Identification model of surface composition of glass relics based on K-means clustering and boosting learning strategy

Hanyun Yin *

University of Science and Technology of China, Hefei, China

* Corresponding Author Email: yhy1043772558@mail.ustc.edu.cn

Abstract. As one of the main commodities circulated on the Silk Road, glass is valuable material evidence of early trade, but glass is vulnerable to external factors leading to weathering, so this paper establishes a comprehensive identification model to study the chemical composition of the surface of glass cultural relics. In this paper, the Chi-square independence test and fuzzy function are used to analyze the relationship between the type of glass, weathering degree and chemical composition. Then the principal component analysis is used to extract the principal components, and the clustering and algorithm analysis is carried out to identify the unknown glass cultural relics. Finally, the variance test method was used to test the difference between the correlation coefficients of glass types and chemical elements, and the correlation and difference between the chemical components of different types of glass cultural relics were analyzed.

Keywords: Difference test, fuzzy mathematical model, cluster analysis, K-means, principal component analysis.

1. Introduction

Due to the different types of fluxes used in glass refining, the main chemical composition of ancient glass products in China is quite different. According to this, modern archaeologists [1] divide it into a lead-barium glass and potassium glass. However, the weathering phenomenon accompanying the burial of cultural relics will lead to changes in their chemical composition, which has an impact on archaeologists' classification and other work. and other work. Accurate analysis and identification of the composition of ancient glass products are of great significance to the study of history and economy.

Therefore, in this paper, the chi-square independence test is used to analyze the significant difference between the weathering degree of the surface and its type, colour, decoration and other types of variables, and the fuzzy mathematical model is used to fit the membership function between the domain and the fuzzy set with the oxygen content of each chemical component before and after weathering as the element. Based on this, we can predict the data produced by known weathered data and types of cultural relics under unweathered conditions. Then the principal component analysis [2] is used to extract the principal components from the original indicators, and the principal components are used for K-means clustering to divide the high-potassium glass into four categories and the lead-barium glass into five categories. Then the data set is constructed, and the boosting learning strategy is selected to classify with AdaBoost, XGBoost and GBDT respectively. Finally, the Pearson coefficient was used to calculate and obtain the correlation between the chemical composition of high-potassium glass and lead-barium glass, and the variance test method was used to test the difference between the correlation coefficient of glass types and chemical elements by using the correlation coefficient matrix.

2. Prediction model based on chi-square independence test

2.1. Chi-square independence test

The Chi-square independence test is suitable for the analysis of enumeration data with two or more factors. Its main calculation parameters are the degree of freedom $df = (C-1)(R-1)$ (C, and R

represent the number of rows and columns of the variable table respectively) and chi-square,as shown in formula (1).

$$X^2=\sum\frac{(O_i-E_i)^2}{E^2_i} \tag{1}$$

Where O_i represents the actual frequency and E_i represents the desired frequency.

Using SPSS software to test the chi-square [3] of the variables, the revised results are as follows: whether the surface of glass relics is weathered or not is significantly related to the type of glass, but not to the colour and surface decoration of cultural relics. In the case of significance, compare the specific values, draw a cross chart, and describe the specific differences. The cross diagram of whether the surface of cultural relics is weathered or not, surface decoration and glass type is drawn as Figure 1.

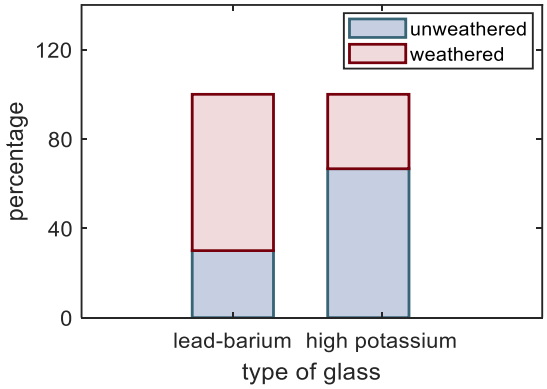

Figure 1. Cross Diagram of Surface Weathering and Glass Type

It can be seen from the above Figure 1, the following conclusions can be drawn. There is a significant difference between glass type and whether it is weathered or not, and lead-barium glass is easier to be weathered than high-potassium glass; there is no significant difference between the colour of cultural relics and whether it is weathered or not; there is no significant difference between the texture of cultural relics and whether they are weathered or not.

