Composition analysis and identification of ancient glass based on K-Means clustering

Keju Dong *

School of Computer Science and Technology, Qilu University of Technology, Jinan, China, 250353

* Corresponding Author Email: kejud1@163.com

Abstract. The analysis and research on the composition of ancient glass is widely used in the reduction and identification of the composition of archaeological relics. In this paper, the relevant data of CUMCM2022 problem C is used to establish a mathematical model, and multiple linear regression, binary logistic regression and K-means clustering algorithms are used to realize the systematic analysis of the chemical composition of cultural relics, and solve the problem of predicting the classification of cultural relics. In the study of ancient glass classification, multiple linear regression and binary logistic regression are used to transform abstract text data into intuitive multiple linear regression equation, and a relatively ideal classification law is obtained. According to the change rate of chemical composition before and after weathering of various kinds of glass, the appropriate chemical composition was selected as the classification index, and the elbow method and K-means clustering were used to obtain the distinctive subclassification results. Finally, the sensitivity of K-means clustering subclassification model is analyzed by numerical perturbation method, and the model shows high stability.

Keywords: Ancient glass, binary logistic regression, K-means clustering.

1. Introduction

The origin of glass in China can be traced back to the Western Zhou Dynasty in 1000 BC, and the real glass appeared roughly in the middle and late Spring and Autumn period of the 6th century BC. Glass art is unique in the development history of Chinese craft. Among them, lead-barium glass is unique in ancient China. It appeared no later than the early Warring States period in the 5th century BC and was still produced in the Western Han Dynasty. Potassium glass mainly appeared in the 4th century BC to the 3rd century AD, equivalent to the middle and late Warring States to the Eastern Han Dynasty [1].

In the process of burying ancient glass for a long time, it is very easy to be affected by the surrounding environment, resulting in dark weathering, reduced transparency, faint color, or forming a weathering product crust outside. In the process of weathering, a large amount of internal elements exchange with environmental elements, resulting in changes in their composition ratio, thus affecting the correct judgment of their categories [2-3]. Therefore, to analyze the chemical composition of glass relics, it is necessary to study the surface weathering law of all kinds of glass relics. Based on the above background and actual data, this paper uses correlation algorithms such as binary logistic regression and K-means to study the composition of ancient Chinese glass, combines the relevant technology of modern data processing and the research results of ancient Chinese glass at home and abroad, and realizes the exploration of the classification law of ancient high-potassium glass and lead-barium glass [4] and the prediction of specific subcategories.

2. Model data and methods

2.1. Data and preprocessing

Data related to CUMCM2022 question C (http://www.mcm.edu.cn/) were used for this study. Two representative types of ancient glass, namely high potassium and lead barium, were selected from the data, including their shape characteristics and chemical characteristics, and their chemical composition was diverse, which can better help us to conduct a more detailed study of ancient glass.

According to the research object and the problem to be solved in this paper, this paper carried out the pretreatment operations of invalid data elimination and vacancy data completion, and finally obtained 55 groups of ancient glass contour feature data and 67 groups of chemical feature data.

2.2. Multiple linear regression forward method

The forward method is a regression analysis method to determine the choice of alternative independent variables [5]. The specific method is as follows: the forward selection process starts with the number of independent variables in the model being zero, and the first one to be included in the model is the independent variable that has the maximum correlation coefficient with the absolute value of the dependent variable and meets the "inclusion criteria". The next independent variable to be included in the model should be the one whose absolute value of partial correlation coefficient is the largest among the remaining alternative variables and conforms to the "inclusion standard". This process continues until no more alternative arguments meet the "inclusion criteria" [6]. Forward method can help us to remove highly correlated independent variables and reduce the amount of data computation.

2.3. Binary logistic regression

Binary logistic regression is an algorithm proposed to solve the binary classification problem. Based on the assumption that the data obey Bernoulli distribution, the algorithm gives conditional probability distribution and uses maximum likelihood estimation to solve the optimal parameters. The schematic diagram of its model is shown in Figure 1 [7]. The algorithm is very suitable for binary classification problem and does not need to scale the input features, which provides advantages for our classification.

