

Ancient glass classification based on data mining methods

Tianhao Cao^{1,#}, Yuying Song^{2,#}, Jiayao Han³, Yong Liu^{4,*}

¹ Software College, Heilongjiang University, Harbin, Heilongjiang, 150080

² School of Physics, Changchun University of Science and Technology, Changchun, China, 130022

³ School of Economics and Business Administration, Heilongjiang University, Harbin, Heilongjiang, 150080

⁴ College of Electronic Engineering, Heilongjiang University, Harbin, Heilongjiang, 150080

* Corresponding Author Email: 1529476947@qq.com

#These authors contributed equally.

Abstract. Ancient glass is highly susceptible to weathering by the burial environment, and a large number of internal elements are exchanged with environmental elements, resulting in changes in its composition ratio, which affects the correct determination of its category. In this paper, we analyze the statistical pattern of the chemical composition content of the surface of artifact samples with and without weathering and predict the chemical composition content of artifacts before weathering by combining the types of glass. A random forest algorithm and a support vector machine algorithm are used to classify the glass, and a sensitivity analysis is performed, and the model is robust.

Keywords: Glass weathering; Glass classification; Random forest model; SVM model.

1. Introduction

The Silk Road was a channel for cultural exchange between China and the West in ancient times, of which glass was valuable physical evidence of early trade exchanges [1]. The early glass was often made into bead-shaped jewelry in the West Asian and Egyptian regions of our country, our ancient glass absorbed its technology and was made in situ locally, so the appearance of glass products with foreign similarities, 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 (SiO₂). Due to the high melting point of pure quartz sand, to reduce the melting temperature, it is necessary to add a flux during refining. The main chemical composition of the added flux differs from one flux to another [3]. Ancient glass is highly susceptible to weathering by the burial environment. During the weathering process, a large number of internal elements are exchanged with environmental elements, resulting in a change in their composition ratio, which affects the correct judgment of their category.

Since ancient glass is highly susceptible to weathering by the influence of the burial environment, a large number of internal elements are exchanged with environmental elements, which affects the correct judgment of its category. Therefore, in response to the above situation, we conducted the following research. Combining the relevant data of ancient glass products, this paper analyzes the relationship between the surface weathering of glass artifacts and their glass types, ornamentation, and color; combining the types of glass, we analyze the statistical patterns of the chemical composition content of artifact samples with and without weathering on their surfaces and predict their chemical composition content before weathering based on the weathering point detection data[4]. Analyze the classification pattern of high potassium glass and lead-barium glass; for each category select the appropriate chemical composition for its subcategory, give specific division methods and division results, and analyze the reasonableness and sensitivity of the classification results. To analyze the chemical composition of unknown categories of glass artifacts, identify the types to which they belong, and analyze the sensitivity of the classification results [5].

2. Methods

The relationship between weathering or not and glass type, ornamentation, and color was tested using the chi-square test and Spearman test [6], where both glass type and ornamentation are unordered categorical variables, and correlation analysis was performed using the chi-square test [7]. And color is an ordered categorical variable, so Spearman's test was used.

Analyze the classification law of high potassium glass and lead-barium glass, each glass has 14 features such as silica, sodium oxide, etc. Combined with the prediction of the change of chemical content of the weathered and unweathered glass, the unweathered chemical content of all data is obtained, after which the random forest model is used to process this multi-featured data for classification [8], and the number of parameter trees is set to 100 in Python, and the depth is 5. The known data were randomly divided into training and acceptance sets, and the training set was substituted into the model for training, and the acceptance set was substituted to obtain an OOB score and accuracy after obtaining the trained model.

This paper uses the K-means clustering algorithm in Python to cluster high potassium glass and lead-barium glass separately, and the contour coefficients are used to find the most suitable category number, and the high potassium glass is divided into 2 subclasses and the lead-barium glass is divided into 3 subclasses. The chemical components that have a greater impact are silica, potassium oxide, and aluminum oxide for high potassium glass, and sulfur dioxide, lead oxide, and sulfur dioxide for lead-barium glass. The data were then extracted according to these characteristics and divided into two tables, the high potassium glass subclass table, and the lead-barium glass subclass table, and then substituted into the SVM model [4,5] for training and prediction, and the error was controlled within 0.001, and the final calculation accuracy was also 100%, which was very effective.

