ISSN 2616-5880 Vol. 4, Issue 1: 54-61, DOI: 10.25236/AJMC.2023.040110

Subclass classification and chemical composition analysis and identification of ancient glass products

Yaxuan Zhang[1], Yilin Chen[2]

[1]School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[2]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China

Abstract: *Glass is one of the first artificial materials invented by mankind, and its development has a long history. Glass production is usually locally sourced, and the glass-making process is similar in different regions, but the chemical composition is different. 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 changes in its composition ratio, which affects the correct judgment of its category. This paper has analyzed the content and correlation of its chemical components, and used principal component analysis and least squares multiple linear regression to realize the classification of glass type subclasses.*

Keywords: *Antique glassware; Multiple linear regression; Hierarchical clustering model*

1. Introduction

In the new era, glass has become a necessity in people's daily production and life, which can provide both convenience and material support for construction. Glass is one of the earliest artificial materials invented by human beings, and its development history has a long history[1]. Our ancient ancestors studied primitive porcelain in the late Shang Dynasty, while glass was invented by Egyptians and Western Asians from 4000 to 3000 BC, i.e., the appearance of glass was about 2000 years earlier than porcelain[2]. In our case, glass appeared later than the world's discovery of glass. Ancient glass in China was made locally by absorbing its technology, so it was similar in appearance to foreign glass products, but the chemical composition was different. The chemical composition of glassware in Xinjiang is soda-lime glass, which is commonly found in Western Asia, and the composition of glassware in late Warring States tombs is lead-barium glass[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 the proportion of its composition, which affects the correct determination of its category.

Glass is susceptible to weathering by environmental factors, leading to changes in the proportion of its chemical composition, while the detection means and other reasons may lead to uncertainty in the proportion of its composition, affecting the judgment of its type. The main work of this paper is as follows.

Analyze the classification rules of high potassium glass and lead-barium glass based on the collected data; select the appropriate chemical composition to classify each category into subcategories, and give the specific classification method and classification results, and conduct the reasonableness and sensitivity analysis on the classification results.

Analyze the chemical composition of glass artifacts of unknown categories in the data to identify the types to which they belong and the sensitivity to the results of the classification analysis.

2. Assumptions and notations

2.1 Assumptions

Use the following assumptions.

1) If the artifact sampling points do not specify whether they are weathered or not, their weathering

ISSN 2616-5880 Vol. 4, Issue 1: 54-61, DOI: 10.25236/AJMC.2023.040110

properties are defaulted to be the same as whether they are weathered or not given in the data.

2) All missing component values are set to 0.001.

3) Since the amount of data for samples tested for weathered high potassium is too small to be predicted using the model, it is assumed that its subsequent data are the same as the pre-test data.

4) Elements containing a large number of missing values are not predicted, and only important features are analyzed

5) If the missing value of a chemical component of a sample > 50% is not considered as an important influencing factor, and individual chemical components are not involved in the regression algorithm.

2.2 Notations

The primary notations used in this paper are listed as Table 1.

Table 1: Notations

D	Sample set
A	Actual number of different classifications
T	Expectation of different classifications
X^2	Cardinality
c	Vector of parameters to be estimated
N	Number of samples
p	Number of features
Y	Label vector
W	Weight vector
ESS	Sum of squared deviations
d(x,y)	Inter-sample distance
p_i	The proportion of samples of class i in the sample set D

3. Model construction and solving

3.1 Analysis of the classification pattern of each chemical composition for high potassium glass and lead-barium glass by principal component analysis, etc.

