Research on Glass Classification and Recognition based on Support Vector Machines

Shimeng Guo *, Jinwei Gan

Northeastern University at Qinhuangdao, Qinhuangdao, China, 066004

* Corresponding author: 1942788964@qq.com

Abstract. Glass artefacts undergo weathering and a significant exchange of internal elements due to burial in the ground. In order to classify the glass types, Principal Component Analysis was used to transform the existing chemical indicators into a small number of principal components. Subsequently, Support Vector Machine approach was used to classify the glass types to be predicted based on the selected principal components. K-means clustering algorithm was then used to classify the two glass types into subclasses on the basis of the selected chemical components and to perform sensitivity analysis. Using this model, the glass types of the samples to be tested can be identified.

Keywords: Principal Component Analysis, Support Vector Machines, K-means clustering analysis, glass artefacts.

1. Introduction

The Silk Road was a gateway for cultural exchange between China and the West in ancient times, of which glass is a valuable physical evidence of early trade exchanges. Our ancient glass absorbed the techniques of West Asian and Egyptian bead-shaped ornaments and was made from locally sourced materials, so the chemical composition would have been different. Ancient glass is highly susceptible to weathering by the environment in which it is buried, and in the course of which internal elements exchange heavily with environmental elements, resulting in changes in the proportions of its composition that can affect judgement. In order to classify and identify the glass, a classification and identification model is developed in this paper, involving methods such as Principal Component Analysis, Support Vector Machines and K-means clustering algorithm.

PCA is typically a feature extraction tool for pattern classification, a method that reduces data complexity by projecting samples from the original high-dimensional space into a low-dimensional feature space without reducing the inherent information contained in the original data. [1-3]

SVM is a very effective machine learning method in statistics, which has successfully solved the persistent problems of high dimensionality and local extremes in machine learning and has achieved great success in many fields. [4-6] Usually the training set is very large and high dimensional, resulting in large memory space requirements and long training times. In order to optimize this problem, the optimal classification hyperplane is only related to the terms with non-zero Lagrange multipliers when solving the corresponding pairwise problems. By extracting the boundary points containing the support vectors in advance and replacing all the training samples with these boundary sample points for Support Vector Machine training, optimization can be achieved, which can solve the problems in terms of long training time and memory consumption of Support Vector Machines.[7]

K-means clustering algorithm is a division-based clustering method and is widely used in data mining, with the advantages of simplicity, ease of use, interpretability and good clustering results. [8]

In this paper, by using PCA, we transformed the 14 chemical indicators of artifacts into a small number of six principal components, which are used as key indicators for measuring glass properties. Based on the key indicators, SVM approach is applied to classify the artefacts to be predicted. Subsequently, by studying the classification pattern of high potassium glass and lead-barium glass, K-means cluster analysis was used to classify the two classes of glass into subclasses, giving specific classification methods and classification results. The chemical compositions of glass artifacts from unknown categories is also analyzed to identify the type that they belong to. Moreover, the sensitivity of the classification results is analyzed.

2. Establishment and solution of glass classification model

2.1. Selection of 14 indicators for glass classification

There are 14 chemical compositions in the artifact component data on which this paper is based, and there are only two types of artifacts, high potassium glass and lead-barium glass. For glass classification, although these 14 chemical indicators are different in nature, there are often correlations between different indicators. We can transform the 14 indicators into a few principal components based on some potential correlations, which are linear combinations of the original indicators and uncorrelated with each other, and they can be used as key indicators to measure the properties of glass. Based on the above analysis, the 14 indicators are selected by using Principal Component Analysis. [9-10]

2.1.1. Principle of PCA

Step1: Standardize the sample matrix.

67 samples with 14 indicators can form a sample matrix ω of size 67×14.

$$\omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1,14} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2,14} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{67,1} & \omega_{67,2} & \cdots & \omega_{67,14} \end{pmatrix} \tag{1}$$

The mean value by column is calculated as $\overline{\omega_j} = \frac{1}{67}\sum_{i=1}^{67}\omega_{ij}^2$, the standard deviation is $S_j = \sqrt{\frac{\sum_{i=1}^{67}(\omega_{ij} - \overline{\omega_j})^2}{67-1}}$. After standardization, $\Omega_{ij} = \frac{\omega_{ij} - \overline{\omega_j}}{S_j}$.