After that, we performed the reasonableness and sensitivity analysis of the models, and the reasonableness reflected that several models described have reasonableness according to the accuracy of prediction. In the sensitivity analysis we mobilized the number of trees in the random forest, depth parameters, and then substituted the data into the model prediction, and found that the result fluctuation is not very obvious, but if we change the number of specific features, such as the content of silica will certainly have an impact on the result, because this feature has a large impact on the result. SVM model, we also changed the kernel function, function dimension, and kernel function constant term, and We found that the SVM model has a greater impact on the results in terms of function dimension and error limit because there are multiple features in this dataset, so a slightly larger function dimension is better, and the error limit directly affects the training results because a larger error model must not be good for classification. After sensitivity analysis, our model has good robustness.

The data to be predicted were substituted into the model and their classification was derived. Also, we introduced other irrelevant features with smaller initial values and put them into the model for classification prediction.

3. Results

3.1. Chemical composition prediction

A chi-square test was performed for glass type and weathering: the data were quantified with weathering set to 1 and unweathered set to 0. High potassium in glass type was set to 1 and lead barium was set to 2, resulting in a final result detection value = $0.009 < 0.05$, indicating a significant correlation between glass type and weathering degree.

The data were quantified with weathering set to 1 and an unweathered set to 0. The final result detection value = $0.862 > 0.05$ for A, B, and C in decoration, indicating that the correlation between decoration and weathering was not significant.

Spearman's rank correlation coefficient to study the correlation between color and weathering: color can be seen as continuous variable observation data transformed to get the rank data, so

Spearman's rank correlation coefficient can be used to conduct the study. The correlation coefficient is $0.067 > 0$, which indicates that there is a certain positive correlation between them. For the correlation coefficient, 0 means no correlation between them, -1 means negative correlation, and 1 means sign correlation, and the closer to -1 or 1, the stronger the correlation. 0.067 is close to 0, not close to -1, and not close to 1, which means the correlation is not strong. It proves that the correlation between color and weathering degree is not significant.

In summary, the ornamentation of the artifacts, color, and weathering are not correlated, and the glass type of the artifacts and weathering are correlated.

Combined with the type of glass, the statistical pattern of the content of chemical components with and without weathering on the surface of the artifact samples was analyzed, and the content of their chemical components before weathering was predicted based on the weathering point detection data". Lead and barium artifacts' weathered and unweathered elements average for comparison is shown in Figure 1.

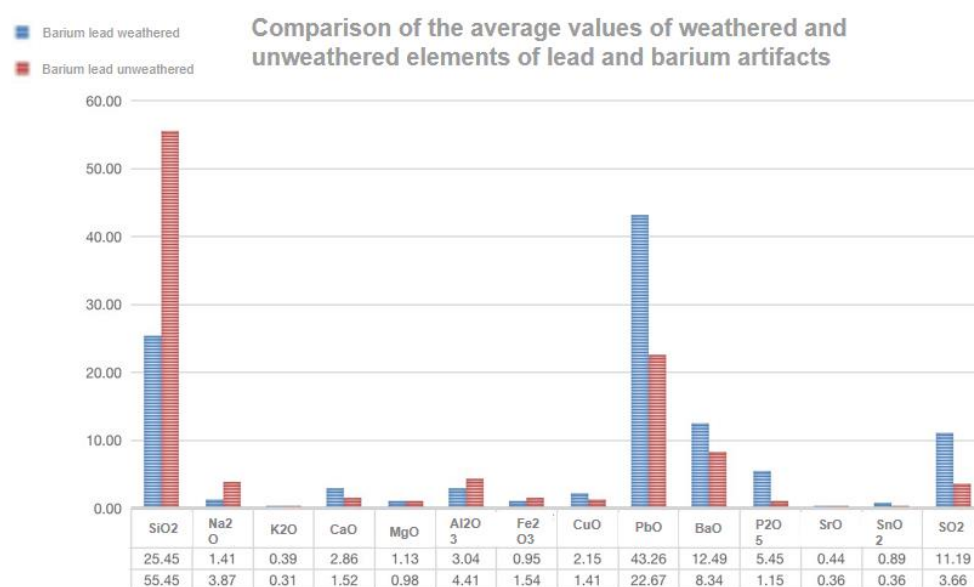


Figure 1. Comparison of the average values of weathered and unweathered elements of lead and barium artifacts