Table 2: Component Matrix

Name	Ingredients	
	1	2
Sulfur dioxide(SO2)	0.066	-0.112
Phosphorus pentoxide(P2O5)	0.128	0.145
Tin oxide(SnO2)	-0.015	0.102
Barium oxide(BaO)	0.067	0.232
Strontium oxide(SrO)	0.121	0.228
Silicon dioxide(SiO2)	-0.183	0.044
Lead oxide(PbO)	0.014	0.203
Iron oxide(Fe2O3)	0.159	0.009
Copper oxide(CuO)	0.09	-0.099
Sodium oxide(Na2O)	0.044	-0.204
Aluminum oxide(Al2O3)	0.172	0.001
Calcium oxide(CaO)	0.126	-0.22
Magnesium oxide(MgO)	0.145	0.125
Potassium oxide(K2O)	0.135	-0.161

In this paper, principal component analysis (PCA) and least squares (OLS) multiple linear regression were used to carry out the analysis of the classification patterns of each chemical component in the annexes for high potassium glass and lead-barium glass[4].

Firstly, Bartlett's sphericity test was performed, and if the significance (P) was less than 0.05 or less than 0.01, the original hypothesis was rejected, indicating that the principal component analysis could be done, and if the original hypothesis was not rejected, it indicated that these variables might provide some information independently and were not suitable for principal component analysis.

ISSN 2616-5880 Vol. 4, Issue 1: 54-61, DOI: 10.25236/AJMC.2023.040110

From the results of Bartlett's spherical test, it is clear that the value of significance p is 0.000***, which presents significance at the level, and the original hypothesis is rejected, and there is correlation between the variables and the principal component analysis is valid.

Using the factor loading matrix heat map generated by SPSS, the importance of the hidden variables in each principal component can be analyzed. The analysis of the hidden variables for each factor can also be performed in the context of specific operations[5].

Finally, the component matrix is derived as Table 2.

The above table illustrates the factor score coefficients (principal component loadings) included in each component, which are used to calculate the component scores and derive the factor formula, which is calculated as: linear combination coefficient * (variance explained / cumulative variance explained), and finally normalized to the factor weight score.

From the above coefficients the final formula of the model is obtained as

F1 = 0.066× sulfur dioxide (SO2) + 0.128× phosphorus pentoxide (P2O5) - 0.015× tin oxide (SnO2) + 0.067× barium oxide (BaO) + 0.121× strontium oxide (SrO) - 0.183× silicon dioxide (SiO2) + 0.014× lead oxide (PbO) + 0.159× iron oxide (Fe2O3) + 0.09× copper oxide (CuO) + 0.044× sodium oxide (Na2O) + 0.172× aluminum oxide (Al2O3) + 0.126× calcium oxide (CaO) + 0.145× magnesium oxide (MgO) + 0.135× potassium oxide (K2O).

From the above, it can obtain.

$$F=(0.369/0.583)\times F1+(0.214/0.583)\times F2. \quad (1)$$

The analysis is then continued with least squares multiple linear regression, where the feature matrix X of the training samples is represented, where N is the number of samples, p is the number of features, each row is represented as each sample, and each column represents each dimension of the feature.

$$X = \begin{pmatrix} x11 & x12 & \dots & x1p \\ x21 & x22 & \dots & x2p \\ \dots & \dots & \dots & \dots \\ xN1 & xN2 & \dots & xNp \end{pmatrix} \quad (2)$$

Then, the label vector Y and the weight vector w of the training samples are represented, where the weight vector refers to the vector formed by the individual coefficients in the linear regression.

$$Y = \begin{pmatrix} y1 \\ y2 \\ \dots \\ yN \end{pmatrix}, \quad W = \begin{pmatrix} w1 \\ w2 \\ \dots \\ wp \end{pmatrix} \quad (3)$$

In order to find the optimal estimate of the parameter w of the function f(x,w), the objective function L(y,f(x,w)) is solved for the given N sets of observations, taking the smallest parameter w. Where L(x) is called the residual function.

$$L(w) = \sum_{i=1}^{N}(xiw - yi)\text{^}2; \quad (4)$$

$$w=\arg\min L(w) = \arg\min\sum_{i=1}^{N}(xiw - yi)\text{^}2. \quad (5)$$

Code validation analysis by the above equation model yielded the following results.

Assuming the content of SiO2 as the basis for discriminating the type, the following example of unweathered high potassium, in unweathered lead-barium glass, P<=0.05 is significantly correlated, such as the following P2O5, PbO, BaO. where the closer the R^2 value is to 1 the more accurate, i.e. the more accurate the model is, but at the same time the more obvious the multiple covariance is.

