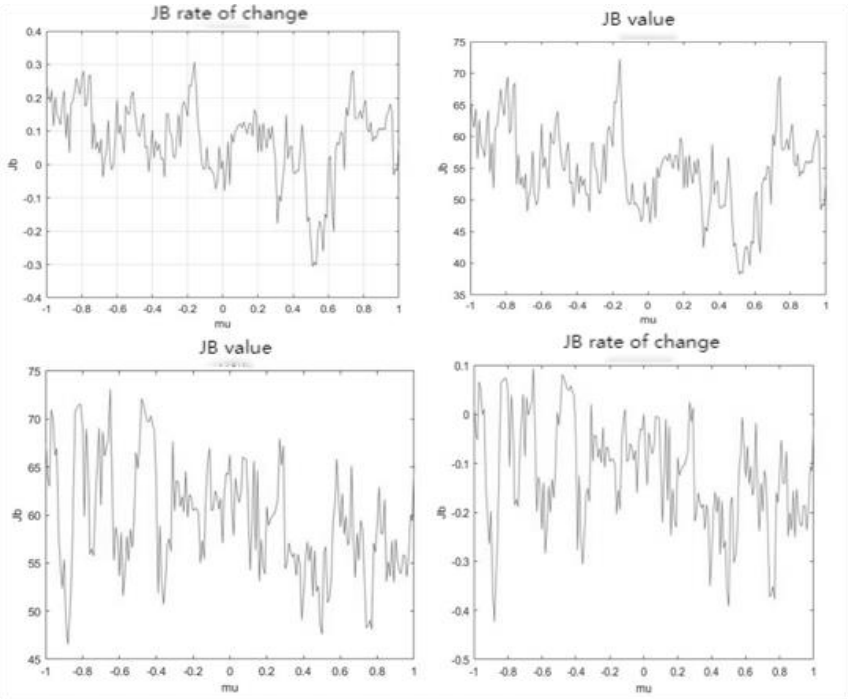

Fig 3. Cluster model sensitivity analysis

According to the sensitivity analysis results of the clustering model shown in fig 3, the JB values of the corresponding high potassium glass fluctuate in the range of [-30%,40%] and the JB values of lead-barium glass fluctuate in the range of [-40%,10%] when the perturbation coefficient changes. It can be determined that the clustering model is little affected by the change of data, 5.2.3 The results of glass category subclass classification are less sensitive, more robust, and better clustering.

3. Glass artifact type prediction

3.1. Prediction of glass artifact types based on random forest algorithm

In order to test the prediction law of glass artifact types based on multiple stepwise regression analysis, this paper uses regression forest algorithm to determine the type classification of unknown glass artifacts.

Firstly, a decision tree classification model is built from the training set data in Form 2 to obtain the basic structure of the decision tree. Then, the decision tree structure is obtained by using the training set data to build a decision number classification model. And based on this decision tree structure, we calculate the importance values of surface weathering and chemical composition of glass features, which are consistent with the importance of respective variables affecting glass artifact types based on multiple stepwise regression analysis. Finally, this paper applied the decision tree classification model to the training and testing data to obtain the model-based classification evaluation results of glass artifact types, as shown in Table 6.

Table 6. Results of glass type assessment for glass artifacts

Artifact number	Predicted result_Y	Glass type assessment result
A1	1	High potassium
A2	2	lead barium
A3	2	lead barium
A4	2	lead barium
A5	2	Lead barium
A6	1	High Potassium
A7	1	high potassium
A8	2	lead barium

3.2. Sensitivity analysis of classification results based on SALib analysis library

The SALib analysis library is an open source library for sensitivity analysis based on Python by providing a decoupled workflow to generate model inputs using the sampling functions in the mathematical model and calculate sensitivity indices for the model outputs using one of the analysis functions [3].

In this paper, the SALib sensitivity analysis library is invoked to set 10,000 sampling points within the parameters of the surface weathering of the glass artifacts and the chemical composition of the glass in the dependent variable, and the sensitivity index of the model output is calculated based on the Soblo variance global analysis method, which analyzes the first-order and second-order sensitivity indices and the total-order index to measure the contribution of the model input to the output variance.

The total-order indices and their corresponding confidence intervals were plotted in the total validity-confidence measure table, as shown in Table 19 and fig 13. Based on the results, we know that the input of variable x_2 (i.e., potassium oxide K2O) has the highest output variance contribution for unknown artifact glass type Y, i.e., the potassium oxide K2O component index has the highest total validity for artifact glass type discrimination. The results are shown in the table 7.

