Research on Composition Analysis of Glass Products Based on Clustering and Random Forest Model

Jiarui Li[*, #], Jinglin Qi[#], Mei Yue[#]

School of Statistics and Big Data, Henan University of Economics and Law, Zhengzhou, China

*Corresponding author: 3173929274@qq.com

[#]These authors contributed equally.

Abstract. Glass cultural relics are important material and cultural heritage of human beings and play an important role in the history of human civilization. This study aims to investigate the relationship between the chemical composition of glass artefacts and the glass type of the artefacts. The chemical composition of glass was analyzed and identified by using K-means clustering, multiple correspondence analysis, SOM model and random forest algorithm. In this paper, based on the given data, the physical properties and chemical composition of the high-potassium glass relics and the lead-barium glass relics were studied respectively, and the classification basis of the two was obtained. It is found that the content of silica and lead oxide can be used as an important basis for distinguishing the two types of glass cultural relics and based on this, a prediction model for further classification of cultural relics and inference of glass cultural relics is designed. This study answered the relationship between the chemical composition of glass relics and the glass type of relics, and further research is needed to obtain more relevant data to train the model to improve its applicability.

Keywords: Clustering algorithm, SOM, random forest model, machine learning.

1. Introduction

Glass is important evidence of early ancient merchant transactions and cultural exchanges. The main material is silicon dioxide. In ancient times, there was mainly lead-barium glass distributed in Chu culture and potassium glass distributed in Southeast Asia and India. Their main chemical components are different, and the weathering conditions are also different due to the influence of the burial environment. Therefore, it is of great significance for the protection and restoration of cultural relics to carry out scientific and reasonable composition analysis and identification of ancient glass products.

In reference[1], for the classification of glass defects, the defect region is constructed, the gain decision tree model is constructed based on the Embedded feature selection method, and the most important features are extracted and combined with the support vector machine model to classify the samples. Finally, the identification time of the model after random optimization is 0. 39 ms faster than that of grid optimization, and the optimization speed of the algorithm is 1 to 2 times faster than that of grid search. In reference[2], for the study of the surface classification of glass cover plates, firstly, aiming at the problem of lack of data, the sub-image division and random defect synthesis algorithm are proposed, and the MPGC-DET data set is constructed. To improve the generalization of the model, based on the modern mature deep convolutional neural network model, combined with transfer learning and SE module, a mobile phone glass cover surface defect classification detection model is built, and the final classification accuracy is 96. 40%.

Because the feature selection of the above research mostly stays at the level of physical features, and the manual design of the feature extraction function also has some subjectivity, the universality of classification[3] may not be high. In this paper, feature selection was carried out from the perspective of chemical composition, and clustering and random forest models were used to classify the types of high-potassium and lead-barium glass, and the SOM algorithm was used to cross-validate the classification of chemical composition. After adjusting the parameters of the random forest model, the final accuracy rate reached 100%. Can play a role in promoting the field of glass analysis.

2. Multiple Correspondence Analysis Model

Correspondence analysis is a visual data analysis method[4] that graphically represents the relationship between categories of categorical variables in a low-dimensional space[5].

Correspondence analysis is to display the proportion structure of each element in the rows and columns of a list in the form of points in a lower dimensional space, which can reveal the differences between the categories of the same variable and the correspondence between the categories of different variables. After inputting the data into Spss26.0, the union diagram of the categories is shown in Figure 1.

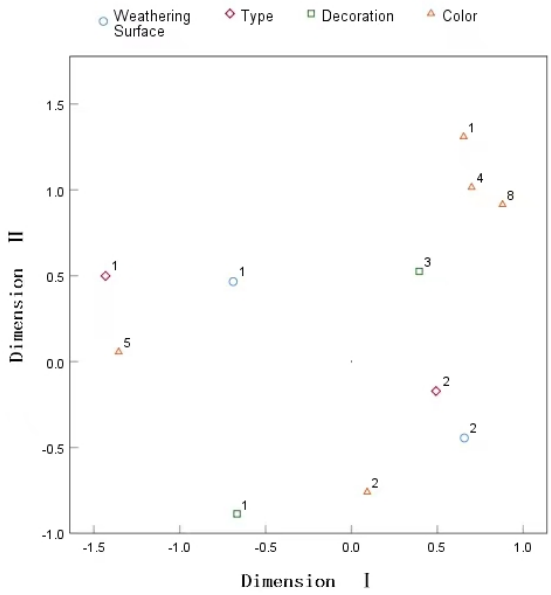

Figure 1. Union Diagram of Category Points

The judgment principle of the corresponding relationship is that starting from the origin (0,0), different categories of the same variable close to the same orientation and the same area of the figure have similar properties, and there may be connections between categories of different variables[6]. Weathering is more likely to occur in light blue lead-barium glass, while it is less likely to occur in blue-green high-potassium glass[7].

