Scientific Research Publishing

Applied Mathematics, 2024, 15, 51-64
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Composition Analysis and Identification of Ancient Glass Products Based on L1 Regularization Logistic Regression

Yuqiao Zhou[1*], Xinyang Xu[2], Wenjing Ma[3]

[1]School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China
[2]School of Accounting, Anhui University of Finance and Economics, Bengbu, China
[3]School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, China
Email: *20202779@aufe.edu.cn

How to cite this paper: Zhou, Y.Q., Xu, X.Y. and Ma, W.J. (2024) Composition Analysis and Identification of Ancient Glass Products Based on L1 Regularization Logistic Regression. *Applied Mathematics*, **15**, 51-64.
https://doi.org/10.4236/am.2024.151006

Received: October 24, 2023
Accepted: January 27, 2024
Published: January 30, 2024

Abstract

In view of the composition analysis and identification of ancient glass products, L1 regularization, K-Means cluster analysis, elbow rule and other methods were comprehensively used to build logical regression, cluster analysis, hyper-parameter test and other models, and SPSS, Python and other tools were used to obtain the classification rules of glass products under different fluxes, sub classification under different chemical compositions, hyper-parameter K value test and rationality analysis. Research can provide theoretical support for the protection and restoration of ancient glass relics.

Keywords

Glass Composition, L1 Regularization Logistic Regression Model, K-Means Clustering Analysis, Elbow Rule, Parameter Verification

1. Introduction

Ancient glass is an important object of interdisciplinary research, ancient glass reveals the development level of the productive forces of the society at that time. It is one of the signs of the development level of handicraft industry, and is the physical evidence of human communication. In archaeology, the protection and restoration of ancient glass relics need scientific basis, usually need to classify ancient glass and carry out composition analysis, but the process is more complicated. The main raw material of glass is quartz sand, the melting point of quartz sand is high, and the addition of flux can reduce its melting point. Different fluxes lead to different compositions of glass, forming different glass clas-

sifications. Different types of glass, with the change of time, its degree of weathering is also different. In the process of weathering, the internal elements of ancient glass exchange a lot with the environmental elements, resulting in a change in its composition ratio, which affects the correct judgment of the type of glass [1]. According to the characteristic data of ancient glass products collected, the identification rules of different types of glass are mined and the composition analysis is carried out. Firstly, based on L1 regularized logistic regression model, the characteristic sample data set of chemical component proportions of classified glass relics is trained, and the over-fitting phenomenon is effectively reduced by introducing regular terms. The weight parameters are obtained by gradient descent method, and specific classification rules are obtained. At the same time, the weight parameters were used for feature selection to reduce the dimension of all chemical elements rationally. On the basis of dimensional reduction, K-Means cluster analysis is carried out to divide three sub-classes. Finally, the evaluation index system of L1 regularized logistic regression model is constructed, and the rationality analysis is carried out, and the hyper-parameter K value is tested by elbow rule.

2. Data Sources and Assumptions

The data in this paper are from the C question of the 2022 National Mathematical Contest in Modeling for College Students. Archaeologists divided a group of ancient glass products into AB two types, and measured the proportion of the main chemical components. Specific data are shown in **Table 1**. For reasons such as detection methods, if the sum of the component proportion is not equal

Table 1. Ancient glass composition content (excerpts) (unit: %).

No.	Type	Sampling site	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO	BaO	P_2O_5	SnO_2	SO_2
01	I	01	69.33	-	9.99	6.32	0.87	3.93	1.74	3.87	-	-	1.17	-	0.39
02	II	02	36.28	-	1.05	2.34	1.18	5.73	1.86	0.26	47.43	-	3.57	-	-
03	I	03 part 1	87.05	-	5.19	2.01	-	4.06	-	0.78	0.25	-	0.66	-	-
		03 part 2	61.71	-	12.37	5.87	1.11	5.5	2.16	5.09	1.41	2.86	0.7	-	-
04	I	04	65.88	-	9.67	7.12	1.56	6.44	2.06	2.18	-	-	0.79	-	0.36
05	I	05	61.58	-	10.95	7.35	1.77	7.5	2.62	3.27	-	-	0.94	-	0.47
54	II	54	22.28	-	0.32	3.19	1.28	4.15	-	0.83	55.46	7.04	4.24	-	-
		54 Severe weathering	17.11	-	-	-	1.11	3.65	-	1.34	58.46		14.13	-	-
55	II	55	49.01	2.71	-	1.13	-	1.45	-	0.86	32.92	7.95	0.35	-	-
56	II	56	29.15	-	-	1.21	-	1.85	-	0.79	41.25	15.45	2.54	-	-
57	II	57	25.42	-	-	1.31	-	2.18	-	1.16	45.1	17.3	-	-	-
58	II	58	30.39	-	0.34	3.49	0.79	3.52	0.86	3.13	39.35	7.66	8.99	-	-