In the same way, the unweathered high potassium glass, weathered high potassium glass, and weathered lead-barium glass were analyzed separately. The significant components of the unweathered high potassium glass are K2O, Fe2O3, and P2O5; the significant components of the weathered high potassium glass are K2O and Fe2O3; and the significant components of the weathered lead-barium glass are CaO, P2O5, PbO, and BaO.

Since the above calculation method is not very accurate for the evaluation of high potassium glass, an unsupervised learning hierarchical clustering model was established by combining the findings of PCA principal component analysis as a basis. For lead-barium glass, barium oxide and strontium oxide were taken as classification elements, and for high potassium glass, iron oxide and potassium oxide were

ISSN 2616-5880 Vol. 4, Issue 1: 54-61, DOI: 10.25236/AJMC.2023.040110

taken as classification elements.

3.2 Unsupervised hierarchical clustering method for subclassification of glass species

In this paper, the unsupervised hierarchical clustering method is chosen to create a hierarchical nested clustering tree by calculating the similarity between data points of different categories, and then to achieve the subcategory classification of glass types.

Clustering is to divide a large number of unknown labeled data sets into multiple categories according to the intrinsic similarity of the data, so that the data within the categories are more similar and the data between the categories are less similar.

Here, the sum of squares of deviations is used as the objective function to decide how to merge small class clusters into large class clusters.

$$\mathrm{ESS} = \sum_{i=1}^{n} x_i^{\,2} - \frac{1}{2}\left(\sum_{i=1}^{n} x_i\right)^2 \qquad (6)$$

The sum of squares of deviations after the merging of class clusters should be the smallest, i.e., the best objective function is to minimize the missing information after the merging of class clusters.

To divide the The data are divided into two classes by type and the distance between samples is calculated using the absolute value distance method.

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|, i = 1,2,\cdots,n \qquad (7)$$

Constitute an n × n distance matrix D. At this time each sample is a class cluster, then the sum of squares of the departures of each sample constituting a class cluster are 0.

The shortest distance method is used to calculate the distance between classes, and then merge them one by one according to the distance between classes. The shortest distance clustering method is to find $d_{pq} = min\{d_{ij}\}$ in the diagonal of the original n×n distance matrix,and to merge the classification objects G_p and G_q into a new class G_r, and then according to the calculation formula.

$$d_{rk} = min\{d_{pk}, d_{qk}\}(k \neq p, q) \qquad (8)$$

The distance between the original classes and the new class is calculated, so that a new distance matrix of order (n-1) is obtained, and the above process is repeated until each classified object is grouped into one class.

Some data are eliminated and the remaining data are renumbered by 0 to 66. Where the vertical coordinate represents the Euclidean distance and the interval shown in the horizontal coordinate is the interval containing the best number of data in each class. Figure 1 below shows the clustering tree obtained by python.

