ISSN 2616-5880 Vol. 4, Issue 1: 42-47, DOI: 10.25236/AJMC.2023.040108

Component Analysis and Identification of Glass Products Based on Support Vector Machine Comprehensive Classification Model

Zibin Bi[#], Zhiqi Zhu[#], Zhenglong Ouyang[#], Yisong Huang[#], Ziheng Liu[#]

Department of Shihezi University, Shihezi, 832003, China
[#]These authors contributed equally.

Abstract: *It is of great significance to study the ancient Silk Road to analyze and identify the types and components of glass products. In this paper, support vector machine (SVM) method is used to study the statistical rule and internal relation between different chemical components of ancient glass products and how to classify ancient glass products. Firstly, the high-potassium glass and lead-barium glass were further divided into subclasses, and then the K-means cluster analysis was carried out. After many fitting repeats, the optimal number of classes was obtained as 5, and the chemical composition characteristics of the classified types were quantitatively analyzed and their ranges were divided. Finally, the five glass types were named as PbO- BaO- SiO_2, PbO (~)- BaO- SiO_2, K_2O- SiO_2- Al_2O_3, PbO- BaO (~)- SiO_2- CuO and K_2O- SiO_2 (~).Then the classification criteria of each category are given on the basis of K-means clustering, and then combined with the support vector machine model, the unknown categories of glass artifacts are classified by machine learning, and finally the final results are obtained by fitting the five ancient glass artifact subcategories based on multiple chemical compositions. And the result probability is predicted from the 5 subclasses, and the maximum probability is taken as the final predicted category.*

Keywords: *Glass Products, K-means Clustering, Machine Learning Classification, Support Vector Machine Model*

1. Introduction

The ancient Silk Road, with its ancient camel-bell paths to blend east and west, led to the introduction of early glass into China, which became a valuable physical evidence of trade exchanges[1]. Its main raw material is quartz sand, so the glass is extremely vulnerable to weathering by the burial environment, the Silk Road is the ancient cultural exchange between China and the West, of which glass is a valuable physical evidence of early trade exchanges. Early glass in West Asia and Egypt were often made into bead-shaped jewelry imported into China, our ancient glass absorbed its technology in the local production of local materials, so the appearance of glass products similar to foreign, but the chemical composition is not the same[2]. The main raw material of glass is quartz sand, the main chemical composition is silicon dioxide (SiO_2). Due to the high melting point of pure quartz sand, in order to reduce the melting temperature, it is necessary to add a flux when refining. In ancient times, fluxes such as wood ash, natural soda ash, saltpeter and lead ore were commonly used, and limestone was added as a stabilizer, which was converted to calcium oxide (CaO) after calcination. Adding different fluxes, its main chemical composition is also different. For example, lead barium glass is added to lead ore as a flux in the firing process, and its content of lead oxide (PbO) and barium oxide BaO is higher, which is usually considered to be our own invented glass species, and the glass of Chu culture is mainly lead barium glass. Potassium glass is made by using substances with high potassium content, such as grass wood ash, as a flux, and is mainly popular in Lingnan, Southeast Asia and India[3].

This study proposes to address the following questions. Based on the data collected in the Annexes, identify the classification patterns of high potassium and lead-barium types of glass; subclassify each category in relation to its chemical composition, and give specific methods and results; and conduct a rational and sensitivity analysis of the classification results. Based on the chemical composition data of the unknown category of glass artifacts, analyze them and identify the type to which they belong; and conduct a sensitivity analysis of the identification results.

2. Model building and testing

2.1 Analysis of the classification law of high potassium and lead barium glass

We first described the high potassium and lead-barium glass data statistically and obtained direct results followed by further in-depth analysis by chi-squared test.

The grouping analysis of the components was performed based on the columnar analysis and some of the results are shown in the following table. From Table 1, it can be seen that the high potassium oxide (K_2O) content in the high potassium glass is high.

Table 1: Potassium oxide (K_2O) content in high potassium and lead barium

Title	Name	Type		Total
		Lead Barium	High Potassium	
Potassium oxide (K_2O)_Outlier handling	0.34 ~ 3.338536585365853	6 (54.5%)	5 (45.5%)	11
	3.338536585365853 ~	30 (73.2%)	11 (26.8%)	41
	~ 0.34	17(100.0%)	0(0.0%)	17
Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively				

From Table 2, it can be seen that lead oxide (PbO) and barium oxide (BaO) contents in lead-barium glass are high. Then we performed chi-square test for each component of high potassium and lead barium to further obtain the variability of their chemical compositions. The results of the chi-square test are shown in the following table.