to 100%, the data between 85% and 105% will be regarded as valid data [2]. In order to facilitate the research, the following assumptions are put forward: 1) All data sources are true and reliable. 2) The steps of data cleaning and data preprocessing are accurate, that is, outliers can be eliminated. For missing values that are not detected chemical components in the data, this paper uses 0 values to fill in, which has no impact on other data. 3) Some undetected trace data did not affect the results. 4) The known type of glass is judged by archaeologists through glass decoration, color and other conditions.

3. The Judgment of Glass Category Based on L1 Regularization Logistic Regression Method

3.1. Research Thought

This part mainly explores the classification law of the two types of glass and the difference between them, and picks out the main chemical components that play an influential role. A glass classification method based on L1 regularization logistic regression was established. Firstly, regular terms were introduced to prevent the occurrence of overfitting phenomenon [3], and then the weight parameters of each chemical component content were calculated by gradient descent method, and logistic regression evaluation indexes were established. After training the characteristic sample data set based on the proportion of chemical components of classified glass relics, the chemical components of the most important influencing factors were determined.

3.2. Research Method

In order to classify the glass correctly, we use a classification method based on logistic regression.

1) Logistic Regression

In machine learning, logistic regression is a widely used classification method to deal with regression problems where the dependent variable is a categorical variable, that is, the binary classification problem or the multi-classification problem, which belongs to a classification algorithm [4] [5] [6] [7] [8]. The output of logistic regression algorithm $y \in \{0,1\}$ is a discrete value. We can consider fitting conditional probabilities because the values of probabilities are also continuous. Logistic regression has predictive function model sigmoid function [9]

$$y = \frac{1}{1+\mathrm{e}^{-\left(w^{\mathrm{T}}x+b\right)}}.$$

The output is in between $[0,1]$. A base value (such as 0.5) is then selected, and it is treated as 1 when the predicted value exceeds the base value, and 0 when it is not.

It is assumed that y represents the type of sample glass and is a random variable obeying the distribution of 0 - 1, $P(Y=1|x)$ represents the probability that

the glass is a Class I glass, and $P(Y=0|x)$ represents the probability that the glass is a Class II glass.

The mathematical expression of logistic regression model is

$$\text{logit}(p)=\ln\frac{P(Y=1|x)}{1-P(Y=1|x)}.$$

When $0<P<1$, there is $-\infty<\text{logit}(p)<+\infty$. So

$$F(x)=w^{\mathrm{T}}x+b=\ln\frac{P(Y=1|x)}{1-P(Y=1|x)}.$$

We get

$$P(Y=1|x)=\frac{1}{1+\mathrm{e}^{-(w^{\mathrm{T}}x+b)}}.$$

Therefore, if $P>0.5$, the test sample is a Class I glass sample, and otherwise it is a Class II glass sample. $\{x_i, i\in[0,1,2,\cdots,14]\}$ is the set of features, $w=\{w_i\in[0,1,2,\cdots,14]\}$ is the weight parameter.

2) Parameter Estimation

In statistics, parameter estimation often uses maximum likelihood estimation, that is, finding a set of parameters under which our data has the greatest likelihood (probability). The maximum likelihood estimation method is to make the function $L(w)$ reach the maximum parameter value "w", as the estimate of the parameter "w", *i.e.*

$$L(w)=\prod\left[p(x_i)\right]^{y_i}\left[1-p(x_i)\right]^{1-y_i}.$$

To make it easier to solve, we take logarithms of both sides of the equation and write them as logarithmic likelihood functions:

$$\begin{aligned}L(w)&=\sum\left[y_i\ln p(x_i)+(1-y_i)\ln\left(1-p(x_i)\right)\right]\\&=\sum\left[y_i\ln\frac{p(x_i)}{1-p(x_i)}+\ln\left(1-p(x_i)\right)\right]\\&=\sum\left[y_i(w\cdot x_i)-\ln\left(1+\mathrm{e}^{wx_i}\right)\right].\end{aligned}$$

The loss function in machine learning is used to measure the degree to which a model's predictions are wrong. If we take the average logarithmic likelihood loss over the entire data set, we get

$$J(w)=-\frac{1}{N}\ln L(w).$$

The logistic regression loss function is as follow:

$$J(w)=-\frac{1}{n}\left(\sum_{i=1}^{n}\left(y_i\ln p(x_i)+(1-y_i)\ln\left(1-p(x_i)\right)\right)\right).$$

There are many ways to solve logistic regression, and we use the gradient descent method here. Before we do that, we introduce firstly the regular term as follows.

