An analytical identification model for ancient glass based on support vector machine and cluster analysis

Yexi Zhang[1, *, #], Rongkai Cheng[1, #] and Jinglei Li[2, #]

[1]School of Electrical and Electronic Engineering, Shijiazhuang Railway University, Shijiazhuang, China

[2]School of Transportation, Shijiazhuang Railway University, Shijiazhuang, China

*Corresponding author: zyx2460926925@163.com

#These authors contributed equally to this work

Abstract. Glass artifacts are easily weathered by environmental influences during burial, and the internal chemical composition is exchanged with the external environment, resulting in changes in its composition ratio. For archaeological interrogation, it is of great significance to study the chemical composition of glass before and after weathering and related issues. In this paper, a support vector machine model is used to reasonably separate the data points of high potassium glass and lead-barium glass, and a linear classification function is derived, i.e., the classification law of the two types of glass is obtained. Then, this paper performs R-type cluster analysis on the chemical composition indexes first, and reduces the dimensionality of several indexes; then performs Q-type cluster analysis on the sample detection points, and performs a one-way ANOVA test on the classification scheme of different numbers of subclasses obtained, and finally divides them into four subclasses. On this basis, the data of the unknown class of glass artifacts are substituted into the trained support vector machine model, and the type to which they belong can be determined by calculating the value of the judgment linear classification function.

Keywords: Support Vector Machine; Cluster Analysis; High Potassium Glass; Lead Barium Glass.

1. Introduction

Ancient glass relics in the burial process, a series of complex physical and chemical changes will occur, resulting in glass weathering, making it darker, less transparent, halo color or outside the formation of weathering products crust, its original appearance is almost unrecognizable, which will affect the determination of its category[3]. The degree of weathering varies from glass artifact to glass artifact and from one part of the same glass artifact to another. This has caused great difficulty for archaeologists in their research and identification of artifacts.

In this paper, by analyzing the information on the chemical composition content contained in each high-potassium glass and lead-barium glass, the classification pattern of both was determined, and for each category, the appropriate chemical composition was selected for subclassification. Based on this classification method, eight unknown glass artifacts were analyzed for chemical composition and identified the respective types to which they belonged.

2. Glass classification research based on support vector classification machine

2.1. Fundamentals of support vector classification machine

Support vector machines are a general machine learning algorithm.[4]The main idea is to find a hyperplane that makes it possible to separate the two classes of data points correctly as possible while keeping the separated two classes of data points farthest from the classification plane [1].
According to the given training set:

$$T = \{[a_1, y_1], [a_2, y_2], \cdots, [a_l, y_l]\} \in (\Omega \times Y)^l \quad (1)$$

Where: $a_i \in \Omega = R^{\Omega}$; Ω is called the input space, and each point a_i in the entry space consists of one *n* attribute of the feature;$y_i \in \mathrm{Y} = \{-1, 1\}, i = 1,2,\cdots, l$.
Find a substantive function $g(x)$ on R^{Ω} to use the classification function.

$$f(x) = sign(g(x)) \tag{2}$$

The problem to infer the value of any pattern *x* corresponding to *y* is a classification problem.

2.2. Solution of linear classification function

Step1.Data pre-processing.

In this paper, the two groups of test points whose summation of chemical composition ratios did not lie between 85% and 105% were excluded, but there were still differences between the summation of chemical composition ratios of the remaining valid data groups, which could bring errors to the subsequent analysis and processing of the data. So the proportion of each group of chemical components will be standardized, so that the sum of the proportion of each group of chemical components in the normalized data is equal to 100%. Then the new percentage of each chemical composition after normalization is calculated as follows.

$$h_{i\,new} = \frac{h_{i\,old}}{\sum_{i=1}^{n} h_{i\,old}} \tag{3}$$

Where $h_{i\,new}$ is the new percentage of a chemical composition and $h_{i\,old}$ is the old percentage of a chemical province.