The standardization process yields a matrix as Ω, Ω is $\begin{pmatrix} \Omega_{11} & \Omega_{12} & \cdots & \Omega_{1,14} \\ \Omega_{21} & \Omega_{22} & \cdots & \Omega_{2,14} \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{67,1} & \Omega_{67,2} & \cdots & \Omega_{67,14} \end{pmatrix}$.

Step2: Calculate the covariance matrix.

The covariance of the sample is r_{ij}, r_{ij} is $\frac{1}{67-1}\sum_{i=1}^{67}(\Omega_{ki} - \overline{\Omega_i})(\Omega_{kj} - \overline{\Omega_j})$.

The covariance matrix is R, R is $\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1,14} \\ r_{21} & r_{22} & \cdots & r_{2,14} \\ \vdots & \vdots & \ddots & \vdots \\ r_{14,1} & r_{14,2} & \cdots & r_{14,14} \end{pmatrix}$.

Step3: Calculate the eigenvalues and eigenvectors of R.

The eigenvalue satisfies $\chi_1 \geq \chi_2 \geq \cdots \geq \chi_{14} \geq 0$, and the feature vector is γ_i, γ_i is $\begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \\ \vdots \\ \gamma_{14,i} \end{pmatrix} (i = 1, 2, \cdots, 14)$.

Step4: Calculate the contribution rate and cumulative contribution rate of principal components.

Contribution rate is w_i, w_i is $\frac{\chi_i}{\sum_{k=1}^{14} \chi_k} (i = 1, 2, \cdots, 14)$.

Cumulative contribution rate is W_i, W_i is $\frac{\sum_{k=1}^{i} \chi_k}{\sum_{k=1}^{14} \chi_k} (i = 1, 2, \cdots, 14)$.

Step5: Represent the principal components.

The components corresponding to the eigenvalues with a cumulative contribution of 80% are noted as principal components, and if there are $m(m \leq p)$ principal components, the principal components can be expressed as follows.

$$\begin{cases} F_1 = \gamma_{11}\Omega_1 + \gamma_{21}\Omega_2 + \cdots + \gamma_{14,1}\Omega_{14} \\ F_2 = \gamma_{12}\Omega_1 + \gamma_{22}\Omega_2 + \cdots + \gamma_{14,2}\Omega_{14} \\ \cdots \\ F_m = \gamma_{1m}\Omega_1 + \gamma_{2m}\Omega_2 + \cdots + \gamma_{14,m}\Omega_{14} \end{cases} \tag{2}$$

2.1.2. Results of PCA

Using SPSS to perform Principal Component Analysis on these 67 samples, the original variables were dimensionally reduced to transform the original 14 variables into 6 composite indicators, which can reflect most of the component information of the glass compositions, so as to achieve a simplifying effect on the data.

The eigenvalues and cumulative contribution rates are shown in Table 1.

Table 1. Eigenvalues and cumulative contribution rates table

Principal Component	Extract The Load Sum of Squares		
	Total	Variance Proportion	Cumulative Percentage
F1	4.055	28.966%	28.966%
F2	2.413	17.233%	46.200%
F3	1.739	12.422%	58.621%
F4	1.124	8.092%	66.651%
F5	1.110	7.931%	74.582%
F6	0.784	5.603%	80.185%

The orthogonal solutions obtained by PCA of SPSS are shown in Table 2.