3) The Introduction of Regular Terms

In this paper, the traditional method without regular terms is easy to cause "overfitting" because of the small sample size, which is not conducive to the establishment of glass classification logistic regression model. The addition of regularization terms can effectively reduce overfitting. To solve this problem, L1 regularization method is adopted in this paper, and regularization term is introduced into the log-likelihood function, which is equivalent to adding such a prior knowledge to the model: w, which follows the zero-mean Laplacian distribution. As an optimization problem, binary logistic regression with regularization terms minimizes the following loss functions

$$J(w) = \min_w C\sum_{i=1}^{n}\left(-y_i \log\left(p(x_i)\right) - (1-y_i)\log\left(1-p(x_i)\right)\right) + r(w),$$

where

$$p(x_i) = \frac{1}{1+\mathrm{e}^{-\left(w^{\mathrm{T}}x+b\right)}},$$

and $r(w)$ is an artificial addition to the prior. The method of gradient descent is used to solve the problem, and the w value is obtained.

3.3. Interpretation of Result

The model was built through Python 3 and the results were obtained through the Scikit-learn library. Choose the test set to be 0.3 and the training set to be 0.7. Get $w_1 = 0.06157854$, $w_9 = 0.49660384$, others are 0, Glass category classification formula is as follow

$$F(x) = 0.06157854 * x_1 - 0.49660384 * x_9.$$

As can be seen from the formula, in the chemical composition of all glass, the classification model of Class I glass and Class II glass established, the coefficient of silicon dioxide (SiO_2) is 0.06157854, and the coefficient of lead oxide (PbO) is −0.49660384, that is, the content index of silicon dioxide (SiO_2) is positively correlated when determining the category. The content of lead oxide (PbO) was negatively correlated. In this classification model, the content index of silicon dioxide (SiO_2) and lead oxide (PbO) describe the proportion of chemical components. If there is more silicon dioxide and less lead oxide, it is marked as 1, that is, Class I glass. Otherwise marked as 0, that is, Class II glass.

4. Subclass Classification of Ancient Glass Components Based on K-Means Clustering

4.1. Research Thought

In this section, based on the data in Table 1, the K-means clustering algorithm model was used to select the appropriate chemical composition for sub-classing the glass and to analyze the reasonableness and sensitivity of the classification results. The remaining 67 samples (different sampling points belong to different samples) after eliminating two invalid samples in Table 1, and for the missing values due to non-detected chemical composition, this paper fills in with 0 val-

ues, making a total of 938 valid data.

4.2. Research Method

The basic idea of K-means clustering algorithm [10] [11] [12] [13] is to cluster K points in space as the center point, and then classify the objects closest to them. Iteration method is used to update each cluster center one by one until the best clustering effect is obtained.

According to the glass classification model based on L1 regularization logistic regression [14] [15] [16] established above, Class I glass is most affected by silicon dioxide (SiO_2), and Class II glass is most affected by lead oxide (PbO). Therefore, SiO_2 and PbO are used as two-dimensional number lines for two-dimensional data clustering. In order to ensure the comparability of data, data of different magnitudes are uniformly converted into the same magnitude. In this problem, z-score standardization is carried out on the SiO_2 and PbO values of each sampling point. Convert two or more sets of data into unit-free z-score scores to standardize data.

$$\text{z-score} = \frac{x - \mu}{\delta},$$

where μ is the mean value of the total data, δ is the standard deviation of the population data, x is the observed value of each sample data.

4.3. Interpretation of Result

The standardized SiO_2 and PbO are taken as Y-axis and X-axis respectively to establish a two-dimensional coordinate system. We use cluster analysis method for classification analysis Cluster analysis refers to the analysis process in which the collection of research objects is grouped into multiple classes composed of similar objects. Data points can be roughly divided into three categories in the scatter plot shown in **Figure 1**. Therefore, the number of clusters can be preliminarily set to 3.