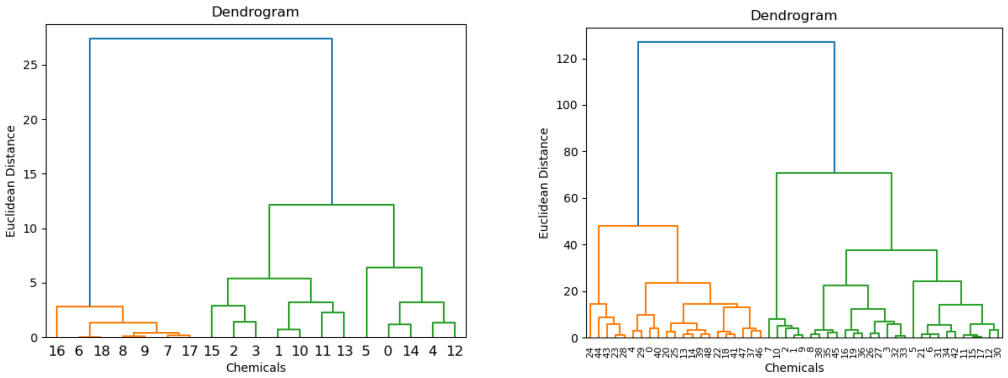

Figure 1: High potassium and lead barium glass clustering tree

3.3 Reasonableness and sensitivity analysis of classification results

In this paper, the Plackett-Burman experimental design method was used for the reasonableness and sensitivity analysis. the Plackett-Burman experimental design method was mainly used to analyze each factor by taking two levels, and the significance of the factors was determined by comparing the

ISSN 2616-5880 Vol. 4, Issue 1: 54-61, DOI: 10.25236/AJMC.2023.040110

difference between the two levels of each factor and the overall difference.

Take high potassium weathered glass as an example: let y be a function about the content of each chemical component, y=xi*Facter n; where xi is the percentage of each chemical element and Factor n is a random variable taking the value of 1 or -1. The ANOVA table is shown in Table 3.

Table 3: The ANOVA table

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	17.5513714	1.25367	2235655	<	0.0001
A-A	17.53562	1	17.5356231271085	<	0.0001
B-B	5.61E-07	1	5.61E-07	1	0.3632
C-C	0.001266	1	0.001266	2257.381	< 0.0001
D-D	0.001446	1	0.001446	2579.189	< 0.0001
E-E	0.000653	1	0.000653	1164.109	< 0.0001
F-F	0.007272	1	0.007272	12968.87	< 0.0001
G-G	0.000158	1	0.000158	281.3607	< 0.0001
H-H	0.00495	1	0.00495	8827.796	< 0.0001
J-J	5.61E-07	1	5.61E-07	1	0.3632
K-K	5.61E-07	1	5.61E-07	1	0.3632
L-L	5.61E-07	1	5.61E-07	1	0.3632
M-M	5.61E-07	1	5.61E-07	1	0.3632
N-N	5.61E-07	1	5.61E-07	1	0.3632

Based on the above ANOVA table, it can be seen that the p-values of significance for the chemical components potassium oxide (C) and iron oxide (G), on which the classification of high potassium glass subclasses is based, are less than 0.0001, and therefore the classification method can be judged to be reasonable and sensitive.

3.4 Chemical composition analysis of glass artifacts of unknown category

The data can be clustered using a clustering model in python. The dataset was first divided into weathered and unweathered, and the dataset was put into the dependent and independent variables as a test set. The two main characteristics of weathered and unweathered were selected based on the findings of the problem two principal component analysis and least squares regression. Two kinds of clustering trees were obtained, and according to the calculation of Euclidean distance both showed the best into two clusters. Among them, cluster1 should be lead-barium glass and cluster2 is high potassium glass. Then the classification results can be known according to the content control data corresponding to the coordinates of their points.

In addition, a decision tree model was used to study the type classification of glass artifacts containing different chemical compositions. From problem 2, it can be concluded that potassium oxide, iron oxide, phosphorus pentoxide, calcium oxide, lead oxide, and barium oxide are the significant components that have a greater influence on the type, and these variables are selected to construct the model in this paper.

The algorithm flow is as follows.

Initialize the attribute set and data set.

Calculate the information entropy of the data set and the information entropy of all attributes, and select the attribute with the greatest information gain as the current decision node.

Update the data set and the attribute set (remove the attributes used in the previous step and divide the data set into different branches according to the attribute values).

Repeat the second step for the subsets in each value case in turn.

If the subset contains only a single attribute, the branch is a leaf node and is labeled according to its attribute value.

The division of all attribute sets is completed.

