# Composition Analysis and Identification of Ancient Glass Products Based on Hierarchical Clustering

Zhixin Lu*

School of Big data, Weifang Institute of Technology, Weifang, China

*Corresponding author: 1179484365@qq.com

**Abstract:** The ancient glass is very easy to be affected and weathered by the burial environment. During the weathering process, there will be a lot of exchange between internal elements and environmental elements, and the composition of glass will also change in proportion, which will affect the judgment of its category; This paper mainly deals with data based on hierarchical clustering analysis and the optimization model of hierarchical clustering analysis.First, the chi square test is used to analyze the data of the four variables. Based on whether the P significance value is less than 0.05, it is further judged that there is no significant difference between the surface weathering and the color and decoration, but only significant difference with the type of glass; Further, the corresponding analysis of glass type and whether there is weathering on the surface is made to draw a more specific conclusion; Count the difference of chemical composition addition of different types of glass with or without weathering, and then systematically describe and count the difference of values to obtain a histogram for observing the difference before and after the change; According to the data of weathered points and unweathered points, the data of unweathered points are extracted, and the average value of each chemical composition of weathered and unweathered points is calculated. The unweathered points do not show weathering of chemical composition on the time axis, so the average value is compared as the content of chemical composition before weathering.

**Key words:** Chi square test; Correspondence analysis (R-Q factor analysis); Correlation analysis; Hierarchical clustering model based on random forest

## 1. Introduction

In ancient times, many decorations and ritual vessels in China were made of a large amount of glass. Through the Silk Road, which is a communication channel, our glassware became more solid and durable. At the same time, we found that the chemical composition of foreign glassware was different from that of our glassware.It is known that the main raw material of glass is quartz sand, and the main chemical component is silicon dioxide. Due to the high melting point of pure quartz sand, in order to reduce the melting temperature, fluxes need to be added during refining. The commonly used fluxes in ancient times included plant ash, natural soda, saltpeter and lead ore. Limestone was added as a stabilizer, and the limestone was converted into calcium oxide after calcination. The main chemical composition varies with the flux added.

## 2. Establishment and solution of model

Chi square test [1] is mainly used to compare the difference analysis between groups of categorical variables, check whether there is significant difference between grouped variables, and determine whether there is significant difference between variables by whether the significant P value is less than 0.05.

In this paper, the interactive software SPSS PRO is used to process the data in the text. The variable X is set as surface weathering, and the variable Y is set as texture, type and color. The following table 1 is obtained.

Table 1 chi square test table

| subject | name | Surface weathering | | total | $X^2$ | correcting $X^2$ | P |
|---|---|---|---|---|---|---|---|
| | | No weathering | weathering | | | | |
| Ornamentation | C | 13 | 15 | 28 | 5.747 | 5.747 | 0.056* |
| | A | 11 | 9 | 20 | | | |
| | B | 0 | 6 | 6 | | | |
| total | | 24 | 30 | 54 | | | |
| type | hyperkalemia | 12 | 6 | 18 | 5.400 | 4.134 | 0.020** |
| | Lead barium | 12 | 24 | 36 | | | |
| total | | 24 | 30 | 54 | | | |
| colour | Blue-green | 6 | 9 | 15 | 6.287 | 6.287 | 0.507 |
| | Light blue | 8 | 12 | 20 | | | |
| | purple | 2 | 2 | 4 | | | |
| | Dark green | 3 | 4 | 7 | | | |
| | Dark Blue | 2 | 0 | 2 | | | |
| | Light green | 2 | 1 | 3 | | | |
| | black | 0 | 2 | 2 | | | |
| | green | 1 | 0 | 1 | | | |
| total | | 24 | 30 | 54 | | | |

Note: * * and * represent the significance level of 5% and 10% respectively

The chi square test table shows that the significance P value of ornamentation and surface weathering is 0.056 greater than 0.05, which is not significant at the level. The original hypothesis is accepted, so there is no significant difference; The significance P value of color and surface weathering is 0.507, which is greater than 0.05. There is no significance on the level. The original hypothesis is accepted, so there is a significant difference. Only the significance P value (0.02) of type and surface weathering is less than 0.05. The original hypothesis is rejected, and there is a significant difference. It is concluded that whether the surface is weathered or not is only related to the type[2].

