Based on Cluster analysis model for glass classification problem

Chenxi Liu*, Zekai Wu, Xin Dong

School of Economics and Management, China University of Geosciences, Beijing, China

* Corresponding Author Email: lcx1007208104@163.com

Abstract. The research on the chemical composition and other physical properties of ancient glass is a very important aspect of ancient glass research, which can provide scientific evidence for archaeological research and help to study the composition system, manufacturing age, preparation process and technical origin of ancient glass. Based on the cluster analysis model, this paper establishes a suitable classification model, which provides a correct classification judgment for weathered glass affected by the external environment. The specific work is to classify the rules of glass products, classify each category, and analyze the rationality and sensitivity.

Keywords: Cluster analysis, logistic regression model, glass artifacts, single objective optimization.

1. Introduction

Glass is one of the earliest man-made materials invented by human beings. It was born in the Lianghe River Valley from the 20th century to the 15th century BC. It has a history of at least 3000 years and has played an important role in promoting the occurrence and development of human civilization, especially modern scientific and technological civilization. In the 1980s, with the strengthening of cooperation between the glass science and technology community and the cultural relics and archaeology community, the scientific research of ancient Chinese glass entered the fast lane, and the "scientific and technological archaeology" of ancient glass was launched on a large scale [1-3]. The classification of glass is of great significance to the study of ancient Chinese glass. In recent years, with the excavation of glass beads cultural relics in previous dynasties, people have used modern science and technology to deeply analyze the chemical composition, shape, structure, use and other aspects of various types of glass beads, and formed the main classification and sub classification of various types of glass.

In this paper, we analyzed the classification rules of high potassium glass and lead-barium glass, performed DBSCAN density clustering analysis, and scored the Rand coefficients of the classification results obtained from their clustering, then ranked the scoring results, selected the top three components, and considered that these components could be used as the basis for the classification of high potassium glass and lead-barium glass, followed by a binary logistic Regression analysis was performed to obtain the component content and the probability of occurrence of high potassium or lead-barium glass[4-6]. When the composition content as a predictor variable is known, its glass type can be inferred from the relational equation. For subclassification of each category, firstly, a single-objective planning model is established with the k-means clustering effect evaluation index contour coefficient as the objective function to determine the subclassification criteria, secondly, iterations are programmed through matlab and their clustering information as well as the contour coefficients are recorded, and finally the optimal clustering scheme is solved for single-component, two-component, and three-component cases [7]. Finally, the sensitivity and rationality analysis was performed, and the maximum number of division classes in the optimal subclass division model established above was gradually adjusted from 2 to 8. The sensitivity of the model was analyzed by changing the value of this important parameter and observing the fluctuation of the size of the optimal contour coefficient for each type.

2. Model assumptions and notation

2.1. Assumptions

It is assumed that the SiO2 content represents the degree of weathering

Assume that the experimental data allow for a certain margin of error

Assume that the content of components with a correlation coefficient less than 0.8 with the degree of weathering does not change with the degree of weathering.

2.2. Notations

Important notations used in this paper are listed in Table 1.

Table 1 Notations

Symbols	Description
i	Sampling point number
k_0	Slope of linear equation
s	Sample contour coefficient
S	Overall contour coefficient
π_i	Probability of occurrence of observed events
k	Number of clusters

3. Model construction and solving

3.1. Glass artifacts classification law

(1) DBSCAN density clustering method

DBSCAN is a typical density-based clustering algorithm, a density-based spatial data clustering method proposed by Martin Ester, Hans-PeterKriegel et al. in 1996, which is one of the most commonly used clustering methods. The algorithm takes a region with sufficient density as a distance center and keeps growing the region, the algorithm is based on the fact that a cluster can be uniquely determined by any core object in it [5].

In the DBSCAN algorithm, there are two most basic neighborhood parameters, which are $\in$ domain and MinPts. where $\in$ neighborhood denotes the samples in the data set D whose distance from the sample point $x_i$is not greater than $\in$, i.e

$$N_{\in} = \{x_j \in D \big| dist(x_i, x_j) \leq \in\} \tag{1}$$

(2) Rand coefficient

Rand index is often used to evaluate the performance of clustering models. The value of Rand index is between [0,1] and 1 when the clustering result is a perfect match [8].

