Analysis and identification of the composition of ancient glass products based on regression models

Zeyao Li *, Yanlin Zeng, Weiting Zhang

School of Chemical Engineering and Technology, Xiangtan University, 411105

* Corresponding Author Email: 2705326072@qq.com

Abstract. The Silk Road was an important channel for economic exchanges between China and foreign countries in ancient times, and glass products were an important physical evidence of trade exchanges. Early glass was introduced to China in the form of bead-shaped products, and on this basis, craftsmen improved it to form the unique ancient Chinese glass. Glass products are important physical evidence of the ancient Silken Road, but they are vulnerable to weathering due to the influence of the burial environment. In this paper, we use a batch of ancient glass artifacts as the research object, and build a machine learning classification model and a clustering model based on the different chemical composition content to analyze the classification basis and the relationship between the chemical composition content of the glass artifacts and optimize the research.

Keywords: Ridge regression; Logistic regression; Cluster analysis.

1. Introduction

The Silk Road was an important channel for economic exchanges between China and foreign countries in ancient times, and glass products were an important physical evidence of trade exchanges. Early glass was introduced to China in the form of bead-shaped products, and on this basis, craftsmen improved it to form the unique ancient Chinese glass. Archaeologists have discovered a number of ancient glass objects, which are susceptible to different degrees of weathering due to the burial environment and therefore have a significant impact on the determination of cultural heritage categories. The chemical composition of the artifacts and professional testing methods were used to classify these glass objects into two types: high potassium glass and lead-barium glass[1]. In this paper, we analyze the relationship between weathering and type, decoration, and color on the glass surface based on relevant data; we analyze the statistical pattern of chemical composition content between unweathered and weathered glass according to different glass types, and predict the chemical composition content of weathered spots when they are unweathered. To investigate the basis of classification of high potassium glass and lead-barium glass, and to classify the two glass types into subclasses, and to give reasonable classification methods and results.

2. Model assumptions and notation

2.1. Assumptions [2]

1. the chemical composition of the glass artifacts in the products only 14 components are considered.
2. the same type of glass artifacts buried in the same environment.
3. The chemical composition content of the glass artifacts is relatively stable and does not change during the measurement period.
4. The testing equipment has no effect on the chemical composition content of the glass artifacts.
5. If the weathered glass artifacts are detected as unweathered points at the detection point, the data are treated as unweathered glass artifacts when processing.

2.2. Notations

Important notations used in this paper are listed in Table 1.

Table 1. Notations

Symbols	Definition of symbols
p_k	Proportion of each substance in the sample
$S(X,Y)$	Confidence level
$C(X \Leftarrow Y)$	Degree of support
x_i	Chemical substances after weathering
x_i'	Unweathered chemical substances
T	Type
S	Surface weathering
O	Decoration
C	Color
$\alpha^{(k)}(t)$	Coefficient before each indicator
z_{ij}	Elemental substance in column i, row j
ρ	Spearman's correlation coefficient

3. Model construction and solving

3.1. Relationship Analysis

3.1.1 Decision tree model building and solving

In this paper, we construct a decision tree classification model [3] by considering whether the surface is weathered as the dependent variable Y and glass type, decoration, and color as the independent variables X, so as to derive the importance ratios of these three characteristics. The impurity-GINI coefficients of the decision tree nodes were calculated as follows.

$$\text{Gini}(p) = \sum_{k=1}^{K} p_k \left(1 - p_k\right) \tag{1}$$

The Entropy entropy is then calculated as

$$H(x) = -\sum_{i=1}^{n} p_i \log_i \tag{2}$$

Using SPSS processing, the results are shown in Figure 1, where the importance of glass type is 53.60%, the importance of ornamentation is 29.60%, and the importance of color is 16.80%.

