A glass classification model for glass artifacts based on PCA and SVM

Jiayi Nie *, Ke Xu, Yifan Chen

Sichuan University Pittsburgh Institute, Sichuan University, Chengdu, China

* Corresponding Author Email: niejiayi@stu.scu.edu.cn

Abstract. The Silk Road was an important channel for trade and commerce between East and West in ancient times, and an important link for cultural exchange between East and West, playing an important role in the development of ancient civilization and history. Glass was one of the important goods carried on the Silk Road in ancient times. In this paper, focus on the problem of component analysis and type identification of ancient glass products. This paper extracted the important factors of the data by PCA principal component analysis, reduced the 14-dimensional data to two dimensions, then built a data point distribution map by visualization, and established a linear classification model based on SVM support vector machine. After that, the model is used to identify the type of samples based on the underlying data. The sensitivity indices of each factor were solved by using Sobol method for the sensitivity analysis of the results.

Keywords: PCA principal component analysis; SVM support vector machine; Sobol sensitivity analysis.

1. Introduction

Glass, whose main raw material is quartz sand, the main chemical composition of silicon dioxide (SiO2). As one of the important goods carried on the Silk Road in ancient times, it was introduced to our country from the West Asian and Egyptian regions, and then modified by our ancient people in situ to have a different chemical composition [1]. In fact, the differences in chemical composition depend to a large extent on the fluxes added during refining to lower the melting temperature. Among the various compositions of glass, lead-barium glass as well as potassium glass were widely popular in China in ancient times [2].

Most of the excavated ancient glass has been weathered by the burial environment and the ratio of internal elements has been changed [3]. In the research about ancient glass products, the change of composition ratio has become a constraint factor affecting the judgment of glass categories [4]. Therefore, it is important to correctly analyze the composition of ancient glass objects and identify them by category [5].

There is a batch of data related to ancient glass products in China, and archaeologists have analyzed the chemical composition of the samples and classified them into two types: high potassium glass and lead-barium glass. In addition, archaeologists have provided basic data based on the information obtained from the tests, recording the basic information of glass artifacts, the percentage of chemical composition of classified glass artifacts and the percentage of chemical composition of unclassified glass artifacts, respectively. This paper will analyze the chemical composition of the unknown category of glass artifacts and perform type identification, based on which the sensitivity of the analysis results will be analyzed.

2. Model assumptions and notation

2.1. Assumptions [6]

1. Assume that each variable in the sample conforms to a normal distribution
2. Assume that all data measurements are true and valid, without considering the effects of human error
3. It is assumed that the small amount of data is also representative

4. the significance level in this paper is 0.05.

2.2. Notations

Important notations used in this paper are listed in Table 1.

Table 1. Notations

Symbols	Meaning
x_{i1}	The decoration of the i-th sample
x_{i2}	Type of sample i
x_{i3}	Color of sample i
x_{i4}	Surface weathering of sample i
$y_{ij}\ j = 1,2,\ldots,14$	Chemical composition of sample i
w_{ij}	Weighting weights of sample i
m	Number of clusters in the clustering model

3. Model construction and solving

3.1. Data dimensionality reduction based on PCA principal component analysis

3.1.1 Principle of PCA Principal Component Analysis.

The PCA principal component analysis method is an unsupervised feature dimensionality reduction method. It derives the covariance matrix from the original data matrix and transforms the high-dimensional data into low-dimensional data by rounding each eigenvalue of this matrix as well as the eigenvectors, while, at the same time, retaining most of the features of the original data[7] .

3.1.2 Specific Implementation of PCA Principal Component Analysis

Based on the principle of PCA principal component analysis method, the original data matrix was constructed for 67 sets of underlying data.

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kn} \end{bmatrix} \tag{1}$$

A total of 67 sets of sample data were available for the study, and each set of sample data contained 14 sets of data variables in addition to the artifact sampling points. Therefore, in equation (20), n = 67, k = 14, and the original matrix Y is a 14 × 67 matrix.

Based on the matrix Y, the covariance matrix C is obtained by solving based on the covariance matrix formula.

$$C = \frac{1}{n} YY^T \tag{2}$$

In this covariance matrix, the element y_{ii} on the diagonal is the variance of the variable y_i itself, while the non-diagonal element y_{ij}

is the covariance between variable y_i and variable y_j.

In this covariance matrix C, there are k eigenvalues λ_i and the corresponding eigenvectors e_i such that.

$$Ce_i = \lambda_i e_i, i = 1,2,\ldots,k \tag{3}$$

Some of the eigenvalues of this covariance matrix and their explained proportions are shown in Table 2 below.

Table 2. Table of partial eigenvalues and their interpretation ratio

Eigenvalue	Initial eigenvalue variance percentage	Cumulative percentage
3335.09	79.64	79.64
732.12	17.48	97.13
69.87	1.67	98.79
21.67	0.52	99.31
9.78	0.23	99.54

The largest eigenvalue is 3335.09, which can explain 79.64% of the data information; the first two eigenvalues can explain up to 97.13% of the data information.