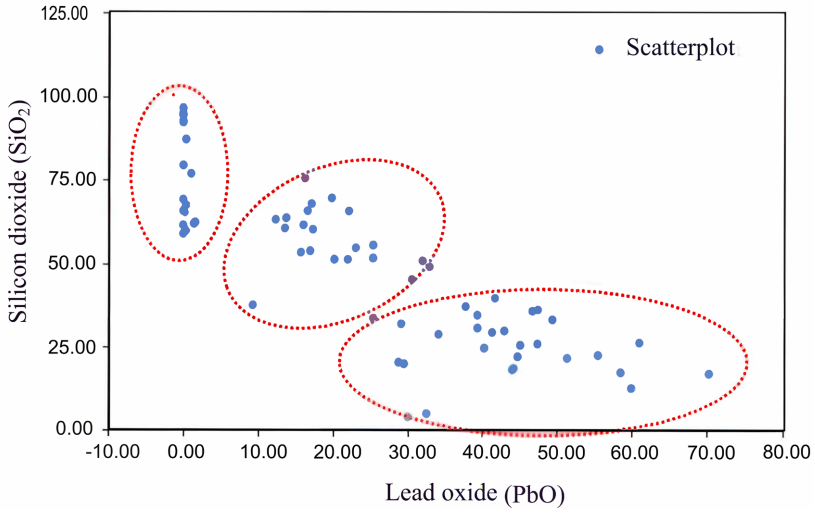

Figure 1. Scatter plot of lead oxide (PbO) and silicon dioxide (SiO_2).

- Variate: {z-score (PbO), z-score (SiO_2)},
- Assume the number of clusters: 3,
- Cluster C = C1, C2, C3.

First, the difference analysis of cluster categories is carried out, and then the frequency of each cluster category is summarized and analyzed to determine whether the amount of various data is balanced. If it is balanced, the process can continue. Then, three clustering centers are preliminarily determined: Z1, Z2, and Z3. After the central coordinate (x_i, y_i) of the cluster is well established, the Minkowski distance between the sample coordinates and each central point can be calculated

$$D_j(k) = \left(\left|x_i - y_i\right|^p\right)^{\frac{1}{p}}, \quad p \geq 1.$$

where k is the secondary sequence number of the iterative calculation. The shortest distance is divided into which category.

In Table 2, for the variable z-score (SiO_2), the P value is 0.000***, showing significance at the significance level. The null hypothesis is rejected, indicating that the variable z-score (SiO_2) has significant differences among the categories divided by cluster analysis. For the variable z-score (PbO), the significance P value is 0.000***, showing significance at the level, rejecting the null hypothesis, indicating that the variable z-score (PbO) has significant differences among the categories divided by cluster analysis. There is a significant difference between the two groups of data on the surface, and the difference can be analyzed by means ± standard deviation.

The frequency of cluster category 1 is 21, accounting for 31.34%. The frequency of cluster category 2 was 27, accounting for 41.79%. The frequency of cluster category 3 is 19, accounting for 26.87%. On the whole, the distribution of the three groups is relatively uniform, indicating that the clustering effect is good. At the same time, the coordinate positions of 67 sampling points in the two-dimensional coordinate system are calculated, and the distance $\alpha_i, \beta_i, \gamma_i$ from the center point of the three types of clustering is calculated. The cluster type corresponding to the minimum distance is the category of the sample. Figure 2 shows the clustering categories.

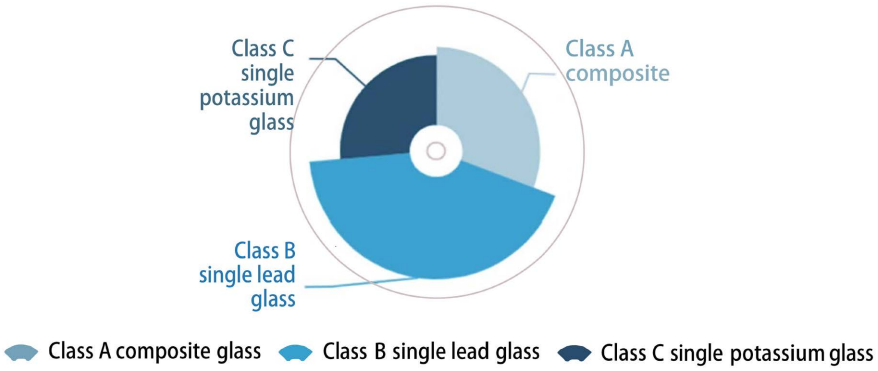

Figure 2. Rose diagram of clustering categories.

Table 2. Clustering category (mean ± standard deviation).

Variable/ categories	category 2 (n = 27)	category 1 (n = 21)	category 3 (n = 19)	F	P
z-score (SiO_2)	−1.012 ± 0.382	0.268 ± 0.396	1.142 ± 0.583	132.335	0.000***
z-score (PbO)	1.015 ± 0.544	−0.214 ± 0.336	−1.206 ± 0.19	170.114	0.000***

Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively.