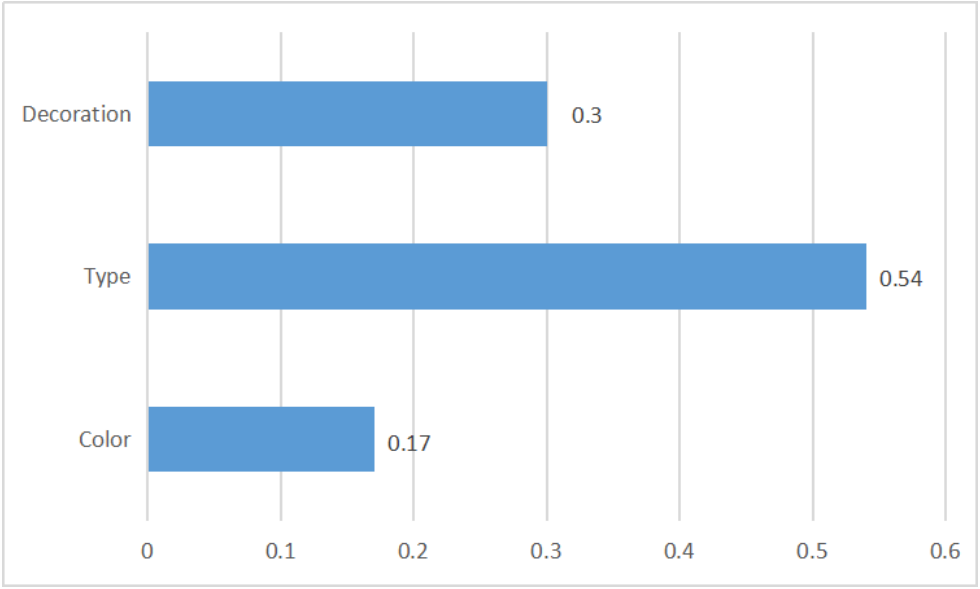

Figure 1. Decision tree classification feature importance

3.1.2 Association rule modeling and solving

After calculating the importance ranking of the above three characteristics on the weathering of the artifacts, the Apriori algorithm model [4] was constructed to further analyze the deep influence of each of the three characteristics of type, color, and decoration on the weathering of the artifacts by mining the data sets that occur frequently in the data values. The confidence and support data of each factor were calculated separately. Confidence level.

$$S(X,Y)=\frac{\text{number }(XY)}{\text{num(all)}} \tag{3}$$

Support level.

$$C(X \Leftarrow Y)=\frac{P(XY)}{P(Y)} \tag{4}$$

After calculating the confidence and support of each index, the frequent n itemsets were searched and filtered by iteration, and the final frequent k itemsets were obtained after several iterations, i.e., the frequent k+1 itemsets could not be obtained. Through the calculation, the combination with a high confidence level was filtered out, i.e., Class C ornamentation, light blue color and lead-barium glass type, which has a confidence level of 100%. Therefore, these combinations have a greater influence on whether the surface of the artifact is weathered or not, and there is a deeper relationship between them [5].

From the combined results of the decision tree classification algorithm and association rule mining algorithm, whether the artifacts will weather or not is most closely related to the type The most weathering-prone combinations are the C motif, light blue, and lead-barium glass type The most weatherable combinations are Class C ornament, light blue and lead-barium glass type.

3.2. Statistical patterns

The pre-processed data were classified according to whether the glass surface was weathered or not, and were divided into four categories: unweathered high potassium glass, weathered high potassium glass, unweathered lead-barium glass, and weathered lead-barium glass. The mean, variance, standard deviation, skewness, and coefficient of variation of the 14 chemical components were obtained by using SPSS [6].

The highest content of both types of glass before and after weathering is SiO2, and the average content of all chemical components decreases after weathering except SiO2, and the variance of each chemical component decreases, i.e. the data distribution is more concentrated. After weathering, the average content of all chemical components decreased to different degrees except for PbO2, and the variance of each chemical component increased to different degrees, with the most obvious changes in SiO2, CuO2 and PbO2.