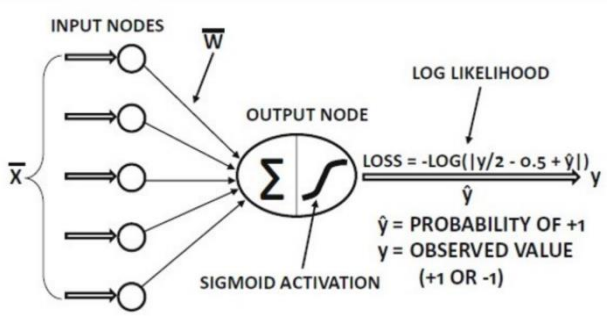


Figure 1. Binary logistic regression model

2.4. K-means clustering and "Elbow Method"

K-means clustering algorithm is a kind of iterative solving clustering analysis algorithm. Its steps are as follows: pre-divide the data into K groups, then randomly select k objects as the initial clustering center, and then calculate the distance between each object and each seed clustering center, and assign each object to the nearest clustering center. For processing large data sets, this algorithm maintains scalability and efficiency, and can quickly obtain effective classification results [8]. The K-means clustering model based on Euclidean distance is adopted in this paper. The distance calculation formula is as follows:

$$d(x,y) := \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad (1)$$

"Elbow method" is a mainstream algorithm used to determine the clustering number k, with the increase of the clustering number k, sample division will be more fine, the degree of aggregation of each cluster will gradually improve, so the error square and SSE will naturally gradually smaller. In addition, when k is less than the real clustering number, the increase of k will greatly increase the degree of aggregation of each cluster, so SSE will decline greatly. When k reaches the real clustering

number, the return on the degree of aggregation obtained by increasing k will rapidly decrease, so the decline range of SSE will decrease sharply, and then it will be gentle with the continuous increase of k value. That is to say, the relation graph between SSE and k is the shape of an elbow, and the corresponding k value of this elbow is the real clustering number of data [9]. Elbow method is an empirical method, which is very suitable for the high dimensional sample data in this paper. the core index SSE(sum of the squared errors, sum of the squared errors) of the elbow method is calculated by the following formula:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \tag{2}$$

3. Model establishment and solution

3.1. Establishment of logistic regression model to analyze the classification rules of high-potassium glass and lead-barium glass

3.1.1 Preliminary screening of variables

To establish a classification model based on binary Logistics regression, it is necessary to select reasonable and effective variables, screen out useless variables, so as to improve classification efficiency and obtain ideal classification rules. Based on the data of Form 1 and Form 2, reasonable variables were selected from 15 variables, including 14 chemical compositions and weathering. Since the sum of each chemical component content of each sampling point in Form 2 is approximately equal to 100%, the model has a serious complete collinearity problem. Therefore, first of all, it is necessary to eliminate some variables with a small amount of data and low classification value to solve the complete collinearity problem. As can be seen from Figure 2, the proportion of samples containing SnO_2, SO_2 or Na_2O in the total samples is much lower than other variables, so it is omitted here.