2.1. Research on the statistical law of chemical content

First of all, in order to understand the statistical law[8] of the chemical composition content on the surface of the sample, it is necessary to understand the general change trend of the compound before and after weathering. Therefore, this paper visualizes the data. The change area diagram of the compound content of the cultural relics before and after weathering is obtained, as shown in Figure 2 and Figure 3.

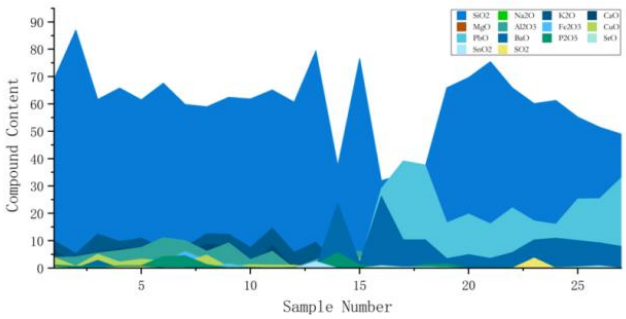

Figure 2. Compound Content before Weathering

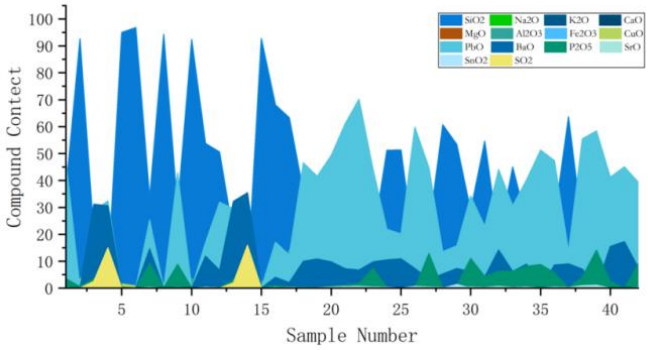

Figure 3. Compound Content after Weathering

It can be seen from Figure 3 that although the compound content of different cultural relics fluctuates slightly before weathering, it is generally stable, with the highest content of silicon dioxide, followed by lead oxide and aluminium oxide. After weathering, the compound content of different cultural relics fluctuated sharply, and the compound content showed a downward trend compared with that before weathering.

Based on determining the approximate variation law of the chemical composition content of the sample, this paper makes a descriptive statistical analysis of the variation trend of the chemical composition content of different glasses based on the type of glass. Most of the chemical content decreases due to weathering, which is in line with the decreasing trend mentioned above. However, for high-potassium glass, its silica value will increase significantly after weathering. For lead-barium glass, after weathering, sodium oxide, magnesium oxide, copper oxide and other compounds show an increasing trend, while the variance increases significantly, and the distribution of compounds in each cultural relic sample becomes extremely uneven, which may be related to the burial environment and the material of the glass itself[8].

3. Glass classification model based on cluster analysis

According to the chemical composition content, select one or more chemical compositions that are most important for the classification, and subclassify the two types of glass again. Cluster analysis and self-organizing feature map (SOM) neural networks are used to solve this problem.

3.1. Classification model

It can be seen from the title that silicon dioxide (SiO_2) is the main component of the glass, and the content of silicon dioxide (SiO_2) is quite different in the two kinds of glass. Generally speaking, if the content of silicon dioxide is more than 75%, it can be preliminarily judged as high-potassium glass. The second is a lead oxide (PbO), whose content directly determines the type of glass.As shown in Figure 4.