In Figure 1 we can see that the weathered artifacts of lead and barium class and unweathered artifacts in silica, sodium oxide, calcium oxide, lead oxide, barium oxide, phosphorus pentoxide, and sulfur dioxide on the content of a large difference: weathered artifacts than unweathered silica content decreased by 54.5%, sodium oxide content decreased by 63.6%, calcium oxide, lead oxide, barium oxide, phosphorus pentoxide, sulfur dioxide The content of weathered artifacts is significantly higher than that of unweathered artifacts, with a large difference of 47.9% increase in the content of lead oxide, 78.9% increase in the content of phosphorus pentoxide, and 67.3% increase in the content of sulfur dioxide.

Comparison of the mean values of weathered and unweathered elements for artifacts in the high potassium category in the same way.

There is a big difference between the weathered artifacts with high potassium and the unweathered ones in the content of silica, potassium oxide, calcium oxide, and aluminum oxide: the weathered ones have 27.7% more silica, 92.2% less potassium oxide, 86.4% less calcium oxide, and 70.8% less aluminum oxide than the unweathered ones. Meanwhile the unweathered has more elements of sodium oxide, iron oxide, lead oxide, and tin oxide than the weathered.

According to the weathering point test data, the average difference between the elemental content of different artifacts by the average data of elements and the same unweathered:

Let the average value of an element be X the content of the element in different artifacts of the same type be X_i then the absolute value of the average difference of n artifacts P_j is.

$$P_j = \frac{\sum_{i=1}^n |X_i - X|}{n} \quad (1)$$

Combined with the above analysis and conclusions can be predicted (calculated value of the last integer, if no single digit takes a decimal).

(1) for lead and barium glass artifacts, 46% to 65% of silica, 16% to 29% of lead oxide, 5% to 12% of barium oxide, 2% to 7% of aluminum oxide, 0 to 5% of sodium oxide, 0 to 3% of copper oxide, 0 to 3% of iron oxide, 0 to 3% of calcium oxide, 0% to 2% of phosphorus pentoxide, 0 to 1% of magnesium oxide, 0 to 0.5% of strontium oxide, 0 ~ 0.5% potassium oxide.

(2) for high potassium class glass artifacts, the artifacts before weathering have 61%~75% silica, 0~3% sodium oxide, 8%~12% potassium oxide, 5%~8% calcium oxide, 0~2% magnesium oxide, 5%~9% aluminum oxide, 0~3% iron oxide, 2%~5% copper oxide, 0~2% lead oxide, 0~2% barium oxide, 0.5%~3% phosphorus pentoxide, 0~0.1% strontium oxide, 0~0.5% sulfur dioxide.

3.2. Glass classification

We started the analysis based on the chemical composition ratios of the classified glass artifacts in the data, firstly, we eliminated the invalid data (composition ratios summed up not between 85% and 105%) to get the valid data, and secondly, we set the high potassium glass as "0", i.e., a negative sample, and the lead-barium glass as "1", i.e., a positive sample, according to the attached Form 1. "The data were then divided into training data and acceptance data. Subsequently divided into training data and acceptance data, the training data were put into the random forest model for training, the type was used as prediction data, and the rest were feature variables. After the training, we get: the OOB score of the random forest: 1.0, Accuracy on the training dataset: 1.0, and Accuracy on the validation dataset: 1.0. The accuracy is calculated by the formula:

$$Accuracy = \frac{TP+TN}{P+N} \quad (2)$$

Where: P (Positive Sample): the number of positive samples. N (Negative Sample): the number of negative samples. TP (True Positive): the number of positive cases correctly predicted. FP (False Positive): the number of negative cases predicted as positive. FN (False Negative): the number of positive cases predicted as negative cases. TN (True Negative): the number of negative cases correctly predicted.

