A study of the composition of ancient glassware based on focal logistic regression models

Hongsheng Wang *, Guihuang Lin, Zanjia Tong
School of Automation, Guangdong University of Technology, Guangzhou, China
* Corresponding Author Email: 3462228700@qq.com

Abstract. Glass has been developed in China for over two thousand years and was an important commodity in the economic and trade exchanges between China and abroad along the Silk Road. Since the Han Dynasty, glassware from the West began to be imported into China via the Silk Road, and Chinese glassmaking techniques have taken on foreign technologies, thus making ancient glass made in China similar in appearance to foreign glass, but the main chemical composition of the two is different due to the different regions and the different materials used. This paper focuses on the types of ancient glass artefacts that make up ancient glass, the patterns of content of each major chemical component and the specific mathematical relationships within them. This paper establishes a correlation analysis model based on information entropy using the information gain and information gain in the decision tree algorithm to correlate the weathering of glass artefacts with their color, decoration and type, and ultimately concludes that only the decoration of high potassium glass has a high correlation with its weathering. Then a nearest neighbor analysis model was established based on the idea of KNN nearest neighbor algorithm, and a joint focal logistic regression model was used to complement the weathering of the sampling points.

Keywords: Correlation Analysis, Nearest Neighbor Analysis Model, Focal Logistic Regression.

1. Introduction

Glass can be differentiated by analyzing its different chemical compositions, for example, lead-barium glass and high-potassium glass, and by analyzing the chemical composition, the origin of the glass can also be distinguished, for example, lead-barium glass is generally considered to have been made in China, while potassium glass is usually found in regions such as Lingnan and South East Asia in China [1]. However, the chemical composition of ancient glass may have changed to varying degrees due to weathering over time, which is a major obstacle to the identification and analysis of ancient glass. There are data on a number of ancient glass objects from China, which archaeologists have classified into two types: high potassium glass and lead-barium glass [2]. This study analyses the relationship between the degree of weathering of these glasses and their corresponding glass type, decoration and color; it examines the statistical patterns of the chemical content of the artefact samples with and without weathering on their surfaces, and predicts the chemical content of the weathered sampling sites prior to weathering [3-6].

2. The basic fundamental of information entropy in correlation analysis models

2.1. The structure of BP neural network

In this paper, we first find out the entropy of each information needed for the data analysis process, S is the total data set, i.e., S= {A, B, C, D}, where A, B, C and D represent the four data sets of decoration, color, glass type and weathering respectively, and calculate the correlation between the three and weathering respectively. For example, take the ornamentation feature as feature A. The probability that the sample belongs to the first ornament type(i=1,2,3) The information entropy of the ornamentation feature for the data set can be found by the following equation [7].

$$E(\mathrm{A}) = -\sum_{\mathrm{i}=1}^{3} \mathrm{p}(\mathrm{i}) \log[\mathrm{p}(\mathrm{i})] \tag{1}$$

The weathering or otherwise of ancient glass is used as feature B, and in this way it can be divided into two categories. B_j For the first j category of the sample collection, the $|B_j|$ is the j is the total number of samples in category $|S|$ is the total number of all samples, and $q_j(i)$ denotes the total number of samples in category j is the probability of belonging to thei the probability of belonging to the first ornamental feature in the $i=1,2,3;\ j=1,2$. The information entropy for the first category is given by the following equation. The information entropy obtained by using feature A to classify the set of samples in the first category[8].

$$E(\mathrm{A\_D}_j) = -\sum_{i=1}^{3} q_j(i)\log\left[q_j(i)\right] \tag{2}$$

with $l(\mathrm{k})$ denotes the probability that the data set belongs to thek The probability of being in category k and from this the information entropy of weathering or not for the dataset is derived as

$$E(\mathrm{D}) = -\sum_{\mathrm{k}=1}^{2} l(\mathrm{k})\log[l(\mathrm{k})] \tag{3}$$

In turn, the information gain is found by the following equation.

$$G(A\_D) = E(\mathrm{A}) - \sum_{j=1}^{2}\frac{|D_j|}{|S|}E(\mathrm{A\_D}_j) \tag{4}$$

where $E(\mathrm{A\_D}_j)$ denotes the set of samples belonging to the j th category of D, at which point A is used to classify them, and finally the resulting information entropy $|\mathrm{D}_j|$ denotes the set of the j th category.