2.2. Statistical analysis and paired sample t-test

Classify that pretreated data according to the type of glass [4], perform preliminary statistical analysis on the content of each chemical component, and calculate the mean value and the standard deviation of each group of data. Visualize the chemical composition content of each group before and after weathering, as shown in Figure 2 below:

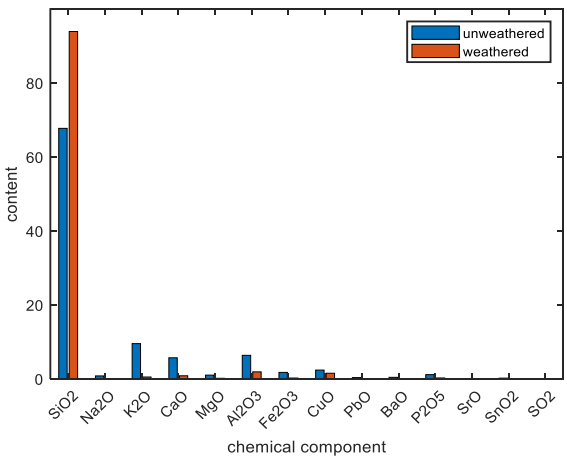

Figure 2. Comparison of chemical composition of high-potassium glass before and after weathering

Paired sample t-test is a kind of double population t-test. The test object is the difference between the data obtained by two matched groups of subjects or the data obtained by the same group of subjects under different conditions. The t-statistic under the paired sample test is calculated as follows:

$$t = \frac{m}{s\sqrt{n}} \tag{2}$$

Where m is the mean of the difference between two samples, n is the sample size, s is the standard deviation of the difference between two samples, and the degree of freedom is DF = n-1. The t-test statistics were calculated for the degrees of freedom, and the data were entered into SPSS to obtain the t-test distribution results as follows:

the content of silica and sodium oxide [5] on the unweathered surface is significantly higher than that on the weathered surface;

Calcium oxide lead oxide and phosphorus pentoxide contents of the unweathered surface are significantly lower than those of the weathered surface.

2.3. Prediction model of chemical composition content

To establish the mapping relationship between the two data sets before and after weathering, we use the theory of fuzzy mathematics to record the data set of chemical composition content after weathering as domain x and the data set of chemical composition content before weathering as fuzzy set A. Our goal is to find the membership degree $u_A(x)$ of the element x in X to the fuzzy set A, that is $2u_A$, the membership function of A. The data after weathering are also treated in the same way, as shown in Figure 3.

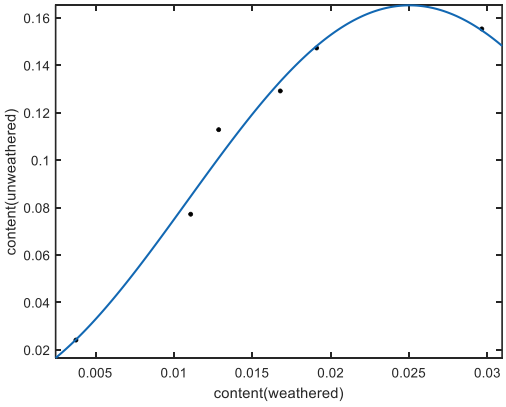

Figure 3. Calcium oxide Fourier series regression (high potassium)

The content of each chemical component before weathering can be obtained by predicting the amount of each chemical component corresponding to the oxygen element before weathering through the membership function. The calculation results are shown in Figure 4.