Table 7. Total validity-confidence measures table

Independent variable	Total validityST	ConfidenceST_conf
x_1(PbO)	0.095118	0.008574
x_2(K2O)	0.541145	0.040419
x_3(Surface weathering)	0.067458	0.005816
x_4(SiO2)	0.298070	0.022992

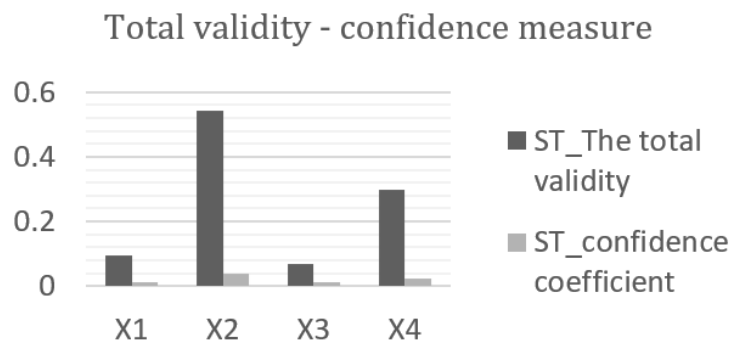

Fig 4. Graph of Total Validity-Confidence Measures

The results are shown in the fig 4. Based on the set 10,000 sampling points, the numerical analysis of all order sensitivity indices in this paper we learns that the lower average total validity value is 0.2504, which indicates that the constructed classification model of glass artifact category has a small sensitivity index to the sampling changes of the input, i.e., the classification relational formula of glass artifact type obtained by multivariate stepwise regression has good adaptability and robustness to the input changes, and the sensitivity test is good.

4. Conclusions

In this paper, the linear combination of high potassium glass and lead-barium glass was mined by linear discriminant analysis, and the expressions of classification laws for high potassium glass and lead-barium glass were obtained according to the operation results. The test results show that the expression can accurately classify high potassium glass and lead-barium glass with an accuracy of 98.37%. Subsequently, principal component analysis was applied to reduce the chemical composition to 3 principal components, and then a clustering algorithm based on genetic simulated annealing algorithm was applied to calculate the clustering centers of glass type subclassifications and the affiliation of the glass artifact samples to be classified for each subclassification, in order to make the algorithm more effective to reach the global optimal solution, and lead-barium and glass were clustered into 2 subclasses. When testing the clustering model, this paper uses the JB objective function to measure the reasonableness of the glass category subcategory classification results, and according to the test results, the JB value of the clustering items in this paper is the largest, indicating that the above glass category subcategory classification results are reasonable. Then this paper adds perturbations to each component of the source data, and by adjusting the perturbation coefficients, it is found that the JB values of the corresponding high potassium and lead-barium glasses fluctuate within the ranges of [-30%,40%] and [-40%,10%], respectively, and it is determined that the clustering model is little affected by the changes of the data and the glass category subclass classification results are less sensitive.

To identify the category of unknown glass, this paper uses decision tree algorithm to test the model results and obtain the results of glass artifact type assessment. Among the unknown categories of glass artifacts, A1, A6, and A7 belong to high potassium glass, and A2, A3, A4, A5, and A8 belong to lead-barium glass. Then this paper calls the SALib sensitivity analysis library to calculate all order sensitivity indicators of the model output, and the lower average total validity value is 0.2504, which indicates that the model has good adaptability and robustness to input changes, and the sensitivity test is good.

References

[1] Zhao Zhiqiang. Research on the composition system and production process of glass beads excavated from the Balikun Shirenzigou site group in Xinjiang [D]. Northwestern University, 2016.

[2] Liu Huating, Guo Renxiang, Jiang Hao. Research and improvement of association rule mining Apriori algorithm [J]. Computer Applications and Software, 2009, 26(1): 4.

[3] Herman, J. and Usher, W. (2017) SALib: An open-source Python library for sensitivity analysis. Journal of Open Source Software, 2(9).

[4] Cai, N., Li, W. B., Huang, Q. H., Zhou, S., Qiu, B. J., He, Z. Q.. Sector-based neighborhood feature engineering for defect detection in glass package insulation terminals [J]. Journal of Electronics and Information, 2022, 44(05): 1548-1553.

[5] He C-C, Liu B-Jin, Ren B-Tong, Jia Jing. Mopping up the glass with a strip wipe, the roll king is actually in my house? --Analysis of factors to improve the overall quality of the domestic industry based on big data mining and machine learning [C]//. 2021 (7th) National Student Statistical Modeling Competition Awarded Papers (I). ,2021:2-56. DOI: 10.26914/c.cnkihy.2021.041762.

[6] Yang X. Detection and classification of glass defects based on machine vision [D]. Fujian Engineering College, 2021. DOI: 10.27865/d.cnki.gfgxy.2021.000021.

[7] Chu, Wangtao. Research and software development of data-driven LCD glass substrate thickness prediction algorithm [D]. South China University of Technology, 2020. DOI: 10.27151/d.cnki.ghnlu. 2020.004533.

[8] Wang WX. Data mining-based analysis and application of product stubborn defect [D]. University of Electronic Science and Technology, 2020. DOI: 10.27005/d.cnki.gdzku.2020.001277.

[9] Wang A-X, Zhang X-J, Cao Y-H. Application of genetic simulated annealing algorithm for parametric inversion of glass and crystal dispersion equations [J]. Infrared and Laser Engineering, 2015, 44(11): 3197-3203.

[10] Luo Jianhong. Particle computing classification knowledge discovery algorithm and its application[D]. Zhejiang University,2010.

[11] Sun Bilang. Research on machine vision-based inspection algorithms and their applications in industry [D]. Huazhong University of Science and Technology, 2006.