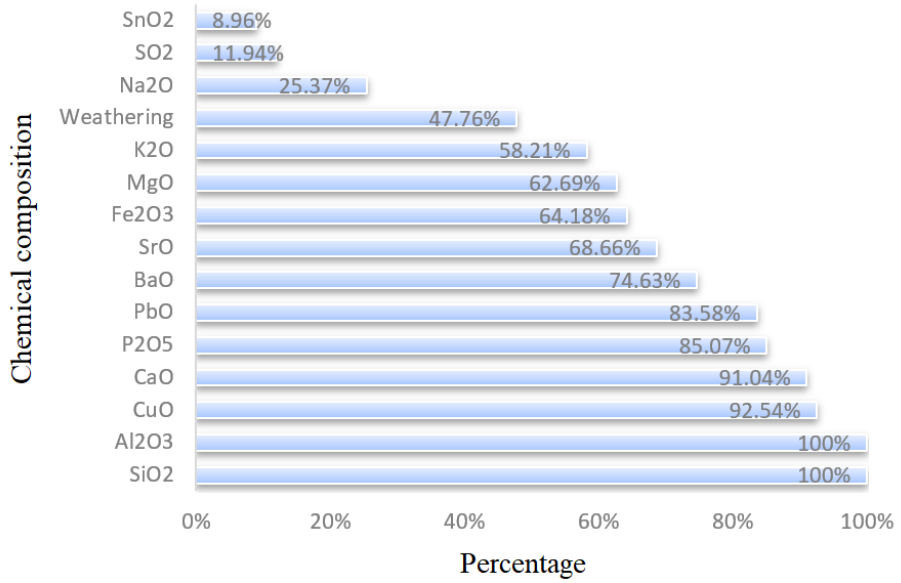


Figure 2. Ranking of the ratio between the number of samples containing this variable and the total number of samples

3.1.2 Use forward method to further screen variables

Due to the small number of initially screened variables, the model still has a serious multicollinearity problem [10], so it is necessary to use other reasonable methods to eliminate more useless variables. In this paper, a linear regression model based on the forward method is established to further screen out reasonable and effective variables.

The model was iterated for five rounds. The results of variable screening by the forward method were shown in Table 1. The significance of the selected variables in each round was <0.05, which had statistical significance. According to the principle of maximum and optimal R2 in statistics, Model 5 should be selected, that is, PbO, K_2O, SiO_2, Al_2O_3 and CuO are selected for binary Logistics regression.

Table 1. Five models produced by the forward method and selected variables

Model number	Predictive variable					The adjusted R^2
1	PbO					0.566
2	PbO	K_2O				0.696
3	PbO	K_2O	SiO_2			0.746
4	PbO	K_2O	SiO_2	Al_2O_3		0.765
5	PbO	K_2O	SiO_2	Al_2O_3	CuO	0.789

3.1.3 Establishment and solution of binary Logistics regression model

The glass type and chemical composition variables PbO, K2O, SiO2, Al2O3 and CuO gradually screened by the forward method were analyzed by binary logistic regression. Results As shown in Figure 3, the significance of chemical component variables SiO2, Al2O3 and CuO is <0.05, which are valid statistical data, and the OR values are all > 1, which has an obvious positive effect on classification and can be used as the independent variable of binary Logistics regression equation.

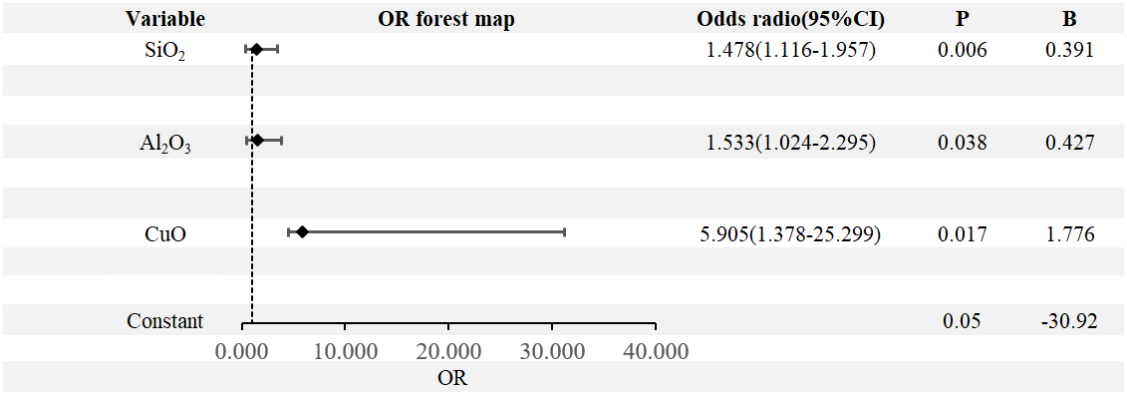


Figure 3. Coefficient, significance, OR value of logistic regression equation

Figure 3 shows the coefficient of the logistic regression equation. Let the glass type be y , and the chemical composition variables SiO_2, Al_2O_3 and CuO are x_1 , x_2 and x_3 respectively. Therefore, the logistic regression equation of glass classification law can be written as follows:

$$y = -0.391x_1 - 1.776x_2 - 0.427x_3 + 30.920 \quad (3)$$

The binary logistic regression equation was used to predict the types of 69 groups of sample data, as shown in Table 2. The accuracy reached 94%, indicating that the model was ideal.

