Research on Glass Classification and Recognition Based on Fisher's Linear Discriminant and Hierarchical Clustering

Tengyuan Rong *, Yu Liang

School of Mathematics and Statistics, Southwest University, Chongqing, China

* Corresponding Author Email: rty2210158714@163.com

Abstract. Buried in the ground, the glass cultural relics weathered, and a large number of internal elements were exchanged. In order to classify the types of glass, this paper first uses Fisher linear discriminant analysis to obtain a binary classification model of glass types, and finds that the prediction accuracy of the model has reached 90%. Considering the possible overfitting of the model, this paper also uses the principal component analysis method to reduce the dimension of the data, thus obtaining the model and determining that its accuracy rate is maintained above 90%. In order to identify the types of unknown glass cultural relics, the multi-category Fisher's Linear Discriminant method based on principal component analysis was used to identify the subcategories they belonged to.

Keywords: Fisher's Linear Discriminant, Principal components analysis (PCA), Hierarchical Clustering, Glass Artifacts.

1. Introduction

In the production of glass, quartz sand is the main raw material of glass and has a high melting point. Therefore, it is necessary to add flux to lower the melting temperature when refining glass, and add limestone as a stabilizer. The added flux is different, and the main chemical composition of the refined glass is also different. For example, the lead-barium glass with lead ore as a flux in the firing process has a relatively high content of PbO and BaO; while the potassium glass fired with plant ash and other substances with high potassium content as a flux has a high content of potassium. Chemical content is high.

According to the relevant data of a batch of ancient glass, and according to the chemical composition and related detection methods of these cultural relic samples, it is known that this batch of ancient glass can be divided into two types: high-potassium glass and lead-barium glass.

2. Data Sources

The data of this research comes from the attachment of Question E of CUMCM in 2022. Among them, Table 1 shows the relevant physical properties of some unearthed glass cultural relics, such as whether they are weathered or not. In order not to cause confusion, we will abbreviate the 43 part 1 as 43.1, and the 54 severely weathered point as 54.

3. Literature review

Due to the large number of glass cultural relics with high value in all kinds of cultural relics protection units at all levels in our country [1], but there has been a lack of in-depth research on the weathering and corrosion of different types of glass cultural relics for a long time, so the glass cultural relics are analyzed in depth according to the chemical components of glass cultural relics. [2] The classification standard will help the protection and restoration of related glass cultural relics.

Here we only divide glass cultural relics into high-potassium glass and lead-barium glass, which is regarded as a two-classification problem, so we can use Fisher linear discriminant analysis to classify samples to analyze the classification rules of high-potassium glass and lead-barium glass.

Discriminant analysis is a statistical method to classify samples that need to be discriminated by training samples of given categories [3]. A classification method in which the data is projected onto a one-dimensional straight line so that the "distance" of the projected points of the same kind of samples on it is as small as possible and the "distance" of the projected points of the heterogeneous samples is as large as possible [4]. Next, we combine Fisher's Linear Discriminant to study the classification rules of two types of cultural relic glass [5].

4. Classification model establishment and solution

4.1. Research on classification laws

Before formally processing the data, we found that there are 14 chemical components listed in Form 2, which is not conducive to data analysis and classification law research, so we consider the following main chemical components: SiO2, PbO, BaO, K2O, CaO , Al2O3, Fe2O3.

We recorded the j th chemical content of the sample of high potassium cultural relic glass i in data as $q_{ij}^{1,}$ j=1,2,..,7. which is corresponding to SiO2, PbO, BaO, K2O, CaO , Al2O3, Fe2O3.

Similarly,We recorded the j th chemical content of the sample of lead barium cultural relic glass i in data as q_{ij}^{2},j=1,2,...,7.

Therefore, in the original sample space, the class-mean vector is

$$m_t = \frac{1}{N_t}\sum q_i^{\,l} \tag{1}$$

$q_i^{\,l} = (q_{i1}^{\,l}, q_{i2}^{\,l}, ..., q_{i7}^{\,l}, 1)^T$, $l = 1$ corresponding to the high potassium cultural relics, $l = 2$ is corresponding to the lead and barium cultural relics. N_l Indicates the number of samples of the corresponding type of artifacts.