(3) Data processing

According to the survey, there are a number of data related to ancient weathered and fresh glass products in China. Archaeologists have classified these cultural relics into high potassium glass and lead barium glass according to their chemical composition and other detection methods, and given the classification information of these cultural relics and the proportion of the corresponding main components. First, the data shall be preliminarily cleaned to remove the data that cause the problem of chemical composition ratio due to external reasons such as detection means. Then the weathering degree of glass is expressed by the content of (silica), and regression analysis is carried out. Finally, the regression equation reflecting the statistical law between the content of each component and the weathering degree is obtained.

In order to analyze the classification mode of high potassium glass and lead barium glass, it is considered that the content of high potassium glass and lead barium glass in the fresh glass is conducive to the classification of high potassium glass and lead barium glass. After preliminary

processing of the original data of cultural relics, the existing data include: the composition content data of the unweathered samples of the unweathered high potassium lead barium glass, the composition content data of the unweathered samples of the weathered high potassium lead barium glass, and the individual composition content of the weathered high potassium lead barium glass before weathering after the weathering point conversion. In the data, the first 18 cases were high potassium glass, and the last 49 cases were lead barium glass.

(4) Modeling

From the observed data, it can be seen that there is a significant difference in the content of some components in high potassium glass and lead-barium glass (The processed data take (potassium oxide) as an example, and its average content in high potassium glass is 0.095 and that in lead barium glass is 0.0018), so it can be concluded that there are some components in the glass that have different content ranges in the two glasses. The DBSCAN is a density clustering technique.

DBSCAN is a density-based clustering method, which requires that the number of objects (points) in a certain region of the clustering space is not less than a certain queue value. This method can discover arbitrarily shaped clusters in spatial data with noise, can link adjacent regions with sufficient density, and can effectively handle outliers. We performed DBSCAN density clustering on each of the 14 components, scored the generated clustering results with Rand coefficient coefficients, ranked the obtained scores, and obtained the components ranked in the top three as the basis for the division of high potassium glass and lead-barium glass.

One or more of the above components were selected as predictor variables for binary logit regression analysis to obtain the relationship between the component content and the probability of occurrence of high potassium or lead-barium glass. It can be seen when the content of the components as predictor variables is known, the glass type can be inferred from the relationship equation.

(5) Model solving

The 14 components were put into the DBSCAN clustering model to obtain their clusters, and some of the results are shown in Table 2 below.

Table. 2 Partial component DBSCAN clustering results

Case Number	(k_2O)	(PbO)	(SrO)	(BaO)	(MgO)
1	1	1	1	1	1
2	-1	1	1	1	2
3	1	1	2	1	1
4	1	1	1	1	3

According to the graph, if we take potassium oxide k_2O as an example, the cases numbered 1, 3 and 4 are classified into cluster 1, and the case numbered 2 is judged as an outlier (Table 3).

Table. 3 Rand coefficient ranking top three components

Ingredient Name	Rand factor
k_2O	0.9719
PbO	0.8649
SrO	0.6115

The clustering effect of DBCSAN for k_2O is shown in Figure 1 below.

