Glass classification and identification based on lasso regression and K-means clustering

Baowei Xu [1,*], Boxun Li [1], Siyu Tian [2]

[1] Department of Mathematics and Information Science, Hebei University, Hebei, China, 071002

[2] Department of Physical Sciences and Technology, Hebei University, Hebei, China, 071002

* Corresponding author: xbwdianbao@163.com

Abstract. Ancient glass is classified into two types, high potassium glass and lead-barium glass, which are highly susceptible to weathering by the burial environment. In order to protect the glass artifacts more safely, based on some data related to the chemical composition ratio of glass artifacts,by the lasso regression model based on the logit transformation, the classification rules of high potassium and lead-barium glass were analyzed, and the glass was divided into four subclasses based on the principal component analysis using K-means clustering, and sensitivity analysis was performed on the classification results. Finally, the validity of the model was verified by determining the type of glass to which a set of test data belonged.

Keywords: Glass classification, lasso regression, K-means clustering.

1. Introduction

The accurate classification of ancient Chinese glass objects is important for the conservation of glass artefacts and the historical tracing of glass artefacts [1]. The chemical composition content of glass artefacts changes before and after weathering [2]. Subclassification of high potassium glass and lead-barium glass is important in reducing the damage to glass artefacts. Lasso regression allows for a more accurate fit to the dataset [3-7]. Based on the dataset with the help of K-means clustering can be more accurate to classify different categories [8-11].

2. Classification model building and solution

2.1. Data preprocessing

There are cases of missing values for individual chemical elements in some samples in the data set. Because of the complex chemical composition of glass, it is difficult to detect the content of some chemicals because they are rarely present. The sample data in the dataset consists of component ratios, and strictly speaking, the sum of the component ratios should be 100%, but considering the unavoidable factors such as testing methods, the component ratios may change. Therefore, the data with the summation between 85% and 105% are considered valid. Since the chemical composition content of glass changes significantly before and after weathering, direct analysis of all samples in the data set will lead to confusion of classification laws. Therefore, based on the principle of controlling a single variable, the data set is divided into two categories: unweathered and weathered, and the classification laws of high potassium glass and lead-barium glass are investigated in these two categories.

2.2. Analysis of the classification pattern of high potassium and lead barium

The content of each chemical component in the data set was used as the independent variable $x_i(i=1,2,...,14)$, Set the type of glass high potassium $(y=0)$ and lead barium $(y=1)$ as dichotomous dependent variables.

If linear regression is used, it is straightforward to obtain:

$$\widehat{y} = \beta_i x_i + \varepsilon \tag{1}$$

But this will not be a number between 0 and 1. From this, for the dichotomous dependent variable in this problem, note that the probability of y taking 1 is $p = P(y = 1|X)$, The probability of taking 0 is $1 - p$, and the ratio of the probability of taking 1 and taking 0 is $p/(1-p)$. For the dichotomous regression target probability will be between 0 and 1, but the value of the dependent variable of the regression equation falls in the set of real numbers, which is unacceptable, so the Logit transformation is performed here to get:

$$\lambda = \ln \frac{p}{1-p} \tag{2}$$

And the dichotomous regression model is a linear regression model built between $\ln\left(\frac{p}{1-p}\right)$ and the independent variable. The equation of the regression model is:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{14} x_{14} + \varepsilon \tag{3}$$

where ε is the random error. Because $\ln\left(\frac{p}{1-p}\right)$ takes values $(-\infty, +\infty)$, such that the independent variable $x_1, x_2, \cdots, x_{14}$ can take values in any range.