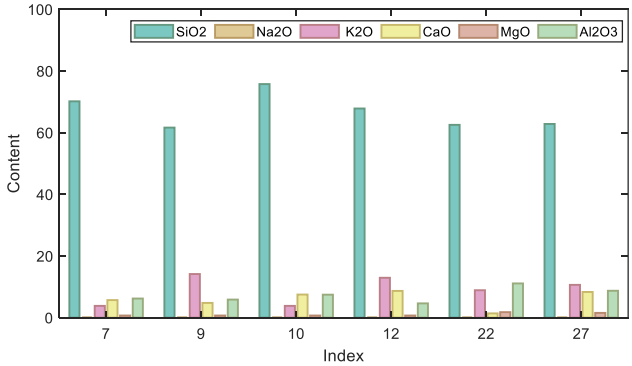

Figure 4. Prediction of chemical composition content of lead-barium glass before weathering

Comparing the predicted values [6] with the actual measured values of the rheumatic points, the vector similarity is calculated to be 0.993, 0.953, and 0.994, respectively, so the prediction model has high confidence.

3. Multi-classification model

3.1. Based on the analysis model of K-clustering

Due to the large number of variables and their correlation, the principal component analysis method is used to process the variables. The principal components are then used as new variables to cluster the targets. K-means clustering algorithm is used in this paper.

Firstly, K clustering centres are initialized, the distance between each object and each clustering centre is compared in turn, and iteration is carried out until the change of the centre meets the convergence condition, and finally, the clustering result is obtained. The SSE-K line graph is drawn to find out the real number of clusters intuitively.As shown in Figure 5.

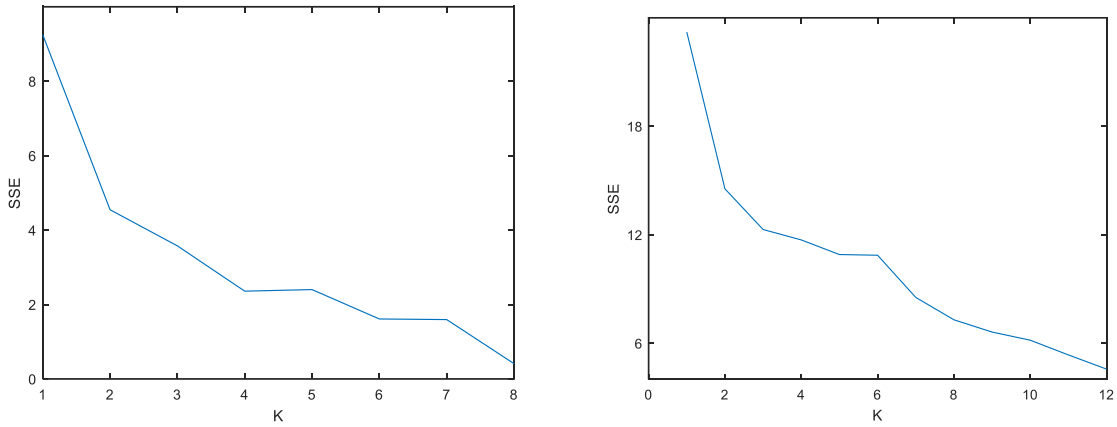

Figure 5. SSE-K Line Chart of High Potassium and Lead-Barium Glass Cluster

The real cluster numbers of high-potassium and lead-barium glasses are 4 and 5, respectively.To test the sensitivity of the model, the principal component parameters used in the model were increased or decreased to varying degrees, and the cluster contour coefficient value was quantitatively evaluated, and the results were shown in the Figure 6 and 7.

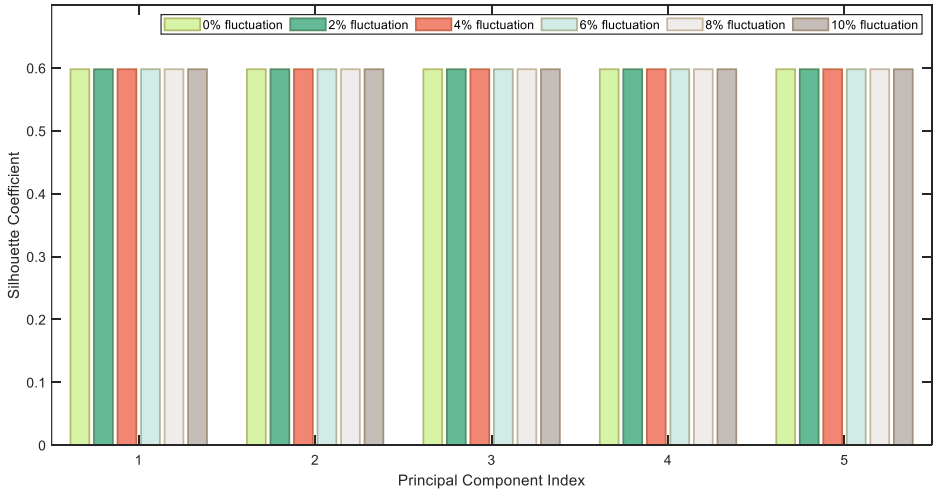

Figure 6. Effect of small fluctuations in the principal components of high-potassium glass on clustering results