In this paper, whether the surface of glass relics is weathered or not is analyzed with different types of glass (high potassium, lead barium), which involves the differences between different categories of the same variable, as well as the corresponding relationship between different categories of different variables. Correspondence analysis (association analysis) model can be used in this paper. The correspondence analysis is mainly based on the combination of the correspondence analysis table and the correspondence analysis chart. The correspondence analysis table shows the dimensions of different variables. The absolute value of the dimension score is obtained from the sum of the values of different variables. The higher the absolute value of the dimension score, the higher the correlation. If different variables tend to be in the same quadrant and on the same side in the corresponding graph of dimensional analysis, the greater the correlation is SPSS PRO software was used to analyze the data[3]. The dimension correspondence table 2 and dimension analysis diagram of Fig 1 surface weathering and type are obtained respectively.

Table 2 Dimensional correspondence between surface weathering and types

| Field Name | term | Dimension 1 | Dimension 2 |
|---|---|---|---|
| Surface weathering | No weathering | 0.41 -0.289 | -1 -1 |
| type | High potassium lead barium | 0.513 -0.231 | 1 1 |

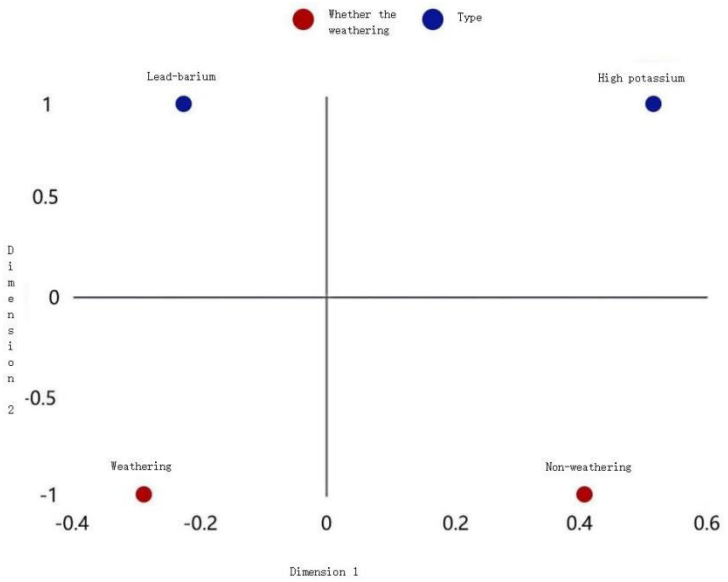



Fig 1 Dimensional Analysis of Surface Weathering and Types

It can be seen from Table 2 that when the dimension is 1, the absolute value of the score of the dimensions of non weathering and weathering combined with high potassium and lead barium respectively can be calculated. It can be concluded that the absolute value of the dimension score of high potassium and non weathering combined (| 0.41+0.513 |=0.923) is greater than the absolute value of the dimension score of high potassium and weathering (| 0.513-0.289 |=0.224), The absolute value (| -0.289-0.231 |=0.52) of the dimension score when lead barium is combined with weathering is greater than the absolute value (| 0.41-0.231 |=0.179) of the dimension score when lead barium is combined with non weathering[4]. When the dimension is 2, no matter how weathering and non weathering are combined with high potassium and lead barium (weathering--high potassium, weathering--lead barium,nonweathering -- high potassium, non weathering -- lead barium), the absolute value of the dimension score is the same as 0,and according to Fig 2 and Fig 3, lead barium and weathering tend to be on the same side, which is located in the second and third quadrants. Therefore, it can be concluded that high potassium is not easy to be weathered, while lead barium is easy to be weathered.

Histograms of main chemical components (the first eight for each) before and after weathering of high potassium and lead barium are listed respectively:

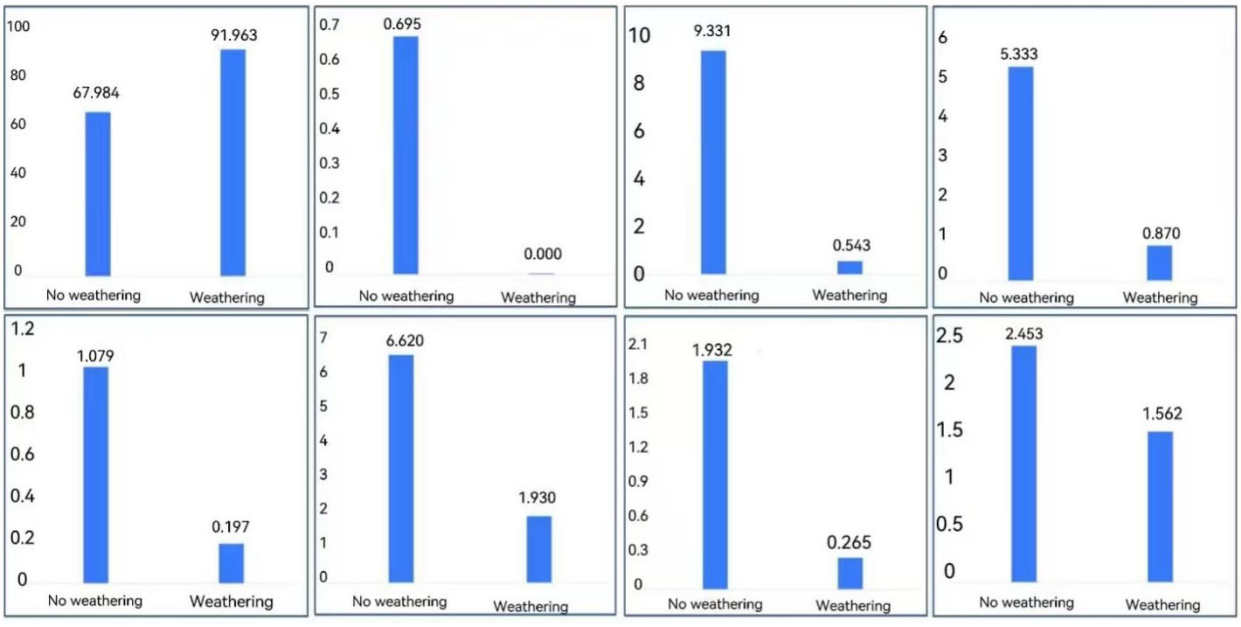



Fig. 2 Histogram of main chemical components before and after high potassium weathering

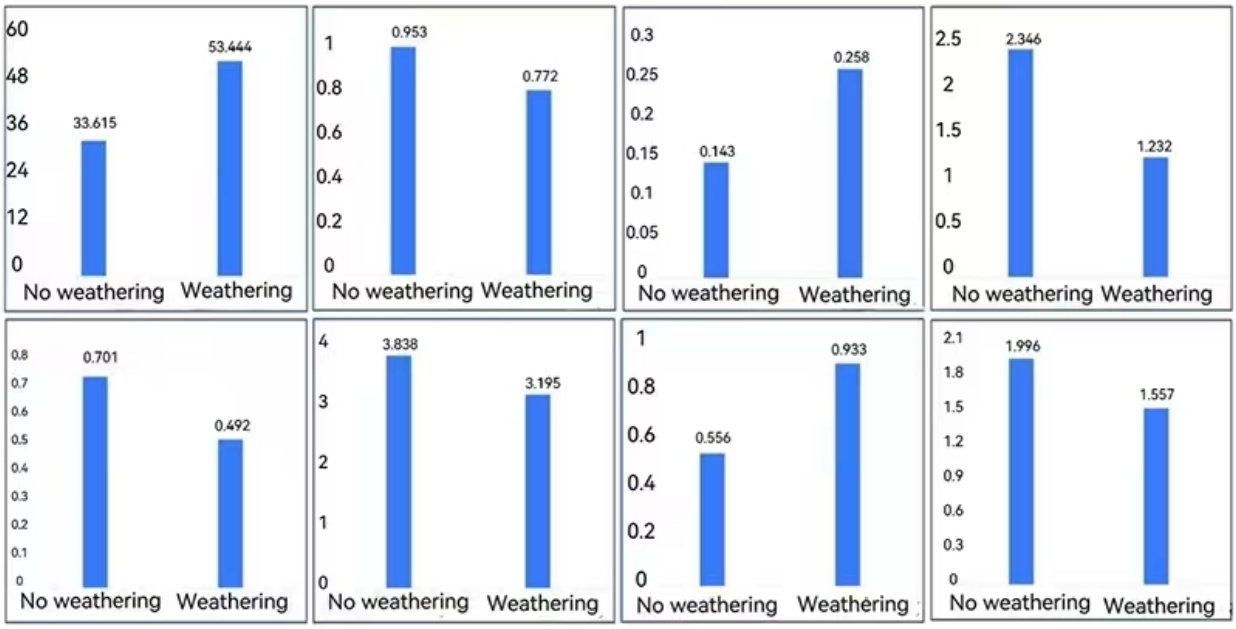



Fig. 3 Histogram of main chemical components before and after Pb Ba weathering

It can be intuitively concluded that the chemical composition of high potassium decreases before and after weathering, and the chemical composition of lead barium also decreases before and after weathering, but the slowing down speed is slower than that of high potassium[5].