The basic method for estimating the parameter vector β_i is the least squares method, the idea being to make the error vector $y_i - \beta_0 - \sum_{i=1}^{n} x_i \beta_i$ as small as possible, i.e.:

$$arg\min_{\beta} \sum \left(y_i - \beta_0 - \sum_{i=1}^{n} x_i \beta_i \right)^2 \tag{4}$$

And the Lasso principle is to add a 1-paradigm number after it:

$$\widehat{\beta}^{Lasso} = arg\min_{\beta} \left\{ \frac{1}{2} \sum \left(y_i - \beta_0 - \sum_{i=1}^{n} x_i \beta_i \right)^2 + \lambda \sum_{i=1}^{n} |\beta_i| \right\} \tag{5}$$

Or write it in another form of expression:

$$\begin{gathered} \widehat{\beta}^{Lasso} = arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{i=1}^{n} x_i \beta_i \right)^2 \\ subject\ to\ \sum_{i=1}^{n} |\beta_i| \leqslant t \end{gathered} \tag{6}$$

The two equations are equivalent in that the first line in the above equation is the objective function of OLS, while the Lasso principle adds a second line of constraints, which corresponds to the second

term in the other form. the smaller the t, the larger the λ. The complexity of the Lasso regression is controlled by the parameter λ. The larger the λ, the stronger the penalty for models with more variables, and the stronger the compression of the estimated parameters. This objective function is minimized so that the coefficients of some less important independent variables will be compressed to zero, thus achieving the effect of screening variables.

The choice of λ is then determined by K-fold cross-validation. For the samples in the dataset, we take K=8 and determine the number of variables and the selection of variables by finding the λ that makes the highest correct prediction rate for the whole sample. In other words, the glass content samples in the data set are randomly divided into 8 equal parts, and the first subsample is kept as the validation set, while the remaining 7 subsamples are used as the training set to estimate the model, and then the first subsample is used to predict the first subsample, and the correct prediction rate of the first subsample is calculated, and so on, and after calculating K=8 times, the adjustment parameters are chosen to make the maximum correct prediction rate of the whole sample, so The variable at this point is the best choice. Take the case of weathered glass as an example, through this process, Figure 1 can be obtained:

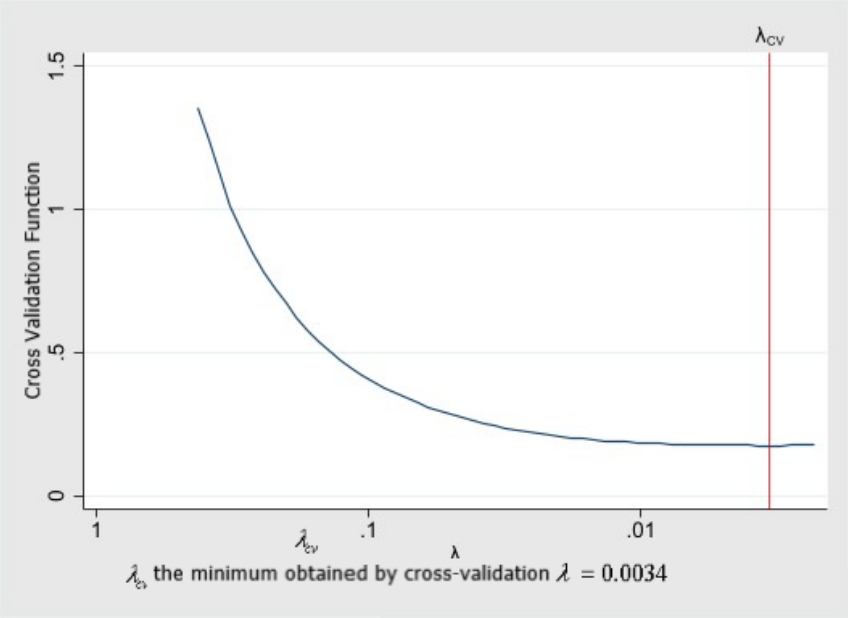

Figure 1. K-fold cross-validation with λ

The optimal values of the conditioning function are given in Figure 1, and the function is very flat around the optimal values, which means that the model has good predictive stability.