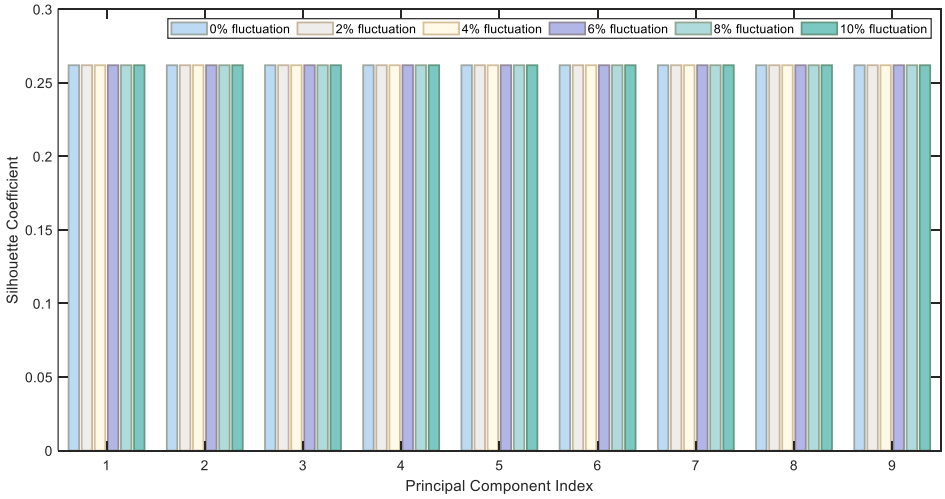

Figure 7. Effect of small fluctuations in principal components of lead-barium glass on clustering results

As can be seen from the figure above, under the condition that each principal component fluctuates within 10%, the cluster outline coefficient does not change, that is, the clustering result does not change. This shows that the multi-classification model has a strong anti-interference ability and good stability.

3.2. Classification model based on boosting learning strategy

To build an effective classification model [7], the boosting algorithm of ensemble algorithms is considered. The Principle of the algorithm is to combine multiple weak classifiers $f_i(x)$ to form a strong classifier, including many concrete algorithms. The prediction accuracy and data volume requirements are compared and analyzed. GBDT is an iterative algorithm based on a decision tree, which is often used in regression analysis. It constructs a group of weak learners [8] (trees) and accumulates the results of multiple decision trees as the final prediction output. xgboost is an improved algorithm based on GBDT, which is more powerful and applicable to a wider range. Referring to the decision tree algorithm, we can choose the greedy algorithm and approximate algorithm to do node segmentation, and finally, select a reasonable number of split points. In this paper, the above three algorithms are used to classify and calculate respectively, and the same results are obtained: A1, A6 and A7 are classified as high-potassium glass, and A2, A3, A4, A5 and A8 are classified as lead-barium glass and the prediction accuracy of the three algorithms is 100%.

4. Associative diversity model

It is necessary to analyze the correlation between the chemical components [9]. Pearson's correlation coefficient is used in the pairwise quantitative correlation analysis of chemical elements. Pearson correlation coefficient is widely used to quantitatively study whether there is a (linear) correlation between two variables. Under the assumption that each chemical component follows a normal distribution within a certain range, there is a formula for calculating Pearson's correlation coefficient (3):

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-u_X)(Y-u_Y)]}{\sigma_X \sigma_Y} \tag{3}$$

For covariance and standard deviation estimation using samples, the above equation can be rewritten as (4):

$$\rho_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma_X}\right)\left(\frac{Y_i-\overline{Y}}{\sigma_Y}\right) \tag{4}$$

It is generally believed that when $|\rho_{X,Y}| > 0.8$, there is a strong correlation between the two variables, while when $|\rho_{X,Y}| < 0.2$, there is a very weak correlation between the two variables, or there is no correlation between them.