3.3. Predicted content

Based on the above-mentioned statistical patterns of chemical content on the surface of the artifacts, two methods were chosen to predict the chemical content of weathered artifacts before weathering. In the case of high potassium glass, the above-mentioned statistics show that the fluctuation of the chemical content after weathering is not significant, and the difference of the mean values can be used to predict the chemical content of high potassium glass before weathering. Let xi be the chemical composition (i=1, 2..., 14) (e.g., x_1: x_1: x_2). (e.g., x_1: silicon dioxide (SiO2), x_2: sodium dioxide (N aO2), and so on), and then calculate the mean value of each chemical composition xi for weathered and unweathered high-potassium glass, and then calculate the difference.

$$Q_i = x_i - x_i^{'} \tag{5}$$

For lead-barium glass, the artifacts with severe weathering points cannot be predicted using the above approach, so a multiple regression equation model is constructed to fit the prediction. As a kind of multiple regression equation, ridge regression can solve the problem of multiple covariance between multiple independent variables X. The data of each chemical composition of lead-barium glass were brought into the matrix of the ridge regression equation, and the ridge regression equation for each chemical substance was obtained as follows.

$$w_i = \left(X^T X\right)^{-1} X^T Y \tag{6}$$

The final solution of this equation leads to the model equation for each type of substance.

$$x_i = c_i^{(0)} + c_i^{(1)} + c_i^{(2)} * s \tag{7}$$

The glass type, the presence or absence of weathering and the weathering point detection data are selected and the ridge regression model is solved to obtain the following metric Equation.

$$\begin{cases} \mathrm{CaO} = 5.852 - 1.423 \times t - 0.526 \times s \\ \mathrm{K_2O} = 14.09 - 4.677 \times t - 2.548 \times s \\ \mathrm{SiO_2} = 114.596 - 31.48 \times t - 6.805 \times s \\ \mathrm{SrO} = -0.279 + 0.272 \times t + 0.043 \times s \\ \mathrm{PbO} = -36.208 + 26.4 \times t + 9.198 \times s \\ \mathrm{P_2O_5} = -2.63 + 1.39 \times t + 1.788 \times s \\ \mathrm{Na}a_2\mathrm{O} = 0.152 + 0.415 \times t - 0.052 \times s \\ \mathrm{BaO} = -8.106 + 9.0 \times t + 0.187 \times s \\ \mathrm{SnO_2} = 0.25 - 0.044 \times t - 0.059 \times s \\ \mathrm{MgO_1} = 1.004 - 0.088 \times t - 0.104 \times s \\ \mathrm{Fe_2O_3} = 2.657 - 0.4 \times t - 0.686 \times s \\ \mathrm{SO_2} = -0.969 + 0.502 \times t + 0.433 \times s \\ \mathrm{CuO} = 2.352 - 0.259 \times t + 0.031 \times s \\ \mathrm{Al_2O_3} = 7.063 - 0.938 \times t - 0.86 \times s \end{cases} \tag{8}$$

For the high potassium glass before weathering and after weathering fluctuations in the content of each chemical composition were taken as the average value, the average value of the two for the difference, and again with the weathering of each chemical composition content value plus the difference, the content of fluctuations in the chemical composition is obviously substituted into the formula 8, you can get the weathering point when the content of each chemical composition without weathering, due to the amount of data, only 10 samples are listed here, the chemical composition content The chemical composition contents are 70.791%, 0.695%, 9.708%, 4.673%, 0.883%, 5.5%, 1.927%, 1.731%, 0.412%, 0.598%, 1.123%, 0.0417%, 0.197%, and 0.102% in order.

For the lead barium glass before weathering and after weathering fluctuations in the content of each chemical composition are not obvious to take the average value, the average value of the two for the difference processing, again with the weathering of each chemical composition content value plus the difference, the content of the chemical composition fluctuations are obvious to substitute into the formula 8, you can get the weathering point when the content of each chemical composition, due to the amount of data, here only 11 samples are listed, chemical composition content The chemical composition content is 61.553%, 0.562%, 0.349%, 2.087%, 0.534%, 3.220%, 0.317%, 4.207%, 3.415%, 14.952%, 4.957%, 0.255%, 0.0442%, 0.084%.