At this time, the intra-class discrete matrix of each class is:

$$S_l = \sum (q_i^{\,l} - m_l)(q_i^{\,l} - m_l)^T, l = 1,2 \tag{2}$$

The total within-class discrete matrix is given as follows:

$$S_\omega = S_1 + S_2 \tag{3}$$

The discrete matrix is:

$$S_b = (m_1 - m_2)(m_1 - m_2)^T \tag{4}$$

According to the Fisher discriminant criterion: $\max J_F(\omega) = \frac{\omega^T S_b \omega}{\omega^T S_\omega \omega}$,The best projection direction is:

$$\omega^* = S_\omega^{\,-1}(m_1 - m_2) \tag{5}$$

Thus, for the samples, there are the following discrimination criteria:

$$\begin{aligned} \omega^*(x^T,1)^T \le 0 &\rightarrow \text{High potassium glass} \\ \omega^*(x^T,1)^T > 0 &\rightarrow \text{Lead barium glass} \end{aligned} \tag{6}$$

Finally, through the spss operation, the results are shown in Table 1 and Table 2:

Table 1. Classification prediction results without weathering samples

Classification results					
No weathering samples		type	Prediction group member information		Total
			Lead barium type	High potassium type	
fact	Count	Lead barium type	23	0	23
		High potassium type	0	14	14
	%	Lead barium type	100.0	.0	100.0
		High potassium type	.0	100.0	100.0
100.0% of the original grouped cases were correctly classified.					

Table 2. Classification and prediction results of weathered samples

Classification results					
weathering samples		type	Prediction group member information		Total
			Lead barium type	High potassium type	
fact	Count	Lead barium type	23	0	23
		High potassium type	0	6	6
	%	Lead barium type	100.0	.0	100.0
		High potassium type	.0	100.0	100.0
100.0% of the original grouped cases were correctly classified					

As can be seen from Table 1 and Table 2, the prediction accuracy of the model is high both for weathered samples and non-weathered samples, which indicates that the model fit is good, the model overfits.

In order to solve the problem of overfitting of the model, the principal component analysis is used to analyze the data.

PCA is a method to convert a set of variables with strong correlation to a set with weak correlation using orthogonal transformations [6]. In addition, by calculating the contribution rate of each newly obtained principal component, the variables can be reduced while maintaining the data relationship as much as possible.

First, we conducted the principal component analysis of the weathered samples, and the calculation found that the cumulative contribution rate of the first four principal components has reached 79.5%. Therefore, we selected the first four principal components for Fisher's Linear Discriminant.

Table 3. Classification and prediction results of weathered samples after PCA

Classification results					
The weathering sample (after PCA)		type	Prediction group member information		Total
			Lead barium type	High potassium type	
fact	Count	Lead barium type	23	0	23
		High potassium type	0	6	6
	%	Lead barium type	100.0	.0	100.0
		High potassium type	.0	100.0	100.0
100.0% of the original grouped cases were correctly classified.					

According to the results in Table 3, the model still has a very high prediction accuracy rate, and the overfitting of the model establishment cannot be ruled out. Therefore, in order to further test the model, the model is cross-verified in this paper.

First, we divided the samples and randomly selected 80% of the samples as training group, leaving 20% as test group. In this extraction, we selected sites 1 and 54 of the sampling sites 07,11,27,48,51 as the prediction group. After testing, only the sampling point 48 made the wrong prediction.

Considering the possible influence of accidental factors in one test, this paper conducted multiple cross-validation and found that only one of the six test groups had the prediction error. All the above show that the model has excellent prediction performance.

Similarly, we also used principal component analysis and dimension reduction, cross-verified the final results, and found that the verification results were good.

In order to further analyze the classification rules of samples, we first find the discrimination function of the Fisher's Linear Discriminant model, and then find the classification function. Among them, the coefficient results of the discrimination function and the classification function of the weathering samples after the principal component analysis are shown in Table 4.

Table 4. The coefficient of the discrimination function and classification function of the weathered samples after PCA

(weathering)discrimination function		(weathering)Classification function		
			Lead barium	High potassium
Principal Component 1	2.283	Principal Component 1	5.171	-17.07
Principal Component 2	0.235	Principal Component 2	0.789	-1.499
Principal Component 3	0.265	Principal Component 3	1.08	-1.502
Principal Component 4	-1.616	Principal Component 4	-3.792	11.958
constant	-0.162	constant	-3.356	-28.607
Unstandardized coefficients		The Fisher's linear discriminant function		

The following division of unknown data. In order to make good use of the above-mentioned classification functions of lead-barium and high-potassium, this paper substitutes unknown data into the classification functions of lead-barium and high-potassium for calculation, and the calculated values are regarded as the possibility of belonging to the types of lead-barium and high-potassium respectively , by comparing the size of the two, take the larger one as the final classification of the unknown data.