Step2. Define variables.

With $i = 1, 2,\cdots, 67$ to indicate 67 sampling and testing points, respectively, the *i*-th adoption testing point number *k* of the indicator to take the value of x_{ik}. $y_i = 1$ indicates high potassium glass class, and $y_i = -1$ indicates lead barium glass class.
Mean vectors were calculated for the 67 sampled detection points:

$$\mu = [\mu_1, \mu_2,\cdots, \mu_{14}] \tag{4}$$

$$= [50.01, 0.80, 1.88, 2.60, 0.70, 4.13, 0.87, 2.00, 25.24, 8.01, 2.78, 0.27, 0.08, 0.61]$$

Step3. Zero-mean normalization of data.

The data for all sample points were normalized using the following equation:

$$\tilde{x}_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}, (k = A, \mathrm{B}\cdots, \mathrm{N}) \tag{5}$$

Correspondingly,

$$\tilde{p}_j = \frac{x_k - x_k}{s_j}, (j = 1,2,\cdots,14) \tag{6}$$

It is the standardized indicator variable. Noted as $\tilde{p} = [\tilde{p}_1, \tilde{p}_2,\cdots, \tilde{p}_{14}]^T$.

Step4.Solve the support vector.

Denote the normalized row vector of data for the 67 classified sample points as:

$$b_i = [\tilde{x}_{i1}, \tilde{x}_{i2},\cdots, \tilde{x}_{i8}], (i = 1,2,\cdots,67)$$

The support vector machine model with linear kernel functions is used for classification, and the proposed support vector is obtained as:

$$b_6, b_{16}, b_{17}, b_{18}, b_{31}, b_{34}, b_{35}, b_{36}$$

Step5.Solve the linear classification function.

The linear classification function is:

$$c(\tilde{p}) = \sum_{i=1}^{8} \beta_i K(b_i, \tilde{p}) + b \tag{7}$$

Where: $\beta_i (i = 1,2,\cdots,8)$ is the weight coefficient of the returned classification function, $\tilde{p} = [\tilde{p}_1, \tilde{p}_2,\cdots, \tilde{p}_{14}], K(b_i, \tilde{p}) = (b_i \cdot \tilde{p})$ is the linear kernel function (to avoid the complex operations

performed in the high-dimensional feature space, the linear kernel function composition algorithm is used in this paper), and b is the constant term of the returned classification function.

The values of $\beta_i(i = 1,2,\cdots,8)$ were solved by Matlab programming in the following order:

0.2043, 0.0954, 1.0000, 0.0081, 0.2486, 0.1446, 0.0655, 0.8492

The obtained b value is: 1.8918.

The final linear classification function is solved as:

$$c(\tilde{p}) = 0.2043K(b_6,\tilde{p}) + 0.0954K(b_{16},\tilde{p}) + 1.0000K(b_{17},\tilde{p}) + 0.0081(b_{18},\tilde{p}) + 0.2486K(b_{31},\tilde{p}) + 0.1446K(b_{34},\tilde{p}) + 0.0655K(b_{35},\tilde{p}) + 0.8492K(b_{36},\tilde{p}) + 1.8918 \quad (8)$$

When $c(\tilde{p}) \geq 0$, it belongs to high potassium glass; when $c(\tilde{p}) < 0$, it belongs to lead barium glass.

All known sample points are substituted back into the classification function and the classification results are all correct, i.e., their misclassification rate is 0.

3. Classification of glass subclasses based on cluster analysis model

3.1. Subclass classification results

Cluster analysis, also known as cluster analysis, is a multivariate statistical method for quantitatively classifying multiple samples (or indicators). The classification of indicators is called R-type cluster analysis[2], and the classification of samples is called Q-type cluster analysis.

Step1.Data pre-processing.

(1) Normalization of chemical composition ratios. Here the experimental data are normalized again as in Equation (3) for the chemical composition ratio so that the proportion of each component sums to 100%. This is not repeated here.

