

Research on Classification and Identification of Glass Relics Based on Pearson Clustering and Machine Learning Models

Lei Gewen*, Wu Zijun, Si Mengze

School of Biomedical Engineering, Southern Medical University, Guangzhou, China, 510515

*Corresponding author: leigw0216@163.com

Abstract. Ancient glass is very susceptible to weathering due to the influence of buried environments. In order to classify and identify glass artifacts, this article uses the Pearson cluster analysis model to classify them into four types of high potassium and three types of lead and barium. The chemical cost composition of the classification types is determined, and then the composition of each chemical component is accumulated and arranged. The proportion of each type's content is used as the basis for subclassing. Finally, multiple machine learning methods were used to identify unknown types of glass artifacts, and the results showed that the accuracy of the support vector machine model reached 100%, with a relative improvement of 2% -11%. Therefore, the support vector machine model was considered for classification.

Keywords: Cluster Analysis, Machine Learning Models, Classification of Glass Cultural Relics.

1. Introduction

Glass played an important role in the ancient Silk Road trade and witnessed the history of East West exchanges. Ancient glass is highly susceptible to weathering due to the influence of burial environments. During the weathering process, elements inside the glass exchange with environmental elements, leading to changes in the chemical composition ratio and increasing the difficulty of determining the type of glass. Nowadays, archaeologists classify these glass artifacts into two categories: high potassium glass and lead barium glass based on relevant data from a batch of ancient glass products, combined with chemical composition and other detection methods[1].

Scholars have conducted extensive research to classify and identify glass artifacts. In 2023, Li Jiakang et al. identified the composition and types of glass using principal component analysis models and K-means clustering as an auxiliary method[2]. Jiang Yongsheng et al. constructed a probability judgment function using the ratio of density function, established an error analysis function using the integration of density function, and conducted sensitivity analysis on the error function by changing the threshold[3]. They established a mathematical model to solve the common problem of glass classification and identification in antique determination. Lu Jiajia uses methods such as random forest, k-nearest neighbor learning, and Naive Bayes classifier to classify ancient glass types[4]. By comparing several different classification models, it was found that the accuracy of the random forest model is higher.

This article is based on Pearson clustering and machine learning models for the classification and identification of glass artifacts. Firstly, based on the basic data of glass relics, the chemical composition of high-potassium glass and lead-barium glass were analyzed using Pearson clustering analysis model, and the classification rules of high potassium glass and lead barium glass were obtained. According to the spectrum obtained using the Pearson clustering analysis model, the glass samples are further classified into 3-5 categories. Finally, analyze the rationality and sensitivity of the classification results. Secondly, using this model to analyze the chemical composition data of glass artifacts of unknown categories, predict the classification results of glass artifacts, and establish a mathematical model to evaluate the accuracy of the given classification results. This article uses the support vector machine model in SPSS to first classify cultural relics samples into two types: high potassium and lead barium. Then, based on the Pearson clustering analysis model established for the two types of high potassium and lead barium, the types and subclasses of unknown glass relics are determined.

This research method is not only professional and rigorous, but also provides strong support for the classification and identification of glass cultural relics, and can play an important role in the classification and identification of other cultural relics

2. Artifact Classification Model

2.1. Data selection and preprocessing

The data of glass cultural relic samples are divided into two types: high potassium and lead barium, the proportion of chemical composition of each type is accumulated, the accumulated sum is set between 85% and 105% as valid data, and invalid data are eliminated.

2.2. Cluster Analysis Model Principle

Hierarchical clustering method (hierarchical clustering method), first of all, each sample is regarded as a class, and the distance between samples and between classes is specified[5]. At the very beginning, since each sample is a class of its own, the distance between samples and the distance between classes are the same. Next, select the two classes with the shortest distance among all the classes and merge them into a new class, and then calculate the distance between the new class and the other classes[6]. Then the two closest classes are merged until all samples are finally merged into one class. This is a continuous process of clustering, which can be represented by a cluster spectrum diagram.

