Glass Classification and Identification based on Decision Tree and Random Forest Classification Models

Zhipeng Zhang †, *, Tong Wang †, Xinpeng Wang †

School of Computer and Communication Engineering, Dalian Jiaotong University, Dalian, Liaoning, 116045, China

* Corresponding author email: 1478606041@qq.com

†These authors contributed equally

Abstract. Glass artifacts are weathered due to age and improper preservation, etc. This paper analyzes the classification laws of high potassium glass and lead-barium glass based on the data; for each category, suitable chemical compositions are selected to classify them into subcategories, and firstly, a decision tree classification model is used to analyze the classification laws of the two categories of glass by substituting the data set for machine learning, and specific classification methods and classification results are given. And the chemical composition of the unknown category of glass artifacts is analyzed, and the random forest classification model is used to predict the training set and test set by.

Keywords: Decision Tree Classification; Machine Learning; Random Forest Classification; Glass Artifacts.

1. Introduction

The Silk Road was an important route for trade between Asian, European and African countries in ancient times. The Silk Road was not only a trade route, but also an important road for the spread of culture and technological exchange. Among them, glass products were loved by various countries because of their special craftsmanship and exquisite shapes. After West Asia and other countries introduced glass products into China, China absorbed their technology and made them with local materials, so our glass products will have different chemical composition compared with other countries [1-4].

The main raw material of glass is quartz sand, which has a melting point of 1750°C. In ancient times, it was not possible to reach this temperature, so it was chosen to add co-solvents such as lead oxide, grass ash and saltpeter to lower the melting point. The co-solvents for making glass products differed from time to time, such as K2O, CaO, and SiO2 from the Spring and Autumn Period to the Warring States Period, BaO, PbO, and SiO2 from the Warring States Period to the Eastern Han Dynasty, and PbO and SiO2 from the Eastern Han Dynasty to the Tang Dynasty [5-7].

Glass products are weathered due to age and improper preservation, etc. Weathering of glass is the loss of crystalline water of glass components in the natural state, which makes glass fragile, reduces light transmission, easily produces cracks, and scale-like flaking. In this process, the chemical composition and proportion of glass products will change, which in turn affects the identification of glass products: such as marked as non-weathered artifacts may also have weathered parts, while marked as weathered artifacts may also exist unweathered parts [8-10].

2. Build a Glass Classification Model

2.1 Machine Learning - Decision Tree Classification

Decision tree classification: Starting from the root node, a feature of the instance is tested, and the instance is assigned to its child nodes according to the test results, at which time each child node corresponds to a value taken for that feature, and so on recursively, the instance is tested and assigned until it reaches the leaf node, and finally the instance is divided into the classes of the leaf nodes.

For analyzing the classification laws of high potassium glass and lead-barium glass, the processed data are integrated, with high potassium type and lead-barium type coded as 1 and 2, and the corresponding chemical composition content (due to the large number of tables, they are placed in the appendix), as the training set and test set, and substituted into the MATLAB decision tree classification program to derive the classification tree graph, and the predicted results are compared with the true values, from which the accuracy rate is derived, while A confusion matrix visualization is performed and used to summarize the classification model prediction results.

2.2 K_Means Clustering Analysis

K_Means algorithm, also known as k-means algorithm, k-means algorithm in the k-means represents the clusters of k clusters means to take the mean of the data values in each cluster as the center of the cluster, or called the center of mass, that is, the description of the cluster with the center of mass of each class.

The idea of the algorithm is as follows: first, k samples are randomly selected from the sample set as cluster centers, and the distance between all samples and these k "cluster centers" is calculated, and for each sample, it is divided into the cluster with the closest "cluster center", and for the new "cluster centers" of each cluster are calculated for each new cluster.

First, by comparing the differences in chemical composition before and after weathering of the same type, we selected the top eight chemical components for each of the two types of high potassium glass and lead-barium glass, respectively, and for the high potassium type we selected the following chemical components: NaO, KO, CaO, MgO, AlO, FeO, BaO, and P2O5; for the lead-barium type we selected the main chemical components. Silicon dioxide, sodium oxide, calcium oxide, copper oxide, lead oxide, phosphorus pentoxide, strontium oxide, sulfur dioxide (see appendix for details of chemical composition difference value data). Then K_Means clustering analysis was performed on the content of the eight components separately.