(2) Elimination of 0-value data. After statistical knowledge analysis, the chemical composition items with a frequency of "0" greater than 60% in each column were eliminated. The eliminated components are sodium oxide, lead oxide, barium oxide, strontium oxide, tin oxide, and sulfur dioxide.

Since the process of the two glass research methods is the same, only the subclass division-solving process for high potassium glass is shown here, and for lead-barium glass, the final results are given in this paper.

Step2.Perform R-type cluster analysis.

(1) Calculate the correlation coefficient and build the correlation coefficient matrix.

In the cluster analysis of variables, the similarity measure of the variables is determined first, and the correlation coefficient is used here as the similarity measure.

Let the value of the variable x_j be $\left[x_{1j},x_{2j},\cdots,x_{mj}\right]^T \in R^m(j = 1,2,\cdots,\mathrm{p})$. Then the sample correlation coefficient between the two variables x_j and x_f can be used as a measure of their similarity.

$$r_{jf} = \frac{\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)(x_{if}-\bar{x}_f)}{\left[\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2\sum_{i=1}^{m}(x_{if}-\bar{x}_f)^2\right]^{\frac{1}{2}}}, (j,f = 1,2,\cdots,p) \quad (9)$$

In the equation, $\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m}x_{ij}$.

The correlation coefficient matrix is then created as.

(2) Clustering of variables using the mean distance linkage method between groups.

The average distance between groups connection method uses the average distance of all sample pairs (two members of a sample pair belong to different classes) of two classes as the distance between the two classes and merges the two classes with the closest distance or the largest correlation coefficient. This method utilizes all the sample information in the two classes so that all the feature information in the samples is represented.

Solve the equation for the average distance between the two types of variables p, q as:

$$D(p,q) = \frac{1}{i \times j} \sum_{i \in G_p} \sum_{j \in G_q} d_{ij} \tag{10}$$

(3) Using Matlab to derive the clustering diagram of each chemical composition index of high potassium glass (similarly, the clustering diagram of each chemical composition index of lead-barium glass can be obtained).

(4)

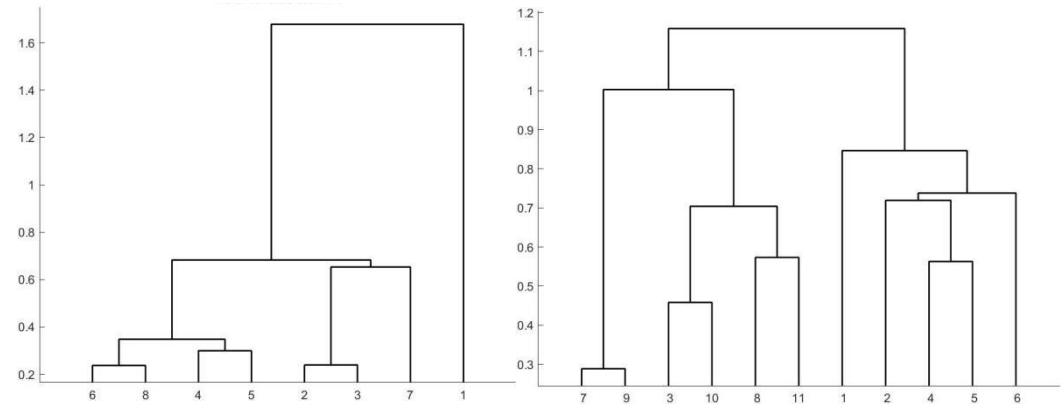


Fig. 1 Cluster diagram of chemical composition of high potassium glass and lead-barium glasss

Step3. Perform Q-clustering analysis.

(1) The eight chemical composition indicators obtained by Step2.R-type cluster analysis were subjected to Q-type cluster analysis for 18 sample points of heritage testing.