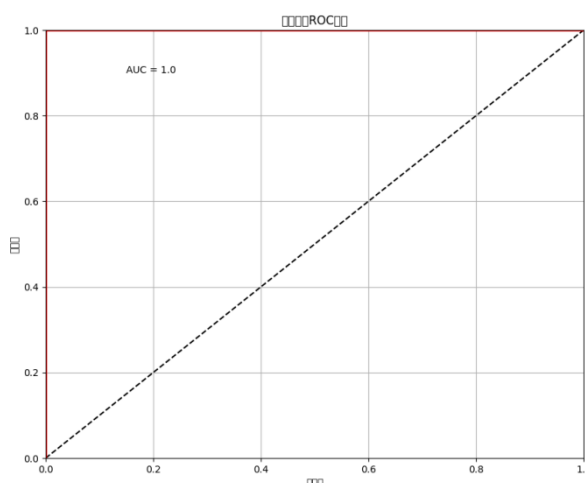


Figure 2. ROC curve

The size of the area enclosed by the ROC curve and the horizontal axis as shown in Figure 2 is called the AUC (Area Under the ROC curve) of the learner, and the closer the value is to 1, the better the algorithm model is, and this time the result ROC is 1, so the model is good.

To select the appropriate chemical composition for each category for its subclassification, we first applied the cluster analysis model to solve the problem. In the first step, Annex 2 data were divided into "high potassium" and "lead-barium" to obtain two data sets. The clustering analysis was

performed separately for the two data sets. Since each indicator in the data set is of the same magnitude, no standardization is required.

With n glasses, the data of p indicators for n glass samples form a data matrix as follows.

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \quad (3)$$

Each row represents a glass sample, each column represents an indicator, and x_{ij} represents the observation of the i th sample about the j th indicator. The basic idea of cluster analysis is to define the distance between samples and the similarity coefficient between indicators, and the similarity coefficient between samples inscribes the similarity between indicators. The samples (or variables) are grouped one by one according to the magnitude of similarity, and the closely related ones are clustered into a smaller category, and the distant ones are clustered into a larger category.

The data are substituted for cluster analysis, and the contour coefficients with different K values are also derived. The following figure 3 shows.

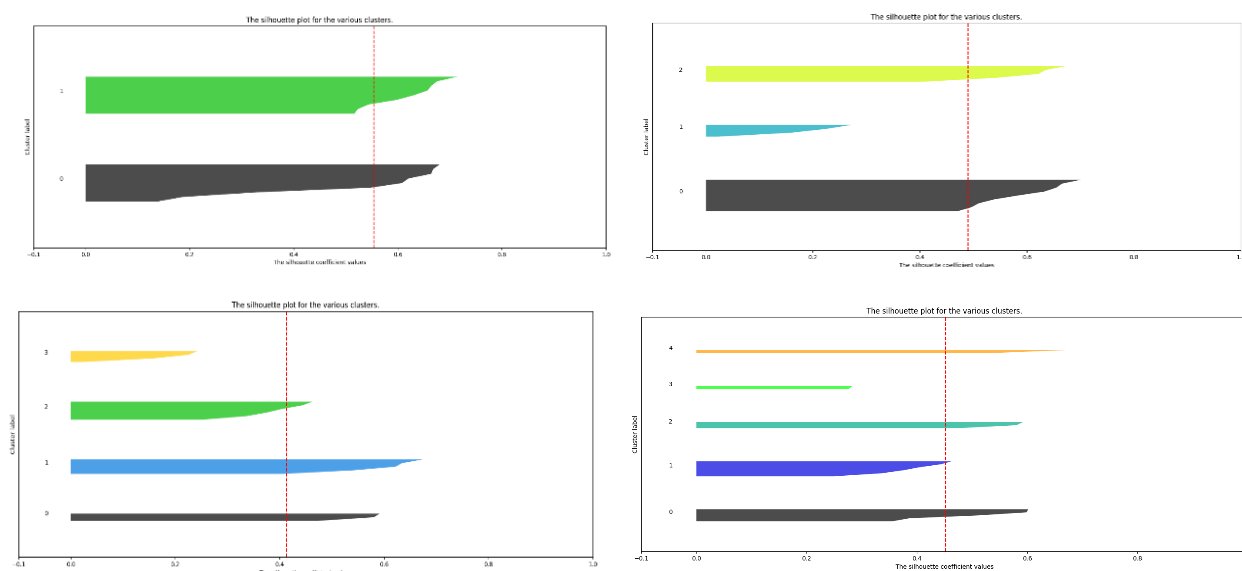


Figure 3. High potassium class profile coefficients

Contour coefficient of 0.55 for high potassium glass K of 2; 0.48 for high potassium glass K of 3. a contour coefficient of 0.41 for high potassium glass K of 4; a contour coefficient of 0.45 for high potassium glass K of 5.