2.3 Elbow Rule and Contour Factor

Elbow Method

We know that k-means takes minimizing the squared error between sample and prime as the objective function, and the sum of squared distance error between prime and sample points within each cluster is called distortion degree (distortions). Then, for a cluster, the lower its distortion degree, the tighter the cluster members are, and the higher the distortion degree, the looser the cluster structure is. The degree of distortion decreases as the category increases, but for data with a certain degree of differentiation, the degree of distortion improves greatly when a certain critical point is reached, and then decreases slowly, and this critical point can be considered as the point with better clustering performance.

Silhouette Coefficient

For a clustering task, we want to obtain clusters that are as close as possible within the clusters and as far as possible between the clusters.

Silhouette Coefficient:

$$S(i) = \frac{b(i)-a(i)}{max\{a(i)-b(i)\}} \tag{1}$$

Among them, a(i) represents the cohesiveness of the sample points and is calculated as follows:

$$a(i) = \frac{1}{n-1}\sum_{j\neq i}^{n} distance(i,j) \tag{2}$$

The mean value of S(i) of all samples is called the contour coefficient of clustering result, defined as S, which is a measure of whether the clustering is reasonable and effective. The value of the contour coefficient of the clustering result is between (-1,1), and the larger the value, the closer the similar

samples are to each other, and the farther the different samples are from each other, the better the clustering effect.

The data of the 8 chemical components picked from the high potassium type and the data of the 8 chemical components picked from the lead-barium type were substituted into the MATLAB elbow rule program respectively, and the best number of clusters could be derived; then the data and the number of clusters were substituted into the contour coefficient program, and the contour coefficient graph was drawn to visually evaluate the good or bad clustering effect.

2.4 Machine Learning - Decision Tree Classification Results

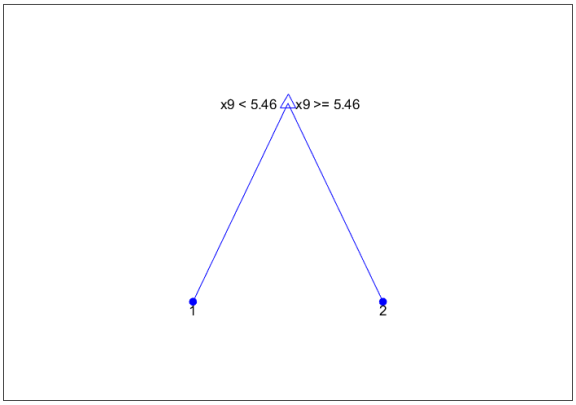

Figure 1. Decision tree classification chart

From Figure 1, the decision variable for the classification of high potassium and lead barium is: x9, which is the lead oxide content, when the lead oxide content is less than 5.46, it is high potassium glass; when the lead oxide content is greater than or equal to 5.46, it is lead barium glass

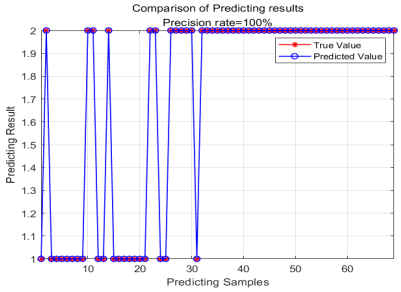

Figure 2. Comparison of prediction results

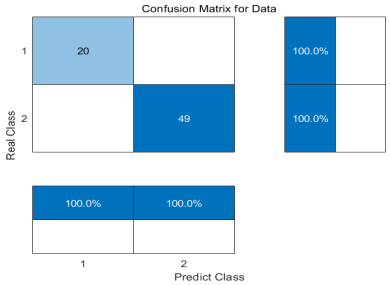

Figure 3. Confusion matrix

From the figure 2 and 3, the prediction results are accurate and the decision tree classification model is accurate. This results in the correct decision variables and a reasonable model.

