ISSN 2616-5775 Vol. 6, Issue 7: 182-187, DOI: 10.25236/AJCIS.2023.060724

Classification Model of Ancient Glass Artifacts Using Fisher Discriminant Analysis

Kaikai Kang[1,#], Weijie Wu[1,#,*], Xianchun He[1]

[1]College of Computer and Data Science/College of Software, Fuzhou University, Fuzhou, 350108, China
[*]Corresponding author: vjjjjjj1@163.com
[#]These authors contributed equally.

Abstract: *Ancient glass artifacts are highly susceptible to weathering in burial environments, which can cause changes in their chemical composition. To address the problem of identifying and analysing the composition of ancient glass artifacts, this paper proposes a classification and identification model based on the Fisher discriminant method. Firstly, considering that the proportion of components in glass can change during the weathering process, which can affect the correct identification of its category, this paper predicts the data before weathering based on the normal distribution law and performs closed operations to ensure that the sum of the component data is 100%. Secondly, through variance analysis and kernel density estimation, the factors affecting the classification are selected, and a Fisher linear discriminant model is established. Finally, cross-validation is performed by applying Fisher linear discriminant to both the training and testing data, and the discriminant results of the two groups of data are found to be consistent with the actual situation, with a discrimination accuracy rate of 100%.*

Keywords: *Identification of Ancient Glass Artifacts, Fisher Linear Discriminant, Kernel Density Estimation*

1. Introduction

Glass was introduced into China through the Silk Road [1]. After absorbing its technology, China used local materials to produce glass with Chinese characteristics. The appearance of this type of glass is similar to that of foreign glass products, but its chemical composition is different. The main raw material for glass is quartz sand, and the main chemical component is silicon dioxide (SiO_2). The main component of glass varies depending on the different flux added during the firing process. For example, lead-barium glass, which is generally considered to be a glass variety invented in China, uses lead ore as a flux and has a higher content of lead oxide (PbO) and barium oxide (BaO). Potassium glass is made by firing high-potassium substances such as plant ash as flux, and is mainly popular in southern China, Southeast Asia, India, and other regions.

The aim of this study is to establish a classification and identification model for ancient glass artifacts based on Fisher's discriminant analysis. By analyzing the compositional characteristics of ancient glass artifacts, a scientific method for identifying glass artifacts has been developed to support cultural heritage preservation and archaeological research. The significance of this study lies in providing researchers with a fast, accurate, and scientific method for identifying glass artifacts, which can facilitate the deepening of cultural heritage preservation and archaeological research. Furthermore, this model can provide technical support for quality control and brand protection of ceramic products, with broad application prospects. The innovation of this study lies in considering the influence of the proportion change of ancient glass artifacts during weathering on the identification results, using a legitimate state allocation rule to calculate the data before weathering and performing a closure operation to ensure that the sum of the compositional data is 100%, which improves the accuracy of the identification model. Secondly, by selecting influential factors through variance analysis and core degree calculation to establish a vertical Fisher linear discriminant model, the influence of different factors on the identification results is fully considered, enhancing the model's robustness and reliability. These innovative aspects provide new ideas and methods for the classification and identification of ancient glass artifacts, promoting further development in this field.

The relationship between the component analysis and identification of the glass artifacts studied in this paper is shown in Figure 1:

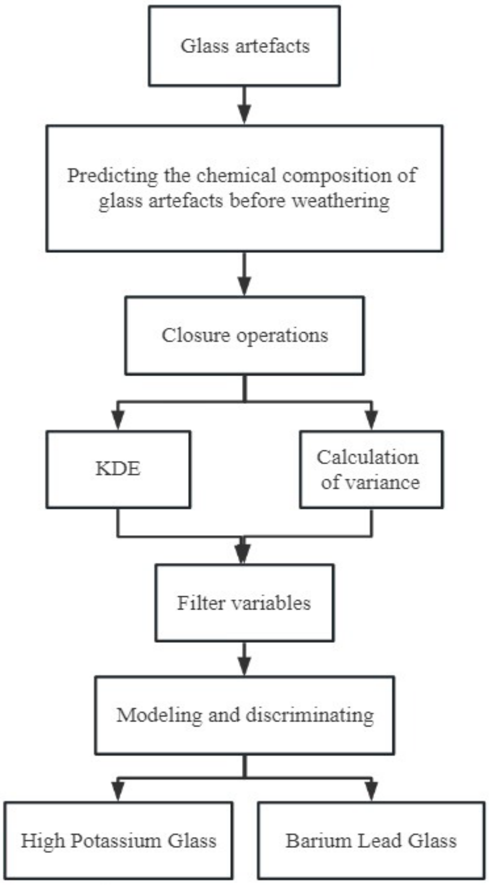

Figure 1: The relationship between component analysis and identification of glass artifacts.

2. Model Establishment

2.1 Predicting the Chemical Composition of Glass Artifacts before Weathering

Ancient glass artifacts are highly susceptible to weathering in burial environments. During the weathering process, there is a significant exchange of internal and environmental elements, which can cause changes in the proportion of its components and affect the correct identification of its category.

Let X be the chemical composition of weathered glass artifacts and Y be the chemical composition of non-weathered glass artifacts. Assume that X and Y both follow a normal distribution, i.e., $X\text{~}N(\mu_x,\sigma_x)$, $Y\text{~}N(\mu_Y,\sigma_Y^2)$

$$Y=\mu_Y+\frac{\sigma_Y}{\sigma_X}(X-\mu_X) \tag{1}$$

Since some samples have been weathered, the above method is used to predict the data before weathering based on the weathered data, which is then used to replace the sample data.

2.2 Closure Operation

To ensure that the sum of component percentages adds up to 100%, relative information can be used to further process the data. Relative information refers to the fact that the component data only reflects information in the ratios between the components, and the absolute values of each component are irrelevant. By multiplying each component of the component data by the same positive constant, the ratios between the components can be maintained. This allows the component data to be considered as an equivalence class, where each component contains the same information and can be represented as a proportionate vector using an appropriate scaling factor. This equivalence class can then be subjected to closure operation, resulting in a closed system of equations that accurately represents the chemical composition of the glass artifacts.

$$C(x)=C(x_1,x_2,\cdots,x_D)^T=\left(\frac{k\cdot x_i}{\sum_{i=1}^{D}x_i},\frac{k\cdot x_2}{\sum_{i=1}^{D}x_i},\cdots,\frac{k\cdot x_D}{\sum_{i=1}^{D}x_i}\right)^T \tag{2}$$

The closing operation involves multiplying the initial vector by an appropriate scaling factor so that the sum of the components after the closing operation is a constant k (which is 100 in this case), for any two vectors $x,y\in R_+^D$, if $C(x)=C(y)$, then x and y are component-equivalent. Therefore, the final data is obtained by multiplying the chemical composition of each glass artifact by the corresponding calculated factor.

2.3 Variance Processing

The type of the sample was taken as the dependent variable, while various chemical components and surface weathering were taken as independent variables. However, due to the small sample size and the large number of independent variables, as well as the phenomenon of overfitting observed in the preliminary linear discriminant analysis, it was necessary to perform preliminary screening of the linear discriminant analysis variables. This study employed the method of variance processing to eliminate variables with small variances and identify factors that have a significant impact on the type of glass as independent variables for the model through data analysis.

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n} \tag{3}$$

2.4 Kernel Density Estimation (KDE)

Kernel Density Estimation [2] (KDE) is a non-parametric estimation method used to describe the dynamic distribution of data. It emphasises the use of kernel density curves to capture the distribution characteristics of data, and can avoid the subjectivity in parameter estimation, thereby improving the accuracy of the estimation results. Assuming that the density function of a random variable W is f(w), the probability density formula for point w is as follows:

$$f(w)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{W_i-\overline{w}}{h}\right) \tag{4}$$

$K(\cdot)$[3]represents the kernel function, n represents the number of observed values, h represents the bandwidth; W_i represents an independent identically distributed random variable, and $\overline{w}$ represents the mean.

Kernel density analysis is an estimation of the probability of an event occurring under certain environmental conditions. It is a method for estimating the probability density function of a population from a sample. In kernel density analysis, each event point within a search radius is weighted differently, with points closer to the "kernel" having a higher weight and points farther away from the "kernel" having a lower weight.