Table 2. Correct rate of Logistics regression prediction

/		Forecast type		Correct percentage
		High manganese	Lead barium	
Measured type	High manganese	47	2	95.9
	Lead barium	2	16	88.9
Overall percentage				94

3.2. Subdivision of ancient glass

3.2.1 Subclassification of high potassium glass

The selection of appropriate chemical components can highlight the characteristics of high potassium glass and lead barium glass, reduce the difficulty of classification, and further screen their respective sub-categories. Since the obvious differences between the same class of glass are caused by weathering, the chemical composition with a higher rate of change before and after weathering is more representative of the differences between the subclasses. The change rates of each chemical composition before and after weathering of high potassium glass were ranked as shown in Table 3, and several chemical compositions with low change rates were discarded, and other chemical compositions were selected as the basis for classification.According to the relevant data in Table 3, SiO2 and CuO, whose change rates were much lower than those of other chemical components, were discarded, and the remaining twelve chemical components were selected as the basis for classification and a K-mean clustering model was established.

Table 3. Average content (%) and change rate of high potassium before and after weathering

Classification	Na_2O	PbO	BaO	SrO	SnO_2	SO_2	K_2O	Fe_2O_3	CaO	MgO	P_2O_5	Al_2O_3	SiO_2	CuO
High potassium without weathering	0.69500	0.41667	0.59833	0.04167	0.19667	0.10167	9.33083	1.93167	5.33250	1.07916	1.40250	6.62000	67.98417	2.45250
High potassium weathering	0	0	0	0	0	0	0.54333	0.26500	0.87000	0.19667	0.28000	1.93000	93.96333	1.56167
Rate of change	1	1	1	1	1	1	0.94177	0.86281	0.83685	0.81776	0.80035	0.70845	0.38213	0.36323

For the selection of the value of K of classification number, this paper adopts the "elbow method". According to the broken line chart in Figure 4, when K=5, the broken line shows an obvious turning point of decline. Therefore, K=5 is selected to carry out k-means clustering, and the clustering results are shown in the classification spectrum chart in Figure 4.

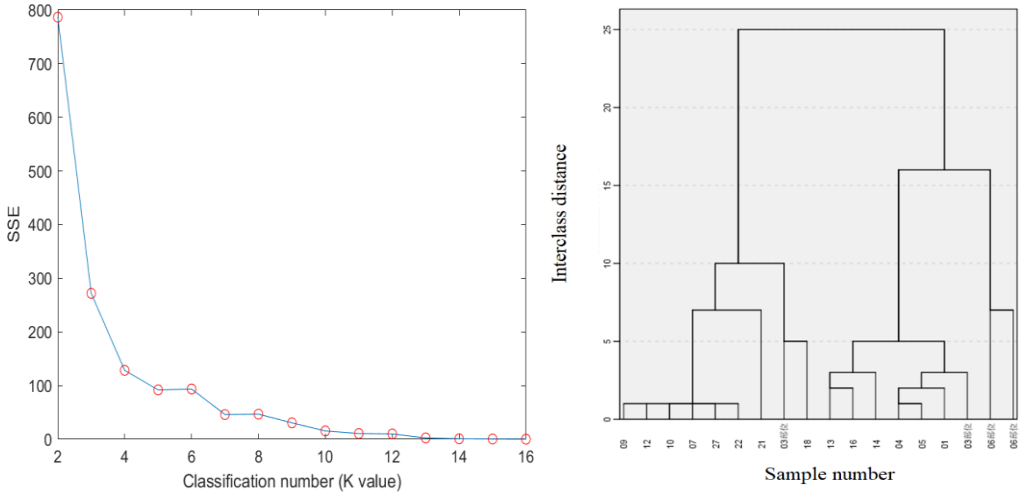


Figure 4. Relationship diagram and classification pedigree diagram of K and SSE of high potassium glass

After K-means clustering, five types of high-potassium glass subclasses were obtained, and the specific information of classification results was shown in Table 4.