The ridge regression model was solved by selecting glass type, decoration, color, presence of weathering and weathering point detection data, and the following equations were obtained.

$$x_i = \alpha^{(0)} + \alpha^{(1)}(t) \times O_2 + \alpha^{(2)}(t) \times O_3 + \alpha^{(3)}(t) \times T + \sum_{i=4}^{11} \alpha^i(t) \times C_{i-3} + \alpha^{(12)}(t) \times B \tag{9}$$

By substituting the data of the three severely weathered points into the equation, the chemical composition content when unweathered can be obtained, and only the data of sample 54 are presented here, with the following chemical composition contents: 1.532%, 30.279%, 3.611%, 1.226%, 0.475%, 42.029%, 5.836%, 1.413%, 5.141%, 0.13%, 0.745%, 1.036%, 0, 4.771%. 0.745%, 1.036%, 0, 4.771%.

3.4. Classification rules

3.4.1 Entropy method modeling and solving

For the classification of high potassium glass and lead-barium glass, the entropy weight method and the logistic regression gradient descent method are combined in this paper. The entropy weighting method assigns weights based on the degree of variation of an index, which in this paper is based on the chemical composition of lead-barium glass and high-potassium glass to find the chemical composition with the higher weight [7-9]. It is necessary to zero the data before use and also to determine whether there are negative values in the input data, which is not necessary since the chemical components are positive. Then, the weight of the i-th sample under the jth indicator in the matrix is calculated, and this weight is used as the probability required for the relative entropy calculation.

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^{n} z_{ij}} \tag{10}$$

Calculation of information entropy.

$$e_j = -\frac{1}{\ln n} \sum_{n=1}^{n} p_{ij} \ln\left(p_{ij}\right) \tag{11}$$

The valid data were processed based on the above equations to obtain the ranking of 14 chemical components in high potassium glass (Table 2) and lead-barium glass (Table 3), respectively.

Table 2. High Potassium Glass Entropy Method Ranking

Ranking	Chemical composition	Ranking	Chemical composition	Ranking	Chemical composition
1	SiO_2	6	Fe_2O_3	11	PbO
2	K_2O	7	P_2O_5	12	SnO_2
3	Al_2O_3	8	MgO	13	SO_2
4	CaO	9	Na_2O	14	SrO
5	CuO	10	BaO		

Table 3. Lead barium glass entropy method ranking

Ranking	Chemical composition	Ranking	Chemical composition	Ranking	Chemical composition
1	SiO_2	6	CuO	11	Mgo
2	PbO	7	CaO	12	SrO
3	BaO	8	SO_2	13	K_2O
4	Al_2O_3	9	Na_2O	14	SnO_2
5	P_2O_5	10	Fe_2O_3		

It can be seen that the four components of SiO_2, K_2O, Al_2O_3 and BaO are more abundant in the high potassium glass and lead barium glass are more abundant.

3.4.2 Logistic regression model building and solving

The above four components were selected to construct a logistic regression gradient descent method [10] to further analyze their effects on the classification patterns of high potassium glass and lead-barium glass to further analyze their effects on the classification laws of high potassium glass and lead-barium glass. Firstly, from the logistic regression model, the function expression can be found as.

$$g(X)=\frac{1}{1+e^{x}} \tag{12}$$

In machine learning the likelihood function of the logistic regression model can be obtained due to the presence of training samples.

$$L(\theta)=\prod\left[g\left(x_i\right)\right]^{y_i}\left[1-g\left(x_i\right)\right]^{1-y_i} \tag{13}$$

This is also the desired objective function. Then in considering the logistic regression model under the influence of gradient descent, since the gradient The gradient descent formula is

$$\theta_j=\theta_j-\alpha\frac{\partial}{\partial\theta_j}\jmath\left(\theta_0,\theta_1,\ldots,\theta_n\right) \tag{14}$$

Therefore, it is known that the objective function L(θ) can be expressed as

$$\frac{\partial L(\theta)}{\partial\theta_k}=\sum_{i=1}^{m}x_{ik}\left[y_i-g(x)\right] \tag{15}$$

Then the solution of the logistic regression gradient descent model can be found by the above equation.

