Component Analysis of Ancient Glass Based on Neural Network

Rong Zeng [*, #], Yuheng Zhang [#]

School of electronics and communication engineering, Sun Yat-Sen university, Shenzhen, China, 528406

* Corresponding Author Email: zengr29@mail2.sysu.edu.cn

[#]These authors contributed equally

Abstract. The weathering of glass relics is easily affected by the environment. In order to protect cultural relics, we have reduced the number of samples per cultural relic and the number of cultural relics sampled, resulting in limited and incomplete data. These incomplete data are used to predict whether the cultural relics have weathered and reveal the weathering law. It is of practical significance to study the weathering process of glass by analyzing the chemical components of glass. In this paper, the coefficient of variation method is used to extract the chemical components with high contribution rate to weathering and the missing values of each group of data are filled by hot card filling to obtain a set of sample data. Based on the known sample data, we fit the density function of random variables by kernel density and expand the sample size by random number function to design the input data of neural network. The expanded sample value serves as the training set of the neural network and the known sample set serves as the validation set of the neural network. Finally neural network is used for determination of weathering type of sampling points. The neural network designed in this paper realizes the high-precision estimation of weathering types of sampling points.

Keywords: Neural Network, Coefficient of Variation, Hot Deck Imputation, Sample Expansion.

1. Introduction

Glass witnessed the cultural exchanges between China and the West along the Silk Road and is a valuable material evidence of early trade. At the same time, the birth and development of the Silk Road promoted the maturity of glass products technology, so the research on glass products has important practical and historical significance [1]. In the long history, the ancient glass is susceptible to the influence of the environment and weathering. During the weathering process, the chemical composition of the glass is exchanged with the external elements in large quantities, resulting in great changes in its chemical composition and appearance. It is difficult to judge the weathering type of a certain region.

Nowadays, scholars have different methods to study this problem, such as using scanning electron microscopy to observe the surface topography, physical property testing and analysis [2-5]. We design a neural network based on few samples and multiple variables for the analysis of weathering conditions of glass relics sampling points. After data preprocessing, the neural network used in this paper can predict the weathering of glass sample points with high resolution.

The current research is mainly aimed at the subjective analysis and judgment of physical properties and our neural network is designed to objectively analyze and judge the weathering type of glass sampling points based on the change of main chemical composition and sample enlargement algorithm. More importantly, the method proposed in this paper is not easily affected by the change of the research object and is suitable for all problems of weathering type analysis.

2. The fundamental of neural network

2.1. Coefficient of variation

According to the change of chemical composition of glass before and after weathering, it can be divided into four types of data, increasing type, decreasing type, stable type and strange type. Our network needs to select the suitable number of chemical components from multiple chemical components as parameters, thus speeding up the training of the network. Coefficient of variation method [6-9] is used to calculate the change degree of each index in the system according to the statistical method and assign an objective value to the index by calculating the relative weight. The specific steps are as follows.

(1) The original data is presented in the form of a matrix, the number of rows in the matrix represents the number of sampling points, and the number of columns represents the number of chemical composition indicators. Matrix X collects the original data, n represents the number of sampling points and m represents the number of chemical composition indicators.

(2) Turn the decreasing indicator forward. K stands for forward coefficient.

(3) Standardize data and eliminate the influence of units.

(4) Calculate the coefficient of variation. V represents the coefficient of variation for each indicator.

(5) Calculate the weight. ω_j Represents the weight of each composition indicator.

(6) Calculate the score for each indicator. $Score_i$ represents the score for each indicator.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \tag{1}$$

$$x_{ij} = \frac{K}{x_{ij}} \tag{2}$$

$$x_{ij=}\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}} \tag{3}$$

$$V_j = \frac{\frac{1}{n}\sum_{i=1}^{n} r_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(r_{ij}-\frac{1}{n}\sum_{i=1}^{n} r_{ij}\right)^2}} \tag{4}$$

$$\omega_j = \frac{V_j}{\sum_{j=1}^{n} V_j} \tag{5}$$

$$Score_i = \sum_{j=1}^{n} \omega_j r_{ij} \tag{6}$$

2.2. Hot deck imputation

The sample data is divided into two types of data: the first is that each indicator has a definite value, and the second is that there is a missing value. Due to the poor training effect caused by missing values in the process of network training, we use hot deck imputation [10] to improve the sample data.

The Euclidean distance between the sample with missing value and the sample of each complete index is calculated, and the parameter of the sample corresponding to the smallest Euclidean distance is filled as the missing value.