The distance between the two classes is defined as the average square distance between the two classes, as shown in (1):

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{i \in G_p} \sum_{j \in G_q} d_{ij}^2 \quad (1)$$

The recurrence formula is:

$$D_{kr}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kr}^2 \quad (2)$$

Where G_p and G_q are classes, and D_{pq} is the distance between G_p and G_q , we get:

$$D_{pq} = \min_{j \in G_p, j \in G_q} d_{ij} \quad (3)$$

2.3. Classification Results of Pearson Cluster Analysis Model

Import the high potassium and lead barium data obtained from data preprocessing into the Pearson cluster analysis model in SPSS for analysis. The results obtained are as follows:

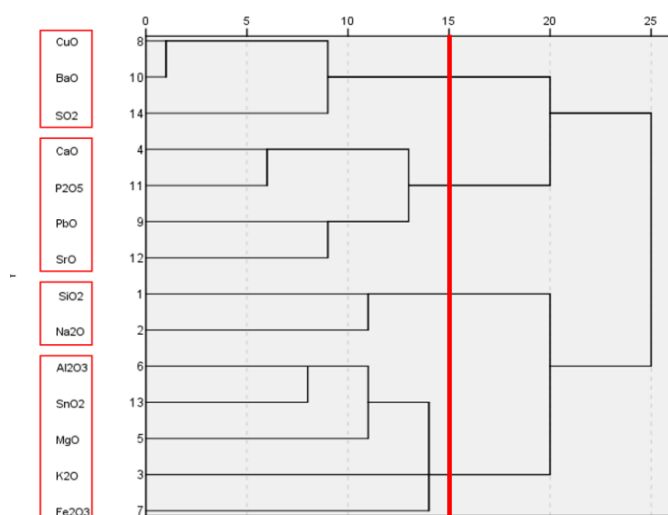


Figure 1. Lead-Barium Pearson Cluster Analysis

As shown in Figure 1, the classification at a relative distance of 20 is selected to classify the lead-barium glass types into four categories. Copper oxide, barium oxide and sulfur dioxide are in the first group, calcium oxide, phosphorus pentoxide, lead oxide and strontium oxide are in the second group, silicon dioxide and sodium oxide are in the third group, and aluminum oxide, tin oxide, magnesium oxide, potassium oxide and iron oxide are in the fourth group.

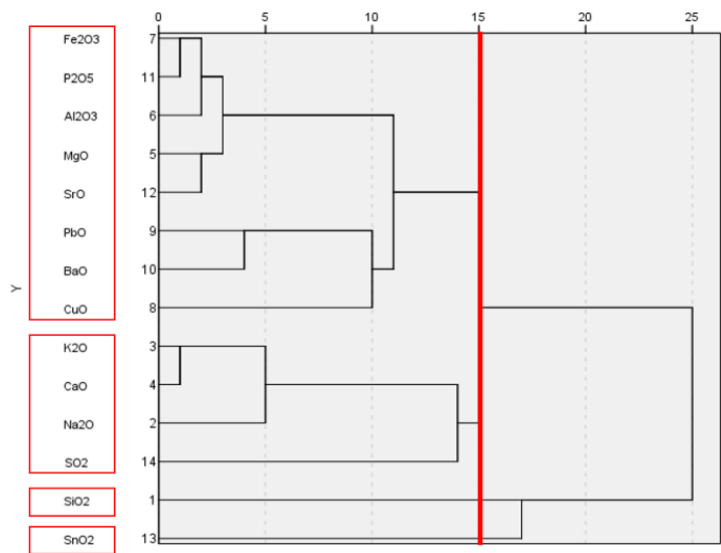


Figure 2. High Potassium Pearson Cluster Analysis

As shown in Figure 2, a relative distance of 15 was chosen to classify the high-potassium glass types into four categories. Iron oxide, phosphorus pentoxide, aluminum oxide, magnesium oxide, strontium oxide, lead oxide, barium oxide and copper oxide belong to the first group, potassium oxide, calcium oxide, sodium oxide and sulfur dioxide belong to the second group, silicon dioxide belongs to the third group and tin oxide belongs to the fourth group.