(2) First, the data of each variable were standardized separately, the Euclidean distance measure was used for the similarity between samples, and the mean distance connection method between groups was chosen for the calculation of the inter-class distance.

(3) Matlab was used to derive the clustering diagram for each testing sample point of high-potassium glass (similarly, the clustering diagram for each testing sample point of lead-barium glass was obtained).

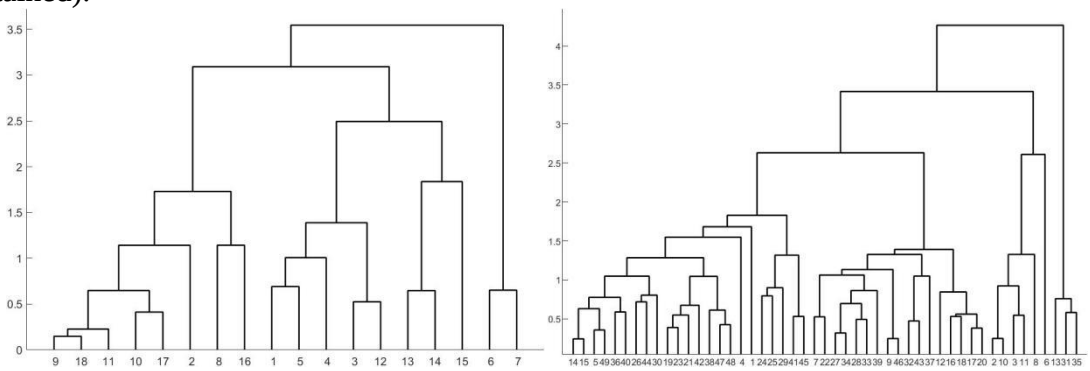


Fig. 2 Cluster diagram of chemical composition of high potassium glass and lead-barium glasss

Step4.Give the final subclass classification results.

Tab 1. High potassium glass subclass classification results

Heritage Sampling Sites	Type	Surface weathering	SiO_2	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	P_2O_5
01	High K	No	71.03	10.23	6.47	0.89	4.03	1.78	3.96	1.20
03Part2	High K	No	62.41	12.51	5.94	1.12	5.56	2.18	5.15	0.71
04	High K	No	68.58	10.07	7.41	1.62	6.70	2.14	2.27	0.82
05	High K	No	63.81	11.35	7.62	1.83	7.77	2.71	3.39	0.97
13	High K	No	60.13	12.77	8.86	0.00	6.28	2.93	4.82	1.29
14	High K	No	63.10	12.40	8.31	0.67	9.32	0.51	0.47	0.16
16	High K	No	66.23	14.75	8.40	0.53	6.28	0.43	1.09	0.00
18	High K	No	81.71	9.69	0.00	1.57	3.14	0.00	0.00	1.40
03Part1	High K	No	87.05	5.19	2.01	0.00	4.06	0.00	0.78	0.66
07	High K	Yes	92.91	0.00	1.07	0.00	1.99	0.17	3.25	0.61
09	High K	Yes	95.24	0.59	0.62	0.00	1.32	0.32	1.55	0.35
10	High K	Yes	96.95	0.92	0.21	0.00	0.81	0.26	0.84	0.00
12	High K	Yes	94.70	1.01	0.72	0.00	1.47	0.29	1.66	0.15
21	High K	No	77.83	0.00	4.78	1.24	6.28	2.41	3.33	1.12
22	High K	Yes	92.35	0.74	1.66	0.64	3.50	0.35	0.55	0.21
27	High K	Yes	93.84	0.00	0.95	0.55	2.54	0.20	1.56	0.36
06Part1	High K	No	68.39	7.45	0.00	2.00	11.27	2.42	2.54	4.23
06Part2	High K	No	60.51	7.77	5.47	1.75	10.17	6.11	2.21	4.55

From the data in the table, it can be seen that for high potassium glass, the subclasses are divided into 4 categories, and the main chemical components that determine the classification results are silica, potassium oxide, copper oxide, and phosphorus pentoxide.

