Analysis of the identification results of the ancient glass products based on the progressive regression model

Hongkai Chen*, Xin Zheng, Yusi Feng

Taiyuan University of Technology, Taiyuan, China

*Corresponding author: hongkai_chen2002@163.com

Abstract. Glass is a precious witness of trade on the Silk Road. It was first introduced into China from West Asia and Egypt. Its production method was also acquired by ancient Chinese craftsmen, and local glass was created. Ancient glass is easily weathered by the influence of buried environment. In the process of weathering, the proportion of chemical composition of cultural relics will change, thus affecting the correct judgment of its category. This paper conducts the composition analysis and identification of ancient glass products by establishing the mathematical model of the correlation before and after different types of glass weathering. This paper first preprocesses the data to generate dummy variables. Then the logic and regression model are established, take the Sigmoid function as the connection function, then calculate the regression coefficient β with the maximum likelihood estimation, and then use the binary Logistic regression to compare the predicted value with the true value, to obtain the prediction success rate and the identification results. A stepwise regression model was then used to perform the sensitivity analysis of the classified results. Finally, the model is optimized, the data is imported for the Fisher linear discrimination, and the results obtained by the Fisher discrimination method are compared with the logistic regression results to make the results more accurate.

Keywords: Glass, Sensibility Analysis, Logic Regression Model.

1. Introduction

Ancient glass was easily weathered by the buried environment. During weathering, the internal elements are greatly exchanged with the environmental elements, and the glass loses its crystalline water, resulting in a change in its composition ratio, thus affecting the correct judgment of its category [1-2].

According to the relevant data of a batch of ancient Chinese glass products, archaeologists divide them into high-potassium glass and lead-barium glass according to professional means [3]. Annex Form 1 gives the classification information of these cultural relics, and Annex Form 2 gives the corresponding proportion of the main components (the blank space is not detected). These data are characterized by the composition, that is, the accumulation of each component ratio should be 100%, but some errors may occur due to the detection means, resulting in the accumulation of its component proportion and non-100% [4-5].

We were asked to identify the type of unknown category glass artifacts by analyzing the chemical composition, and to analyze the sensitivity of the classification results. It can be transformed into the properties of known samples, requiring the determination of the type of samples and the analysis of the sensitivity of the results. The idea for the following question: First, the chemical composition data of all samples is imported into Spass and virtual variables are generated, and then a logical regression model (Logistic Regression) is established to conduct logical regression on the known sample data to determine whether the classification is reasonable [6]. Then, the overbinary Logistic regression identified the type of glass relics, and the data was fitted to analyze the sensitivity of the classification results. Finally, the model is optimized using the Fisher discrimination method [7].

2. Model building and solution

2.1. Data preprocessing

The chemical composition of the unclassified glass relics in form 3 was accumulated, and the cumulative sum was found to be between 85% and 105%, so they are all valid data. The chemical composition data of all the collated samples is then imported into Spass and its corresponding glass category, and the virtual variable t is generated to represent the classification of the existing data. The t=1 indicates that the relic sample belongs to high potassium glass and t=0 belongs to lead barium glass.

2.2. Establish a logistic regression model

Since the types of sample cultural relics are high-potassium glass and lead-barium glass, that is, the dependent variable is the categorical variable, and only two cases are 0-1. Thus, a logistic regression is used for processing. Defining y as the probability of the event, y 0.5 indicates the occurrence, that is, the relic belongs to high potassium glass; y <0.5 indicates no occurrence, that is, the relic belongs to lead-barium glass. The random variable Y can only take two values of 0 and 1, and its probability distribution is:

$$P\{Y=1\}=p, P\{Y=0\}=1-p \tag{1}$$

Consider the two-point distribution probability of the explained variable y given the explanatory variable x:

$$\begin{cases} P\{y=1 \mid x\}=F(x,\beta) \\ P\{y=0 \mid x\}=1-F(x,\beta) \end{cases} \tag{2}$$

F (x,β) is a connection function, used to connect x and y, and β is the regression coefficient. It is necessary to guarantee that the value domain of the connection function is [0,1] because of

$$E(y \mid x)=0\times P(y=0 \mid x)+1\times P(y=1 \mid x)=P(y=1 \mid x) \tag{3}$$

According to the knowledge of probability theory and mathematical statistics, take the Sigmoid function as the connection function F (x,β):

$$F(x,\beta)=S(x_i,\beta)=\frac{\exp(x_i,\beta)}{1+\exp(x_i,\beta)} \tag{4}$$

The Sigmoid function is a monotonic increasing function where the value domain is [0,1], whose function image is shown in Figure 1.