For the data detected at weathered points and fresh points, the data are divided into two categories, and the average value of the two categories is calculated through the function in Excel, as shown in the following table 3 :

Table 3 Comparison of Weathered Points and Unweathered Points

| Cultural relics sampling point | Weathering point | Unweathered point |
|---|---|---|
| silicon dioxide | 48.215 | 55.74 |
| Sodium oxide | 0.45 | 25.61 |
| Potassium oxide | 2.3 | 0.1633 |
| calcium oxide | 2.63 | 1.3578 |
| magnesium oxide | 0.666 | 0.926 |
| alumina | 3.61 | 5.363 |
| ferric oxide | 0.92 | 0.449 |
| Copper oxide | 2.04 | 1.362 |
| Lead oxide | 24.38 | 20.85 |
| Barium oxide | 7.64 | 7.259 |
| Phosphorus pentoxide | 2.85 | 1.376 |
| Strontium oxide | 0.258 | 0.228 |
| Stannic oxide | 0.084 | 0.02556 |
| sulfur dioxide | 0.685 | 0 |

According to the image, the chemical composition content of weathering point before weathering is shown in the following table 4 :

Table 4 Chemical Composition Content of Weathering Point before Weathering

| Cultural relics sampling point | Weathering point |
|---|---|
| silicon dioxide | 48.215 |
| Sodium oxide | 0.45 |
| Potassium oxide | 2.3 |
| calcium oxide | 2.63 |
| magnesium oxide | 0.666 |
| alumina | 3.61 |
| ferric oxide | 0.92 |
| Copper oxide | 2.04 |
| Lead oxide | 24.38 |
| Barium oxide | 7.64 |
| Phosphorus pentoxide | 2.85 |
| Strontium oxide | 0.258 |
| Stannic oxide | 0.084 |
| sulfur dioxide | 0.685 |

## 3. Stochastic forest analysis model

In this paper, before data processing, the data parameters are set. The data of all chemical composition types are used as variable X, and the cultural relic glass type is used as variable Y. The data are analyzed by SPSS PRO[6]. A random forest analysis model is built through training set data. The characteristic importance of chemical composition is calculated in the random forest model. For the sub classification of chemical components of each category, a hierarchical clustering model is established for the preprocessed data, and the characteristics of the data are further divided based on the hierarchical clustering model. Then, the main chemical components of all chemical components are obtained through the clustering tree.

The random forest model and hierarchical clustering model used in this paper have errors and advantages in results. The feature importance data obtained from the random forest construction model is obtained through multiple model training, and the data is reliable and stable; In the process of data processing, the amount of data is too large, and the post-processing data will reduce the accuracy of the data due to the increase of the decision tree model. The singular value of AHP has a great influence on the data results, and the result data will deviate, but the data operation has relatively little influence and the hierarchy is clear.

### 3.1. Classification rule

Analyze the classification rules of high potassium glass and lead barium glass, establish a random forest model with the pretreated data, first take the chemical composition variable as the independent variable X, then select the glass type as the variable Y, disrupt the sample data, use the confused sample data for model training, determine the proportion of sample data to participate in model training, and divide the data set into k mutually exclusive subsets with similar size, Each time, the union of k-1 subset is used as the training set, and the remaining 1 subset is used as the test set. Conduct k times of training and testing, and finally return the mean value of k evaluation results. It can avoid the over fitting problem on training set caused by unreasonable data set division[6]. Set the number of decision trees in the random forest to 100, and put back the sample data to participate in the data training test. Set the maximum depth of the tree to 10, and the maximum value of the leaf node of the decision tree to 50. Because of the randomness of the random forest model, the results obtained by each training model are not unique. The original sample data will be used for the calculation of this training model, and then the characteristic importance of chemical components will be obtained,as shown in Table 5.