Table 2: Content of lead oxide (PbO) and barium oxide (BaO) in high potassium and lead barium

Title	Name	Type		Total
		Lead Barium	High Potassium	
Lead Oxide (PbO)_Outlier Handling	16.98 ~ 28.26603448275862	11(100.0%)	0(0.0%)	11
	28.26603448275862 ~ 40.24	13 (56.5%)	10 (43.5%)	23
	40.24 ~	17 (94.4%)	1(5.6%)	18
	~ 16.98	12 (70.6%)	5 (29.4%)	17
Barium oxide (BaO)_Outlier handling	10.29 ~	18(100.0%)	0(0.0%)	18
	6.1 ~ 9.834150943396228	16(100.0%)	0(0.0%)	16
	9.834150943396228 ~ 10.29	6 (33.3%)	12 (66.7%)	18
	~ 6.1	13 (76.5%)	4 (23.5%)	17
Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively				

Table 3: Results of chi-square test

	Type (standard deviation)		F	P
	Lead barium(n=53)	High potassium (n=16)		
Silicon dioxide (SiO_2) Outlier handling	21.076	17.328	1.794	0.185
Sodium oxide (Na_2O)_Outlier handling	1.053	0.345	0.810	0.371
Potassium oxide (K_2O)_Outlier handling	2.335	4.934	32.285	0.000***
Calcium oxide (CaO)_Outlier handling	1.613	3.084	26.363	0.000***
Magnesium oxide (MgO)_Outlier handling	0.384	0.339	0.019	0.890
Aluminum oxide (Al_2O_3)_Outlier handling	3.189	2.446	0.061	0.806
Iron oxide (Fe_2O_3)_Outlier handling	0.995	0.935	1.495	0.226
Copper oxide (CuO)_Outlier handling	2.343	1.344	1.406	0.240
Lead Oxide (PbO)_Outlier Handling	16.658	14.842	0.129	0.721
Barium oxide (BaO)_Outlier handling	7.888	4.201	1.273	0.263
Phosphorus pentoxide (P_2O_5)_Outlier handling	3.577	1.122	13.797	0.000***
Strontium Oxide (SrO)_Outlier Handling	0.207	0.134	0.883	0.351
Tin Oxide (SnO_2)_Outlier Handling	0.128	0.456	5.255	0.025**
Sulfur Dioxide (SO_2)_Outlier Handling	2.114	1.872	1.554	0.217

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively.

Table 3 demonstrates the results of the chi-square, including standard deviation, F-test results, and significance P-values. We mainly based on the P-value to develop further analysis, as shown in Table 10, the P-value of potassium oxide (K_2O) is 0.000***, which indicates that there is a significant difference between high potassium and lead-barium glass in this component, while the P-value of barium oxide (BaO) is 0.263, which indicates that there is also a more significant difference in this component. In addition, the results indicate that there are also significant differences in the content of calcium oxide (CaO), phosphorus pentoxide (P_2O_5), and tin oxide (SnO_2) components.

2.2 Subclassification of high potassium and lead-barium glass based on cluster analysis

After reviewing the relevant literature, cluster analysis provides a more scientific method for classifying ancient glass types. Therefore, we solved this problem by using the idea of cluster analysis to cluster each sample with each chemical composition in the high potassium and lead-barium glass types as variables.

We used SPSS software class category variability analysis to analyze the frequency of each cluster category according to the cluster summary, and the cluster set cluster annotation to determine which category each sample was assigned to, and finally we chose five categories whose results were better (Figure 1).