Table 4. Results of high potassium subclassification

Subclass	Sample number	Characteristic
S1	07,09,10,12,22,27	It's all high potassium weathered glass.
S2	03(1),18	Calcium, iron, copper and other elements less A ornamented unweathered glass.
S3	01,03(2),04,05,13,14,16	Unweathered glass with a high content of potassium and calcium.
S4	06(1),06(2)	Unweathered glass with high content of magnesium, aluminum and other elements.
S5	21	A special type of potassium-free glass.

3.2.2 Subclassification of lead barium

The change rate of each chemical composition before and after weathering of lead barium glass was ranked as shown in Table 5, and several chemical compositions with lower change rates were discarded, and other chemical compositions were selected as the basis for classification. Based on the relevant data in Table 5, Fe2O3 and MgO, which had much lower change rates than other chemical components, were discarded, and the remaining twelve chemical components were selected as the basis for classification and a K-mean clustering model was established.

Table 5. The average content (%) of Pb and barium before and after weathering and the rate of change

Classification	SO_2	P_2O_5	CaO	PbO	Na_2O	CuO	SrO	SiO_2	SnO_2	K_2O	Al_2O_3	BaO	Fe_2O_3	MgO
Lead barium is not weathered	0.15 9130	1.04 9130	1.32 0435	22.0 8478	1.68 2609	1.43 1739	0.26 8261	54.6 5957	0.04 6522	0.21 8696	4.45 6087	9.00 1739	0.73 6522	0.64 0435
Lead barium weathering	1.36 6154	5.27 3088	2.69 5385	43.3 1385	0.21 6154	2.27 5769	0.41 8462	24.9 1629	0.06 8462	0.13 4622	2.97 0000	11.8 0741	0.58 4615	0.65 0000
Rate of change	7.58 5120	4.03 0173	1.04 1286	0.96 1253	0.87 1536	0.58 9514	0.55 9905	0.54 4221	0.47 1603	0.38 9738	0.33 4936	0.31 1670	0.20 6248	0.01 4936

As for the selection of the value of K of classification number, according to the "elbow method", it can be seen from the broken line diagram in Figure 5 that when K=5, there is an obvious turning point of decline in the broken line. Therefore, K=5 is selected to conduct k-means clustering, and the clustering results are shown in the classification spectrum diagram in Figure 5.

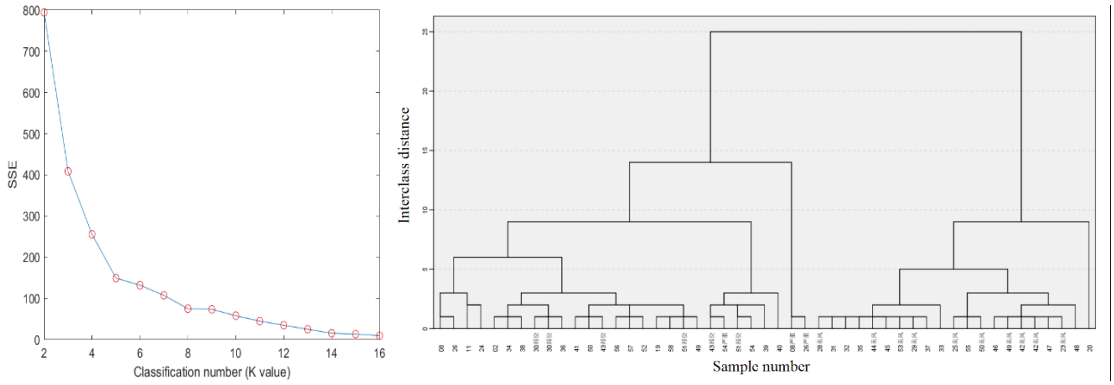


Figure 5. Relationship diagram and classification pedigree diagram between K and SSE of lead-barium glass

After K-means clustering, five types of high-potassium glass subclasses were obtained, and the specific information of classification results was shown in Table 6.