Similarly, the contour coefficient of lead-barium glass K for 2 is 0.38; the contour coefficient of lead-barium glass K for 3 is 0.43. the contour coefficient of lead-barium glass K for 4 is 0.39; the contour coefficient of lead-barium glass K for 5 is 0.39

From the above analysis, it can be seen that high-potassium glass can be divided into 2 subclasses and lead-barium glass can be divided into 3 subclasses. After dividing the subclasses, we averaged the chemical composition of all samples of each subclass.

According to the chemical composition table of the lead-barium class, we can see that we can continue to classify lead-barium glass according to sulfur dioxide, lead oxide, and sulfur dioxide, and those that do not contain sulfur dioxide and the ratio of silica to lead oxide is about 1:2 will be classified as lead-barium A, those that contain sulfur dioxide and the ratio of silica to lead oxide is 3:1 will be classified as lead-barium B, and those that contain sulfur dioxide and the ratio of silica to lead oxide is about 1:1 will be classified as lead-barium C. Of course, this is only a general analysis based on the chart. Of course, this is only a general analysis based on the chart, and now we will

analyze the three features based on silica, lead oxide, and sulfur dioxide by substituting them into the SVM support vector machine model. The error is controlled within 0.001, and the final prediction accuracy Accuracy is 100%

$$Accuracy = \frac{Y}{N} \quad (4)$$

where Y is the number of correctly predicted samples and N is the total number of samples.

According to the chemical composition table of the high potassium class, it can be seen that the high potassium glass can be classified according to silica, potassium oxide, and alumina, and the ratio of silica to potassium oxide to alumina is about 9:1:1 for high potassium class A, and the ratio of silica to potassium oxide to alumina is about 30:1:1 for high potassium class B.

We substituted the composition table of the chemical subclasses of lead-barium glass and the chemical subclasses of high-potassium glass into the model to obtain an accuracy Accuracy of 100%.

The plausibility analysis has been given above and we used a random forest model with an accuracy of up to 100% and a subclass division model of up to 95%. Now that we have a data matrix with behavioral glass types, listed as chemical components, we vary the number of trees in the random forest model and the maximum depth of the trees and observe the change in accuracy as the Figures 4 and 5.