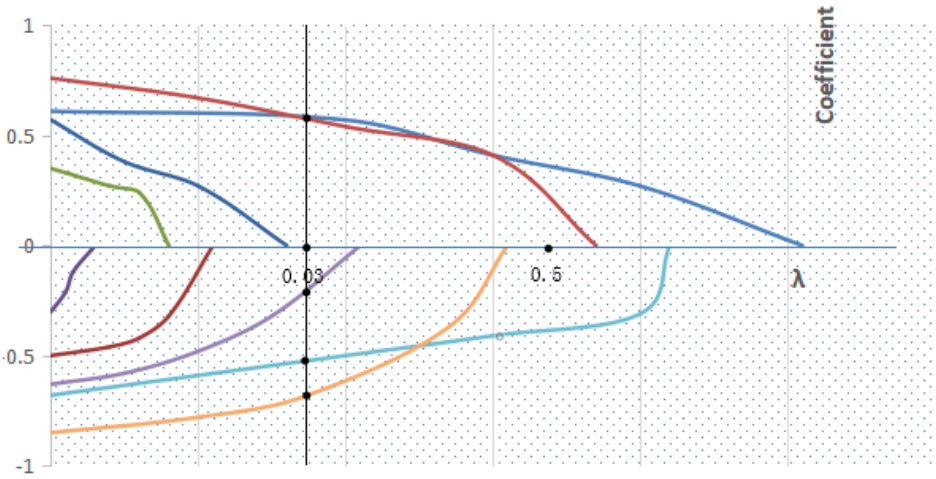

Figure 2. Variable filtering

It can also be seen from Figure 2 that the number of variables with a coefficient value of 0 increases as λ increases, and the number of variables is 5 when the optimal λ obtained by cross-validation is obtained. After combining the results of the analysis, we can get:

Table 1. Lasso regression coefficient

Unweathered glass	Standardized Lasso regression coefficients	Weathered glass	Standardized Lasso regression coefficients
(Constant)	3.880	(Constant)	8.390
x_1	-0.232	x_1	-7.955
x_3	-0.649	x_9	1.024
x_4	-0.791	x_{10}	1.673
x_9	6.233		
x_{10}	0.509		

Table 1 shows that the main basis for the distinction between high potassium glass and lead-barium glass is SiO_2、BaO、PbO. In order to further discuss the specific classification law, the following Figure 3 are made:

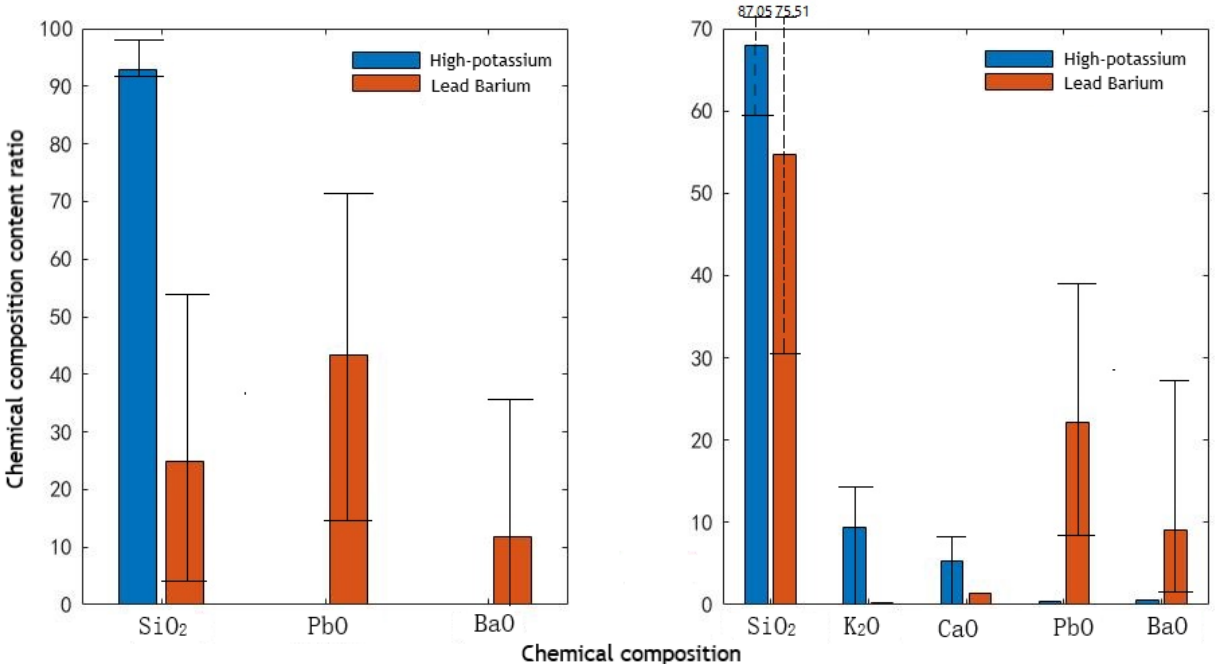

Figure 3. High potassium, lead barium main component content comparison

(Note: The upper and lower bounds indicate the maximum and minimum values in the sample)