The above four chemical components are selected and solved by logistic regression gradient descent method using SPSSPRO, as shown in Figure 2.

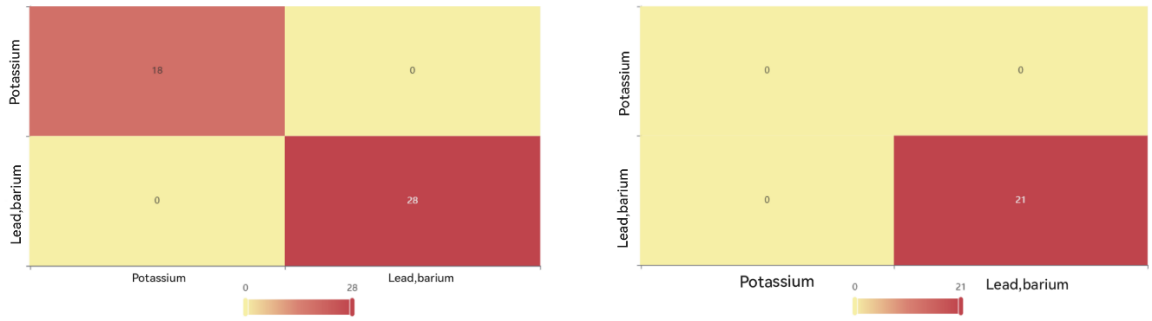

Figure 2. Logistic regression classification

From the above figure, it can be seen that the logistic regression prediction results are better, i.e., the chemical composition selected by the entropy method is reasonable of the chemical composition.

Combining the entropy weight method and logistic regression method, it can be seen that the classification of high potassium glass and lead-barium glass is based on the chemical composition The classification of high potassium glass and lead-barium glass is based on the amount of chemical composition, i.e., those with high BaO content are lead-barium glass and those with low BaO content are high potassium glass.

4. Conclusion

In this paper, a machine learning classification model was developed to determine the relationship between glass weathering and its type, color, and decoration; after that, descriptive statistics were

conducted on the chemical composition content of the glass surface, and a ridge regression model was developed to predict the type of glass, the presence or absence of weathering on the surface of the artifact, and the chemical composition content of each type as the independent variables, and the dependent variables. The chemical composition content of the weathered sites was predicted without weathering. A logistic regression model was developed and the glass was classified according to the above ranking to verify whether the entropy model was reasonable.

References

[1] Lv Xiaoling, Song Jie. Big data mining and statistical machine learning [M]. People's University of China Press: Big Data Analytics and Statistics Applications Series, 2016, 07:2394.

[2] Zheng JG. Data mining and its application research [M]. Yunnan University Press: West Yunnan Academic Series, 2014, 05.113.

[3] Fei Zhicong. Entropy-Hierarchy Analysis and Gray-Hierarchy Analysis [D]. Tianjin University, 2009.

[4] Joseph M. Hilbe. Logistic Regression Models [M]. Taylor and Francis;CRC Press:2011- 03-23.

[5] Fang Xiangzhong. Cardinality distribution and cardinality test[J]. China Statistics,2022(05):29-31.

[6] Li Hongcheng, Zhang Maojun, Ma Guangbin. SPSS data analysis practical tutorial [M]. People's Post and Telecommunications Publishing House, 2017, 03: 338.

[7] Xiuli He. Research on Multivariate Lineaar Model and Ridge Regression [J]. Huazhong University of Science and Technology, 2006.

[8] Hu Hongxiao, Xie Jia, Han Bing. Comparative Study on Missing Value Processing Methods [J]. School of Statistics, Southwestern University of Finance and Economics, 2007.

[9] Lu Chun. On the Treatment of Data Missing in Building Model with Grey Theory[J]. Journal of Liaoning Provincial College of Communications 2013.

[10] Li Fang, Li Dongping. Combination evaluation model based on entropy weight method [J] Information Technology and Informatization, 2021.