Using SPSS software for data processing [10], the Pearson correlation coefficient matrix of all chemical compositions of high potassium glass and lead-barium glass was obtained. To sum up, the correlation coefficient of the chemical composition of the same type of glass is related to the chemical reaction between chemical compositions during weathering. However, there is no significant difference in the correlation coefficient of chemical composition between different types of glasses, that is, the change of chemical composition caused by chemical reaction during weathering does not reflect the selectivity of specific elements or compounds macroscopically. Apart from the strong correlation between individual chemical elements caused by specific chemical reactions in specific environments, most of the chemical components do not show an obvious correlation, but the correlation coefficients of the same group of chemical components of different types of glass are different because the content of chemical components of different types of glass is different. The chemical reactions and products of elements with different concentrations are also different. Thereby reflecting that the correlation coefficients of the same group of chemical compositions of different types of glass are different.

5. Conclusions

In this paper, the significant difference analysis is carried out firstly, and the fuzzy mathematical model is used to fit the membership function between the domain and the fuzzy set, which takes the oxygen content in each chemical composition before and after weathering as the element, and then the prediction is carried out. Then the principal component analysis is used to extract the principal components from the original indicators, and the principal components are used for K-means clustering to divide the high-potassium glass into four categories and the lead-barium glass into five categories. Then the data set is constructed, and the boosting learning strategy is selected to classify with AdaBoost, XGBoost and GBDT respectively. Finally, the Pearson coefficient was used to calculate and obtain the correlation between the chemical composition of high potassium glass and lead-barium glass. In this model, the prediction model of chemical composition content adopts the concept of fuzzy mathematics, and the membership function is fitted by clustering and screening of data, which not only establishes the correlation between two sets without mapping relationship, but also improves the prediction accuracy of the model; However, due to the small amount of data given in the problem, the anti-interference ability of the chemical composition prediction model is weak, and the follow-up work should improve the training set to increase accuracy. The model established in this paper can perfectly solve the problems of analyzing the relationship between the chemical components of glass cultural relics, predicting the content of chemical components before weathering, and identifying phenols in unknown cultural relics. It is an important contribution to solving this kind of problem in the future, especially after the improvement of the model, it can be widely used in the archaeology of glass cultural relics, and can also be applied to similar problems.

References

[1] Geological Dictionary Office, Ministry of Geology and Mineral Resources. Geological Dictionary: Geological Publishing House, 2005

[2] Wei Jie. Research on improved K-means clustering algorithm [J]. Information Communication, 2018 (5): 14-15.

[3] Chen Min. Factors affecting the weathering of lead glass [D]. Dalian University of Technology, 1987.

[4] An Shengli, Chen Pingyan. The difference between equivalence test and difference test and its simulation validation. China Health Statistics, 2007,24 (3): 226-228. DOI: 10.3969/j.issn.1002-3674.2007.03.001.

[5] Hu Chunjian. Significance test of Kendall τ correlation coefficient in small sample [J]. Control Engineering, 2013.

[6] Gao Yunlong, Luo Sizhe, Pan Jinyan, et al. Robust adaptive probabilistic weighted principal component analysis. Acta Automatica Sinica, 2021.

[7] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model [J]. Systems Engineering-Theory and Practice, 2020, 30(1): 158-160.

[8] Shao Hui, Zhou Hongqing, Liu Min, Zhu Haikui, Shen Xiaodong. Effect of alumina addition on barium silicate composite glass ceramics. Chinese Journal of Ceramics, 32 (1): 17 – 20, 2011.

[9] Wang Shengtie, Liu Yutang. Summary of Chu Culture. BEIJING BOOK CO. INC. 2013.

[10] Liu Song, Lv Liangbo, Li Qinghui, Xiong Zhaoming. Glass Beads Unearthed from Han Tombs in the South of the Five Ridges and the Sino-Foreign Exchange on the Maritime Silk Road of the Han Dynasty. Conservation and Archaeological Science, 31 (4): 18 – 29, 2019.