Table 2. Orthogonal solution table

Chemical Composition	Principal Component					
	F1	F2	F3	F4	F5	F6
SiO2	0.784	-0.517	0.002	-0.164	0.063	0.135
Na2O	0.098	-0.332	-0.026	0.813	-0.317	-0.099
K2O	0.665	0.255	0.451	0.085	0.022	-0.384
CaO	0.349	0.67	0.364	0.006	-0.222	-0.394
MgO	0.396	0.625	-0.305	0.242	0.148	0.242
Al2O3	0.603	0.375	-0.038	0.386	0.081	0.294
Fe2O3	0.464	0.575	0.122	-0.153	-0.13	0.265
CuO	-0.296	0.039	0.737	0.041	-0.036	0.327
PbO	-0.763	0.283	-0.426	-0.012	-0.179	-0.201
BaO	-0.766	-0.06	0.438	0.219	0.202	0.176
P2O5	-0.402	0.671	-0.11	-0.208	0.036	0.08
SrO	-0.679	0.341	-0.195	0.281	-0.011	0.014
SnO2	0.21	0.047	-0.236	0.154	0.839	-0.188
SO2	-0.454	0.071	0.521	0.049	0.361	-0.133

The orthogonal solution table illustrates that principal component 1 is a mixture of SiO2, PbO, BaO and SrO. Principal component 2 is a mixture of CaO, MgO and P2O5. Principal component 3 is SO2. Principal component 4 is Na2O, principal component 5 is SnO2, and principal component 6 is a mixture of K2O and CuO.

2.2. Support vector machine classification method

2.2.1. Basic ideas of support vector machines

Support vector machines, models that perform classification functions, is a technique for extracting hidden information from large amounts of data by constructing two classes of hyperplanes that make it possible to classify data into two classes and thus achieve good discrimination. The points that fall on the boundary plane are called support vectors, and these points play a key role in a support vector machine. Using the Support Vector Machine approach, the six indicators obtained from the principal component analysis are trained, and then it is possible to achieve species prediction for eight types of glass whose components are known and whose categories are unknown.

2.2.2. Support vector machine theoretical foundations

For the linearly divisible case, we can transform the original more complex problem into a quadratic programming problem as follows.

$$\begin{aligned} &\min \frac{1}{2}\|\xi\|^2 \\ &s.t. y_i[(\xi \cdot m_i)+n] \geq 1(i=1,2,\cdots,t_o) \end{aligned} \tag{3}$$

By using the Lagrange multiplier method we have:

$$L(\xi,n,\mu)=\frac{1}{2}\|\xi\|^2+\sum_{i=1}^{t}\mu_i\left\{1-y_i[(\xi\cdot m_i)+n]\right\} \tag{4}$$

Where $\mu=[\mu_1,\mu_2,\cdots,\mu_t]^T$ is the Lagrangian Multiplier.

Considering the couplet, we can have the equation that sets the partial derivatives of each variable to zero, i.e. $\frac{\partial L}{\partial \xi}=0, \frac{\partial L}{\partial n}=0$.

It is possible to obtain:

$$\begin{cases} \xi = \sum_{i=1}^{t} \mu_i y_i m_i \\ \sum_{i=1}^{t} \mu_i y_i = 0 \end{cases} \tag{5}$$

By substituting the Lagrangian equation into the dual of the original problem, we can get:

$$\max_{\alpha} - \frac{1}{2} \sum_{i=1}^{t} \sum_{j=1}^{t} y_i y_j \mu_i \mu_j (\mu_i \cdot \mu_j) + \sum_{i=1}^{t} \mu_i \tag{6}$$

$$\begin{cases} \sum_{i=1}^{t} y_i \mu_i = 0 \\ \mu_i \geq 0, i = 1, 2, \cdots, t \end{cases} \tag{7}$$

By solving the above equation, we can obtain:

$$\xi^* = \sum_{i=1}^{t} \mu^*{}_i y_i m_i \tag{8}$$

From the KKT complementarity condition we can obtain:

$$\mu^*{}_i \left\{ 1 - y_i [(\xi^* \cdot m_i) + n*] \right\} = 0 \tag{9}$$

The corresponding m_i^* is positive only if m_i satisfies as a support vector, otherwise its value magnitude is 0. Choosing one of the positive components for calculation gives:

$$n^* = y_j - \sum_{i=1}^{t} y_i \mu_i^* (m_i \cdot m_j) \tag{10}$$

From this, the hyperplane is constructed and the decision function is obtained as follows