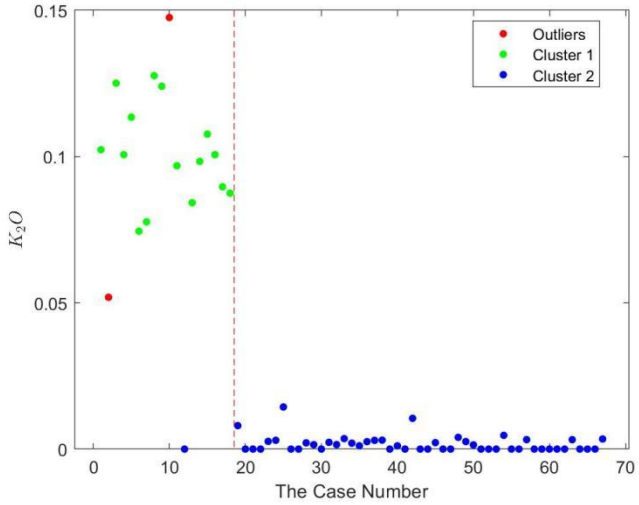

Figure. 1 DBCSAN clustering effect

As can be seen from the figure, excluding the two outliers, the first 18 points of case number are clustered as cluster 1 and the points of case number 19-67 are clustered as cluster 2. In the original data data, the first 18 cases are high potassium glass and the last 49 cases are lead-barium glass. The distribution of cases from this cluster is highly similar to that of high potassium glass and lead-barium glass in the original data, thus demonstrating that the difference in the range of K2O content in high potassium glass and lead-barium glass can be the most important basis for differentiating the two types of glass.

The clustering effect of DBCSAN generated by some other components as feature variables is shown in Figure 2 below.

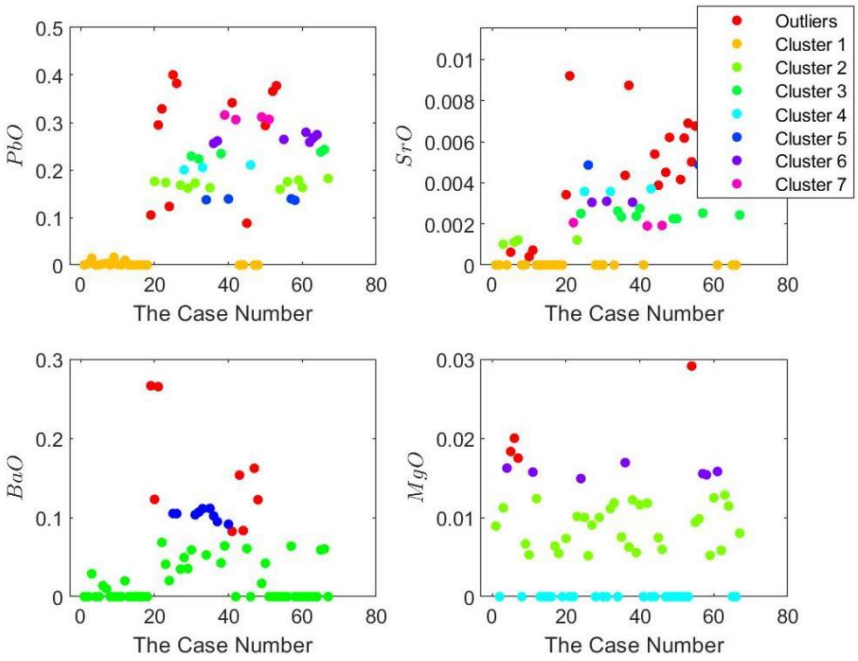

Figure. 2 The effect of DBCSAN clustering with other components as feature variables

(7) Logistic regression analysis [6]

Principle of Logistic Regression Analysis

Suppose the conditional probability of the observed event occurring under the action of the independent variable is $P(y_i = 1|x_i) = \pi_i$, then the conditional probability of the observed event not occurring under the action of the independent variable is $P(y_i = 1|x_i) = 1 - \pi_i$.

$$\pi_i = \frac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\alpha\beta_i x_i)}} \tag{2}$$

$$1-\pi_i = \frac{1}{1+e^{(\alpha+\sum_{i=1}^{n}\beta_i x_i)}} \tag{3}$$

The event occurrence ratio is abbreviated as Odds, nd the natural logarithm of the event occurrence ratio is taken to obtain the logistic regression analysis linear model as

$$ln(\frac{\pi_i}{1-\pi_i}) = \alpha + \sum_{i=1}^{n}\beta_i x_i \tag{4}$$

Solution of the binary logistic regression model

From the DBSCAN clustering model above, it can be seen that the case distributions derived from the PbO and k_2O clustering are most similar to the case distributions of high potassium glass and lead-barium glass in the original data, so PbO and k_2O were selected as predictor variables for binary logistic regression, respectively.