Table 5 Importance of Chemical Composition Characteristics

| Characteristic name | Importance of characteristics |
|---|---|
| silicon dioxide | 15.40% |
| Sodium oxide | 2.90% |
| Potassium oxide | 10.60% |
| calcium oxide | 6.70% |
| magnesium oxide | 3.00% |
| alumina | 12.10% |
| ferric oxide | 8.80% |
| Copper oxide | 5.80% |
| Lead oxide | 14.40% |
| Barium oxide | 4.10% |
| Phosphorus pentoxide | 6.60% |
| Strontium oxide | 7.60% |
| Stannic oxide | 1.40% |
| sulfur dioxide | 0.70% |

The data in Table 5 shows that, in addition to the main chemical component silicon dioxide (SiO2), the most important chemical component of characteristic importance is lead oxide (PbO), aluminum oxide ($Al_2O_3$), potassium oxide ($K_2O$) and other chemical components are far less important than lead oxide, so the impact of other components on lead oxide is not considered[7].According to the importance of chemical composition, the types of glass can be divided into glasses containing lead oxide or not.

The two kinds of glass can be divided into lead oxide high potassium glass, lead oxide free high potassium glass, lead oxide lead barium glass and lead oxide free lead barium glass,as shown in Fig 4.

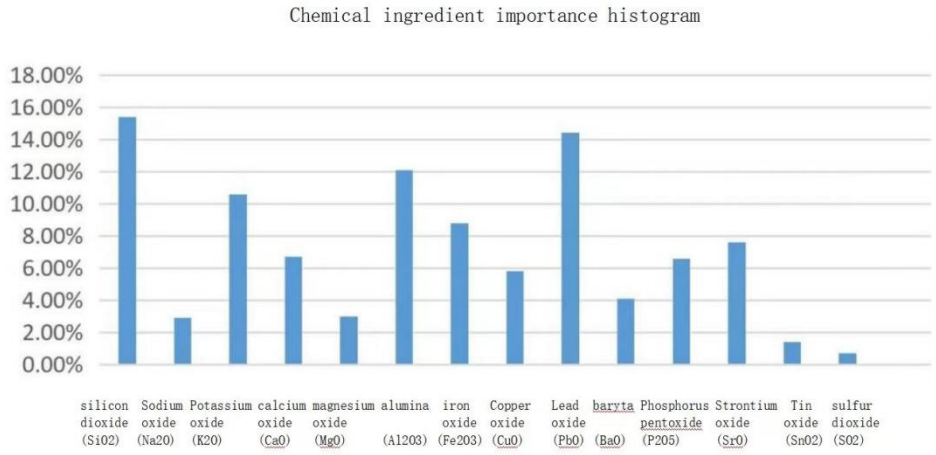



Fig. 4 Importance Bar Chart of Chemical Composition

The cultural relics No. 15 and 17 were removed from the data. Because they did not meet the requirements in data pre-processing, they were not considered in data processing. The filtered data is used to obtain a histogram through Excel.

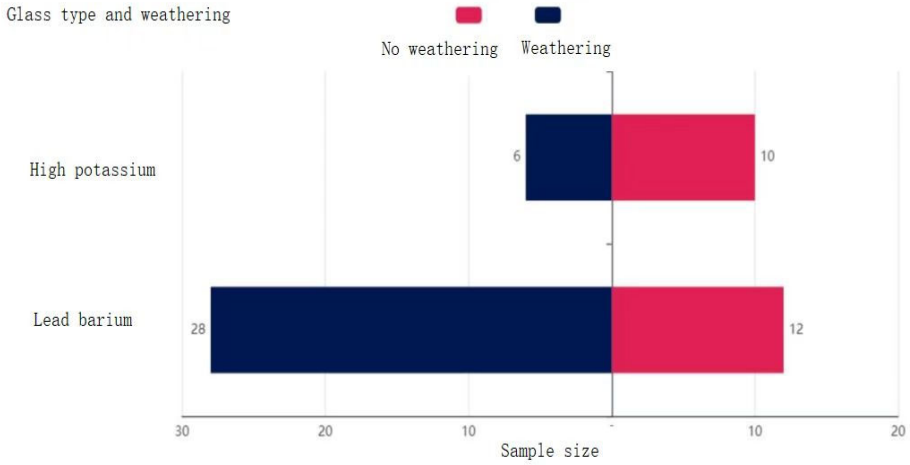



Fig. 5 Glass Type and Weathering Histogram

According to the glass type and weathering image, it can be intuitively analyzed that the weathering of glass is mainly lead barium glass, the weathering proportion in lead barium glass is 0.7, the non weathering proportion is 0.3, and the weathering proportion of lead barium glass in weathered high potassium and weathered lead barium is about 0.82; The proportion of weathering in high potassium glass is 0.38, the proportion of non weathering is 0.67, and the proportion of weathering in high potassium glass and weathered lead barium is about 0.18.