The sample points after classification are substituted back to Table 1 for testing, and it is found that Class 1 is mainly composed of Class II glass, but the content of SiO_2 and K_2O is very high, and the properties of Class I glass are similar, and the quality is high, and the weathering proportion is low. After consulting the relevant literature, it was identified as Class II glass with similar properties to Class I glass, so it was named "Class A composite glass" (not composite glass in the chemical sense); Category 2 is mainly Class II glass, but the content of each component is less, the quality is inferior, and it is named "Class B single lead glass"; Category 3 is based on Class I glass, the lead content is basically 0, and it is susceptible to weathering, and the quality is inferior, and it is named "Class C single potassium glass". Complete the subdivision of various types of glass. Table 3 shows the coordinates of cluster center points.

5. Model Test

5.1. Evaluation Indexes

The values in the logistic regression evaluation index are formed based on the parameters of the confusion matrix. The confusion matrix (see Table 4) is used to predict the performance of the results. TP indicates that both the test sample data and the predicted result are positive quantities. TN indicates the number of negative test sample data and predicted results. FP indicates the number of negative test sample data and positive prediction results. FN indicates the number of positive test sample data and negative prediction results [17].

1) Accuracy: represents the proportion of the number predicted to be correct in the model to the total:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

2) Precision: represents the proportion of predicted correct results in the model:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FF}}.$$

3) Recall rate: represents the proportion of the number of predictions in the model that are true, and the ratio of the true value to the true data:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Table 3. Cluster center point coordinates.

Cluster type	central value_z-score (SiO_2)	central value_z-score (PbO)
Class A composite glass	0.2681484506373271	−0.21377443548874983
Class B single lead glass	−1.0122076137090077	1.014815740759344
Class C single potassium glass	1.142025689829439	−1.2058295713283438

Table 4. Confusion matrix.

Total population	Condition positives	Condition negative
Test outcome positive	TP	FP
Test outcome negative	FN	TN

5.2. Rationality Analysis

The confusion matrix and the confusion matrix diagram are generated according to the parameters of the experimental results, as shown in **Figure 3**, and the parameters of the experimental results are generated according to the main evaluation parameters. In the glass classification model, there are 21 test data, 15 Class II glasses in the sample data, 0 of which are predicted to be Class I glasses, and 6 Class I glasses, 0 of which are predicted to be Class II glasses.

The data returned by Python shows that the Accuracy, Precision and Recall of the logistic regression classification model are 100%, 100% and 100% respectively. Meet the expected effect.

5.3. Optimal Cluster Number Test

In order to further determine the optimal clustering number—K value, the elbow rule is used in this topic. The vertical axis represents the mean of the distance from all samples to the cluster center under the current K value. If the mean value is regarded as a loss, the K value corresponding to the minimum loss point is the optimal cluster number. In this problem, the mean clustering scatter plot is drawn when K = 1, 2, 3, 4, 5, 6, as shown in **Figure 4**.

In machine learning, hyper-parameters are parameters whose values are set before the learning process begins, rather than parameter data obtained through training. The K value is the hyper-parameter in the cluster analysis. In order to verify the stability of the model, K value is determined by elbow rule and contour system.

1) The elbow rule determines the k value by finding the inflection point where the loss value declines smoothly [18]. First calculate SSE:

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} |p - m_i|^2,$$

where m_i is the center of mass of cluster *i*.

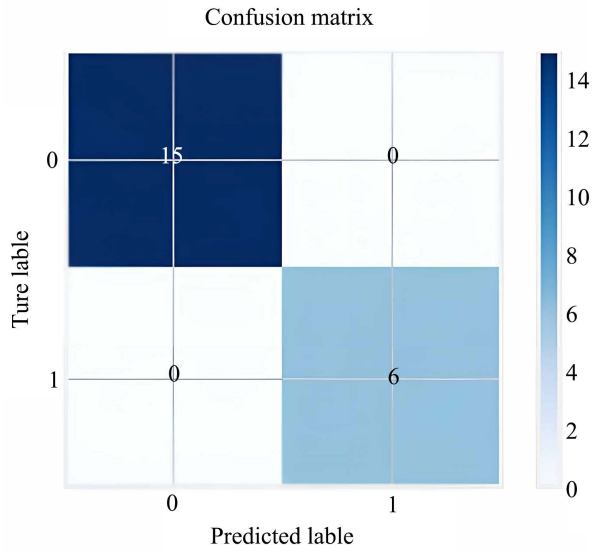

Figure 3. Confusion matrix graph.