In summary, the degree of correlation between the discrete features A (ornamentation) and D (weathering or not) can then be found according to the following formula

$$C = \frac{G(A\_D)}{\sqrt{E(\mathrm{D})\times E(\mathrm{A})}} \tag{5}$$

where C is the degree of association.

Similarly, the correlation between discrete feature glass type, color and weathering can be calculated using the above method [9].

In the absence of a classification of the glass, it is possible that the effects of the characteristics of the two types of glass may cancel each other out and therefore lead to smaller correlation results. Therefore, if the final calculated correlation results are small, the model can be further improved on the basis of the above correlation analysis model to calculate the correlation between weathering and decoration and color for high potassium and lead barium glasses respectively. In this case, the full set in the above model $S_1=\{A_1,B_1,D_1\}$ .The full set in the above model is then calculated. $S_2=\{A_2,B_2,D_2\}$ Again, using the above correlation model, the two sets of correlations are calculated separately and the results are analysed again.

The normal distribution is a common continuous probability distribution that is often used in the natural and social sciences to represent an unspecified random variable. The normal distribution is a convenient model for quantifying phenomena in the natural and behavioral sciences, so this paper assumes that the chemical composition of each weathering site before weathering obeys a normal distribution. The probability density function of the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

Firstly, with the premise of completing the weathering of the sampling points, the artefact sampling points were characteristically clustered according to their decoration, type and color, and the groups of artefacts containing weathered sampling points were taken for study. This resulted in five groups of artefacts, which were taken from the data of unweathered glass artefacts in each of the five groups, and the individual chemical composition of the five groups of unweathered glass artefacts was studied separately. The proportion of each chemical composition in each group is used to define the corresponding normal distribution curve. For each proportion of chemical composition in each group, the mean $\bar{x}$ and standard deviation σ .The mean and standard deviation of each chemical composition in each group were determined and the interval of the proportion of unweathered chemical composition at the sampling site in that group was determined as $\left[\bar{x}-\sigma, \bar{x}+\sigma\right]$ [10].

3. Results

3.1. The establishment of simulation model

This paper analyses the pattern of chemical content on the surface of cultural relics, completes the information on sampling points, clusters the relics according to their decoration, type and colour, and takes out groups containing 'weathered' sampling points for study, here there are five groups after screening, and the statistical pattern of chemical composition before and after weathering of sampling points is studied.

3.2. Analysis of experimental results

The correlations between decoration, glass type and color and weathering are all small, but it is clear from the previous analysis of the data that for different types of glass, decoration is strongly related to weathering or not, but the correlations are reduced by the offsetting effect of the same features in different glass types. The logistic regression model is the classic dichotomous model and can be used to predict probabilities due to the way it is calculated. The main weights used in making predictions are the weight vector $\vec{w}=\left[w_1, w_2, \cdots, w_n\right]$.For a sample feature $\vec{x}=\left[x_1, x_2, \cdots, x_n\right]$ it first calculates the score for that sample $z_i=\vec{w}\vec{x}_i^T+b$ and then uses a sigmoid activation function to transform the z_i. The value domain is transformed to [0, 1] and this value is used as the probability to bifurcate the sample with a cut-off of P = 0.5.

$$P_i=\frac{1}{1+e^{-z_i}}=\frac{1}{1+e^{-\vec{w}\vec{x}_i^T-b}} \tag{7}$$

and $\vec{w}, b$ are solved by the gradient descent method. The loss function used in the solution is cross-entropy, and the binary Boolean value is noted as $y_i \in \{0,1\}$, there are.

$$\begin{aligned} L &= -\frac{1}{n}\sum_{i=1}^{n} y_i \ln P_i + \left(1-y_i\right)\ln\left(1-P_i\right) \\ \frac{dL}{dP_i} &= \frac{1-y_i}{1-P_i}-\frac{y_i}{P_i} \end{aligned} \tag{8}$$

By the chain rule of derivation, it follows that $\frac{\delta L}{\delta \vec{w}}$ with $\frac{\delta L}{\delta b}$. Denote the learning rate as h , the logistic regression model is solved by $\vec{w}=\vec{w}-h\frac{\delta L}{\delta \vec{w}}, b=b-h\frac{\delta L}{\delta b}$ to solve for the optimal weight vector and bias.