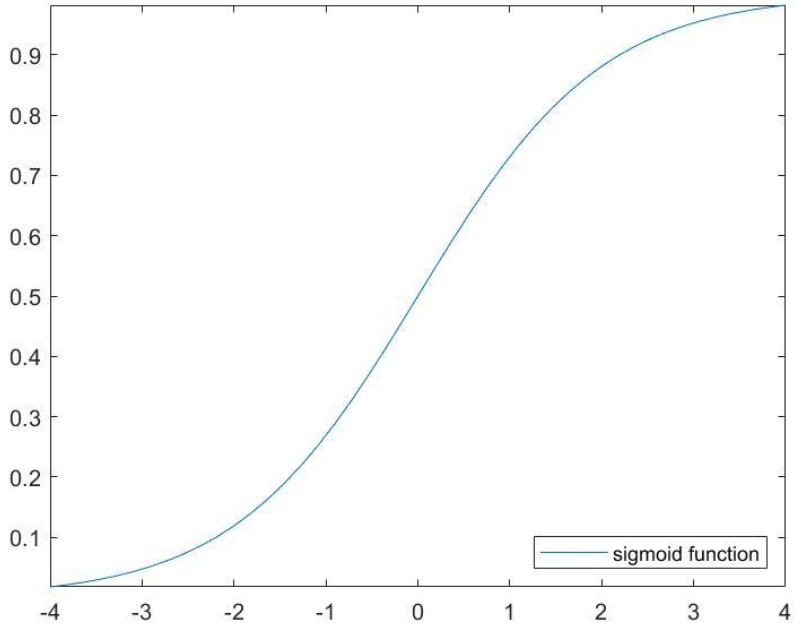

Figure 1. Figure 1. Function Image of Sigmaid.

For the imported data to be detected, when the yˆ0.5 is obtained after the logistic regression, the predicted y=1 is considered, which belongs to the high-potassium glass, or otherwise, the lead-barium glass.

Since the classification data of a large number of samples have been detected, a logical regression of the known sample data is used to compare the predicted y value with the virtual variable t, and when the two values are equal, the prediction is considered a success, or otherwise a failure[8-10]. The data from form 2 were imported into Spass, and the binary Logistic regression was selected to compare the regression predicted value with the real value, and the prediction success rate is shown in Table 1.

Table 1. Prediction success rate

The actual category		forecast		
		High potassium glass		Correct rate
		0.00	1.00	
High potassium glass	0.00	49	0	100
	1.00	0	18	100

In other words, logistic regression has a prediction success rate of 100% for 18 high potassium glass samples and 49 lead-barium glass samples of 100%, so it is reasonable and accurate to classify the classified samples by means of logistic regression.

2.3. Identify the category

Import the 8 sets of data of A1-A8 in form 3 into Spass, select the binary Logistic regression, and the resulting yˆvalues and the prediction results are shown in the table:
It is worth noting that the yˆvalues obtained from the logistic regression are all very close to 1 and 0, indicating that the above predicted values have a very high confidence level and can classify the unknown categories very accurately.

2.4. Sensibility analysis

From the above analysis, we can find that using logistic regression enables very accurate classification and prediction of samples of unknown categories. However, as the prediction power

improves, the overfitting phenomenon also becomes more and more obvious, and the overfitting linear nature is simulated using matlab, as shown in Figure 2 and Figure 3.