For lead-barium glass, the subclasses are divided into 4 categories, and the main chemical components that determine the classification results are silica, aluminum oxide, lead oxide, and barium oxide.

Step5.Analysis of the reasonableness and sensitivity of subclass classification results.

(1) Rationalization analysis

In the Q-type cluster analysis, 2, 3, 4, 5, and 6 subcategory number division schemes were generated in this paper, and then the SPSSR software was used to conduct one-way ANOVA for these different subcategory number division schemes, and the optimal number of subcategory division was determined by comparing their significance p-values.

Tab 2. Significant P-values for determining the results of high potassium glass classification under different subclass classification conditions

Number of indicator categories k_j / Significance between groups P_i	j=2	j=3	j=4	j=5	j=6
P_a	0.168	<0.001	<0.001	<0.001	<0.001
P_c	0.774	<0.001	<0.001	<0.001	<0.001
P_h	0.862	0.475	0.002	0.005	<0.001
P_k	<0.001	<0.001	<0.001	<0.001	<0.001

Tab 3. Significant P-values for determining the results of high potassium glass classification under different subclass classification conditions

Number of indicator categories k_j / Significance between groups P_i	j=2	j=3	j=4	j=5	j=6
P_a	0.61	0.055	<0.001	<0.001	<0.001
P_c	<0.001	<0.001	<0.001	<0.001	<0.001
P_h	0.017	0.018	<0.001	<0.001	<0.001
P_k	0.218	<0.001	<0.001	<0.001	<0.001

When $P < 0.05$, that is when the number of subcategories is 4, the error is considered to be within a reasonable range and the classification result is reasonable.

(2) Sensitivity analysis

It is known that the more the number of indicators after dimensionality reduction is set, the more specific the category will be, which is reflected in the fact that the less relevant indicators will be separated to form a new category.

In this paper, the number of indicators after dimensionality reduction is 4, and the selected indicators are 1, 2, 7, and 8. These 4 indicators are not separated by the increase in the number of indicators after dimensionality reduction, which indicates that the 4 selected indicators have a strong correlation in these 4 categories, and the R clustering analysis for high potassium glass has good stability within a certain range.

The sensitivity analysis of the cluster analysis for lead-barium glass is similar to that of high-valent glass, and the number of indicators after the final set of dimensionality reduction is 4, and the selected indicators are 1, 5, 8, and 9. Using the above analysis method, these indicators have a strong correlation in these 4 categories, which indicates that the R cluster analysis for lead-barium glass has good stability within a certain range.

3.2. Classification and subclassification of unknown glass artifacts

Step1. Data pre-processing.

Here, the experimental data are normalized to the chemical composition ratio as in Eq. (3), so that the proportion of each component sums to 100%, to restorers arising from subsequent data processing and to facilitate the adaptation to the support vector machine model trained in Section 2.2. This is not repeated here.

Step2. Zero-mean normalization of the data.

After pre-processing, the data of the glass artifacts to be classified are normalized to zero-mean as in Eq(5). To obtain the normalized indicator variables for the support vector machine. It is not repeated here.

Step3. Calling the support vector machine model.

Since the support vector machine model trained in Section 2.2 has good discrimination accuracy, the saved support vector machine model is invoked directly for the identification of unknown glass artifacts to classify them. After calculating the values of its linear classification function, the results of the identification types are obtained as follows.

Tab 4. Unknown glass artifact identification results

Affiliation Type	High Potassium	Lead Barium	Lead Barium	Lead Barium	Lead Barium	High Potassium	High Potassium	Lead Barium
Artifact Number	A1	A2	A3	A4	A5	A6	A7	A8

Step4. sensitivity analysis test of classification results.

For the sensitivity of the prediction results of this model, scientific research methods such as control variables and transformation methods are used in this paper. The model was tested using the original data from the training model.