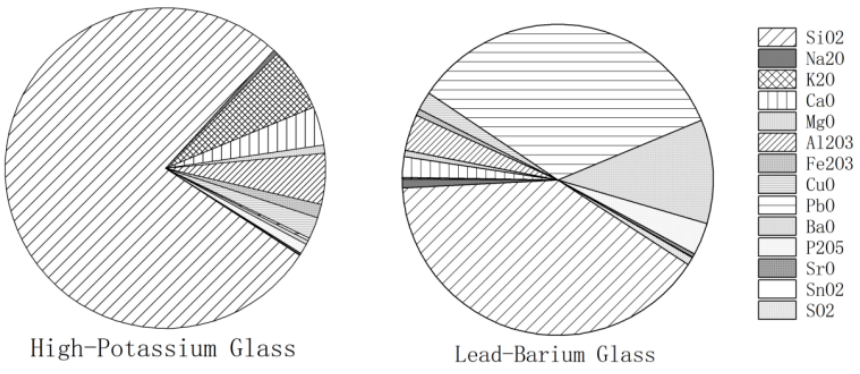

Figure 4. Chemical composition of different types of glass

Because there are many chemical compositions and different glass types[9] in the samples of cultural relics, to ensure the correctness and rationality of the classification results, this paper uses the elbow rule to determine the number of clusters, the core idea of which is that the larger the number of clusters K is, the finer the classification of samples will be, and the degree of aggregation of each cluster will naturally increase. The elbow diagram obtained from the input data is shown in Figure 5.

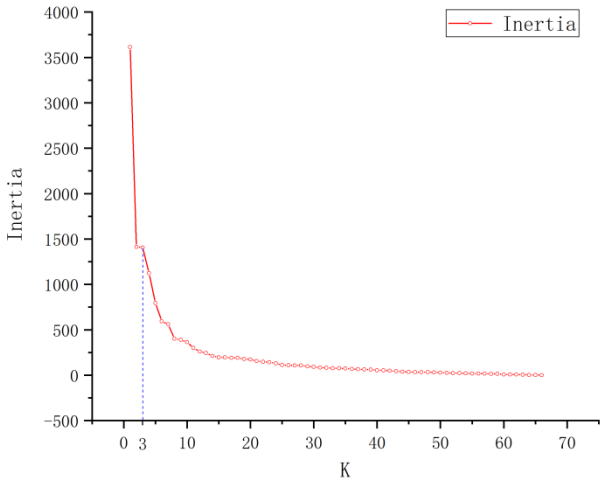

Figure 5. Elbow diagram of hierarchical clustering

From the elbow graph, we can see that there are obvious inflexion points at 2 and 4, indicating that the number of clusters is better at 2 or 4. Based on this number of clusters, we further carry out K-means clustering.As shown in Table 1.

Table 1. Clustering results

Number	Type	Clustering results	Number	Type	Clustering results
1	High potassium	1	52	Lead-barium	2
2	Lead-barium	2	53	Lead-barium	1
3	High potassium	1	54	Lead-barium	2
4	High potassium	1	55	Lead-barium	2
5	High potassium	1	56	Lead-barium	2
6	High potassium	1	57	Lead-barium	2
	…		58	Lead-barium	2

The k-means clustering algorithm is based on the Euclidean distance, through the iterative update of the cluster centre, and finally achieves the minimum sum of the distance from all sample points to the cluster centre in a class. This algorithm relies on the number of categories defined by the elbow rule as the value of K. According to the systematic clustering of elbow images, this paper chooses K = 2 as the initial number of clustering centres, and uses Matlab R2018b software to code and solve. The final clustering results are shown in Table 2. After comparing the clustering results with the types of glass cultural relics, the accuracy of clustering to the types of glass cultural relics is about 84.48%, which is basically in line with its expected function.

By analyzing the final cluster centers of the above clustering scheme[10], silica (SiO2) is extracted as the most important influential component of class 1, and lead oxide (PbO) is extracted as the most important influential component of class 2, and a threshold value is set to distinguish glass subclasses.On this basis, according to the inflection point determined by the elbow rule, the high-potassium glass and lead-barium glass were divided into four subcategories, and the threshold value of their main influencing components was determined. The judgment rules are as follows: For high-potassium glass, the content of silicon dioxide (SiO2) is more than 75%, which is classified as high-

potassium A category, including 9 sample points.This type of glass is basically weathered; the content of silicon dioxide (SiO2) is < 75%, which is classified as high-potassium type B, including 8 sample points, and this type of glass is characterized by no weathering; for lead-barium glass, the situation is slightly more complicated: the content of lead oxide (PbO) is > 40%, which is classified as lead-barium type A, including 13 sample points, and this type of glass are characterized by weathering;Lead oxide (PbO) content < 40%, classified as lead-barium type B. This type of glass cannot be directly judged whether it has been weathered or not, and contains 28 sample points.

3.2. Self-organizing feature map (SOM)

Considering the possible inaccuracy of clustering analysis, this paper takes another method to verify the clustering results, carry out sensitivity analysis, and illustrate the rationality of class division[11].