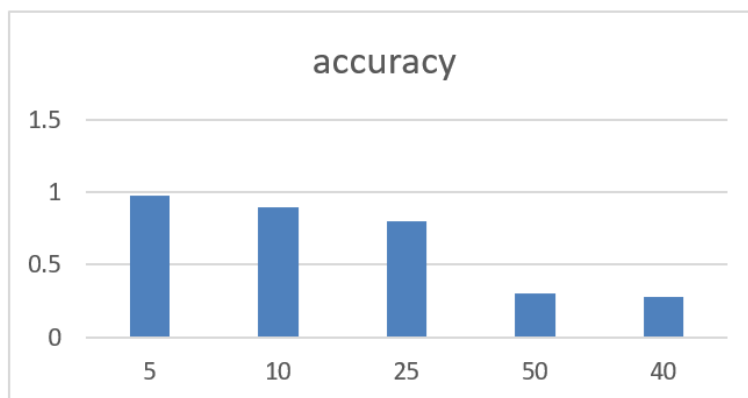


Figure 4. Changing the number of trees accuracy change

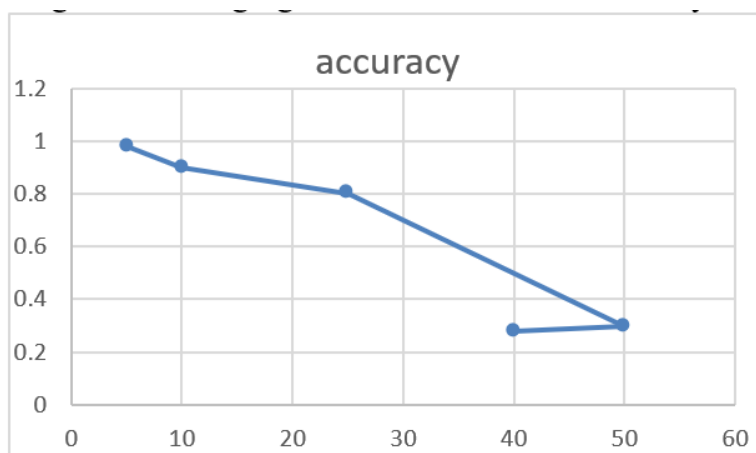


Figure 5. Changing the number of trees accuracy changes the graph

It can be seen that the model is insensitive to changes in the number of trees and the depth of trees, but if the chemical components of any glass type are changed, it produces different prediction results because the model itself is a model based on these characteristics.

In the subclass partitioning model: we adjusted the main adjustable parameters C, kernel, degree, gamma, and coef0 and found that the predicted results were the same for a small range of mobilization, for the obvious reason that we controlled the error to 0.001, and if we changed the limiting error to a larger value, we would find that the accuracy would also decrease.

3.3. Identification of glass types

First, the data are substituted into the random forest model to obtain the initial classification such as in figure 1.

Table 1. Initial classification

Artifact sampling points	SiO ₂	Na ₂ O	K ₂ O	CaO	MgO	Al ₂ O ₃	Fe ₂ O ₃	CuO	PbO	BaO	P ₂ O ₅	Sr	SnO ₂	SO ₂	Surface weathering	Initial classification(0 is high potassium,1 is lead barium)
A1	78.45			6.08	1.86	7.23	2.15	2.11			1.06	0.03		0.51	No weathering	0
A2	37.75			7.63		2.33			34.3		14.27				weathering	1
A3	31.95		1.36	7.19	0.81	2.93	7.06	0.21	39.58	4.69	2.68	0.52			No weathering	1
A4	35.47		0.79	2.89	1.05	7.07	6.45	0.96	24.28	8.31	8.45	0.28			No weathering	1
A5	64.29	1.2	0.37	1.64	2.34	12.75	0.81	0.94	12.23	2.16	0.19	0.21	0.49		weathering	1
A6	93.17		1.35	0.64	0.21	1.52	0.27	1.73			0.21				weathering	0
A7	90.83		0.98	1.12		5.06	0.24	1.17			0.13			0.11	weathering	0
A8	51.12	0.00	0.23	0.89	0.00	2.12	0.00	9.01	21.24	11.34	1.46	0.31	0.00	2.26	No weathering	1

Subsequent substitution into the SVM model yields subclasses as in Table 2.

Table 2. Subcategory classification

Artifact sampling points	SiO ₂	Na ₂ O	K ₂ O	CaO	MgO	Al ₂ O ₃	Fe ₂ O ₃	CuO	PbO	BaO	P ₂ O ₅	Sr	SnO ₂	SO ₂	Surface weathering	Initial classification(0 is high potassium,1 is lead barium)	Subclass division
A1	78.45			6.08	1.86	7.23	2.15	2.11			1.06	0.03		0.51	unweathered	0	High potassium class B
A2	37.75			7.63		2.33			34.3		14.27				weathered	1	Lead and barium class B
A3	31.95		1.36	7.19	0.81	2.93	7.06	0.21	39.58	4.69	2.68	0.52			unweathered	1	Lead and barium class A
A4	35.47		0.79	2.89	1.05	7.07	6.45	0.96	24.28	8.31	8.45	0.28			unweathered	1	Lead and barium class B
A5	64.29	1.2	0.37	1.64	2.34	12.75	0.81	0.94	12.23	2.16	0.19	0.21	0.49		weathered	1	Lead and barium class B
A6	93.17		1.35	0.64	0.21	1.52	0.27	1.73			0.21				weathered	0	High potassium class B
A7	90.83		0.98	1.12		5.06	0.24	1.17			0.13			0.11	weathered	0	High potassium class B
A8	51.12	0.00	0.23	0.89	0.00	2.12	0.00	9.01	21.24	11.34	1.46	0.31	0.00	2.26	unweathered	1	Lead and barium class B

Analysis of tuning parameters: Other irrelevant features were introduced into the data, and a small initial value was set and put into the model for classification, and it was found that the classification results were consistent, and the resulting error variation was also very small, only about 0.5, which shows that the model has good robustness.