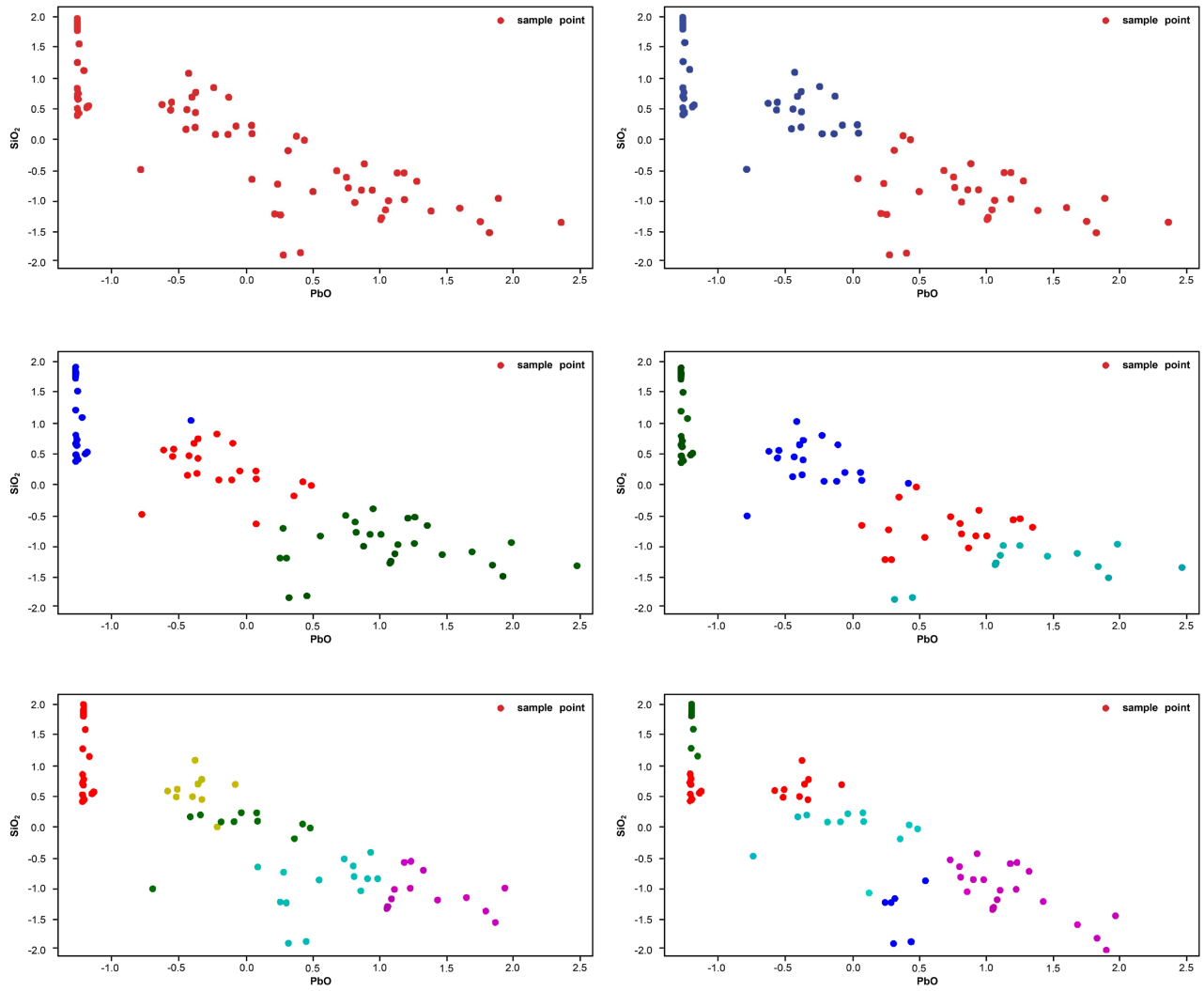

Figure 4. Mean clustering scatter plot.

SSE reflects the sum of squares of the distance between the coordinates of each sample point and the cluster center, which can be seen as the "loss value". The smaller SSE is, the less loss is, and the point is closer to the cluster center, which is more in line with this kind of characteristics, and the classification is more effective.

The elbow diagram takes K value as the X-axis, SSE value as the Y-axis, coordinates of each point are marked, and curves are connected to generate the elbow relationship diagram, as shown in **Figure 5**.

In the process of K value from 1 traversal to 2 and then to 3, SSE value decreased significantly, indicating that the increase in the number of clusters significantly improved the classification effect and the model sensitivity was high. In the process of K value traversing from 3 to 4 and after, the curve gradually flattens out, the return of polymerization degree obtained by K increase will quickly become smaller, and the decline of SSE is also sharply decreasing, which means that the loss progress is reduced. Therefore, from the point where K value is 3, the significance of the next classification is not obvious, and when K value reaches the real cluster number, 3 is the most appropriate K value, and the data should be divided into 3 categories. It is consistent with the scatterplot analysis above.