The following is the construction of the self-organizing feature map neural network (SOM) in this paper. This paper will implement the following steps through Matlab R2018b.In this paper, the type, colour, surface weathering degree and 14 chemical components are used as the input layer to exclude the less influential ornamentation attributes. To improve the fineness of the competition results and facilitate the observation and perception classification, a two-dimensional triangular array with a competition layer of 10 × 10 is set. A total of 23 variables constitute the input layer and are projected to 100 grid cells.The weight is initialized to a value that is randomly close to 0, a group of input samples are randomly taken, and the samples are completely traversed on the competitive layer to calculate the Euclidean distance (i.e., similarity) between the samples and each neuron. At the end of the traversal, a neuron with the smallest similarity is selected as the winning neuron, and then the weights of other neurons in the neighbourhood of the winning neuron are updated to complete a round of iteration. To ensure coverage, the number of iterations is set to 3000. First, state the node name, as shown in Table 2.

Table 2. Names in SOM corresponding to 23 nodes

Ornamentation A	Ornamentation B	Ornamentation C	Type	Blue	Green
Input 1	Input 2	Input 3	Input 4	Input 5	Input 6
Purple	Black	Surface weathering	Silica (SiO2)	Sodium oxide (Na2O)	Potassium oxide (K2O)
Input 7	Input 8	Input 9	Input 10	Input 11	Input 12
Calcium oxide (CaO)	Magnesium oxide (MgO)	Alumina (Al2O3)	Iron oxide (Fe2O3)	Copper oxide (CuO)	Lead oxide (PbO)
Input 13	Input 14	Input 15	Input 16	Input 17	Input 18
Barium oxide (BaO)	Phosphorus pentoxide (P2O5)	Strontium oxide (SrO)	Tin oxide (SnO2)	Sulfur dioxide (SO2)	
Input 19	Input 20	Input 21	Input 22	Input 23	

After the self-organizing feature mapping, a neural network (SOM) is used to solve the problem, the output is shown as the weight perception diagram in Figure 5-6 and Figure 5-7. In the diagram, the inputs with regular distribution weights are Input 10 and Input 18. According to Table 5-9, the two input factors are silicon dioxide (SiO2) and lead oxide (PbO).

The results of Figure 6 show that the image reflects the distance (i.e., similarity) between neurons. It can be seen that the darker colour forms a dividing line to divide the data set into two similar parts, but because of the high dimension of the input data, it is impossible to show the fitting degree with the whole picture of the data on the two-dimensional plane, so we continue to analyze with the help of Figure 7.

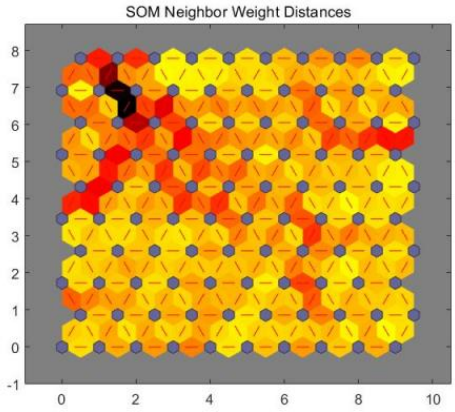

Figure 6. SOM weight perception diagram (data set)

The results in Figure 7 show that the influence of SiO2 and PbO on clustering is almost opposite from the weight trend of Input 10 and Input 18; According to the distribution of Input 10 and Input 18 grid perception maps, SiO2 and PbO have a relatively uniform influence weight distribution for all data sets, which well explains that these two factors have a good influence on all samples. The conclusion of the K-means cluster centre is that the content of SiO2 in high-potassium glass is much higher than that in lead-barium glass, while the content of PbO in lead-barium glass is much higher than that in high-potassium glass, which indicates that the above cluster analysis method is highly reasonable.

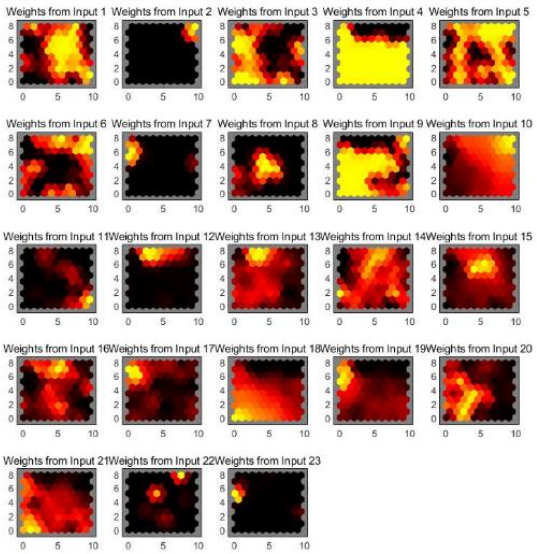

Figure 7. SOM Weight Perception Diagram (Factor)

3.3. Random forest algorithm

Supervised learning is an important branch of machine learning, which requires both input variables and output variables. The supervised learning method is used in this paper. In order to avoid the overfitting phenomenon in the process of cultural relics prediction, this paper uses the random forest algorithm of the Bagging optimization algorithm to identify the categories of cultural relics.