2.5 Discriminant Analysis

FISHER'S linear discriminant analysis [4] (FLDA) is a standard technique for dimension reduction in pattern recognition. It projects the original high-dimensional data onto a low-dimensional space, where all the classes are well separated by maximizing the Raleigh quotient, i.e., the ratio of between-class scatter matrix to within-class scatter matrix. Assume there are n training sample vectors given by $\{r_i\}_{i=1}^{n}$ for p classes: C_1, $C_2,\dots,C_P$, and there are n_j samples for the jth class, i.e., $\sum_{j=1}^{p}n_j=n$. Let μ be the mean of the entire training samples, i.e., $(\frac{1}{n})\sum_{i=1}^{n}r_i=\mu$, and μ_j be the mean of the jth class. i.e., $(\frac{1}{n_j})\sum_{r_i \in C_j} r_i = \mu_j$. Then, the within-class scatter matrix $\boldsymbol{S}_w$ and the between-class scatter matrix $\boldsymbol{S}_B$ are defined. respectively, as

$$\mathbf{S}_W = \sum_{\mathbf{r}_i \in C_j} (\mathbf{r}_i - \mu_j)(\mathbf{r}_i - \mu_j)^T \tag{5}$$

$$\mathbf{S}_B = \sum_{j=1}^{p} n_j (\mu_j - \mu)(\mu_j - \mu)^T \tag{6}$$

The goal is to find a transform vector w such that the Raleigh quotient is maximized, which is defined as

$$q = \frac{w^T S_B w}{w^T S_W w} \tag{7}$$

$\mathbf{w}$ can be determined by solving a generalized eigenproblem specified by $\boldsymbol{S}_B\boldsymbol{w} = \lambda\boldsymbol{S}_W\boldsymbol{w}$, where λ is a generalized eigenvalue. Since the rank of $\boldsymbol{S}_B$ is $p-1$ there are $p-1$ eigenvectors associated with $p-1$ nonzero eigenvalues. Therefore, an $L \times (p-1)$ matrix $\mathbf{W}$ can be found to transform the original L-dimensional data into a $(p-1)$-dimension space. In this low-dimensional space, it is expected that the p classes can be well separated.

2.6 Cross-validation

The data is divided into a training set and a testing set, with a ratio of 8:2. A Fisher linear discriminant equation is then established based only on the training set, and the model is tested using the testing set to evaluate its robustness. This process is repeated several times to eliminate the effect of randomness, and the average accuracy of the model is calculated to achieve cross-validation [5].

3. Model Application

This study included a total of 69 sampling points from 58 cultural relics. The classification information for each cultural relic and the proportion of the major chemical components for each sampling point are known, as shown in Table 1. The data is sourced from http://www.mcm.edu.cn/html_cn/node/5ed58964e849cd9deb3917bc257b2654.html. The objective is to discriminate unknown glass types based on the available data.

3.1 Data Preprocessing

Initially, for the known and unknown weathered detection points, the chemical composition before weathering was predicted for a total of 43 samples by computing their variance and mean.

Subsequently, a closure operation was applied to all transformed data, ensuring that the sum of the proportion of each chemical component in each sample equaled 100%.

Table 1: Training Sample

Sampling Point	SiO_2	K_2O	CaO	…	PbO	BaO	P_2O_5
01	69.33	9.99	6.32		0	0	1.17
02	36.28	1.05	2.34		47.43	0	3.57
03 location 1	87.05	5.19	2.01		0.25	0	0.66
03 location 2	61.71	12.37	5.87		1.41	2.86	0.7
04	65.88	9.67	7.12		0	0	0.79
…							
54	17.11	0	0		58.46		14.13
55	49.01	0	1.13		32.92	7.95	0.35
56	29.15	0	1.21		41.25	15.45	2.54
57	25.42	0	1.31		45.1	17.3	
58	30.39	0.34	3.49		39.35	7.66	8.99

3.2 Variable Selection for Linear Discriminant Analysis

By computing the variance of each chemical index in the sample dataset, the differences between each variable and the overall mean were evaluated. It was found that the variances of K_2O, SiO_2, PbO, and BaO were significantly higher than those of the other chemical components. Using kernel density analysis, it was determined that only two chemical indices, silicon dioxide and lead oxide, were needed to accurately identify the type of glass artifact. Therefore, K_2O, SiO_2, PbO, and BaO were chosen as independent variables, while the type of sample was chosen as the dependent variable for linear discriminant analysis.