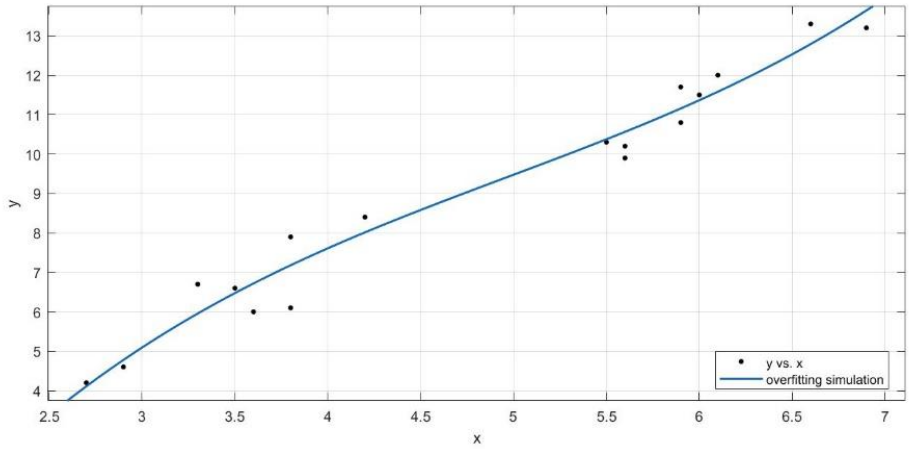

Figure 2 Normal fitting of the function images

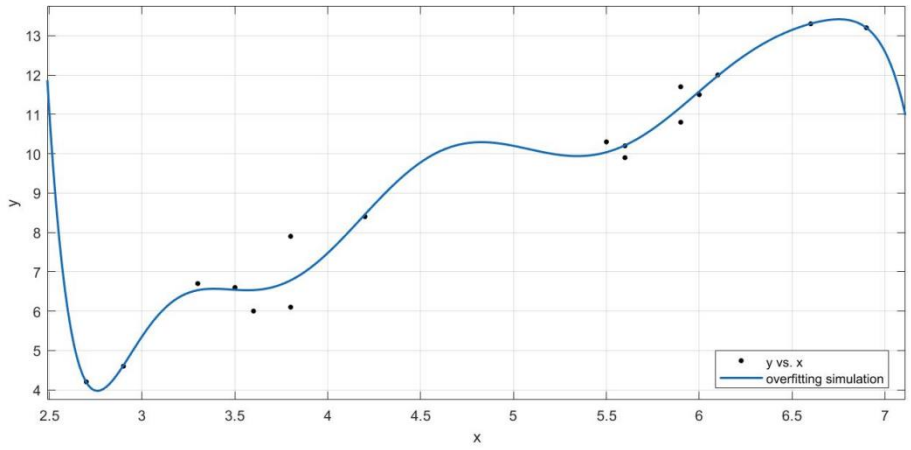

Figure 3 Overfitting of the normal fitting function images

From the comparison of normal fit and overfit function images, we can find that overfitting prediction is very good for sample data, and the prediction effect of out-sample data alone may be greatly weakened. At the same time, when the sample is affected by the error data or noise, it will have an obvious impact on the prediction results, and has too high noise sensitivity.

2.5. Model optimization of the Fisher linear discriminant analysis

The essence of the Fisher discrimination method is a projection method that projects points of high-dimensional spaces into low-dimensional spaces and makes all points separated on both sides of the hyperplane ω T x = 0. The points projected on the normal vector of ω satisfy the difference between the populations (marked in B) is as large as possible, and the difference within the same population (marked in E) is as small as possible. Make Fisher linear discrimination of existing imported data in Spass, and classify cultural relics of unknown categories:

The resulting Bayesian discriminant function coefficient table and the parameters of the sample are brought into two Bayesian discriminant functions to compare the function values Value1 and Value2, where Value2 indicates the probability that the sample category belongs to 0 or high potassium glass, and Value1 indicates the probability that the sample category belongs to the lead-barium glass. Type the samples into a category with large function values, such as Table 2.

Table 2. Sample input into categories with large function values

Number of cultural relics	Surface weathering	Value 1	Value 2	category
A1	No weathering	0.00887	0.99113	High potassium glass
A2	weathering	0.99973	0.00027	lead barium glass
A3	No weathering	0.99999	0.00001	lead barium glass
A4	No weathering	1.00000	0.00000	lead barium glass
A5	weathering	0.99994	0.00006	lead barium glass
A6	weathering	0.00001	0.99999	High potassium glass
A7	weathering	0.00029	0.99971	High potassium glass
A8	No weathering	0.06533	0.93467	High potassium glass

Comparing the prediction results obtained by Fisher discrimination and logistic regression, we find that the deviation probability from 0 and 1 is greater than the logistic regression, which can slow down the influence of overfitting phenomenon on the results to some extent.