Information entropy is the most commonly used metric to measure the purity of a sample set. Suppose the current sample set D has a total of k classes of samples in which the proportion of the ith class of samples is pi, then the information entropy of D is defined as

ISSN 2616-5880 Vol. 4, Issue 1: 54-61, DOI: 10.25236/AJMC.2023.040110

$$\mathrm{Ent(D)} = -\sum_{i=1}^{k} p_i \log_2 p_i \quad (9)$$

Suppose the attribute a of the sample set has n different values, sort these values from smallest to largest to get the set of attribute values $a_1, a_2, \cdots, a_n$, take the median $\frac{a_i+a_{i+1}}{2}$ of the interval a_i, a_{i+1} as the candidate division points, so this paper get the set of division points $T_a = \left\{\frac{a_1+a_{i+1}}{2} | 1 \le i \le n-1\right\}$ containing n-1 elements., based on each division point t, the sample set D can be divided into subsets D_t^- and D_t^+

For each division point t, its information gain value is calculated according to Eq.

$$\mathrm{Gain(D,a,t)} = \mathrm{Ent(D)} - \sum_{\lambda\in\{-,+\}} \frac{|D_t^{\lambda}|}{|D|} \mathrm{Ent}(D_t^{\lambda}) \quad (10)$$

Gain(D,a,t) is the information gain of sample set D after dichotomization based on division point t. The division point should maximize Gain(D,a,t).

The following is an example of calculating the information gain of the attribute "potassium oxide content".

The set of candidate division points for this attribute consists of 34 values.

T={0，0.11，0.14，0.15，0.2，0.21，0.23，0.25，0.26，0.29，0.3，0.32，0.34，0.35，0.4，0.44，0.59，0.71，0.74，0.92，1.01，1.05，1.41，5.19，7.37，7.68，9.42，9.67，9.99，10.95，12.28，12.37，12.53，14.52}

$$\mathrm{Gain(D,a,t=0)} = \mathrm{Ent(D)} - \sum_{\lambda\in\{-,+\}} \frac{|D_0^{\lambda}|}{|D|} \mathrm{Ent}(D_0^{\lambda})$$

$$=0.860-\frac{1}{67}(-1\log_2 1)-\frac{34}{67}\left(-\frac{17}{34}\log_2\frac{17}{34}-\frac{17}{34}\log_2\frac{17}{34}\right) \quad (11)$$

$$=0.860-0-0.507=0.353$$

$$\mathrm{Gain(D,a,t=0.11)} = \mathrm{Ent(D)} - \sum_{\lambda\in\{-,+\}} \frac{|D_{0.11}^{\lambda}|}{|D|} \mathrm{Ent}(D_{0.11}^{\lambda})$$

$$=0.860-\frac{2}{67}\left(-\frac{2}{2}\log_2\frac{2}{2}\right)-\frac{33}{67}\left(-\frac{17}{33}\log_2\frac{17}{33}-\frac{16}{33}\log_2\frac{16}{33}\right) \quad (12)$$

$$=0.860-0-0.492=0.368$$

$$\mathrm{Gain(D,a,t=0.14)} = \mathrm{Ent(D)} - \sum_{\lambda\in\{-,+\}} \frac{|D_{0.14}^{\lambda}|}{|D|} \mathrm{Ent}(D_{0.14}^{\lambda})$$

$$=0.860-\frac{3}{67}\left(-\frac{3}{3}\log_2\frac{3}{3}\right)-\frac{32}{67}\left(-\frac{17}{32}\log_2\frac{17}{32}-\frac{15}{32}\log_2\frac{15}{32}\right) \quad (13)$$

$$=0.860-0-0.473=0.387$$

The information gain of all division points can be calculated by the above method, and the information gain of each attribute can be calculated, so that the optimal division attribute can be selected and the decision tree can be constructed(Figure 2 and Figure 3).