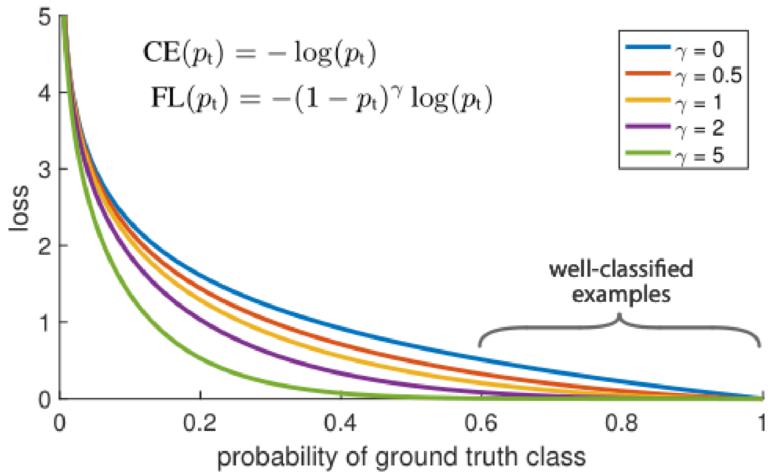

Figure 1. Focal cross-entropy Focal Loss

From the image of the cross-entropy function in Figure 1 (corresponding to $\gamma = 0$) it can be seen that positive samples with probabilities above 0.8 still have large loss values, which will prevent the model from learning from hard-to-classify samples. Therefore, using the state-of-the-art focal cross-entropy Focal Loss (when $\gamma = 0.8$ the loss value decays rapidly to near 0 when the probability of positive samples is above 0.8) as a loss function to make the model more focused on hard-to-classify to improve the performance of the logistic regression model.

$$L = -\frac{1}{n}\sum_{i=1}^{n} y_i\left(1-P_i\right)^{1.5}\ln P_i + \left(1-y_i\right)P_i^{1.5}\ln\left(1-P_i\right) \tag{9}$$

$$\frac{dL}{dP_i} = \begin{cases} 1.5\sqrt{1-P_i}\ln P_i - \dfrac{\left(1-P_i\right)^{1.5}}{P_i}, y_i = 1 \\ -1.5\sqrt{P_i}\ln\left(1-P_i\right) - \dfrac{P_i^{1.5}}{1-P_i}, y_i = 0 \end{cases} \tag{10}$$

By the chain rule of derivation, the same can be derived $\frac{\partial \mathcal{L}}{\partial \vec{w}}$ with $\frac{\partial \mathcal{L}}{\partial b}$ and then the weight vector and bias of the model are solved using the gradient descent method. Using the sampling points of the lead and barium artefacts with known weathering conditions as the training set, the weight vectors and biases obtained by solving the focal logistic model can be used to make a judgement on the weathering conditions of the lead and barium artefacts. The box plots of the pre-weathering coefficients of variation for each group, as derived from Python, are shown in Fig 2.

As can be seen from Figure 2, the coefficients of variation for the five groups are generally small, thus indicating that the five groups selected for this paper are relatively stable before weathering and that the grouping in this paper is somewhat reasonable. In addition, the fluctuations in the data for these five groups before weathering are small, so these five groups can be used as the basis for the chemical composition content before weathering.

The chemical composition of the weathered artefacts, before weathering, was predicted based on the assumption of normal distribution. The means of the proportions of each chemical composition of the unweathered artefacts in each of the five groups were first found $\bar{x}$ and their standard deviations σ The mean values and standard deviations of the proportions of each chemical composition in each of the five groups were obtained.