3. Conclusion

Using progressive way, first get the sample principal component analysis, and then using hierarchical clustering method found sample deep hierarchical relationship, then according to the principle of elbow to sample data classification, and then use scatter plot and gradually regression method to test the rationality and sensitivity of classification results, make it more scientific and reliable. A logistic regression model was used to classify the data and verify their accuracy. Stepwise regression was then used to test the sensitivity of the classification, and finally Fisher linear discrimination was used to optimize the model. For the multiple linear regression model, the prediction results may differ greatly from the actual results because of ignoring the interaction effects and the non-linear causality. Although Spearman's correlation coefficient model has a wide range of applications, the accuracy decreases compared with Pearson's correlation coefficient model, which does not well represent the correlation between the two variables. Through the composition analysis, identification and prediction of glass products, the important influencing factors of ancient glass weathering can be found, which can be used for more targeted protection and restoration of unearthed cultural relics. The obtained prediction model can be used as a basis for the investigation of ancient glass age.

References

[1] Liu Chengxin. The Application of Logic Regression Model in Risk User Detection in Banking Financial Institutions [J]. Fintech Era, 2022 (09): 71-73.

[2] Bao Fu, Ma Wen, Gao Yudou, Yang Tianjun, Li. Blackout sensitivity prediction of power users based on a logistic regression model [J]. Microcomputer applications, 2022,38 (07): 67-68 + 72.

[3] Tian Dongxia, Cao Jiucai. Apple yield prediction based on the stepwise regression method and the BP neural network model [J]. Modern agricultural Science and Technology, 2022 (14): 131-133 + 142.

[4] Yuan Ke, Huang Yabin, Du Zhanfei, Li Jiabao, Jia Chunfu. Grouping password algorithm identification scheme based on mixed gradient boosting decision tree and logistic regression model [J]. Engineering Science and Technology, 2022,54(04):218-227.DOI:10.15961/j.jsuese.202100341.

[5] Pan Xiaojun, Zhang Youchun. Research on the logic regression model of "Computer Network" course based on xMOOC + SPOC [J]. Journal of Tonghua Normal University, 2022,43(02):141-144.DOI:10.13877/j.cnki.cn22-1284.2022.02.022.

[6] Zhang Yuyu. Calibration and statistical analysis of air quality data based on a stepwise regression model [J]. Journal of Heilongjiang Ecological Engineering Vocational College, 2021,34 (05): 9-11 + 30.

[7] Xu Fei, Huang Hua, Wang Zhuo, Luo Qing. Study on Conformation DeType Based on Fisher [C] / /. Summary set of the 9th Academic Conference of Geophysical Technical Committee of Chinese

Geophysical Society- -Global Geophysical Exploration and Intelligent perception Seminar. ,2021:9-10.DOI:10.26914/c.cnkihy.2021. 005909.

[8] Shan Qiufu, Zhang Tao, Li Chao, Luo Lin, Chen Fangrui, Zhang Haitao. Construction of the mainstream cigarette flue gas quality prediction model based on the Fisher linear discriminant analysis method [J]. Food and Machinery, 2021,37(02):78-84+92.DOI:10.13652/j.issn.1003-5788.2021.02.014.

[9] Liang Lu Fang. Solving Algorithm for Linear Discriminant Analysis Problem of Fisher [D]. Yunnan Normal University, 2020.DOI:10.27459/d.cnki.gynfc. 2020.000254.

[10] Gosterislioglu, Y. A. , et al. "Investigation the effect of weathering on chemically strengthened flat glasses." Journal of Non-Crystalline Solids: A Journal Devoted to Oxide, Halide, Chalcogenide and Metallic Glasses, Amorphous Semiconductors, Non-Crystalline Films, Glass-Ceramics and Glassy Composites 544-(2020):544.