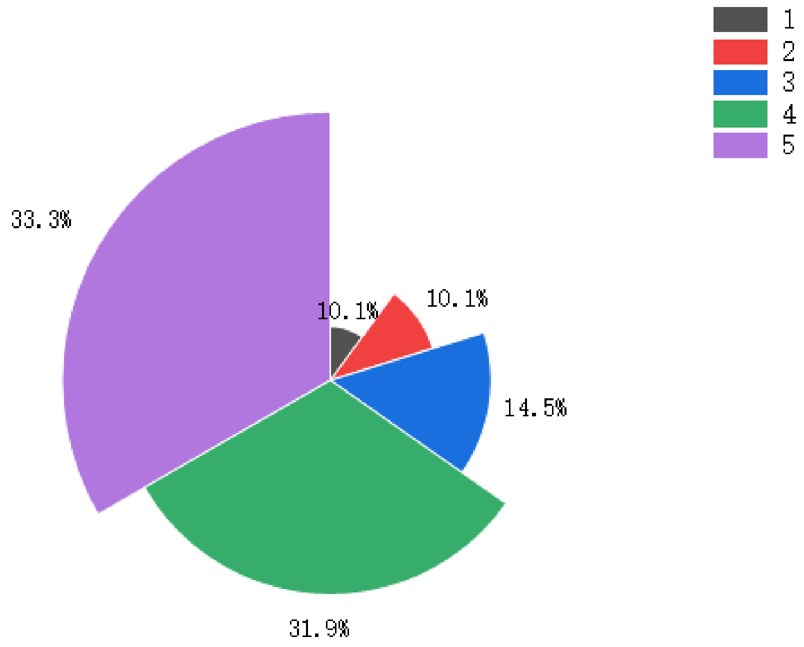

Figure 1: Clustering category summary 3D pie chart

Figure 1 shows the results of the model clustering in a visualized form, i.e., the percentages accounted for. Finally, we obtained the data set clustering annotation table and the clustering centroid coordinates table (both given in the supporting material), and we derived the centroid values of each chemical composition and combined them with the relevant literature information. The names for the five clustering categories mentioned above are as follows.

Z1: PbO-BaO-SiO_2

Z2: PbO (~)-BaO-SiO_2

Z3: K_2 O-SiO_2 -Al_2O_3

Z4: PbO-BaO (~)-SiO_2 -CuO

Z5: K_2O-SiO_2(~)

Z1, Z2, and Z4 categories are indicated, and in the lead-barium type glass, the glass is classified by the relative content of PbO, BaO, and SiO_2 Z1 indicates that in addition to PbO and BaO components, SiO_2 Z2 has a significantly higher proportion of PbO content compared to Z1, while the proportion of SiO_2 The percentage of PbO content in Z2 increases significantly compared to Z1, while the percentage of Si content decreases relatively. In Z4, the BaO content is significantly increased and the proportion of CuO content is also relatively increased. Z3, Z5 category indicates that in high potassium type glass, Z3 in addition to K_2O composition, SiO_2 and Al_2O_3 are both higher in proportion to the content of In contrast, Z5, compared to Z3, has a significantly higher SiO_2 content is significantly higher in Z5 compared to Z3.

3. Sensitivity and Accuracy Analysis of Cluster Analysis Method

We used cluster analysis to consider the various chemical compositions of high potassium and lead-barium type glasses as indicator variables, in order to further determine whether cluster analysis is reliable as a method for solving this problem. Therefore, here we do the sensitivity and accuracy analysis for each indicator variable. As can be seen from Figure 2, each indicator is in its vicinity relative to the reference line, and the sensitivity of each indicator variable can be considered average.

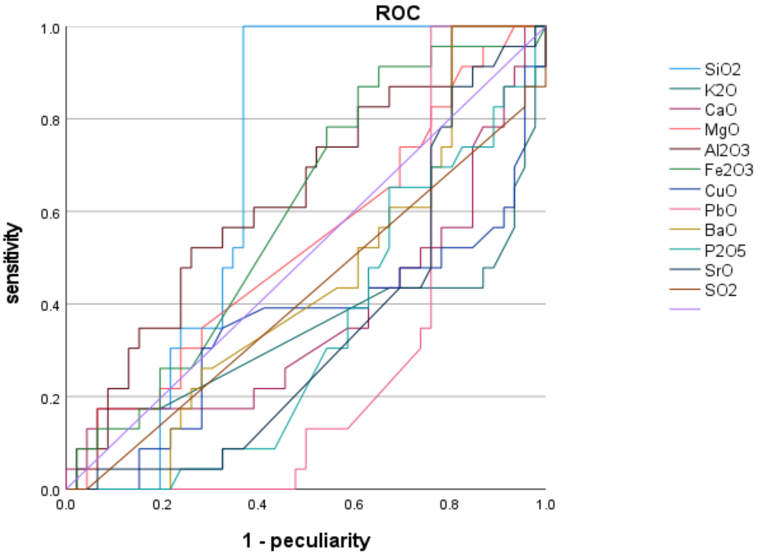

Figure 2: Sensitivity ROC curve for each indicator variable

As can be seen from Figure 3, each indicator is above the reference line and close to 1. This can indicate that the accuracy rate of each indicator variable is high, and the cluster analysis method can be considered reliable.