It can be seen from Figure 3 that for weathered glass, a glass with a silica content level around 93.96%, with a specific interval range of [92.35, 96.77], and not containing lead oxide and barium oxide can be considered as a high potassium type glass; a glass with a silica content level around 25.66%, with a specific interval range of [3.72, 53.33], and the presence of barium oxide and The glass is considered as lead-barium glass with the specific intervals of [15.71, 70.21] and [0, 35.45] for lead oxide and barium oxide, respectively. Similarly, for unweathered glass, the silica content has a certain degree of differentiation if it falls between the small interval [31.94, 59.01] or the large interval [75.51, 87.05], and if it falls in the middle interval, it needs to be analyzed for other components, and if it appears to have a high potassium content as well as calcium content, the specific intervals are [0, 14.52] and [0, 8.7], respectively, it can be If the high potassium and barium contents

are in the intervals of [9.3, 39.22] and [2.03, 26.23], then the glass is considered to be a high potassium glass.

2.3. Subclass division

Subclassifications were made for each category, and the appropriate chemical composition was first selected for each type of glass artifact sample. Through preliminary analysis of the data in the dataset, it was observed that the contents of the 14 chemical components showed different degrees of variation, both for different glass types and before and after weathering, but there were more variables with mainly 0 values, and their individual influence was small. In order to consider the influence of all independent variables comprehensively, so this paper first used principal component analysis to extract factors for the 14 chemical components, and each type of glass to obtain the corresponding principal components, and then use the K-means clustering algorithm to perform cluster analysis on the sample principal component data in each type to obtain the classification of subclasses.

According to the above analysis, the flow chart of the whole process is shown in Figure 4:

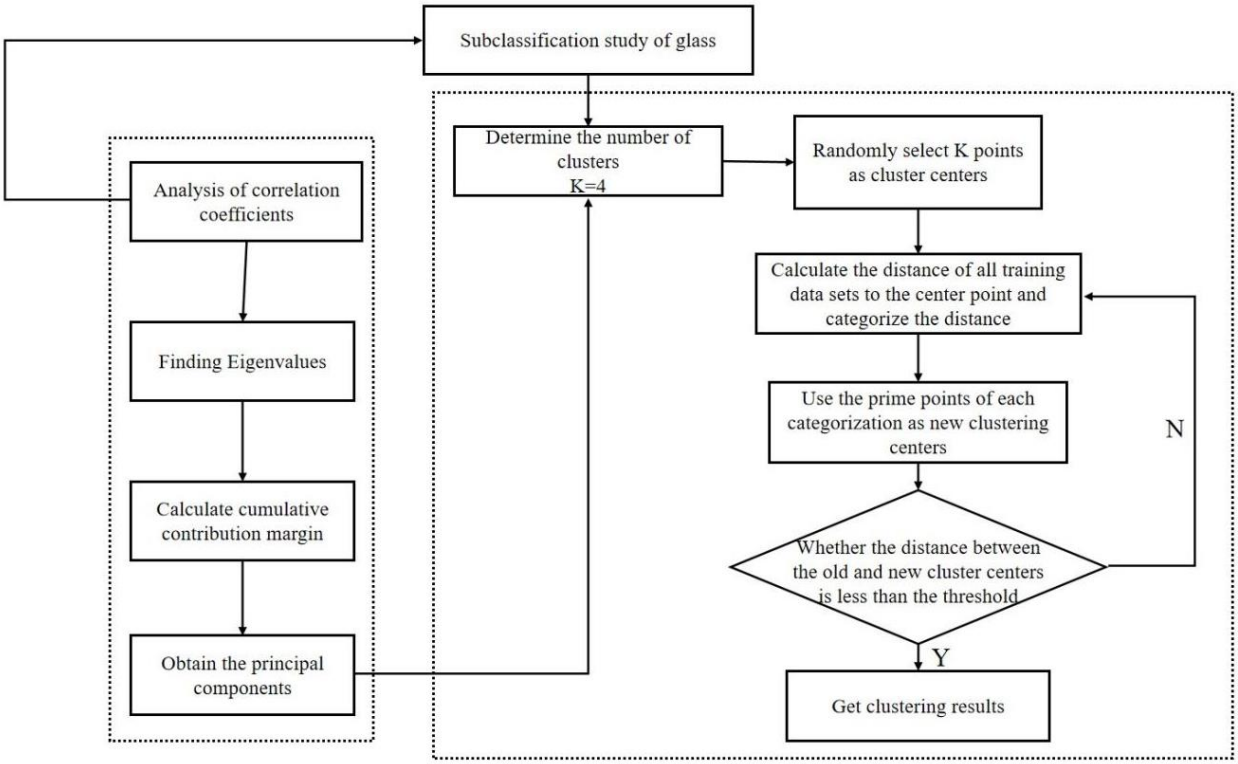

Figure 4. Flow chart of subcategory division

In order to ensure the reliability of the results of cluster analysis using principal components, the cumulative contribution of principal components was ensured to be above 90% when extracting principal components. The final number of principal components we selected was 6 for the high potassium category and 9 for the lead-barium category. Because there are already two types of high potassium and lead barium, weathered and unweathered, the classification of high potassium glass and lead barium glass should be divided into four categories, and the number of cluster centroids chosen is 4. The results of the principal component extraction and cluster analysis of the two glass types, taking lead barium glass as an example, are shown in Table 2.

Table 2. Lead barium glass subclass classification results

No.	Category	No.	Category
Weathered		Unweathered	
2	IV	20	IV
36	IV	23	IV
8(1)	III	25	IV
8(2)	III	28	IV
26(1)	III	29	IV
26(2)	III	30(1)	IV
11	II	31	IV
19	II	32	IV
34	II	33	IV
38	II	35	IV
39	II	37	IV
40	II	42(1)	IV
41	II	42(2)	IV
43(1)	II	44	IV
43(2)	II	45	IV
49	II	46	IV
50	II	47	IV
51(1)	II	49	IV
51(2)	II	53	IV
52	II	55	IV
54(1)	II	24	III
54(2)	II	30(2)	II
56	II	50	II
57	II		
58	II		
48	I		

Analysis of Table 1, for lead barium divided into class IV and class II mainly due to the influence of unweathered and weathered, and for class I and class III mainly in the weathered area of the division, respectively, to observe the location of the final clustering center of class 2 and class 1 and class 3, it can be seen that the class 2 and class 1 with the greatest degree of difference is the main component 1 and main component 6, where the coefficient of the largest variable in the main component 1 is SiO_2、Al_2O_3、BaO.And these three oxides have gain effect on the melting point temperature of glass, so the main component 1 of lead barium can be called the melting point property component; in the main component 6, Na_2O、K_2O、SnO_2 have the largest coefficient, indicating that it has more influence on the main component 6, and these three oxides all have a greater impact on the gloss of glass, so the main component 6 can be called the gloss property component; so the subclass 1 The classification is based on the melting point and glossiness of the glass, which are mainly influenced by SiO_2, Al_2O_3、BaO、Na_2O、K_2O、SnO_2.

The biggest difference between class II and subclass III is the main component 2. For the variables CaO、MgO, which are the largest coefficients in the main component 2, both have a greater impact on the thermal stability of the glass, and MgO magnesium oxide can substantially improve the thermal stability while CaO calcium oxide has a more obvious negative impact on the thermal stability, so the main component 2 can be called the thermal stability quality component, so the division of subclass III is based on the thermal stability of the glass, which is mainly affected by CaO、MgO.

After the same analysis we can also obtain the subclassification of high potassium glasses, for both subclass II and subclass IV are divided from the unweathered high potassium (class I), the

classification of subclass II is based on the color characteristics of the glass, mainly influenced by CuO、SnO_2. The classification of subclass 4 is based on the thermal stability of the glass, which is mainly influenced by tNa_2O、CaO.