When the predictor variable was PbO, the coefficients of each predictor variable obtained are shown in Table 4 below.

Table 4 The binary logistic regression coefficients of PbO

item	Regression coefficient	P°	OR	OR 95% CI
(PbO)	50.276	0.048	6.83e+21	1.417 -3.29e+43
intercept distance	-1.599	0.005	0.202	0.067-0.613

$$ln(\frac{\pi_{PbO}}{1-\pi_{PbO}}) = 50.276*（PbO）-1.599 \tag{5}$$

When the predictor variable is k_2O, the coefficients obtained for each predictor variable are shown in Table 5 below.

Table 5 k_2O binary logistic regression coefficients

item	Regression coefficient	P°	OR	OR 95% CI
(k_2O)	-151.557	0.036	0.000	0.000 - 0.000
intercept distance	4.201	0.000	66.744	8.193-543.752

$$ln(\frac{\pi_{K2O}}{1-\pi_{K2O}}) = -151.557*（k_2O）+4.021 \tag{6}$$

3.2. Subclass Analysis

(1) Silhouette coefficient

The silhouette coefficient [9] (silhouette coefficient) uses the similarity measure between objects in a dataset to assess the quality of clustering, and is an evaluation indicator of the density and dispersion of clusters. The silhouette coefficient is suitable for cases where the actual category information is unknown. The silhouette coefficient s of a sample is calculated as follows:

$$S = \frac{\mathrm{b-a}}{\max(\mathrm{a,b})} \tag{7}$$

The range of the contour coefficient S is [-1,1], and the larger the value of S, the more reasonable the clustering result. The overall contour coefficient S of the clustering results can be obtained by averaging the contour coefficients of all samples, i.e.:

$$S = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{8}$$

(2) Single-objective planning

The objective of this paper is to find the best subclass division scheme, the size of the contour coefficient can measure the effect of clustering, the larger the contour coefficient, the better the effect of clustering, so the objective is converted to the maximum contour coefficient, that is

$$max\ S_{mnk} \tag{9}$$

When performing cluster analysis, the single-dimensional clustering basis has certain limitations. Appropriately increasing the dimensionality of the clustering basis can dig deeper into the data information, but too high a dimensionality of the clustering basis will lead to complex clustering conditions that are difficult to interpret. For the subclass classification of glass artifacts, too complex classification conditions are obviously inappropriate, so it is limited to select at most three chemical components as the clustering basis. At the same time, the number of subclasses of glass artifacts should not be too many, and the number of subclasses is limited to less than or equal to five species, i.e.

$$1 \le m \le 3 \tag{10}$$

$$2 \le k \le 5 \tag{11}$$

Objective planning model for this problem.

$$max\ S_{mnk}\ s.t. \begin{cases} 1 \le i \\ 2 \le k \\ i \le 3 \\ k \le 5 \end{cases} \tag{12}$$

(3) Matlab solution

It is known that there are 14 chemical components, the number of chemical components based on clustering i=1,2,3, the number of clusters k=2,3,4,5, there are two types of glass need to be subclassified. Therefore, a total of $(C_{14}^1 + C_{14}^2 + C_{14}^3) \times 4 \times 2 = 3709888$ clustering analysis should be completed. Through matlab programming, we traversed 3709888 cases and recorded their clustering information and contour coefficients. Finally, we ranked the contour coefficients to obtain the overall optimal clustering and the optimal clustering when the clustering was based on single component, double component, and triple component, and recorded the information of these optimal clustering parameters and results to develop the subclassification criteria. Sub-category classification criteria are shown in Table 6.