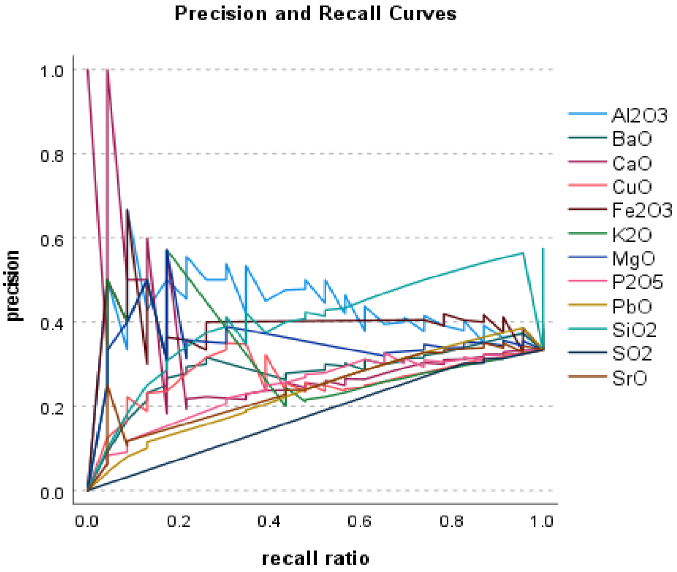

Figure 3: Accuracy recall curve of each index

3.1 Support vector machines perform classification regression on data

Support vector machines are based on the statistical structural risk minimization principle[4] .The optimal partitioning hyperplane between positive and negative classes is obtained by maximizing the soft interval. For the nonlinear case, kernel techniques can be used[5] For the nonlinear case, the kernel technique can be used to map the data in the low-dimensional space to the high-dimensional space, using the quadratic gauge.The technique perfectly solves the dual problem of the original problem to obtain the partition hyperplane, and this new machine learning method based on statistical theory successfully solves the problem of high dimensionality and local extrema that has always existed in machine learning.

By analyzing each chemical component in Annex Form 3, they were classified according to the five results (Z1, Z2, Z3, Z4, and Z5) obtained from the cluster analysis in 5.2.2. With Z1, Z2, Z3, Z4, and Z5 as the y-axis and the chemical composition of each sampling point as the x-axis, the predicted probabilities of the five types were finally given by the support vector machine model combined with the content of each chemical composition. The identification results are given in the support material (Form 3 Identification Results.xlsx). Table 4 shows some of the results.

Table 4: Identification results

Predicted Results_Y	Predicted outcome probability_1	Predicted outcome probability_2	Predicted outcome probability_3	Predicted outcome probability_4	Predicted outcome probability_5
Z2	0.317578482	0.445772003	0.047556679	0.151882917	0.037209919
Z1	0.438897614	0.040255633	0.391421312	0.063415123	0.066010317
Z1	0.455799738	0.04169112	0.234489319	0.044884792	0.22313503
Z1	0.458212367	0.039869095	0.258703045	0.040809996	0.202405498

3.2 Sensitivity and accuracy analysis of support vector machine methods

We use the same method to further analyze the support vector machine method, and finally obtain the indicator sensitivity ROC curve as well as the indicator precision recall curve as shown below.

From Figure 4 and Figure 5, we can see that the sensitivity and accuracy of each indicator variable is good. In addition, we obtain the following model review plots for the overall review of the support vector machine model. The value of a good model should be around 0.5, and it can be seen from Figure 4 that the support vector machine approach seems to be closer to a good model in general. It can indicate that the model is more reliable.