Firstly, we specify a progressive percentage from 100%, and in 14 cycles, each time a different chemical component is selected and expanded according to the given percentage, while the other indicators remain unchanged.

The expanded data set is then placed into the classification model and the misclassification rate is observed and analyzed for misclassification, and after 14 iterations are completed, the current percentage is incremented and the process continues.

After enough iterations, the indicator that was expanded when the first misclassification occurred was identified as silica, and it could be determined that this chemical component was the most sensitive in the model and was affected by data fluctuations that affected the results and triggered the greatest likelihood of misclassification. Thereafter, indicators such as lead oxide, magnesium oxide, copper oxide, iron oxide, barium oxide, and strontium oxide were also found in order. The chemical

components that are more advanced in the above ranking are more likely to cause misclassification by the model after being perturbed by the corresponding data.

Tab 5. Table of the order of each chemical composition

	Misjudgment rate									
component	pre=2	pre=2.5	pre=2.8	pre=3.0	pre=3.2	pre=3.5	pre=3.8	pre=4	pre=4.3	pre=4.5
(SiO_2)	0	0.029851	0.044776	0.059701	0.089552	0.089552	0.089552	0.104478	0.104478	0.104478
(Na_2O)	0	0	0	0	0	0	0	0	0	0
(K_2O)	0	0	0	0	0	0	0	0	0	0
(CaO)	0	0	0	0	0	0	0	0	0	0
(MgO)	0	0	0	0.014925	0.014925	0.014925	0.014925	0.014925	0.029851	0.029851
(Al_2O_3)	0	0	0	0	0	0	0	0	0	0
(Fe_2O_3)	0	0	0	0	0	0	0	0.014925	0.014925	0.014925
(CuO)	0	0	0	0	0	0	0.029851	0.044776	0.044776	0.044776
(PbO)	0	0	0.014925	0.044776	0.044776	0.074627	0.074627	0.074627	0.089552	0.089552
(BaO)	0	0	0	0	0	0	0	0	0.014925	0.014925
(P_2O_5)	0	0	0	0	0	0	0	0	0	0
(SrO)	0	0	0	0	0	0	0	0	0	0.029851
(SnO_2)	0	0	0	0	0	0	0	0	0	0
(SO_2)	0	0	0	0	0	0	0	0	0	0

4. Summary

In the analysis of unknown glass artifacts, this paper uses a support vector machine classification model to classify the data to be classified accurately and efficiently, which is good at the task of predictive classification in the case of a small amount of known data. However, in the problem of subclassification of glass categories, the determination of the chemical components that cause influence during the classification is derived from the analysis of R-cluster analysis and the analysis of existing data, which is somewhat subjective.

This paper establishes an excellent judgment classification model with data subclassification model, which has important promotion value in the fields of management classification of college students, identification of excavated cultural relics, judgment of potential customers for enterprise marketing, and division of factory product grade.

References

[1] Shoukui Si. Mathematical modeling algorithms and applications (3rd edition) [M]. Beijing: National Defense Industry Press, 2022 Reprint: 256.

[2] Yu Junbo,Song Yuntao,Guo Zhijuan,Xu Renting,Wang Qiaolin,Han Wei.Exploration of the role of R-type cluster analysis in regional chemical prospecting element grouping[J]. Computational Techniques of Physical and Chemical Exploration,2014,36(06):771-776.

[3] Wang Jie, Li Mo, Ma Qinglin, et al. A weathering study of an octagonal lead-barium glass vessel from the Warring States period[J]. Glass and Enamel,2014,42(2):6-13. doi:10.13588/j.cnki.g.e.1000-2871.2014.02.002.

[4] Chen X. H., Liu D. L., Tian Y. J., et al. Topologizable support vector classifier[J]. Journal of Intelligent Systems,2018,13(1):147-151. doi:10.11992/tis.201610019.