3.4. Discussion

The accuracy of the random forest model is up to 100%, and the subclass division model is up to 95%; the sensitivity analysis of the model has been conducted, and the error variation derived is also very small, only about 0.5, which shows that the model has good robustness; the elemental content of unweathered artifacts predicted by the mean deviation method cannot predict more accurate values, and the artifacts of different processes differ greatly in some elements, and such elements are difficult to predict.

4. Conclusion

(1) this paper analyzes the relationship between the surface weathering of glass artifacts and their glass type, decoration, and color; combined with the type of glass, analyze the statistical law of the surface of artifact samples with and without weathering chemical composition content, and based on the weathering point detection data, predict their chemical composition content before weathering.

(2) Analyze the classification pattern of high potassium glass and lead-barium glass; for each category, select the appropriate chemical composition to classify them into subcategories, use SVM

and random forest algorithm to give specific classification methods and classification results, and analyze the rationality and sensitivity of the classification results.

(3) The chemical composition of unknown categories of glass artifacts was analyzed to identify the types to which they belonged, and the sensitivity of the classification results was analyzed.

References

- [1] Yang Jing. Cultural Exchange between China and the West from the Perspective of Foreign Glass Products on the Silk Road[J]. Ethnic Art Forest, 2016(1):6.
- [2] Hou Xiumin. Interpretation of Two Kinds of Glass Products Unearthed in Luoyang: Warring States Glass Beads and Eastern Han Dynasty Glass Bottles[J]. Wealth Management: Collection, 2018(7):4.
- [3] Tan Shengguang, Qin Chunlei. Introduction: Colorful Colors: The Art of Glass in the Communication of Ancient East and West^Ming[J]. Art Collection, 2022(3):4.
- [4] Su Tianming, Sun Qiang, Zhang Weiqiang. Sandstone weathering and its engineering geological effects[J]. Geological Science and Technology Information, 2015, 34(1):6.
- [5] Ren Yubo, Wen Rui, Xian Yiheng, et al. Study on chemical composition and production process of glass earring excavated from the tomb of Zihan in Yingcheng[J]. Journal of Cultural Relics Conservation and Archaeological Sciences, 2022, 34(3):10.
- [6] Hao SHL, Gao YQ, Liu LK, Wang HX, Li YF, Li J. Study on the correlation between Chinese medical evidence of radiation proctitis and stereoscopic images based on Spearman's rank correlation analysis [J]. Shanxi TCM. 2020(03):51-53.
- [7] YANG De-Hai, ZHAO Wei-Jin, XIE Yi-Yan, PENG Ren, PU Song, WANG Tian-Long, FU Xi-Hua, ZHENG Shi-Fang, ZHANG Xiao-Long, LI Xiao-Ting. Correlation analysis of physical and chemical properties of tobacco planting soil and chemical composition of tobacco leaves in Dali Prefecture[J]. China Soil and Fertilizer, 2022(01):97-103.
- [8] Zhong Q, Luo Z, Qi Shuhua. Sensitivity analysis of sample size for extraction of citrus orchards by random forest classification algorithm[J]. Jiangxi Science, 2019, 37(05): 664-669. DOI: 10.13990/j.issn1001-3679.2019.05.003. Journal (Natural Science Edition), 2013, 25(04):85-89.
- [9] Chen Mei-Gu, Lin Xing-E, Li Xin-Guo, Liu Saki-Di, Gao Hong-Mao, Ming Jian-Hong, Dai Min-Jie, Zhou Zhao-Xi. A comprehensive evaluation of durian quality based on principal component analysis and cluster analysis. Food Industry Science and Technology
- [10] Chen N., Shen Z. Min. Tea quality identification based on chemical composition detection and SVM classification [J]. Anhui Agricultural Science, 2010, 38(15): 7851-7852. DOI: 10.13989/j.cnki.0517-6611.2010.15.110.