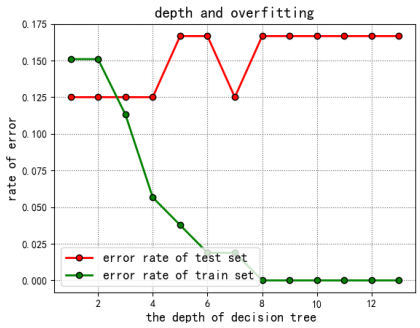

Figure 2: The depth and overfit of the decision tree

ISSN 2616-5880 Vol. 4, Issue 1: 54-61, DOI: 10.25236/AJMC.2023.040110

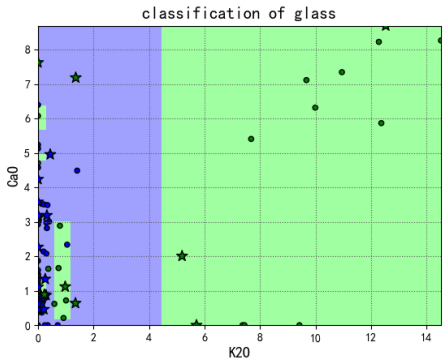

Figure 3: The classification of the glass

Firstly, the missing values in the form are made up to 0, and then the decision tree algorithm model is trained using python

The accuracy of the obtained trained model is high. The model evaluation results are shown in Table 4.

Table 4: Model evaluation results

	Accuracy	Recall Rate	Accuracy rate	F1
Training set	1	1	1	1
Test set	0.952	0.952	0.96	0.954

The above table shows the classification evaluation metrics for the training and test sets, and the quantitative metrics are used to measure the classification effectiveness of the decision tree on the training and test data.

● Accuracy: the proportion of correctly predicted samples to the total samples, the greater the accuracy, the better.

● Recall rate: The proportion of predicted positive samples out of the actual positive samples, the larger the recall rate, the better.

● Accuracy rate: The proportion of the predicted positive samples to the actual positive samples, the larger the accuracy rate, the better.

F1: The sum of precision and recall. Precision and recall affect each other, although a high level of both is the desired ideal situation, in practice, it is often the case that a high precision rate results in a low recall rate, or a low recall rate results in a high precision rate. If both are needed, then the F1 metric can be used.

From the evaluation results it is clear that the prediction model is relatively good.

4. Conclusion

In this paper, Bartlett sphericity test is first conducted to determine whether principal component analysis can be carried out, and hidden variables of principal components are analyzed to obtain factor score coefficients contained in each component used for model construction. Then, least square multiple linear regression is conducted to continue the analysis, and SiO2 content is used as the basis for discrimination type, and it is concluded that the important component of fresh high potassium glass is K2O. Based on the above research, this paper uses barium oxide and strontium oxide as the classification elements of lead barium glass, and uses iron oxide and potassium oxide as the classification elements of high potassium glass for classification, and finally obtains the cluster tree of two types of glass, which can be selected according to the number of classification required. Then the Plackett Burman experimental design is used for global sensitivity analysis, and the classification method is more reasonable according to the significance p value.

In addition, this paper also uses the decision tree model to study the type classification of glassware with different chemical components, selects important components that have a greater impact on the type, such as potassium oxide, iron oxide and phosphorus pentoxide to build the model, assumes that the samples are high potassium glass and lead barium glass, and puts the data into the test set that accounts

ISSN 2616-5880 Vol. 4, Issue 1: 54-61, DOI: 10.25236/AJMC.2023.040110

for 30% of the total data set for training, obtains relatively high accuracy, and completes the classification of unknown glass categories, The sensitivity analysis of the results shows that the model has good accuracy and stability.

References

[1] Chen Shuyu. The origin and development of ancient glass in China [J]. Cultural Identification and Appreciation, 2019(4): 44-45.
[2] Zhou Jing. Glass trade on the Silk Road and the eastern transmission of glass manufacturing technology [J]. Journal of Suzhou College of Arts and Crafts, 2017(4): 15-18.
[3] Gao Shan. Silk Road glass was once considered a treasure by the Chinese [J]. World Culture, 2018(8): 40-41.
[4] Jing Zhongquan, Jiang Xiuhui et al. Research on the index system of coal mine safety production capacity based on hierarchical analysis (APH) [J]. China Journal of Safety Science. 2006,16(9): 74-79
[5] Ma Xiuhong, Song Jianshe, Dong Shengfei. Exploration of decision trees in data mining [J]. Computer Engineering and Applications, 2004, (2): 133-135.