In this paper, Matlab R2018b is used to implement the random forest algorithm, and the identification results of the 8 cultural relics are shown in Table 3.

In order to improve the prediction accuracy of the model and ensure that the model can better identify unknown cultural relics, we need to adjust the maximum split number of the random forest and the number of learners. The F1 score is the harmonic average of recall and precision, which is widely used to measure the accuracy of binary classification models. Therefore, this paper uses the

Table 3. Identification Results of Unknown Cultural Relics

Cultural relic number	Type of cultural relic	Cultural relic number	Type of cultural relic
A1	High potassium	A5	Lead and barium
A2	Lead and barium	A6	High potassium
A3	Lead and barium	A7	High potassium
A4	Lead and barium	A8	Lead and barium

F1 score to measure the prediction accuracy of the random forest model.The mathematical model of the binary classification problem F1 score is as follows (1):

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{1}$$

According to the confusion matrix, the F1 score is 0.91, which is close to 1, indicating that the accuracy of the model is high. The accuracy of the model is 0. 86, which indicates that the model can meet the requirements.

4. Conclusion

In this paper, multiple correspondence analysis models are used to explore the statistical law of chemical content. At the same time, in the random forest model, the prediction accuracy reaches 100% after adjusting the parameters, and the prediction results are reliable. The model has been tested scientifically and reasonably and has strong stability. For the sensitivity analysis of the second and third models, SOM (Self-Organizing Feature Map) and random forest are used for cross-validation, which can solve the over-fitting phenomenon in the prediction model and make the classification results have good stability. The model in this paper is systematic and practical: for the classification of glass types, the elbow rule is used to objectively determine the number of clusters, which avoids the influence of certain subjective factors, and at the same time, the relevant literature in recent years is referred to in the classification of subcategories, which is scientific and reasonable. However, in the analysis of model 2, the clustering results obtained by the K-means clustering algorithm can only converge to the local optimal solution and can not get the global optimal solution. In the analysis of model 3, the running speed of the random forest algorithm is significantly lower than that of a single decision number, and the program runs slowly when dealing with more variables.

Due to the small scale of the data given in this paper, the model in this paper may produce large errors in the identification when applied to high-potassium or lead-barium glass relics unearthed in other areas. In the future, more data should be used to train the model to improve its applicability. As an important material cultural heritage, glass cultural relics should be protected after excavation, so the factors such as temperature, humidity and light resistance of chemical substances contained in cultural relics should be taken into account when building models in the future.

References

[1] Yang Xin. Glass defect detection and classification based on machine vision [D]. Fujian Institute of Engineering, 2021.

[2] Wu Chuang, Yu Dayong. Research on surface defect classification detection of mobile phone glass cover based on deep convolutional neural network [J]. Software Engineering, 2021, 24 (12): 6.

[3] Jiang Qiyuan, Xie Jinxing,. Mathematical model. Version 3 [M]. Higher Education Press, 2003.

[4] Cai Linlin. Model selection and parallelization of random forest [D]. Harbin Institute of Technology. 2012.

[5] Ma Yunlong. RBF neural network prediction algorithm based on principal component analysis and its application [D]. Jilin University.

[6] Zhang Kun, Shen Haibo, Zhang Hong, et al. Comprehensive evaluation method of node importance in complex network based on grey relational analysis [J]. 2022(4).

[7] Zhou Jianhui, Meng Lei, Wang Lijuan, et al. Multiple correspondence analysis of pathogen distribution characteristics of fever with rash syndrome in Gansu Province from 2009 to 2019 [J]. China Public Health, 2022, 38 (3): 4.

[8] Qi Mengsha, Li Wei, He Ping, et al. Obstacle survey and multiple correspondence analysis in pulmonary rehabilitation practice of chronic obstructive pulmonary disease [J]. Chinese Journal of Respiratory and Critical Care, 2021, 20 (2): 4.

[9] Liu Zhen. Analysis on the Methods of Cultural Relics Identification. 2021.

[10] Wang Chengyu, Tao Ying. Weathering of silicate glass. Journal of Silicate, 2003, 31 (1): 8.

[11] Xue Xinju. Python-based K-means algorithm and its application. Science and Technology Horizon, 2018 (24): 2.