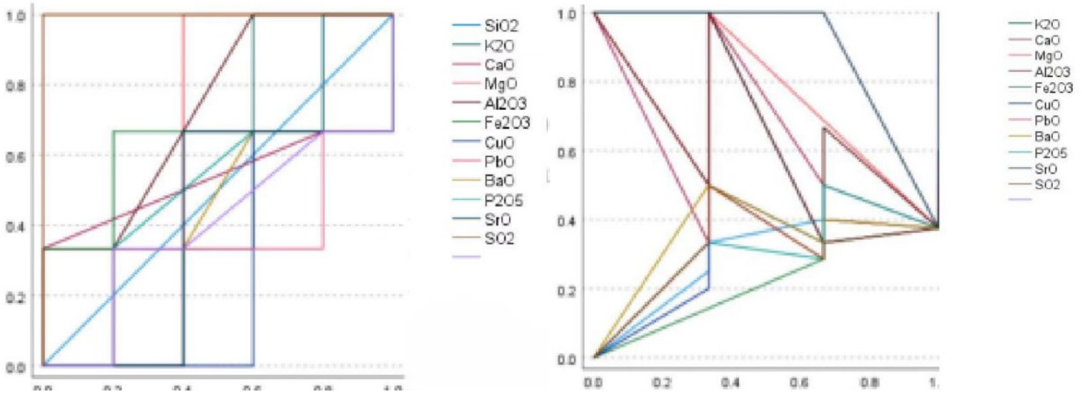

Figure 4: Sensitivity ROC curve of each indicator variable (Left) and Accuracy recall curve of each indicator (Right)

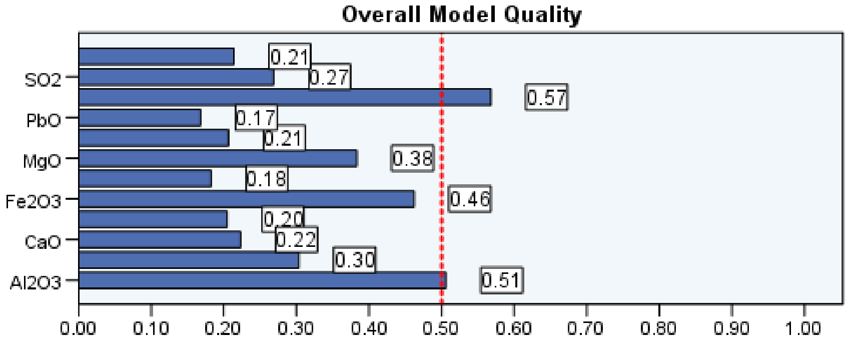

Figure 5: Model evaluation table

4. Conclusions

The model in this paper can be an effective model for evaluating systems with a large amount of unknown information, and is a comprehensive evaluation model combining qualitative and quantitative analysis. The model can better solve the problem that evaluation indexes are difficult to quantify and count accurately, and can exclude the influence brought by human factors, so that the evaluation results are more objective and accurate. The whole calculation process is simple, easy to understand, and easy for people to grasp; the data does not need to be normalized, and the original data can be used for direct calculation, which is reliable; the evaluation index system can be increased or decreased according to the specific situation; there is no need for a large number of samples, as long as a small number of representative samples can be. However, in the processing of data, the choice of standard normal distribution function as the weight function may be slightly inadequate. The machine learning integrated classification regression prediction model using K-means cluster analysis combined with support vector machine model is promoted, and the overall evaluation index is improved by 23.61% compared with the traditional support vector machine model. The model accuracy can be increased by combining independent sample t-test, method analysis, and chi-square test.

References

[1] Stanimirova I, Walczak B, Massart D L. Multiple factor analysis in environmental chemistry. Anal Chim Acta, 2005, 545: 1-12.

[2] Kieft IE, Jamieson D N, Rout B.PIXE cluster analysis of ancient ceramics from North Syria [J]. Nucl Instr Meth. Phys Res B, 2002, 190: 492-496.

[3] ZHANG Bin, CHENG Huan -- sheng,MA Bo.PlXE and ICP - AESanalysis of early glass unearthed frorm Xinjiang(China) [J]. Nucl InstrMeth, Phys Res B, 2005, 240: 559 -564.

[4] Harman G, Kulkarni S. Statistical learning theory and induction [M]//Learning from data: concepts, theory, and methods. 2nd ed. [S.l.]: John Wiley & amp; Sons, Inc, 2012: 3186-3188.

[5] Frank M, Wolfe P. An algorithm for quadratic programming [J]. Naval Research Logistics Quarterly, 1956, 3(1/2): 95-110.