Considering that the two-dimensional images are more figurative, the two-dimensional data are easy to calculate and still retain most of the information in the original data, the two largest eigenvalues, e_m and e_n, are selected and constructed into a base matrix P of 214.

$$P = [e_m : e_n]^T \tag{4}$$

With this basis matrix P , the dimensionality reduction transformation from the original 14 × 67 data matrix Y to the 2 × 67 data matrix Y′ can be achieved, i.e.

$$Y' = PY \tag{5}$$

According to the above steps for the dimensionality reduction process, the new two-dimensional data is obtained as shown in Figure 1 below.

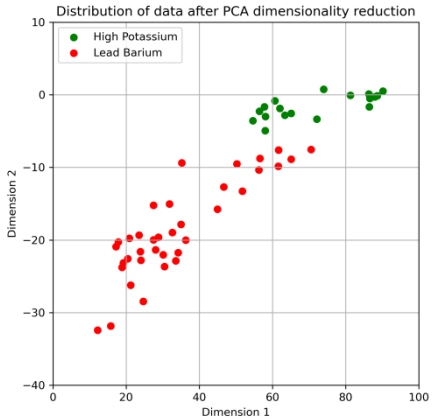


Figure 1. Distribution of data after PCA dimensionality reduction

3.2. Type identification based on SVM support vector machine

SVM support vector machine is a young machine learning algorithm that defines a linear classifier that classifies all data points in two classes on that feature space with maximum interval in n-dimensional data points.SVM performs well the principle of structural risk minimization, has greater advantages in areas such as small sample decision making and classification, and can be generalized to other applications of machine learning problems [8].

According to the SVM support vector machine algorithm [9], there exists a hyperplane in the Rn space determined by its normal vector W and intercept b with the equation.

$$Y^T W + b = 0 \tag{6}$$

The distance ρ between two hyperplanes can be obtained by generalizing the formula for parallel linear distance as

$$\rho = \frac{2}{\| W \|} \tag{7}$$

The objective of the SVM support vector machine is to find two hyperplanes in this feature space so that the distance ρ between them is maximized, and the optimization equation is established according to this objective

$$\max_{W,b} \rho \Leftrightarrow \max_{W,b} \rho^2 \Leftrightarrow \min_{W,b} \frac{1}{2}||W||^2 \tag{8}$$

$$s.t.\, k_i({y_i}'^T W + b) \geq 1, i = 1,2, \ldots, n. \tag{9}$$

This objective equation also has the constraint

$$\begin{cases} {y_i}'^T W + b \geq +1, k_i = +1 \\ {y_i}'^T W + b \leq -1, k_i = -1 \end{cases} \tag{10}$$

3.2.1 Constructing the Lagrangian function

To further solve this equation, construct the Lagrangian equation as

$$L(W,b,\alpha) = \frac{1}{2}\| W\|^2 - \sum_{i=1} \alpha_i \left[k_i \left(y_i'^T W + b \right) - 1 \right] \tag{11}$$

Further solution.

$$\nabla_W L(W,b,\alpha) = W - \sum_{i=1}^n \alpha_i k_i Y_i = 0 \Rightarrow W = \sum_{i=1}^n \alpha_i k_i Y_i \tag{12}$$

$$\nabla_b L(W,b,\alpha) = -\sum_{i=1}^n \alpha_i k_i = 0 \Rightarrow \sum_{i=1}^n = 0 \tag{13}$$

$$\underset{W,b}{L(W,b,\alpha)} = \frac{1}{2}||W||^2 - \sum_{i=1}^n \alpha_i [k_i({y_i}'^T W + b) - 1] = -\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k_i k_j {Y_i}'^T {Y_j}'^T + \sum_{i=1}^n \alpha_i \tag{14}$$

On this basis, proceed to solve the optimal solution of the objective equation.

$$\widehat{W} = \sum_{i=1}^n \widehat{\alpha_i} k_i Y_i \tag{15}$$

$$\hat{b} = k_j - \sum_{i=1}^n \widehat{\alpha_i} k_i {Y_j}'^T Y_i' \tag{16}$$

In summary, the optimal solution of the objective function is found as

$$\widehat{W} = \begin{bmatrix} -0.04634 \\ 0.6878 \end{bmatrix}, \hat{b} = 7.075 \tag{17}$$

3.2.2 Classification by SVM Support Vector Machine

After the above process has been found the optimal solution of the objective function and its value, on the basis of which its discriminant function can be further obtained as[10].

$$f\left(Y'\right) = \sum_{i \in SV} \hat{\alpha}_i k_i Y'^T Y_i' + \hat{b} \tag{18}$$

The existence of different solutions of the discriminant function, the size of the solution represents its corresponding glass type. The specific judgment conditions are shown in Table 3 below.

Table 3. Judgment Conditions

f(Y')	Glass Type
≤-1	Lead Barium Glass
≥1	High potassium glass
Other	Unidentifiable

3.2.3 Glass type identification based on discriminant function

According to the data of the proportion of each chemical composition inside the unknown artifact (variables y1, y2, . . . , y14) to obtain their glass types.