Table 6. Lead barium subclassification results

Subclass	Sample number	Characteristic
S1	39,40,43(2),51(2),54,54*	Weathered lead barium glass with C ornamentation, weathering degree is more serious.
S2	02,08,11,19,24,26,30(1),30(2),34,36,38,41,43(2),49,50,51(2),52,56,57,58	For glass with normal weathering degree and unweathered glass with weathering characteristics.
S3	08*,26*	Most heavily weathered category.
S4	20	A special type of lead barium glass with very low lead content.
S5	23^,25^,28^,29^,31,32,33,35,37,42(1)^,42(2)^,44^,45,46,47,48,49^,50^,53^,55	Unweathered glass or unweathered points, unweathered features are significant.

* stands for severe weathering point

^ stands for unweathered point

3.3. Sensitivity analysis of classification results

In order to test the reliability of classification results, this paper uses random interference items that obey 0-1 normal distribution to interfere with the numerical part of the chemical composition content of some ancient glass, that is, add the generated random numbers to the values of some chemical components in form 2, and then classify the samples of high-potassium glass and lead-barium glass after interference. By comparing the sample classification results before and after interference, the sensitivity was analyzed.

Although the random interference items obeying 0-1 positive distribution are weak interference items, the average content of some chemical components is less than 1%. Therefore, the application of interference items to these chemical components may cause serious deviation in classification and cannot play a normal interference role. Therefore, chemical components with an average content of more than 1% and a high content change rate before and after weathering should be selected to ensure their normal interference effect.

For high-potassium glass, 4 chemical components including K_2O, CaO, Al_2O_3 and Fe_2O_3 are selected for interference based on Table 3 and the above selection criteria. The specific interference values of corresponding chemical components in all samples are shown in Figure 6.

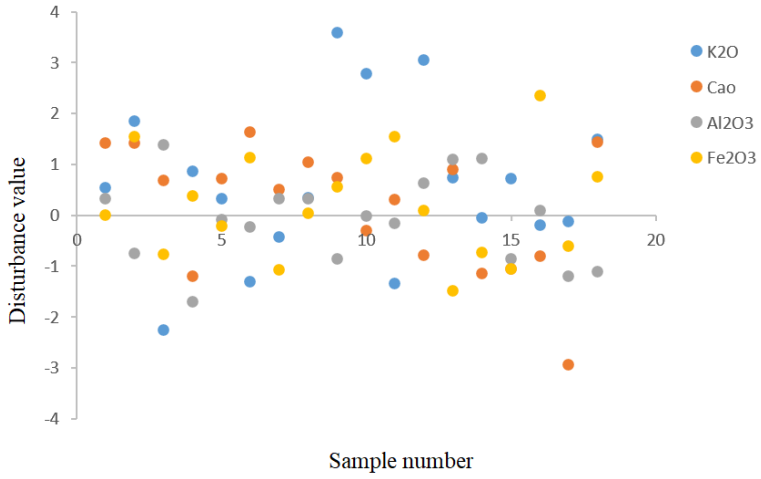


Figure 6. Disturbance value of each sample of high potassium glass

For lead-barium glass, the same treatment is done, and the chemical components of interference are SiO_2, P_2O_5, CaO and PbO. The specific interference values of corresponding chemical components in all samples are shown in Figure 7.