2) The silhouette coefficient [19] is calculated by finding the maximum value of the silhouette coefficient S_i.

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}.$$

- If $a_i > b_i$, then $S_{(i)} = \frac{b_{(i)} - a_{(i)}}{a_{(i)}} \in [-1, 0)$;
- If $a_i = b_i$, then $S_{(i)} = 0$;
- If $a_i < b_i$, then $S_{(i)} = \frac{b_{(i)} - a_{(i)}}{b_{(i)}} \in [-1, 0)$.

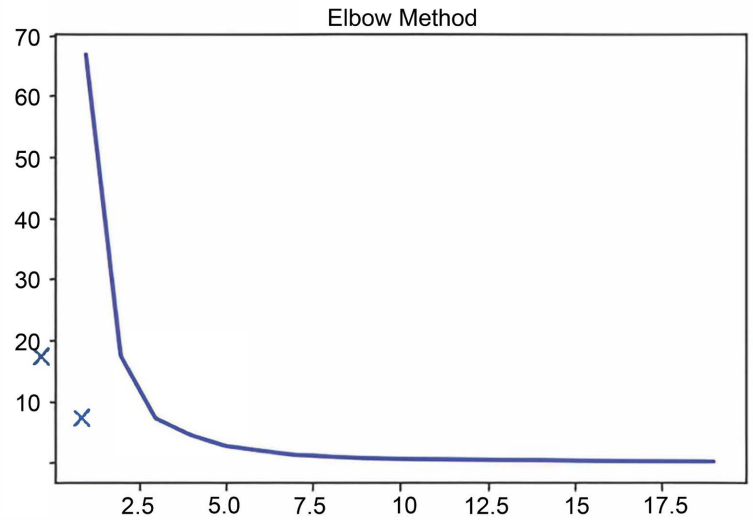

Figure 5. Elbow diagram & silhouette coefficient diagram.

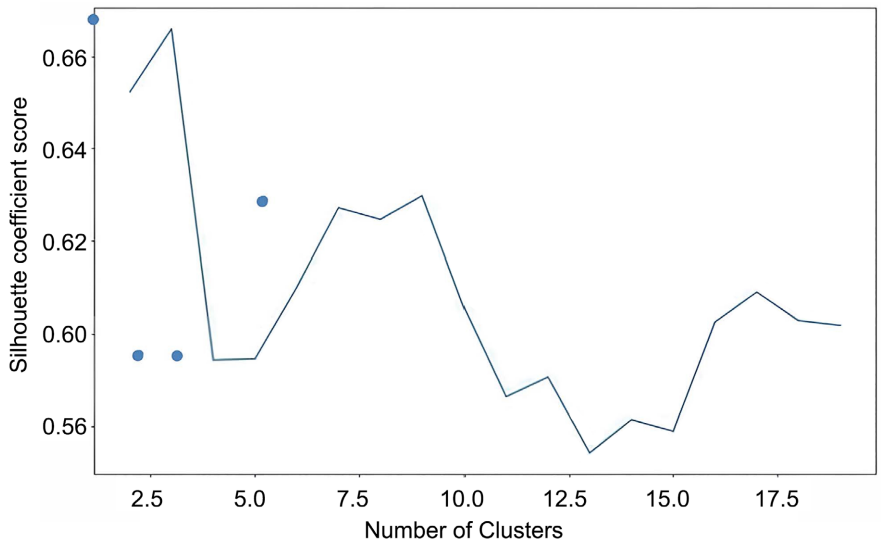

Figure 6. Silhouette coefficient diagram.

Where a_i is the average distance between sample i and other points in the same category, and the average distance between sample i and samples in the nearest different categories. First, the data is standardized, and then the graph is drawn to find the most suitable K value in order to find the most suitable K value. It can be seen from the silhouette coefficient diagram (**Figure 6**) that when K = 3, the silhouette coefficient value is the highest, and the clustering effect is the best at this time, which is consistent with the judgment result of elbow rule.

6. Conclusions

In this paper, a mathematical model based on L1 regularization logistic regression is established to explore the classification rules of two kinds of glass. With the help of regular term, the overfitting phenomenon is effectively reduced. On the basis of dimensionality reduction, K-means cluster analysis is carried out, and three subclasses are divided. Finally, the rationality of the model is verified, and the optimal cluster number is tested. Spss, Python and other mathematical software are applied to the calculation of the model, which makes our calculation results more accurate.

The regularized logistic regression model can effectively classify linear separable problems with fast training speed and strong interpretability. In the K-Means clustering model, the center point and the number of categories are analyzed scientifically by means of elbow rule and contour system. The analysis and identification of ancient glass components will be more conducive to the protection and restoration of glass cultural relics, so as to provide more scientific data for archaeology and better understand the evolution and change of human history.