Table 6 Sub-category classification criteria

Category		Contour factor	Classify the components	Divide the number of groups
High potassium	Ⅰ	0.970	SO_2	2
	Ⅱ	0.932	Na_2O. SrO	2
	Ⅲ	0.894	Na_2O. SrO. SO_2	2
Lead Barium	Ⅰ	0.985	SO_2	3
	Ⅱ	0.968	$SO_2.S_nO_2$	2
	Ⅲ	0.956	$K_2O.SO_2.S_nO_2$	2

For high potassium glass, the contour coefficient of the optimal clustering decreases with the increase of the division basis, so it is suitable for single-component subclassification, with SO2 as the division basis and the number of classes is 2. For lead-barium glass, the contour coefficient of the optimal clustering decreases with the increase of the division basis and considering that the multidimensional division basis will aggravate the subjectivity of the division, so it is more suitable for single-component subclassification, with SO_2 as the division basis and the number of classes is 3. The number of subclasses is 3.

The optimal clustering results are shown in Figure 3.

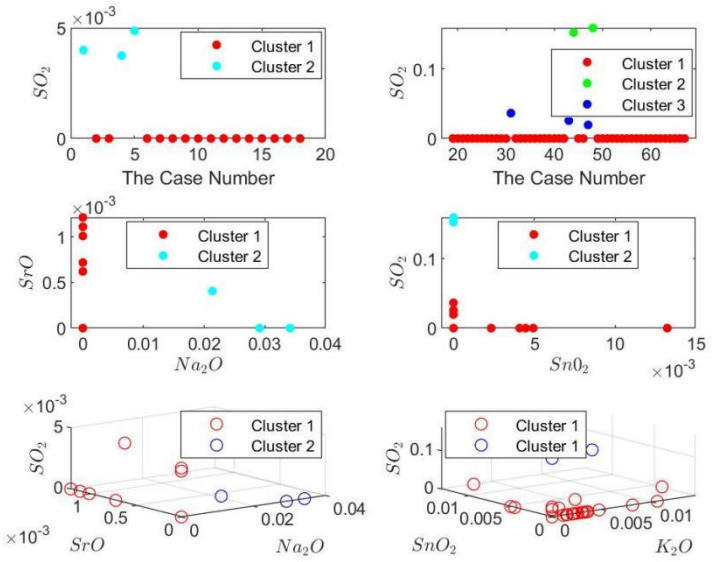

Figure. 3 Subclass optimal clustering results

(4) Sensitivity analysis [10]

In the optimal subclass delineation model established in single-objective planning above, the maximum number of delineated classes is limited to five and the maximum number of chemical components on which clustering is based is limited to three. The model sensitivity was analyzed by changing the values of these two important parameters and observing the fluctuation of the optimal contour coefficient size for each type.

By gradually increasing the maximum number of classifications from 2 to 8, the left side of Figure 4 can be obtained, and the best classification contour coefficients are high for each classification case, and overlap occurs in six-line segments. The lines in the figure are horizontal, indicating that the parameter maximum classification number has minimal effect on the model and the model is not sensitive to this parameter.

Adjusting the maximum classification as well as the number of components from 1 to 5 yields the above Figure 4 right. Overall, the model is more sensitive to the number of components based on the maximum classification of the parameter than to the maximum classification tree of the parameter. With the change of this parameter, the optimal classification contour coefficients of high potassium glass and lead-barium glass fluctuate with an amplitude of about 10%, and the model is sensitive to this parameter. In the future, the model stability can be improved and the sensitivity can be reduced by improving the setting of this parameter.