2.3. Sample expansion

Because the training set of the neural network requires a large sample size, but in order to protect the glass relics from further weathering, we can only get a limited number of sampling points. Through limited sampling points, to improve the training effect of neural network appropriate, algorithms are adopted to expand the sample size.

We divide the data into two groups according to the weathering type. The fitlist function of MATLAB software is used to carry out nuclear fitting of finite sample points for each chemical component of each group of data, to construct the probability density function of each chemical

component and generate large groups of simulation samples through random number function. Each chemical component generates the same number of simulation samples, which are combined to form a set of chemical components of sampling points of the same type as the set group.

2.4. Model assumptions

(1) There is no weathering sampling point for non-weathered glass.

(2) When the expanded sample function is implemented, the random sample generated in the weathered group is weathered, and the random sample generated in the non-weathered group is non-weathered.

(3) When the coefficient of variation method is used, due to the need to highlight the weathering of sampling points through the change of chemical composition, the contribution rate of stable type and strange type of data to weathering is low, and no data processing is needed.

(4) The weathering type corresponding to the average value of the sampling points of weathered glass is weathering.

2.5. The determination of the number of network layers and the structure of neural network

Our neural network architecture follows the Multilayer Perceptron (MLP) model [11], which is renowned for its efficiency and conciseness. The adoption of the MLP (Multilayer Perceptron) model is justified by several key factors.

Firstly, the MLP exhibits remarkable capabilities in learning and representing nonlinear and intricate patterns, enabling it to approximate arbitrary continuous functions effectively. Consequently, it demonstrates exceptional performance in addressing complex problem domains.

Secondly, the MLP model possesses the capacity for feature extraction and representation learning. This property enables each hidden layer to extract increasingly sophisticated and abstract features by leveraging the outputs of preceding layers. Consequently, the MLP framework provides robust support for processing high-dimensional data and tackling intricate tasks.

Lastly, data preprocessing plays a pivotal role in rendering the MLP model feasible for glass weathering prediction. As glass consists of diverse chemical components, an initial impurity screening process is conducted to identify the principal constituents. Subsequently, the coefficient of variation method is employed to ascertain the main chemical components significantly influencing weathering. This meticulous selection and screening procedure enhances the MLP model's generalization capability when working with limited data samples, while facilitating efficient network training.

2.6. The structure of neural network:

The neural network architecture comprises three fundamental components: the input layer, the hidden layer, and the output layer. The input layer serves as the initial entry point for the data. The hidden layer, which consists of four linear layers and incorporates three activation functions, plays a crucial role in feature extraction and transformation. The activation function employed in this model is the Rectified Linear Unit (ReLU) function, introduced by (6):

$$\mathrm{ReLU}(x) = \max(x, 0) \tag{7}$$

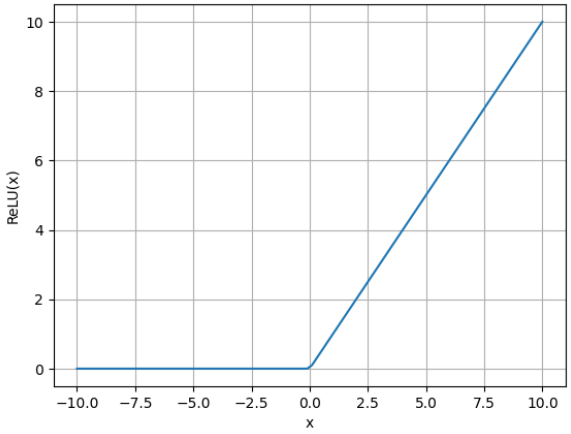

Figure 1. ReLU function

The ReLU function maps the input value to either zero or the original input, depending on whether it is negative or non-negative. This non-linear activation function enhances the network's capability to capture complex patterns and learn non-linear relationships within the data [12]. Figure 1 shows that the relationship between independent and dependent variables of the ReLU function.

The Cross-Entropy Loss function, denoted as Loss, was utilized in our study to train a Multilayer Perceptron (MLP) model for the classification task [12]. Figure 2 shows that the relationship between independent and dependent variables of the cross-entropy function. Given the challenge of working with limited sample sizes, we employed simulation techniques to generate synthetic data that adequately represented the underlying distribution. The Cross-Entropy Loss is a widely adopted loss function in machine learning for its effectiveness in handling classification problems. It measures the dissimilarity between the predicted class probabilities $\hat{y}$ and the true labels y, capturing the information loss during the classification process. Mathematically, the Cross-Entropy Loss is defined as (7):