2.4. Division method and division result

According to the chemical composition information obtained from pretreatment, Pearson cluster analysis model was used to analyze the chemical composition data of high potassium and lead barium. Next, accumulate and sum the chemical component content in each classification type obtained, and arrange the proportion with excel. Among them, because the cumulative content of chemical components in the fourth category in the lead barium data is very small, it indicates that the lead barium glass contains a very small amount of this chemical component, so the fourth category is excluded. Arrange the data in each attachment separately, and the division results obtained are shown in Table 1 and Table 2:

Table.1. Classification of High Potassium Glass

The First Category	High Fe ₂ O ₃ +P ₂ O ₅ +Al ₂ O ₃ +MgO+SrO+PbO+BaO+CuO content
The Second Category	High CaO+K ₂ O+NaO+SO ₂ content
The Third Category	SiO ₂ content greater than 90%
The forth Category	comprising SnO ₂

Table.2. Classification of Lead-Barium Glass

The First Category	High BaO+CuO+SO ₂ content
The Second Category	High CaO+P ₂ O ₅ +PbO+SrO content
The Third Category	High SiO ₂ +NaO content

3. Artifact Identification Model

3.1. Data selection and pre-processing

Based on the chemical composition data of the unclassified artifacts, they were classified into two types, high potassium glass and lead-barium glass, using the support vector machine model in SPSS.

3.2. Principles of Support Vector Machine Modeling

The theory of Support Vector Machines (SVMs) originally came from the treatment of the problem of classifying two classes of data. SVMs consider finding a hyperplane such that the points in the training set belonging to the different classifications lie on exactly different sides of the hyperplane and, furthermore, such that the points are as far away from that hyperplane as possible[7].

In other words, a hyperplane is found such that the blank area between its two clubs is maximized. As shown in Figure 3 for the two-dimensional two-class linear separable pattern, the circle and triangle in the figure represent two types of training samples, H for the two types of correct separation of the classification line, H1, H2 are the closest point in each type of sample from the classification line and parallel to the classification line of the straight line, then, the distance between H1 and H2 that is, the classification of the two types of intervals. The so-called optimal classification line requires that the classification line not only separates the two types of training samples without error, but also maximizes the classification interval between the two types. Extended to higher dimensional spaces, the optimal classification line becomes the optimal hyperplane[8].

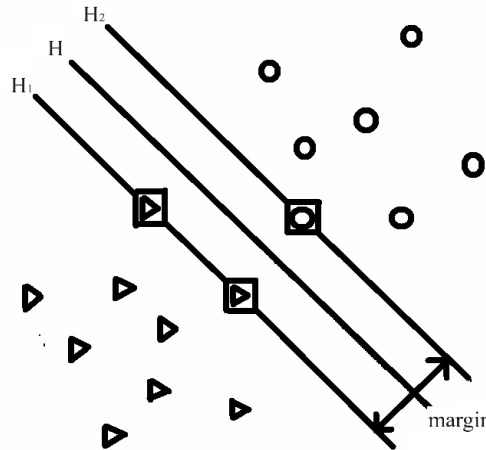


Figure 3. Schematic of the optimal hyperplane

Let the training sample input be $x_i, i = 1, \dots, l$, and the corresponding desired output be $y_i \in \{+1, -1\}$. Where +1 and -1 represent the category identifiers of the two classes, respectively. The goal of the support vector machine is to construct an objective function, based on the principle of minimizing structural risk, so as to distinguish the two classes of patterns as correctly as possible.

For the above training samples in the linearly divisible case, there is then a hyperplane that completely separates the two classes of samples, which can be described as $\omega \cdot x + b = 0$. Where “ $\cdot$ ” is the vector dot product, ω is an n-dimensional vector, and b is the offset.