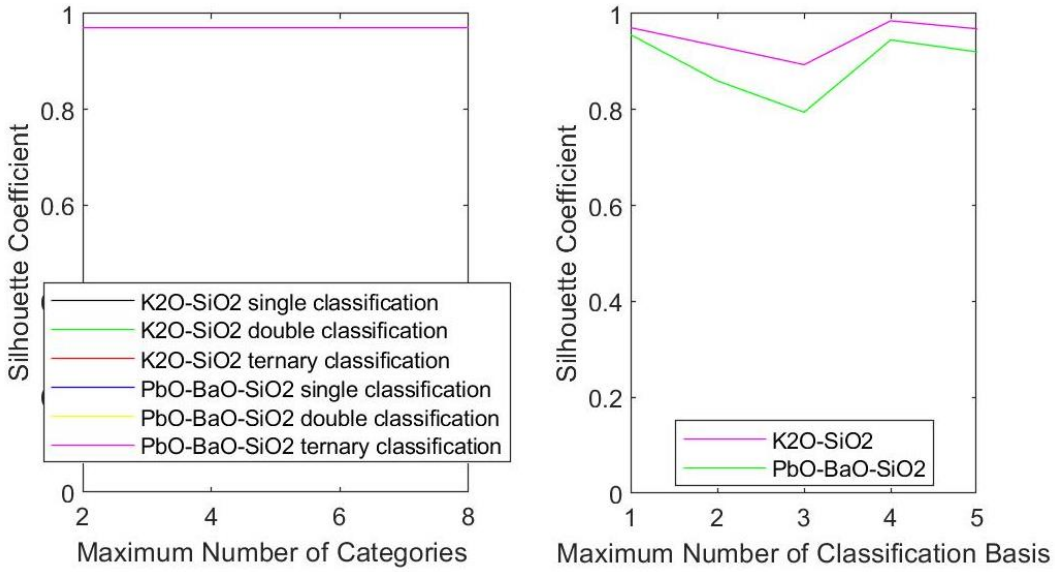

Figure. 4 Fluctuation of the profile coefficient with the maximum number of categories

4. Conclusion

In this paper, the glass classification law was studied, and the DBSCAN density cluster analysis method was used to score the clustering results with Rand coefficients, and finally PbO and K_2O were selected as the classification basis, and a binary logistic regression model was established. The glass subclass division was studied, and k-means cluster analysis was introduced to transform the problem into a single-objective planning problem by taking the contour coefficient of the clustering effect evaluation index as the objective function. All combinations of decision vectors under the constraints are traversed by matlab programming, and finally the optimal clustering partitioning scheme is solved for single-component, two-component, and three-component times. Finally, for sensitivity and plausibility analysis, the maximum number of division classes in the optimal subclass division model established above is gradually adjusted from 2 to 8. The model sensitivity is analyzed by changing the value of this important parameter and observing the fluctuation of the optimal contour coefficient size for each type.

The research advantage of this paper is that when establishing the glass classification model, the single objective programming model is adopted, the classification basis under various circumstances is considered, and the thinking and modeling of depth are set. At the same time, the evaluation methods of RAND coefficient and profile coefficient increase the reliability and rationality of the model. Logistic probabilistic nonlinear regression model classification model also provides a certain fault tolerance rate for glass classification. This modeling idea can be widely used in various classification models in the future.

References

[1] He Qiang Review of Gan Fuxi et al.'s History of the Development of Chinese Ancient Glass Technology [J] Chinese Journal of Science and Technology History, 2017, (3): 371-376

[2] Li Fei, Li Qinghui, Gan Fuxi, Zhang Bin, Cheng Huansheng. Proton excited X-ray fluorescence analysis of chemical composition of a batch of ancient Chinese glasses [J]. Journal of Silicate, 2005 (05): 581-586

[3] Wang Chengyu, Tao Ying. Weathering of silicate glass [J]. Journal of Silicate, 2003 (01): 78-85

[4] Huang Yi, Wang Sitong, Zhang Tingting, et al Identification of Middle East crude oil by logistic regression analysis based on diagnostic ratio [J] Environmental Science and Technology, 2017, 40 (10): 66-70

[5] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//kdd. 1996, 96(34): 226-231.

[6] Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied logistic regression [M]. John Wiley & Sons, 2013.

[7] Likas A, Vlassis N, Verbeek J J. The global k-means clustering algorithm [J]. Pattern recognition, 2003, 36(2): 451-461.

[8] Santos J M, Embrechts M. On the use of the adjusted rand index as a metric for evaluating supervised classification [C]//International conference on artificial neural networks. Springer, Berlin, Heidelberg, 2009: 175-184.

[9] Dinh D T, Fujinami T, Huynh V N. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient[C]//International Symposium on Knowledge and Systems Sciences. Springer, Singapore, 2019: 1-17.

[10] Saltelli A. Sensitivity analysis for importance assessment [J]. Risk analysis, 2002, 22(3): 579-590.