3. Sensitivity analysis

The sensitivity analysis of the classification model is particularly important when the chemical composition content of glass samples fluctuates within a certain range due to inaccuracies in the measurement of chemical composition content that may be caused by testing methods. Therefore, to address this issue, this paper randomly selects representative glass sample data (each with medium chemical composition content level and few chemical compositions with 0 composition) in high potassium weathering, high potassium unweathered, lead-barium weathering, and lead-barium unweathered. In this paper, the glass sample data with numbers 22, 14, 44, and 36 are randomly selected, and their chemical compositions are processed to fluctuate, and the interval of variation is set as the given content The interval of variation is set as 5%~10% of the given content, but the sum of all contents after fluctuation is ensured to be between 85%~105%. After the fluctuation treatment, each sample was subjected to principal component extraction and cluster analysis again, and if the fluctuating sample belonged to a category that changed, the model was considered to be more sensitive to the fluctuating chemical composition of the sample. On the contrary, it is not sensitive to it.

Taking sample No.36 as an example, its silica content was set to fluctuate within the specified range to obtain the results in Figure 5.

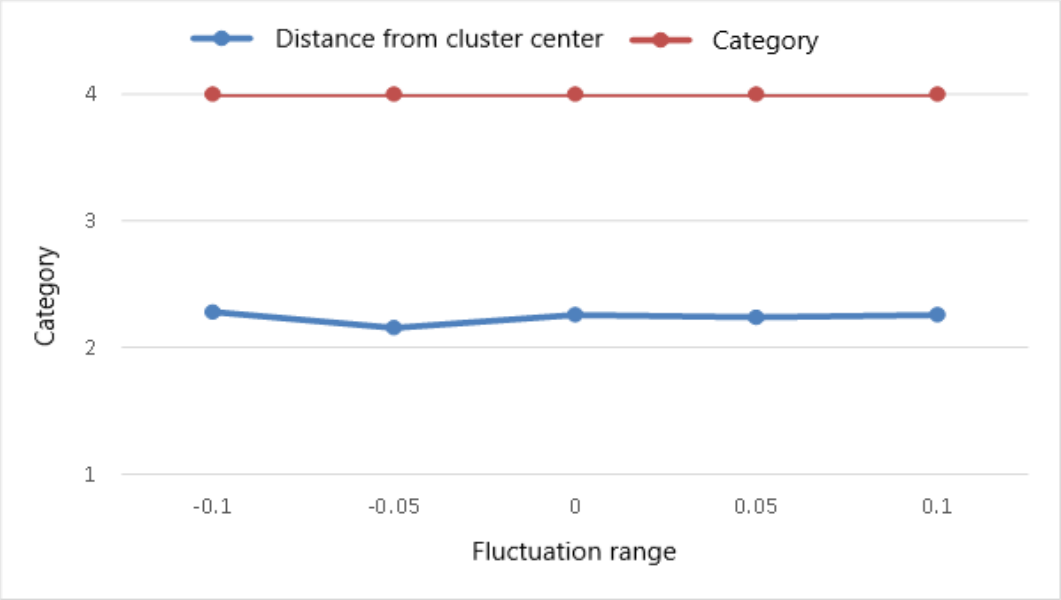

Figure 5. Sensitivity analysis of sample 36

The horizontal coordinate indicates that the chemical composition content fluctuates within 10% above and below, the integer value of the vertical coordinate indicates the category to which this sample data belongs, while the remaining continuous data indicates the distance of the sample points from the final clustering center. It can be seen from Figure 5 that the glass after lead-barium weathering is more stable and less sensitive to the classification results. After the same treatment of the other samples, after the treatment of fluctuations in chemical content, it was found that none of the classification results changed, so the classification results of all four samples were not sensitive to data fluctuations, so the model can be considered to have a strong stability.

4. Applying the model

First classify the glass data in the test set according to the classification law of high potassium and lead-barium types obtained in Section 2. The classification results in Table 3 were obtained:

Table 3. Preliminary division results

Glass Category	Artifact Number
High Potassium	$A1$, $A6$, $A7$
Lead Barium	$A2$, $A3$, $A4$, $A5$, $A8$

After that, the principal components of each glass sample were extracted according to the high potassium and lead-barium categories, and then the distances to the four clustering centers of the category were calculated separately to finally obtain the specific classification of the samples.