Based on this, the data of the unknown glass type is substituted to identify its glass type, and the visualization results are shown in the Figure 2 below.

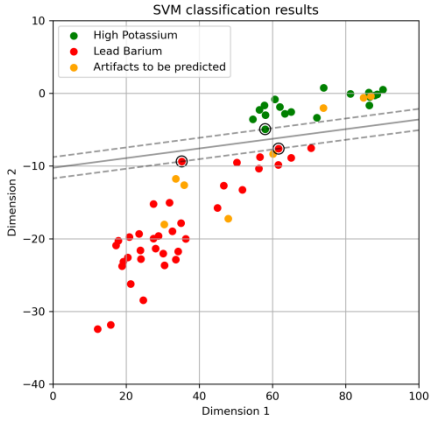


Figure 2. Unknown type of cultural relics classification results

All eight groups of artifacts fell within the predictable range after the PCA principal component analysis dimensionality reduction process and SVM classification process, and the specific results are shown in Table 4 below.

Table 4. Unknown artifact type identification results

Artifact Number	A1	A2	A3	A4	A5	A6	A7	A8
Calculation result	2.26	-3.28	-6.74	-2.57	-1.45	2.77	2.72	-7.01
Glass Type	High Potassium	Lead Barium	Lead Barium	Lead Barium	Lead Barium	High Potassium	High Potassium	Lead Barium

3.3. Sensitivity analysis based on Sobol method

The Sobol method was selected for the sensitivity analysis of the model for the identification model of glass types, and the results are shown in Table 5 below [11].

Table 5. First-order sensitivity index for each chemical component

Chemical composition	First-order sensitivity index
Silicon dioxide	0.0139
Sodium oxide	0
Sulfur dioxide	-0.0021

Among all chemical components, lead oxide (PbO) has the highest first-order sensitivity index of 0.5197, which far exceeds the other components. In addition, copper oxide (CuO), aluminum oxide (Al_2O_3), and silicon dioxide

(SiO_2) also have some influence on the identification results.

It can be seen that the classification model is very sensitive to lead oxide (PbO), and the classification model is more sensitive to copper oxide (CuO), aluminum oxide (Al_2O_3), and silicon dioxide (SiO_2). (CuO), aluminum oxide (Al_2O_3) and silicon dioxide (SiO_2).

Therefore, the model mainly focuses on the content of lead oxide (PbO) in the samples for classification, while other chemical components are considered to complete the type identification.

4. Conclusion

In this paper, based on the results of PCA principal component analysis, the original high-dimensional data were reduced to a suitable dimension and visualized. On this basis, the appropriate classification method is selected by observing the distribution of each data point in combination with the data characteristics and internal laws. Considering that the total data volume is small and each

data point has obvious demarcation after the PCA dimensionality reduction process, the linear division of the dimensionality reduction data is chosen by SVM support vector machine. The SVM-based classification model was constructed based on the SVM principle and the existing data. After that, the accuracy of the model was tested with the existing chemical composition ratio data. Based on this, a sensitivity analysis was performed using the Sobol method to identify the sensitivities of the factors in the model and to determine the main influencing factors of the model. At the same time, the sample data were brought into the classification model to identify the types of unknown glass artifact samples.

References

[1] Takuya Iwanaga, William Usher, and Jonathan Herman. Toward SALib 2.0: Ad- vancing the accessibility and interpretability of global sensitivity analyses. Socio- Environmental Systems Modelling, 4:18155, May 2022.

[2] Jon Herman and Will Usher. SALib: An open-source python library for sensitivity analysis. The Journal of Open Source Software, 2(9), jan 2017.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blon- del, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[4] Organizing Committee of Gaojiao Society. 2022 National University Student Mathematical Modeling Competition. The National Mathematical Modeling Competition for University Students, 2022.

[5] Wang Jie; Li Mo; Ma Qinglin; Zhang Zhiguo; Zhang Meifang; Wang Julin;. Weathering of an octagonal lead-barium glass vessel from the Warring States period. Institute of Metallurgy and Materials History, University of Science and Technology Beijing; Beijing Key Laboratory of Electrochemical Processes and Technologies, Beijing University of Chemical Technology; Chinese Academy of Cultural Heritage, 2014.

[6] . Bjelland, Tor Kristian. "Classification: assumptions and implications for conceptual modeling." (2005).

[7] . Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

[8] Cervantes, Jair, et al. "A comprehensive survey on support vector machine classification: Applications, challenges and trends." Neurocomputing 408 (2020): 189-215.

[9] Chauhan, Vinod Kumar, Kalpana Dahiya, and Anuj Sharma. "Problem formulations and solvers in linear SVM: a review." Artificial Intelligence Review 52.2 (2019): 803-855.

[10] He, Qing, et al. "A novel classification method based on hypersurface." Mathematical and computer modelling 38.3-4 (2003): 395-407.

[11] Nossent, Jiri, Pieter Elsen, and Willy Bauwens. "Sobol'sensitivity analysis of a complex environmental model." *Environmental Modelling & Software* 26.12 (2011): 1515-1525.