Therefore, high potassium glass and lead barium glass can be divided into weathered high potassium glass, non weathered high potassium glass, weathered lead barium glass and non weathered lead barium glass[8].

The following classification of high potassium glass and lead barium glass can be obtained by summarizing the conclusion of the importance of chemical composition characteristics and the conclusion of glass type and weathering,as shown in Table 6.

Table 6 Division of high potassium glass and lead barium glass

| | | | | |
|---|---|---|---|---|
| Classification | Weathered lead oxide containing high potassium glass | Non weathered lead oxide containing high potassium glass | Weathered lead-free high potassium glass | Non weathering lead oxide free high potassium glass |
| | Weathered lead barium glass containing lead oxide | Non weathered lead barium glass containing lead oxide | Weathered lead free lead barium glass | Non weathering lead oxide free lead barium glass |

### 3.2. Classification of chemical constituents

Before the formal data processing, the data shall be preprocessed. Empty data shall be classified as missing values for processing and assigned a value of 0. The valid data meeting the requirements of the title shall be screened out. The cumulative proportion of each component between 85% and 105% is the valid data. The chemical composition of cultural relics numbered 15 and 17 can be obtained. The obtained data shall be standardized, and a hierarchical clustering model shall be established for the data[9]. The hierarchical clustering shall divide the chemical components according to the data characteristics obtained from the random forest model. The chemical components shall be set as variables. First, each chemical component to be clustered shall be divided independently, and two independent individuals with the smallest distance shall be selected for merging into a new individual, and then the distance between the new individual and other individuals shall be calculated, Then, the two individuals close to each other are merged until all the individuals are merged into one class, and the clustering tree is finally obtained,as shown in Fig 6.

The minimum distance is calculated by hierarchical cluster analysis

$$D(x,y)=\sum_{k=1}^{p}x_k-y_k[2]^{\frac{1}{2}} \tag{1}$$

The shortest distance is adopted between samples:

$$D(G_1,G_2)=\min_{\substack{x_i\in G_1\\ y_i\in G_2}}\{d(x_i,y_i)\} \tag{2}$$

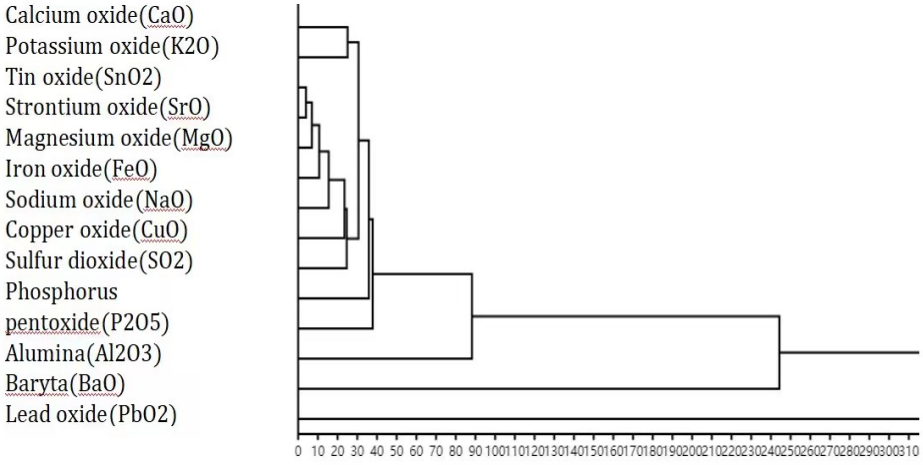



Fig. 6 Clustering tree of chemical composition

Those greater than 0 but less than or equal to 20 in the tree graph are classified as the first category, those greater than 20 but less than or equal to 40 are classified as the second category, and those greater than 40 are classified as the third category. The 14 chemical components can be divided into 3 categories. The first category is divided into tin oxide, strontium oxide, magnesium oxide, iron oxide and sodium oxide; The second category is divided into copper oxide, calcium oxide, potassium oxide, sulfur dioxide, phosphorus pentoxide, aluminum oxide; The third category is divided into barium oxide, lead oxide and silicon dioxide,as shown in Table 7.