Table 4. Final division results

Artifact Number	Glass Category
$A1$	High Potassium class I(Unweathered class)
$A2$	Lead Barium class II(Weathered class)
$A3$	Lead Barium class III(Unweathered subclasses)
$A4$	Lead Barium class II(Unweathered subclasses)
$A5$	Lead Barium class I(Weathered subclasses)
$A6$	High Potassium class III(Weathered class)
$A7$	High Potassium class III(Weathered class)
$A8$	Lead Barium class II(Unweathered subclasses)

5. Conclusion

In order to analyze the classification laws of high potassium glass and lead-barium glass, this paper established a lasso regression model based on Logit transformation, completed the screening of independent variables for the dichotomous variables of glass types, and obtained the conclusion that the classification laws of high potassium and lead-barium glass were mainly based on K_2O、BaO、PbO. Then the K-means clustering method was used based on principal component analysis to divide each type into four subclasses, and the classification results were justified by exploring the meaning of principal components. Finally, the classification results were analyzed after the fluctuation processing of sample data, and it was argued that the model has certain stability and low sensitivity.

Then using the glass artifact data of the unknown category of the test set substituted into the classification model in Section 2, the numbers A1-A8 were obtained as belonging to high potassium calss I, lead barium class II, lead barium class III, lead barium class II, lead barium class I, high potassium class III, high potassium class III, and lead barium class II, respectively.

References

[1] Lu Shoulin. Ancient glass and its conservation [C]//. Conservation techniques for cultural relics (1981-1991)., 2010: 299 - 309.

[2] Hu ZZ, Li P, Jiang LUMAN, Wang TONYANG, Du GU, Yang BO. Analysis of LA-ICP-MS fractions of ancient glass materials and study of their production sources [J]. Rock and mineral testing, 2020, 39 (04): 505 - 514.DOI: 10.15898/j.cnki.11-2131/td.201909210134.

[3] Wu, Shu-Chen, Qi, Zong-Feng, Li, Jian-Xun. Intelligent global sensitivity analysis based on deep learning [J]. Journal of Shanghai Jiaotong University, 2022, 56 (07): 840 - 849. DOI: 10.16183/j.cnki.jsjtu.2021.191.

[4] Ke, Zhenglin. Application of Lasso and its correlation method in multiple linear regression models [D]. Beijing Jiaotong University, 2011.

[5] Li Xueyang, Shao Xigao. Research on the factors influencing the quality of university students based on multiple linear regression and Lasso regression[J]. Journal of Ludong University (Natural Science Edition), 2022, 38 (04): 350 - 356.

[6] Wang Lu, Sun Jubo. Application of Lasso regression method in the selection of characteristic variables [J]. Journal of Jilin Engineering and Technology Teacher's College, 2021, 37 (12): 109 - 112.

[7] Hou Da Ke. Comparative study and empirical analysis of Lasso class variable selection methods [D]. Shandong University, 2021. DOI: 10.27272/d.cnki.gshdu.2021.002132.

[8] Wang Qian,Wang Cheng, Feng Zhenyuan, Ye Jinfeng.A review of K-means clustering algorithm research [J]. Electronic Design Engineering, 2012, 20 (07): 21 - 24. DOI: 10.14022/j.cnki. dzsjgc.2012.07.034.

[9] Gao Tian, Fan Xiong. Research on lithology classification based on k-means algorithm [J]. Microcomputer Applications, 2022, 38 (08): 113 - 115.

[10] Sun Qizong, Huaer Tian, Sun Liying. A product custom feature classification method based on.

[11] K-means algorithm [J]. Jiangxi Science ,2022, 40 (03): 423 - 428+433.DOI: 10.13990/j. issn1001-3679.2022.03.003.

[12] Liu X., Wang X. Li. Research on airline customer value classification based on k-means and neighborhood rough sets[J]. Operations Research and Management, 2021, 30 (03): 104 - 111.