Author Contribution

Yu-Qiao Zhou: Methodology; Supervision and Leadership; Xin-Yang Xu: Con-

ceptualization; Visualization; Software; Validation; Writing manuscript; Method design; Data analysis; Wen-Jing Ma: Data collation; Visualization; Verification; Investigation; Writing review and editing. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] Li, Y.L. (2021) A Study of Chinese Excavated Glass from the 10th to 12th Centuries. MA Thesis, Liaoning Normal University, Dalain.

[2] Gan, F.X., Zhao, H.X., Li, Q.H., *et al.* (2010) Scientific and Technological Analysis and Research on the Excavated Warring States Glassware in Hubei Province. *Jianghan Archaeology*, **2010**, 108-116.

[3] Feng, M.H., Zhang, T.L., Wang, L.H., *et al.* (2019) Application of Improved Elastic Network Model in Deep Neural Networks. *Computer Applications*, **39**, 2809-2814.

[4] Jiang, Q.Y., Xie, J.X. and Ye, H. (2011) Mathematical Model. Higher Education Press, Beijing.

[5] Zou, Y. (2021) A Study of Several Types of First-Order Approximate Knife-Cut Estimation in Binary Logistic Regression Models. MA Thesis, Guizhou University for Nationalities, Guiyang.

[6] Yin, J.J. (2011) Review and Application of Logistic Regression Model Analysis. MA Thesis, Heilongjiang University, Harbin.

[7] Zhu, J.M., Geng, Y.G., Li, W.B., *et al.* (2022) Fuzzy Decision-Making Analysis of Quantitative Stock Selection in VR Industry Based on Random Forest Model. *Journal of Function Spaces*, **2022**, Article ID 7556229. https://doi.org/10.1155/2022/7556229

[8] Wang, J.J., Liang, Y., Su, J.T., *et al.* (2021) An Analysis of the Economic Impact of US Presidential Elections Based on Principal Component and Logical Regression. *Complexity*, **2021**, Article ID 5593967. https://doi.org/10.1155/2021/5593967

[9] Zhou, W.B., Huang, D.B. and Li, R. (2021) An Improved Fuzzy Hierarchical Clustering Algorithm. *Journal of Beijing Union University*, **35**, 29-34.

[10] Jin, J.B. (2020) A Method for Analyzing Massive Data Based on K-mean Clustering Algorithm. *Journal of Jiujiang College* (*Natural Science Edition*), **35**, 53-55.

[11] Liu, R. (2022) Data Analysis of Educational Evaluation Using K-Means Clustering Method. *Computational Intelligence and Neuroscience*, **2022**, 1-10. https://doi.org/10.1155/2022/3762431

[12] Nie, L., Yang, T., Zhang, J.J., *et al.* (2023) An In-Vehicle Safety Message Relay Selection Method Based on Multi-Attribute Decision Making and K-means Clustering. *Journal of Wuhan University* (*Science Edition*), **69**, 609-616.

[13] Zeng, J.B., Zhang, Y.Y., Zhang, Z., *et al.* (2023) Identification of Power Battery Voltage Inconsistency Faults in Electric Vehicles Based on K-Means++ Clustering with Dynamic K-Values. *Scientia Sinica Technologica*, **53**, 28-40. https://doi.org/10.1360/SST-2022-0194

[14] Niu, Z.H., Chen, B. and Bu, C.Y. (2022) Large Velocity Pulse Prediction and Influencing Factors Analysis Based on L1 Regularized Logistic Regression Model. *Jour-*

nal of Earthquake Engineering, **44**, 306-320.

[15] Liu, Z.X., Zhen, S.J., Qin, B., *et al.* (2012) L1-Regularized Logistic Regression Modeling for Financial Distress Prediction. *Economic Mathematics*, **29**, 106-110.

[16] Zhang, H., Qin, B. and Xu, J.F. (2011) Regularized Logistic Regression Model for Financial Early Warning of Listed Companies. *Journal of East China Jiaotong University*, **28**, 42-47.

[17] Wang, X.Y. (2020) Research on Discovery of Drug-Related Criminals Based on Phone Bill. MA Thesis, People's Public Security University of China, Beijing.

[18] Wang, C.W. (2022) Study on Typical Day Selection Based on Improved K-means Clustering. MA Thesis, Shanghai Institute of Electrical Engineering, Shanghai.

[19] Yin, A.Y., *et al.* (2018) Improved K-Means Algorithm Based on MapReduce Framework. *Application Research of Computers*, **35**, 2295-2298.