$$\varphi(x) = \sum_{i=1}^{t} \mu_i^* y_i (m_i \cdot x) + n^* \tag{11}$$

The classification function is further obtained as:

$$f(x) = \mathrm{sgn}[\sum_{i=1}^{t} \mu_i^* y_i (m_i \cdot x) + n^*] \tag{12}$$

2.2.3. Application of support vector machines and prediction results

Using a support vector machine, it is possible to make predictions about the type of glass. The chemical composition of eight glass artefacts is known and the values of their corresponding six classes of principal components F1-F6 can be obtained from the calculations. Substituting these values into the support vector machine, noting the high potassium category as 1 and the lead-barium category as 2, the algorithm is performed and the results are shown in Table 3.

Table 3. Prediction table of artifact categories obtained by SVM

Artifact Number	F1	F2	F3	F4	F5	F6	Category Serial Number	Category
A1	1.24	0.58	0.17	-0.24	-0.05	1.06	1	High Potassium
A2	-0.27	0.98	-0.38	-1.68	-0.79	-1.91	2	Lead Barium
A3	0.22	1.91	-0.11	-0.75	-1.21	-0.18	2	Lead Barium
A4	0.29	1.77	-0.2	-0.62	-0.39	1.86	2	Lead Barium
A5	1.11	0.45	-1.06	1.71	1.43	1.67	2	Lead Barium
A6	0.64	-1.45	0.07	-1.29	0.07	0.21	1	High Potassium
A7	0.77	-1.30	0.02	-0.97	0.08	0.35	1	High Potassium
A8	-0.79	-0.74	1.58	-0.40	0.20	1.18	2	Lead Barium

2.3. K-means clustering algorithm for classifying subclasses

2.3.1. Steps of K-means clustering algorithm

The steps of K-means clustering algorithm are shown in Figure 1.

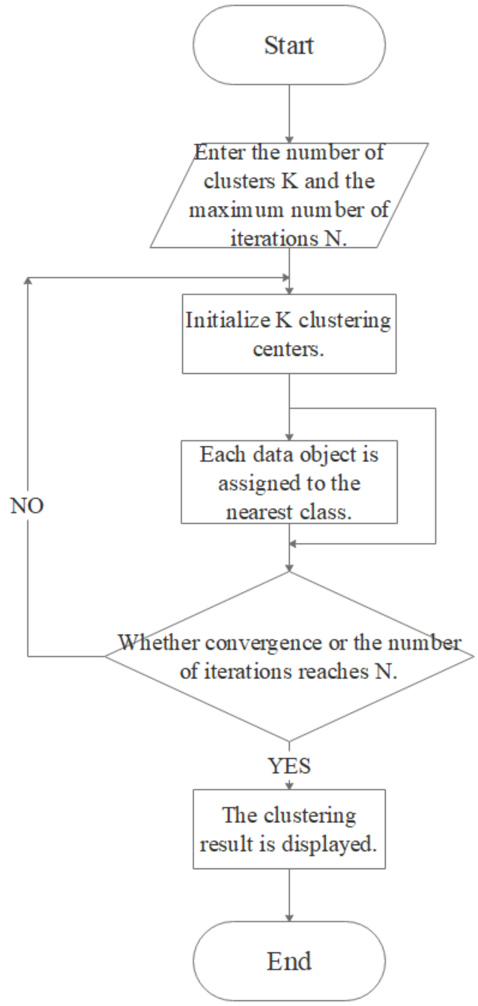

Figure 1. K-means clustering algorithm flow chart

2.3.2. Preparation for cluster analysis

Firstly, the number of clusters that should be divided should be determined. Combining the results of PCA with SVM, it is obvious that the most obvious feature that distinguishes lead-barium glass from high-potassium glass is the content of lead oxide and barium oxide, which can be used as a criterion for classification. For the analysis of lead-barium glass, we decided to specify the number of clusters K=3 after reviewing the relevant information. On the other hand, the distinctive feature of high potassium glass is that the potassium oxide content is high and the calcium oxide is also relatively high, which can be measured for this glass. The number of clusters K=3 is specified here as well.