Table 7 Subcategory

| first kind | Category II | Category III |
|---|---|---|
| Stannic oxide | Copper oxide | |
| Strontium oxide | calcium oxide | Barium oxide |
| magnesium oxide | Potassium oxide | Lead oxide |
| ferric oxide | sulfur dioxide | silicon dioxide |
| Sodium oxide | Phosphorus pentoxide | |
| | alumina | |

### 3.3. Rationality and sensitivity of classification results

Rationality: The importance of the data characteristics processed by the random forest model is used to randomly extract data from multiple sample data for model training. The results obtained from multiple processing of data by a large number of decision tree models are very stable and reliable. When the model is used multiple times, each application will not affect each other; Each random selection of sample data is very practical, and there is no single data, making the data more flexible[10]. The similarity between the minimum distances of each element of the hierarchical clustering model is high, the data clustering for many times is reasonable, the recoverability between each level is high, and the data is arbitrary.

Sensitivity: when the random forest model is used, the decision tree model is used many times to process data. With the increase of the number of decision trees, the accuracy rate will tend to be stable, but it will fluctuate in a stable range; The singular value generated by hierarchical clustering analysis

in the process of processing data has a great impact on the data, and the clustering algorithm will make the shape of the preprocessed data chain, which has a great impact on the classification results.

## 4. Conclusion

The random forest model is trained by randomly selected data in the sample data, which greatly improves the fitting of data results to the actual situation. The data is stable and reliable. The chi square test is a commonly used model for data fitting optimization in statistics. The fitting of multiple expected and actual frequencies is beneficial to the authenticity of the data.factor analysis can be used together to judge the degree of correlation of a certain type based on the model of factor analysis and correlation analysis. In the factor load matrix thermodynamic diagram, common factors are extracted from the variable group. According to whether some variables can be reduced to a factor, it indicates that the variables are highly correlated, or equivalent to how many factors are required to completely cover all variables. If fewer factors are required, it indicates that the relationship between different variables is closer, so that the correlation between chemical components can be directly visualized, It is not necessary to characterize the strength of the correlation between chemical components by comparison.

## References

[1] Fang Xiangzhong. Chi Square Distribution and Chi Square Test [J]. China Statistics, 2022, (05): 29-31

[2] Wang Qing. Automatic Classification Method of Library Collection Documents Based on Random Forest [J]. Automation Technology and Application, 2022,41 (07): 51-53+72. DOI: 10.20033/j.1003-7241. (2022) 07-0051-04

[3] Zhang Yonghua. Design and Implementation of Telecom Network Alarm Correlation Analysis Based on Big Data Technology [J]. Telecommunications Engineering Technology and Standardization, 2016, 29 (4): 18-23

[4] Li Long, Liu Zhenyi, Li Haoran, Yao Yonghong. Establishment of a predictive model for AMI patients complicated with AKI based on the random forest model [J/OL]. Chongqing Medical Science: 1-6 [2022-10-27]

[5] Ye Congzhou, Xiao Penglin, Qin Jun, Zhang Chengxiong, Chen Lie, Wang Ke. Research on prediction model of energy consumption load of super large buildings based on clustering and random forest regression [J]. Green Building, 2022-14 (05): 48-51+55

[6] Zhao Junmei, Zhang Liping, Liu Dan, Yi Feng. Research on regression analysis and prediction technology of power data based on machine learning [J/OL]. Journal of Testing Technology, 2022 (06): 525-529+536 [2022-10-27]

[7] Li Meng, Bao Lei, Hu Yi, Cheng Song, Hu Xiaobo, Gao Ying. Implementation of online random number detection method based on chi square test [J]. Microelectronics, 2022,52 (03): 388-392. DOI: 10.13911/j.cnki.1004-3365.210329

[8] Zhao Chenfei Research on feature selection method based on spectral feature analysis and chi square test [D]. Tianjin Normal University, 2019

[9] Sun Haolin Correlation analysis of driving behavior based on correlation analysis and frequent pattern mining [D]. Liaoning University of Engineering and Technology, 2019. DOI: 10.27210/d.cnki.glnju.2019.000471

[10] Li Cuiling, Li Yukang, Tan Haoran, Wang Xiu, Zhai Changyuan. Detection and grading method of grape downy mildew based on K-means clustering and random forest [J/OL]. Journal of Agricultural Machinery: 1-14 [2022-10-27]