2.5 K_Means Clustering Analysis Results

The results are shown in the figure 4 and 5, table 1 and 2.

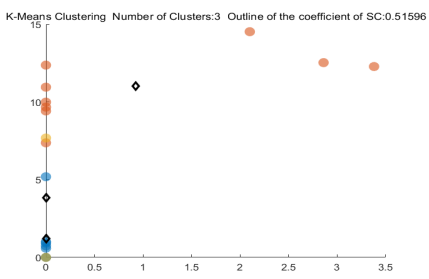

Figure 4. High potassium clustering

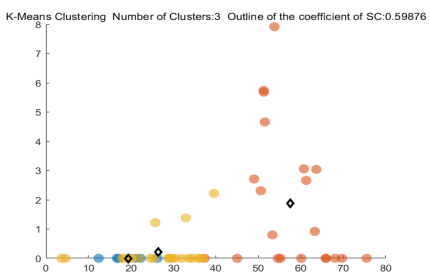

Figure 5. Lead-barium clustering

Table 1. Cluster labeling of high potassium class data sets

Clustering type	Na2O	K2O	CaO	MgO	Al2O3	Fe2O3	BaO	P2O5
1	0	9.99	1.07	0.87	3.93	1.74	0	1.17
1	0	5.19	0.62	0	4.06	0	0	0.66
1	0	12.37	0.21	1.11	5.5	2.16	2.86	0.7
1	0	9.67	0.72	1.56	6.44	2.06	0	0.79
1	0	10.95	1.66	1.77	7.5	2.62	0	0.94
4	0	7.37	0.94	1.98	11.15	2.39	1.38	4.18
4	0	7.68	6.32	1.73	10.05	6.04	0.97	4.5
2	0	0	2.01	0	1.98	0.17	0	0.61
2	0	0.59	5.87	0	1.32	0.32	0	0.35
2	0	0.92	7.12	0	0.81	0.26	0	0

Table 2. Cluster labeling of lead-barium class data sets

Clustering type	SiO2	Na2O	CaO	CuO	PbO	P2O5	SrO	SO2
1	36.28	0	2.34	0.26	47.43	3.57	0.19	0
1	20.14	0	1.48	10.41	28.68	3.59	0.37	2.58
3	4.61	0	3.19	3.14	32.45	7.56	0.53	15.03
1	33.59	0	3.51	4.93	25.39	9.38	0.37	0
1	29.64	0	2.93	3.51	42.82	8.83	0.19	0
2	37.36	0	0	4.78	9.3	5.75	0	0
2	53.79	7.92	0.5	2.99	16.98	0	0.33	0
1	31.94	0	0.47	8.46	29.14	0.14	0.91	0
2	50.61	2.31	0.63	1.12	31.9	0.19	0.2	0
1	19.79	0	1.44	10.57	29.53	3.13	0.45	1.96

2.6 Elbow Law and Contour Factor Results

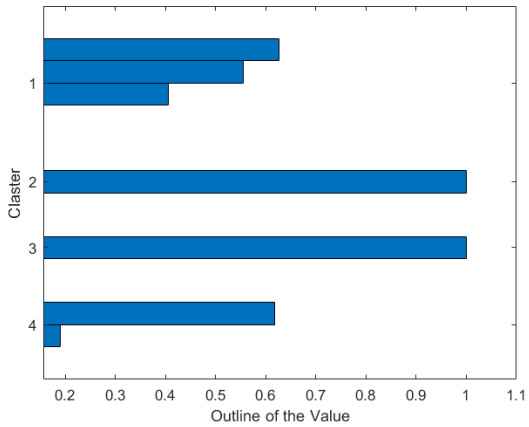

Figure 6. High potassium clustering profile coefficients

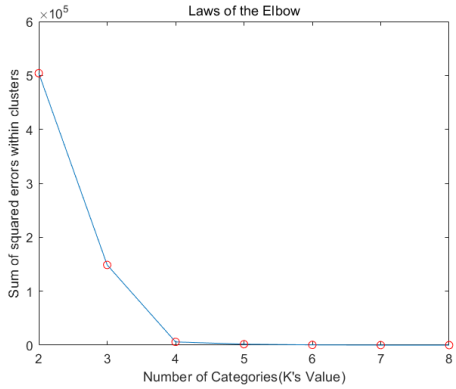

Figure 7. High potassium clustering elbow rule

From the above figures 6 and 7, the optimal number of clusters for high potassium clustering is 4 and the values of contour coefficients are greater than 0. Therefore, the results of high potassium clustering are reasonable.