Secondly, the number of iterations should be determined, for which we may let the number for both types of glass be N = 10.

2.3.3. Results of K-means clustering algorithm

a. Lead barium glass

The initial clustering centers of lead oxide and barium oxide obtained by K-means clustering algorithm are shown in Table 4, and the final clustering centers are shown in Table 5.

Table 4. Initial Clustering Centers

Chemical Composition	Center1	Center2	Center3
Lead Oxide	70.21	12.31	29.92
Barium Oxide	6.69	2.03	35.45

Table 5. Final Clustering Centers

Chemical Composition	Center1	Center2	Center3
Lead Oxide	52.21	18.91	36.08
Barium Oxide	5.77	8.37	16.97

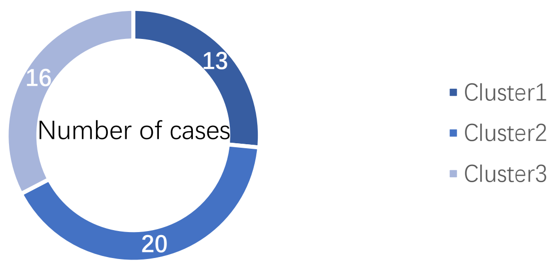

Figure 2. A doughnut plot of the number of cases in each cluster

The iteration history shows that convergence has been achieved here up to 5 iterations as the cluster centers no longer change or change slightly. It can be concluded that the subclasses can be divided into three categories based on the contents of lead oxide and barium oxide, and the details are shown in Figure 2.

There are 13 pieces of low lead-barium subclass, 20 pieces of medium lead-barium subclass, and 16 pieces of high lead-barium subclass.

b. High potassium glass

Table 6. Initial Clustering Centers

Chemical Composition	Center1	Center2	Center3
Potassium Oxide	0.00	9.42	14.52
Calcium Oxide	4.71	0.00	8.27

The initial clustering centers of potassium oxide and calcium oxide obtained by K-means clustering algorithm are shown in Table 6, and the final clustering centers are shown in Table 7.

Table 7. Final Clustering Centers

Chemical Composition	Center1	Center2	Center3
Potassium Oxide	0.47	7.42	11.76

Calcium Oxide	1.42	1.86	7.41

Figure 3. A doughnut plot of the number of cases in each cluster

The iteration history shows that convergence has been achieved here iteration to 2 since the cluster centers no longer change or change slightly. It can be concluded that the subclasses can be divided into three categories based on the contents of potassium oxide and calcium oxide, and the details are shown in Figure 3.

There are 7 pieces in the low potassium-calcium subclass, 4 pieces in the medium potassium-calcium subclass, and 7 pieces in the high potassium-calcium subclass.

2.3.4 Sensitivity analysis of K-means clustering algorithm

i. Change the cluster number K, while the iteration number N remains the same

Taking lead-barium glass as an example, adjust the parameters so that K=2 while n=10 remains unchanged. The initial clustering centers are shown in Table 8, and the final clustering centers are shown in Table 9.

Table 8. Initial Clustering Centers

Composition	Center1	Center2
Lead Oxide	70.21	9.30
Barium Oxide	6.69	23.55

Table 9. Final Clustering Centers

Composition	Center1	Center2
Lead Oxide	52.21	18.91
Barium Oxide	5.77	8.37

Figure 4. A doughnut plot of the number of cases in each cluster

The iteration history shows that convergence has been achieved for iterations up to 3. It can be concluded that the subclasses can be divided into two categories based on the content of lead oxide and barium oxide, and the details are shown in Figure 4. As K diminishes, the number of iterations decreases and the number of pieces in each class in the clusters increases.

ii. Change the iteration number N, while the cluster number K remains the same

Still discussing lead barium glass, at this point adjust the parameters so that n=3 while K=3 remains unchanged. The initial clustering centers and the final clustering centers are shown in Table 10.