The optimal hyperplane can be obtained by solving the following convex quadratic optimization problem.

$$\min \Phi(\omega) = \frac{1}{2} |\omega|^2 \quad (4)$$

Satisfaction of constraints:

$$y_i[(\omega \cdot x_i) + b] - 1 \geq 0, i = 1, \dots, l \quad (5)$$

To solve the constrained optimization problem, the Lagrange function shown in equation (6) is introduced:

$$L = \frac{1}{2} |\omega|^2 - \sum_{i=1}^l a_i y_i (\omega \cdot x_i + b) + \sum_{i=1}^l a_i \quad (6)$$

Where $a_i > 0$ is the Lagrange coefficient, then the optimal obtained after solving the above problem classification function is:

$$f(x) = \text{sgn}\{\sum_{i=1}^l y_i a_i^* (x \cdot x_i) + b^*\} \quad (7)$$

An assumption needs to be made here: the data must be 100 percent separable. However, none of the actual data is so coherent and there will be noise. However, due to the small amount of sample data used in the artifact categorization process, following the above steps will give a better prediction.

3.3. Model building and testing

For the preprocessed data, a support vector machine model was built. Take the unknown category of glass in the data artifacts' chemical compositions as inputs and their belonging types as outputs. The support vector machine model is tested to be 100% correct, which is a very high rate of correctness, and it can be seen that the model is very reliable.

3.4. Application of the model

The artifact identification model was solved using the Lagrange multiplier method, based on SPSS, using a support vector machine model. The glass artifacts were classified by analyzing the known chemical composition of the data. The classification results are as shown in Table.3:

Table.3. Results of artifact identification

Artifact Number	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	A ₈
Type	high-potassium	lead-barium	lead-barium	lead-barium	lead-barium	high-potassium	high-potassium	lead-barium

Finally, based on the artifact classification model for the unknown type of glass artifacts of subclasses, the two types of high potassium and lead-barium obtained using the support vector machine model will be divided again, and the results obtained are as shown in Figure 4:

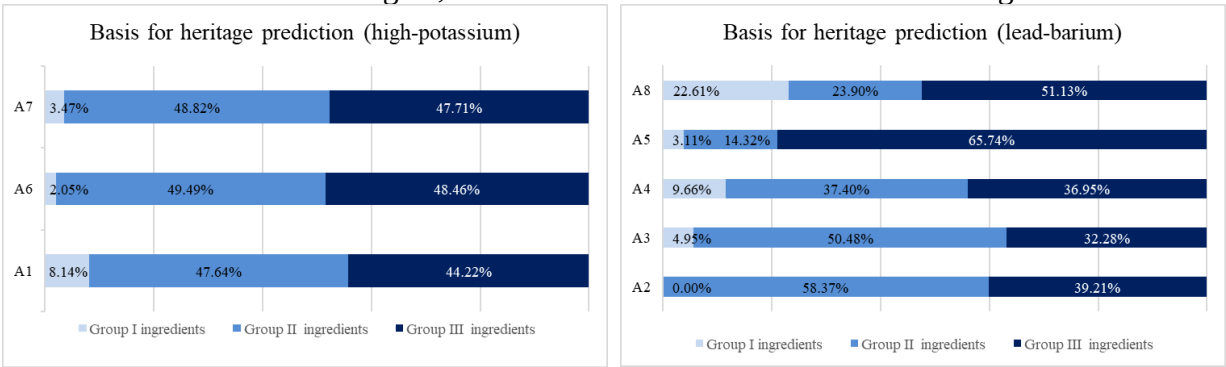


Figure 4. Basis for heritage prediction

The final results were obtained as shown in Table 4.

Table. 4. Results based on Artifact Classification Models

Artifact Number	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	A ₈
Type	high-potassium	lead-barium	lead-barium	lead-barium	lead-barium	high-potassium	high-potassium	lead-barium
Subtype	II	II	II	II	III	III	III	III

4. Model Evaluation

In this paper, we mainly use the methods of establishing random forest and cluster analysis model to maximize the advantages of the two models. We have basically solved the problems of chemical composition content before weathering, subclassification of different glass types, unknown cultural relics, and chemical composition correlation between different categories. The chemical composition content before weathering, the subclassification of different glass types, the type of unknown artifacts, and the chemical composition between different categories are basically solved, which provides a theoretical analysis and identification of glass artifacts. analysis and identification of glass artifacts provides a theoretical basis.