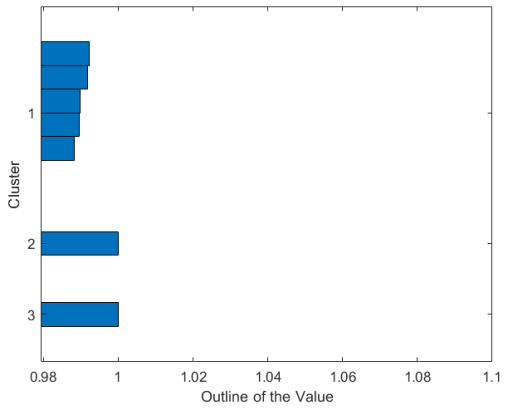

Figure 8. Lead-barium clustering profile coefficient

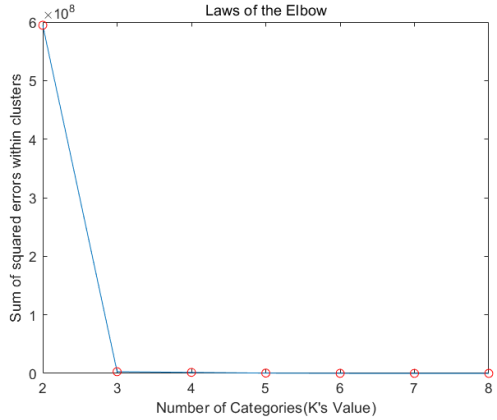

Figure 9. Lead-barium clustering elbow rule

From the above two figures 8 and 9, the best number of clusters for lead-barium clustering is 3 and the values of contour coefficients are greater than 0. Therefore, the results of lead-barium clustering are reasonable.

3. Build a Glass Identification Model

3.1 Machine Learning - Random Forest Classification Model

Random forest is composed of multiple decision trees, each of which is not the same. In building the decision tree, we randomly select a portion of the samples from the training data that we put back, and we do not use all the features of the data, but randomly select some of the features for training. Each tree uses different samples and features, and the training results are not the same.

Our data were substituted into MATLAB as samples, and the confusion matrix and chemical composition-importance graphs for the training and test sets were plotted to visualize which chemical components were more important, and predictions were made for the training and test sets, and the predicted results were plotted against the true values, along with the accuracy rates, and the number of decision trees-error curves were plotted to show that as the number of decision trees increased, the trend of error magnitude.

Finally, the data required to be predicted by the problem are substituted into the program for prediction and the classification results are derived.

3.2 Machine Learning - Random Forest Classification Model Results

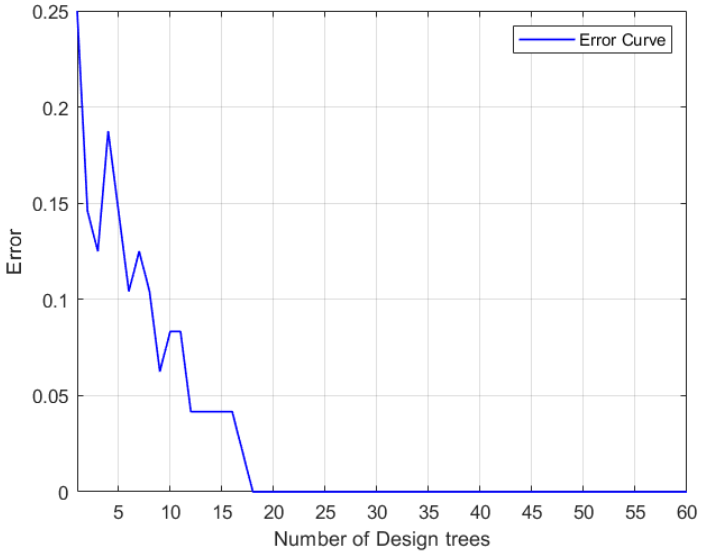

Figure 10. Number of decision trees - error

From the above figure 10, it can be obtained that the random forest prediction model has a high accuracy rate for the prediction result of this question, and the model is reasonable, so the prediction of the sample in the question can be carried out, and the prediction result is as follows table 3:

Table 3. Sample prediction results

Artifact Number	A1	A2	A3	A4	A5	A6	A7	A8
Category	High potassium glass	Barium lead glass	Barium lead glass	Barium lead glass	Barium lead glass	High potassium glass	High potassium glass	Barium lead glass

4. Conclusion

In this paper, firstly, we used decision tree classification model to analyze the classification laws of two types of glass by substituting the data set for machine learning, and the results were obtained as high potassium glass when the lead oxide content was less than 5.46, and lead-barium glass when it was greater than or equal to 5.46. Finally, the eight chemical components with significant changes before and after weathering in the two types of glass were screened according to the table of chemical component difference values, and the eight chemical components were analyzed by K_Means clustering, and the reasonableness of the model was evaluated by the elbow rule and the contour coefficient, and the results were obtained as the best number of clusters for the high potassium class was 4 and for the lead-barium class was 3. The sensitivity was detected by changing the number of screened chemical components.

To identify the glass types, a random forest classification model was subsequently used, and the accuracy was 100% by predicting the training and test sets, followed by substituting the samples in the problem for prediction, and the results were: A1, A6, and A7 for high potassium glass; A2, A3, A4, A5, and A8 for lead-barium glass. At the same time, the accuracy of the BP neural network classification model optimized by the particle swarm algorithm was lower than that of the random forest classification model for the training and test sets.

References

[1] Wang Xiaofu. The Silk Road: The Early Exchange between Oman and China: A Reply to the Question of the Silk Road[J]. Journal of Tsinghua University (Philosophy and Social Science Edition), 2020, 35 (04): 1-14+211. DOI: 10.13613/j.cnki.qhdz.002956.

[2] Tan X. The transmission and influence of Islamic glass on the Silk Road (8th-16th centuries) [D]. Jinan University, 2020. DOI: 10.27167/d.cnki.gjinu.2020.000412.

[3] Qi Haoyue. Exploration of foreign cultural factors cultural relics in Liao Dynasty in the context of Silk Road [D]. Inner Mongolia Normal University, 2020. DOI:10. 27230/ d. cnki. gnmsu.2020.000051.

[4] Ye Liqun. The historical role of the Liaoxi section of the Grassland Silk Road from foreign artifacts and other perspectives[J]. Journal of Tonghua Normal College,2020,41(03):80-86. DOI: 10.13877/ j. cnki. cn22-1284.2020.03.013.

[5] Zhou Renqin. The discovery and research of foreign ornaments in Hepu[J]. Cultural Identification and Appreciation, 2020(03):34-37.

[6] Cheng Yajuan. The evolution of Buddhism in the process of glass transmission on the Silk Road[J]. Journal of Nanjing Art Institute (Art and Design),2018(05):84-93+210.

[7] Tian Ye. The "Silk Road on Ice" sets sail from Yamal [J]. China Oil Enterprise,2018(08):27-32+2.

[8] Xiong Zhaoming. Archaeological discoveries of Hepu Port on the Han Dynasty Maritime Silk Road[J]. Democracy and Science,2018(01):25-28.

[9] Anle. Explaining the important role of the Northern Grassland Silk Road from the glass bowl excavated from Feng Sufu's tomb in Northern Yan [C]//. Proceedings of the First Eastern Collections Conference, 2018: 66-74.

[10] Cheng Qian, Yu Chun, Xilin, He Wei. Detection and preliminary study on the composition of excavated glass from the Lobtso lagoon site in Tibet--Also on the Ali section of the Silk Road in Western Tibet[J]. Journal of Tibetan Studies,2017(02):264-274+319.