Using one of the groups of silica as an example, the calculation shows that the mean and standard deviation is finally given by the equation $\bar{x}+\sigma$. It was determined that the proportion of silica composition of the same group of artefacts before weathering was in the $\left[\bar{x}-\sigma, \bar{x}+\sigma\right]$ between. Similarly, the above solution method can be used to find the range of values for the proportion of other chemical components in the group and the range of values for the proportion of each chemical component in the other groups

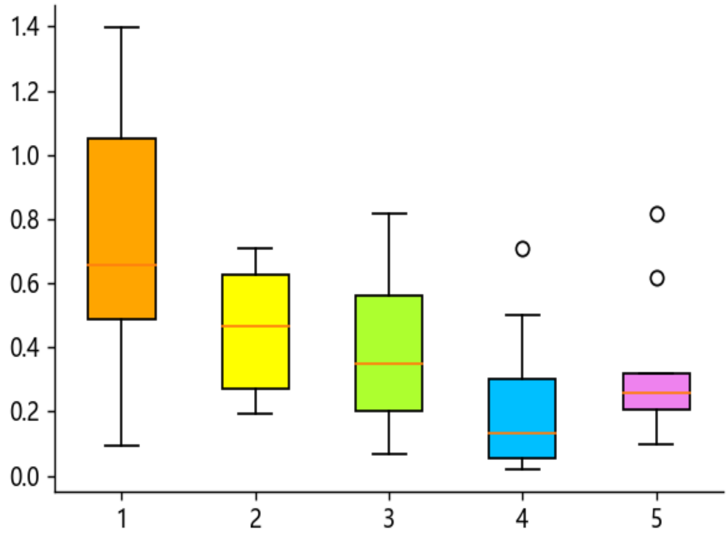

Figure 2. Coefficient of variation graph

4. Conclusions

In this paper, the information entropy correlation model based on the decision tree algorithm is examined and the results of the above discrete correlation model are combined to show that the weathering of high potassium glass is highly correlated with its grain and color, while the weathering of lead-barium glass is very poorly correlated with its grain and color.This paper uses a logistic model to dichotomies glass artefacts using 'weathered or not' as a boundary. The results indicate that the accuracy of the model is close to 100% for high potassium glass. In addition, the model has implications for the study of glass classification.

References

[1] Zhang Jian, Hao Nannan. Gemological classification and identification of meteorite glass[J]. Shandong Chemical Industry,2021,50(17):160-161.

[2] Liu Z. Xiang, Deng P. F., Pan S. Hao, Luo W. D., Zhang S. L., Li T. Ming. Classification and research status of fireproof glass[J]. Guangzhou Chemical Industry,2021,49(15):16-18.

[3] Yang X. Detection and classification of glass defects based on machine vision[D]. Fujian College of Engineering,2021.

[4] Huang JY. Digital image processing-based defect detection and classification of mobile phone panel glass [D]. Xi'an University of Electronic Science and Technology, 2020.

[5] Lv Shaoguo, Sun Cong,Wang Chao. Classification of fireproof glass and its application in architecture[J]. Science and Technology Information,2009, (25):38.

[6] Li Xianfeng. Classification of point-supported glass curtain wall support systems and analysis of component performance[J]. Journal of Building Structures, 2007, (S1):274-279.

[7] Jiang Lilin. Classification and research status of double-glazed curtain walls in China[J]. Guangdong building materials, 2007, (07):8-9.

[8] Han Helei, Li Jianzhong, Feng Gang. Online classification and recognition of glass products based on computer vision[J]. Journal of Instrumentation,2006, (S1):777-778.

[9] Liu Liping. Experimental culture of glass sea squirt (Ciona intestinalis), classification of blood cells and its immune response study[D]. Graduate School of Chinese Academy of Sciences (Institute of Oceanography), 2005.

[10] Ai Jieyan, Zhu Xuefeng. Improved algorithm-based ART2 network for colour classification of microcrystalline glass[J]. Journal of South China University of Technology (Natural Science Edition),2003, (01):74-78.