Observing Table 4, it can be found that the classification functions of lead-barium and high-potassium are mainly affected by their corresponding principal components 1 and 4. When the principal component 1 increases, the sample is more likely to belong to the lead-barium type; when the principal component 4 increases, the sample is more likely to belong to the high-potassium type.

4.2. Classification of subcategories of glass cultural relics

Observing the data, it is found that the cultural relics are divided into weathered and unweathered states. Considering that glass cultural relics are divided into lead-barium glass and high-potassium glass, we conducted a comparison of the four categories of weathered lead-barium glass, unweathered lead-barium glass, weathered high-potassium glass and unweathered high-potassium glass. Subcategories of cultural relics.

For weathered lead and barium artifacts. Considering that there are too many variables in the data table and the correlation between variables is strong, this paper considers the method of principal component analysis, and thus reduces the dimensionality of the data variables. After conducting principal component analysis on the weathered lead-barium type glass cultural relics, it was found that the contribution rate of the first five principal components reached 85.79% [7], which already contained most of the information of the model. Therefore, we choose the first 5 principal components for clustering

For the research on the analysis rules of glass cultural relics, since it is difficult to obtain the classification results intuitively from the data table, here we use the method of systematic hierarchical clustering to classify the weathered lead-barium glass cultural relics [8].

The merging algorithm of system clustering is a combination of the closest two types of data points by calculating the distance between the two types of data points, and repeating this process until all the data points are combined into one type and a cluster is generated. Clustering method for class pedigree graph [9].

In this paper, spss is used for systematic hierarchical clustering, and the pedigree diagram in Figure 1 is obtained under the average link distance between groups

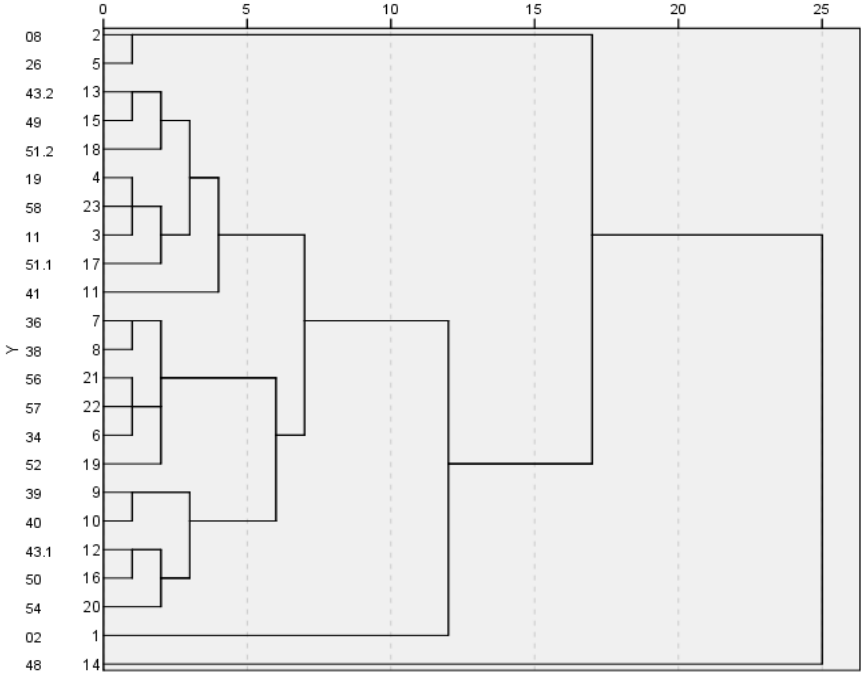

Figure 1. Genealogy of weathered lead-barium glass cultural relics

At the same time, in order to determine the number of clusters, we use the elbow rule to ensure the scientificity and reliability of the final results.

The elbow rule is a method to achieve a reasonable estimate of the number of clusters by calculating the degree of change in the distortion of each class (measured by the aggregation coefficient). Using spss, the aggregation coefficient line graph is shown in Figure 2：