3.3 Discriminant Analysis

Fisher linear discriminant analysis was used for classification, and the typical discriminant function coefficients are shown in Table 2 and Table 3, and the discriminant function is as follows.

$$Y_0 = 1.693\mathrm{SiO_2} + 2.062\mathrm{PbO} + 3.808\mathrm{K_2O} + 2.283\mathrm{BaO} - 79.558 \tag{8}$$

$$Y_1 = 1.793\mathrm{SiO_2} + 1.916\mathrm{PbO} + 4.853\mathrm{K_2O} + 2.158\mathrm{BaO} - 85.564 \tag{9}$$

Y_0 represents the probability of the discriminant result being barium lead glass, while Y_1 represents the probability of the discriminant result being high potassium glass.

Table 2: Coefficients of Typical Discriminant Functions

	Function
SiO_2	0.025
PbO	-0.036
K_2O	0.261
BaO	-0.031
const	-0.569
unstandardized coefficients	

Table 3: Coefficients of Classification Functions

Variable	Type	
	0	1
SiO_2	1.693	1.793
PbO	2.062	1.916
K_2O	3.808	4.853
BaO	2.283	2.158
const	-79.558	-85.564

3.4 Discriminant Result

The model achieved 100% accuracy in predicting the training set data and also 100% accuracy in predicting the testing set data. Additionally, it can be observed that most of the high-potassium glass and lead-barium glass samples are clustered around their respective class centroids, indicating a good classification performance.

The significant effect of the model achieving 100% accuracy in prediction can be attributed to the following reasons:

1) Kernel density estimation revealed that only two chemical elements, silicon dioxide and lead oxide, were needed to accurately identify the type of glass artifact.

2) Silicon dioxide had a high proportion (far exceeding other chemical components) and the largest variance, thus having a significant impact on the model.

3) There were significant differences in the chemical composition of lead oxide, potassium oxide, and barium oxide between the two types of glass artifacts.

4. Conclusion

Using Fisher discriminant function and considering K_2O, SiO_2, PbO, and BaO as the four indicators, a Fisher linear discriminant model was established. The model achieved 100% accurate discrimination for known types of ancient glass artifacts, indicating high reliability. When the model was applied to select unknown types of artifacts, the predictive accuracy was also 100%. The main chemical composition variables used in the model were consistent with the different flux elements required for refining various types of glass in practice, indicating that the model is reasonable and feasible for artifact discrimination. The method is simple, efficient, and has high classification accuracy, demonstrating strong applicability. However, the model still has some limitations. Firstly, its establishment is based on ancient glass artifacts with known types, which may have limitations in classifying and identifying artifacts with unknown glass types. Secondly, the influence of other factors, such as manufacturing techniques, age of use, etc., has not been thoroughly studied yet.

References

[1] Yin Yulong. Analysis of ancient glass composition through correlation prediction [J]. Contemporary Chemical Research, 2023, (126, 1): 122-126.

[2] Jia Dawei, Wu Ziyan, He Xiang. Multivariate correlated kernel density estimation method for probabilistic seismic demand analysis under multidimensional performance limit states [J]. Journal of Vibration Engineering, 2022, 35(6): 1299-1310.

[3] Mao Shanshan, Ma Shengguo, Wei Benhai, et al. Delay estimation of power line communication network based on kernel density [J]. Electronic Design Engineering, 2023, 31(1): 133-137+142.

[4] Gui Lin, Yang Jianbo, Huang Yuanshuai. Value of logistic regression and Fisher linear discriminant analysis models in differential diagnosis of ovarian tumors [J]. Chongqing Medicine, 2018, 47(6): 800-802.

[5] Madaki U Y, Singh V V, Chiwa M D. Assessment of Diabetic Patients Among Adults in Maiduguri, Using Multivariate Discriminant Model[J]. Biostatistics and Biometrics Open Access Journal, 2019, 9.