The Random Forest model uses an integrated algorithm that introduces sample randomization and feature randomization, and is able to handle a large number of input variables and high-dimensional data, and produce highly accurate classifiers. Random forest model has better accuracy and faster training speed, while it is less prone to overfitting and has a higher tolerance for outliers and noise. Higher[9].

Pearson cluster analysis is a correlation analysis method that uses the Pearson correlation coefficient as a measure of the clustered variables. analysis by using the Pearson's correlation coefficient as a measure of the clustered variables[10]. It is used to measure the degree of linear correlation between two continuous variables. The Pearson correlation coefficient is the best measure of correlation for normally distributed data. Pearson's correlation coefficient is the most appropriate and efficient way to solve for normally distributed data[11].

Support vector machine (SVM) is a new type of machine learning method based on SLT. SVM is a new machine learning method based on SLT, which has become a hot research topic in the current machine learning community due to its excellent learning performance. SVM model shows many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition[12].

5. Conclusions

In this paper, we first use the Pearson cluster analysis model to classify the glass artifacts, and use its chemical composition as input to establish the artifact classification model. Then we used the support vector machine model to solve the problem and established the cultural relic identification model. After analyzing and applying the data, the cultural relics classification model achieved high accuracy and stability. At the same time, the accuracy of the cultural relic identification model in predicting the type of each sampling point is close to 1, which is very accurate; and the folding line is relatively stable, so it can be seen that the established model is highly reliable. Therefore, these two models have high application value in the field of cultural relics analysis and identification.

References

- [1] Peng Ruilan, Xiao Jirong. Current status and development prospect of the application of artificial neural network in nursing [J]. General Practice Nursing, 2023, 21(08): 1060-1062.
- [2] Li Jiakang, Duan Jianzhou, Feng Jiawen. Research on the classification law of ancient glass based on principal component analysis model [J]. Popular Standardization, 2023(20): 91-93.
- [3] Jiang Yongsheng, Zhang Junfen, Xue Tingting et al. Exploration of classification based on probabilistic judgment function--taking glass classification as an example [J]. Practice and Understanding of Mathematics, 2023, 53(10):116-125.
- [4] Lu Jiajia. Ancient glass classification model based on integrated feature selection and random forest[J]. Journal of Silicate, 2023, 51(04): 1060-1065.
- [5] Xie Hui, Zhang Jianguo, Liu Yongxing et al. Application of BP neural network in classification and subclassification of glass artifacts [J]. Journal of Zhoukou Normal College, 2023, 40(05): 23-27.
- [6] Kang Glimmer. Research on subclassification of glass artifacts based on improved GDBT and K-means clustering [J]. Information Technology and Informatization, 2023(07): 149-152.

- [7] Hui Jiaojiao, Cao Hongbao. Identification of glass artifacts based on machine learning classification algorithm [J]. Modern Information Technology, 2023, 7(13): 101-104. DOI: 10.19850/j.cnki.2096-4706.2023.13.021.
- [8] Chang JK, Zhao JH, Ren XH, et al. A review of support vector machines [J]. Optical disk technology, 2007(2): 2.
- [9] Fang Kuangnan, Wu Jianbin, Zhu Jianping, Xie Bangchang. A review of research on random forest methods [J]. Statistics and Information Forum, 2011, 26(03): 32-38.
- [10] Hu Leifang. Five commonly used systematic cluster analysis methods and their comparison [J]. Zhejiang Statistics, 2007(04): 11-13.
- [11] Zhang Hua, Long Cheng, Hu Siyang, et al. Distribution network topology calibration based on hierarchical clustering method and Pearson's correlation coefficient method [J]. Power System Protection and Control, 2021, 49(21): 88-96.
- [12] Zhang Xuegong. On statistical learning theory and support vector machines [J]. Journal of Automation, 2000(01): 36-46.