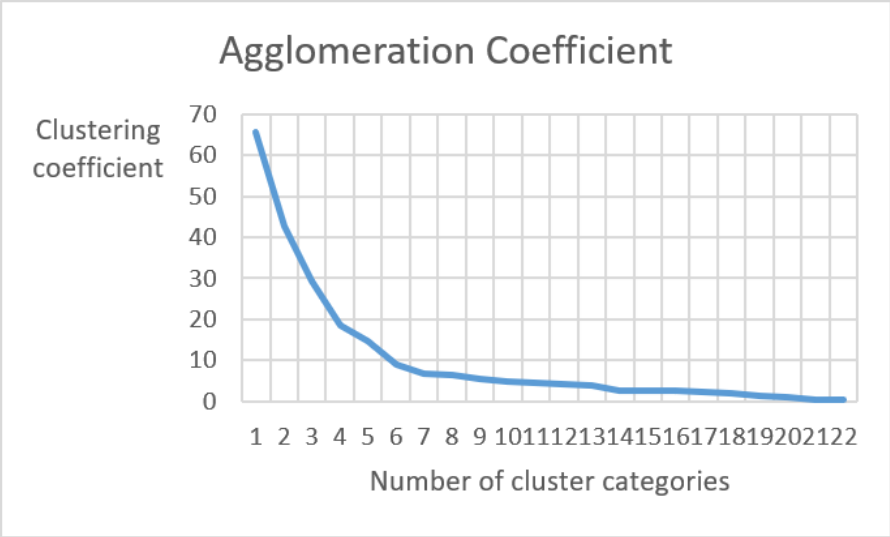

Figure 2. Line chart of aggregation coefficient of weathered lead-barium glass cultural relics

According to Figure 2, we observed that when the number of cluster categories is less than 4, the change trend of the degree of distortion (aggregation coefficient) of the samples is more gentle than that of the samples when the number of cluster categories is greater than 4, that is, the aggregation The "elbow" of the coefficient line chart is 4. Therefore, combined with the elbow rule, this paper believes that it is most reasonable to set the number of clustering categories to 4, and some classification results are shown in Table 5 [10].

Table 5. Classification results of weathered lead-barium glass cultural relics

Sample number	2	8	11	19	26	…	58
Classification results	I	II	I	I	II	…	I

5. Conclusions

In the classification model, this paper first uses Fisher linear discriminant analysis to obtain a binary classification model of glass types, and found that the prediction accuracy of the model has reached 90%. Considering the possible overfitting of the model, this paper also uses the principal component analysis method to reduce the data dimension, thus obtaining the model and determining that its accuracy rate is maintained above 90%, and can obtain excellent results in cross-validation. With the help of the classification model, this paper finds that the classification functions of lead-barium and high-potassium cultural relics are mainly affected by principal components 1 and 4 obtained by principal component analysis.

References

[1] Siqin Bilig, Li Qinghui, Gan Fuxi. Laser Ablation-Inductively Coupled Plasma-Atomic Emission Spectroscopy/Mass Spectrometry Analysis of Ancient Chinese Potassium Glass Components[J]. Analytical Chemistry, 2013,41(09):1328-1333 .

[2] Cao Caixia, Guo Hong. Classification of disease types of stone, ceramics and glass cultural relics collected in museums with fuzzy mathematics [J]. Journal of Beijing Union University (Natural Science Edition), 2009,23(04):58-60.DOI :10.16255/j.cnki.ldxbz.2009.04.024.

[3] Tian Bing.Fisher Discriminant Analysis and Its Application[J].Journal of Weinan Normal University,2014,29(23):8-11+24.DOI:10.15924/j.cnki.1009-5128.2014.23.002.

[4] Ding Xueli, Qi Changsheng, Fang Li. Classification and identification of traditional Chinese medicinal materials based on Fisher discriminant analysis[J]. Journal of Chifeng University (Natural Science Edition), 2021,37(11):19-22.DOI:10.13398/j.cnki.issn1673 -260x.2021.11.006.

[5] Huang Liwen.Fisher stepwise discriminant analysis method based on variable selection[J].System Science and Mathematics,2021,41(08):2338-2348.

[6] Fang Hanguo.A Study on Macroeconomic Prosperity Index Based on Principal Component Analysis[J].Contemporary Economy,2022,39(01):26-31.

[7] Shen Guifang. Research on Personal Credit Risk Assessment Based on PCA and SVM [J]. Journal of Yichun University, 2022,44(10):49-52+120.

[8] Li Luyao. Short-term prediction system for college students' employment based on hierarchical clustering[J]. Journal of Jilin University (Information Science Edition), 2022,40(01):64-70.DOI:10.19292/j.cnki.jdxxp.2022.01.007 .

[9] Jiang Haifeng. Application of Hierarchical Clustering Algorithm in Higher Vocational Teaching Recommendation System [J]. Science and Education Journal, 2021(29):36-38.DOI:10.16400/j.cnki.kjdk.2021.29.012.

[10] Yang Zhenzhen, Li Hongyao, Yang Keyi, Liu Lin. Cluster Analysis of Consumption Level of Residents in Provinces and Cities Based on System Clustering Algorithm and Elbow Criterion [J]. China Science and Technology Information, 2021(12):121-122.