Table 10. Initial and Final Clustering Centers

Composition	Initial Clustering Centers			Final Clustering Centers		
	Center1	Center2	Center3	Center1	Center2	Center3
Lead Oxide	70.21	12.31	29.92	52.21	18.91	36.08
Barium Oxide	6.69	2.03	35.45	5.77	8.37	16.97

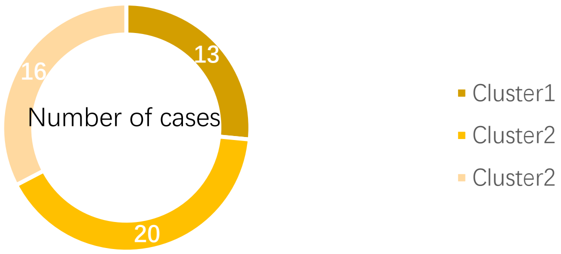

Figure 5. A doughnut plot of the number of cases in each cluster

The iteration history shows that the iteration stops at 3 and no convergence is achieved. However, this scenario is still satisfied with 13 pieces for the low lead-barium subclass, and the details are shown in Figure 5. There are 20 pieces for the medium lead-barium subclass, and 16 pieces for the high lead-barium subclass.

3. Conclusions

For the classification of high potassium and lead-barium glass, considering that the 14 categories of chemical compositions, although different in nature, often have correlations between different indicators, the 14 indicators can be transformed into 6 principal components according to some potential correlations.

On this basis, we applied the Support Vector Machine method well for the prediction of the unknown glass artifact category to get the results.

At the end, K-means clustering algorithm was applied based on selected chemical compositions. For lead-barium glass, it can be divided into three glass subclasses of low lead-barium, medium lead-barium, and high lead-barium. For high potassium glass, it can be classified into three subclasses of low potassium-calcium, medium potassium-calcium, and high potassium-calcium, with excellent classification results known by sensitivity analysis.

References

[1] HUANG Yufei, SHI Xinfa, HE Shizhong, et al. A fault diagnosis method for wind turbine gearbox based on PCA and SVM [J]. Journal of Engineering for Thermal Energy and Power, 2022, 37 (10): 175 - 181.

[2] CHEN Meigu, LIN Xing'e, LI Xinguo, LIU Xiaodi, GAO Hongmao, MING Jianhong, DAl Minjie. Comprehensive evaluation of durian quality based on principal component analysis and cluster analysis [J]. Science and Technology of Food Industry, 2022.

[3] WEI Xueqin, LI Wen, LI Bingguo, GENG Lufei, LIN Yushi, PANG Jie. Quality evaluation of 12 plant Jiaosu: Based on principal component analysis and cluster analysis[J]. Food Research and Development, 2022, 43 (17): 41 - 48.

[4] GAO Z, FANG S C, LUO J, et al. A Kernel-Free Double Well Potential Support Vector Machine with Applications [J]. European Journal of Operational Research, 2020, 290 (1): 248 - 262.

[5] DING S, ZHANG N, ZHANG X, et al. Twin Support Vector Machine: Theory, Algorithm and Applications [J]. Neural Computing and Applications, 2017, 28 (11): 3119 - 3130.

[6] LIU Y, DING H, HUANG Z, et al. Distributed and Robust Support Vector Machine [J]. International Journal of Computational Geometry and Applications, 2021, 30 (3): 213 - 233.

[7] UTAMI N A, MAHARANI W, ATASTINA I. Personality Classification of Facebook Users According to Big Five Personality Using SVM (Support Vector Machine) Method [J]. Procedia Computer Science, 2021, 179 (1): 177 - 184.

[8] ZHOU Chenglong, CHEN Yuming, ZHU Yidong. Granular K-means clustering algorithm [J]. Computer Engineering and Applications, 2022.

[9] FENG Jianying, SHI Yan, WANG Bo, MU Weisong. Cluster analysis in data mining and its application in agriculture[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53 (S1): 201 - 212.

[10] Chen Yan, Zheng Jian, Chen Shimiao, Yu Jiangmin, Pan Zujian, He Jiang, Gan Weitang. The main traits, principal component analysis and cluster analysis of 10 papaya germplasms [J]. South China Fruits, 2021, 50(04): 69 - 74.