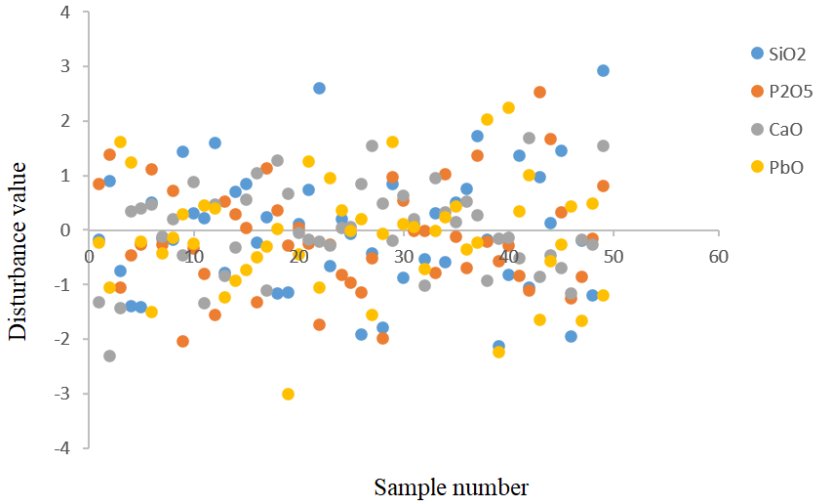


Figure 7. Disturbance values of Pb barium glass samples

Conduct K-means clustering on the samples after interference, and the classification results before and after interference are shown in Table 7:

Table 7. Sensitivity test results

Category	Same classification	Different classification
High potassium	18	0
Lead barium	47	2

As can be seen from Table 7, classification results of high-potassium glass after interference are still the same as those before no interference, so it can be judged that this model has good stability and good classification effect. Although there are two different classifications of lead-barium glass, the sample size of lead-barium glass is large, and a slight error in classification accuracy does not mean that its model is unstable, and the error is within the acceptable range. To sum up, the classification effect of this model is good and stable.

4. Conclusions

In this paper, firstly, the multicollinearity problem of the model is eliminated by combining data screening and "forward method", so as to give full play to the advantages of binary logistic regression, so that we can get a strong interpretation, intuitive and easy to understand, high accuracy of the classification law. Secondly, we deeply analyzed the chemical composition changes of high-potassium glass and lead-barium glass before and after weathering, and selected the appropriate chemical composition as the classification basis. At the same time, the "elbow method" is used to obtain the scientific classification quantity, so that K-means clustering can obtain the subclassification results with distinct characteristics and good effect. Finally, we use the numerical perturbation method to test the sensitivity of the model, and the experiment proves that our model has high stability and accuracy.

References

[1] J. Henderson,J. An,H. Ma. The Archaeometry and Archaeology of Ancient Chinese Glass: a Review[J]. Archaeometry,2018,60(1).

[2] Fuxi Gana,b a Shanghai Institute of Optics and Fine Mechanics, P.O.BOX 800-211, Shanghai, P.R China.b Fudan University, 220 Handan Road, Shanghai,=P.china. Development of Chinese ancient glass and ancient Silk Roads[C]//.2005 International Symposium on Chinese Ancient Glass and Ancient Silk Roads.2005 :329.

[3] Wang Chengyu,Tao Ying,Chen Min,Huang Ming. Weathering of Soda-Lime and Lead Glasses[J]. Transactions of the Indian Ceramic Society,2014,50(6).

[4] Dimitar Dimitrov,Miroslava Nedyalkova,Sergio Madurga,Ludmila Naneva,Vasil Simeonov. Multivariate analysis for the classification of copper–lead and copper–zinc glasses[J]. Open Chemistry,2020,18(1).

[5] Sreenivasa Rao Jammalamadaka. Introduction to Linear Regression Analysis[J]. The American Statistician,2003,57(1).

[6] Kenneth N. Berk. Forward and backward stepping in variable selection[J]. Journal of Statistical Computation and Simulation,1980,10(3-4).

[7] Adeleke Kazeem Adedayo,Dawud Adebayo Agunbiade. PARAMETER ESTIMATION AND DETERMINATION OF SAMPLE SIZE IN LOGISTIC REGRESSION[J]. Journal of Mathematics and Statistics,2013,8(4).

[8] Ikotun Abiodun M.,Ezugwu Absalom E.,Abualigah Laith,Abuhaija Belal,Heming Jia. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data[J]. Information Sciences,2023,622.

[9] Sammouda Rachid,El Zaart Ali. An Optimized Approach for Prostate Image Segmentation Using K-Means Clustering Algorithm with Elbow Method[J]. Computational Intelligence and Neuroscience,2021,2021.

[10] Zhao Naifei,Xu Qingsong,Tang Man-Lai,Jiang Binyan,Chen Ziqi,Wang Hong. High Dimensional Variable Screening under Multicollinearity[J]. Stat,2020,9(1).