$$\text{Loss} = -\sum y * \log(\hat{y}) \tag{8}$$

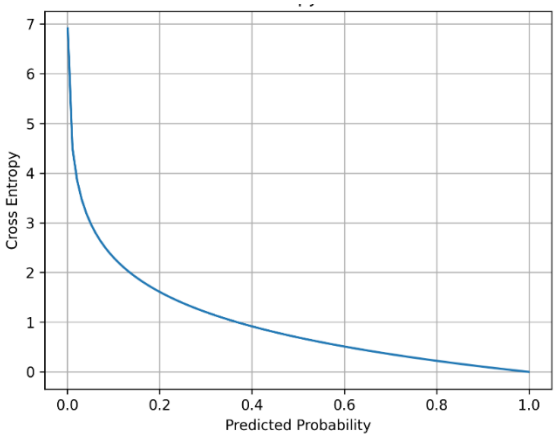

Figure 2. Cross entropy function

Where *y* represents the true label distribution, and $\hat{y}$ denotes the predicted probability distribution produced by the MLP model. The logarithmic operation accentuates the impact of incorrect predictions, enabling the model to learn from the discrepancies. During the training phase, the MLP model optimized its parameters by minimizing the Cross-Entropy Loss through backpropagation and stochastic gradient descent. This approach allowed the model to adapt to the specific characteristics of the synthetic dataset while generalizing well to unseen samples. In our experiments we obtain a final Cross-Entropy Loss value with a final loss value of 1.65e-03 for high potassium glass and 6.21e-02 for lead barium glass. It shows that our training process is successful in reducing information loss and improving the classification accuracy of the model. The choice to employ a cross-entropy loss function is consistent with previous research in the field of deep learning [13]. Its applicability to

small sample analysis and classification tasks has been well demonstrated, making it a suitable choice for our study.

3. Module construction and problem analysis

3.1. The processing of data

First, we obtain the chemical composition list of random sampling points of some cultural relics (This data comes from question C of the CUMCM 2022). There are two types of glass cultural relics, the first is high-potassium glass, and the second is lead-barium glass. The data collected at the sampling site includes the following chemical components: silicon dioxide, sodium oxide, potassium oxide, magnesium oxide, alumina, iron oxide, copper oxide, lead oxide, barium oxide, phosphorus pentoxide, strontium oxide and sulfur dioxide. Each chemical composition is marked with the proportion of the total composition at the respective sampling point.

The weathering type of the cultural relics in the table has been confirmed. Due to the assumptions and the analysis of the actual situation, it can be known that there may be non-weathering sampling points and weathering sampling points for weathered glass, and the sampling points for non-weathering glass are non-weathering, but there is an outlier. The existence of this outlier will not cause a large error to the training of neural network and will improve the anti-interference ability of neural network. If the weathered glass indicates weathering and non-weathering sampling points, the type of sampling point can also be confirmed.

After completing the confirmation of each data type, we verify the rationality of the data and add the chemical composition ratio of each data set. If the total proportion is between 85% and 115%, the data of the sampling point is reasonable.

We need to reduce the number of parameters and select parameters with obvious changing trends, to improve the efficiency of neural network operation. First, we directly remove the chemical components that most samples did not have, and the content was low.

The remaining chemical components are analyzed by coefficient of variation method, and the chemical components could be classified into increasing type, decreasing type, stable type and strange type. Since the stable and strange data do not have an obvious trend of change, these parameters are not conducive to the training of neural networks, and it is not necessary to carry out the forward conversion in the coefficient of variation method. After the decreasing type of data in the chemical composition is turned forward, the results obtained by the coefficient of variation method are shown in the figure below. Figure 3 shows the scores of each component of the high-potassium glass and Figure 4 shows the scores of each chemical component of the lead-barium glass. In the weathering process, the chemical components of high-potassium glass silica, potassium oxide, calcium oxide and alumina play an important role, and the chemical components of lead-barium glass silica, alumina, barium oxide and lead oxide play an important role.

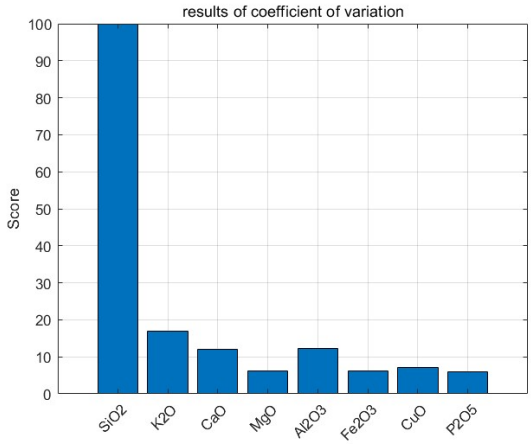

Figure 3. the scores of each component of the high potassium glas

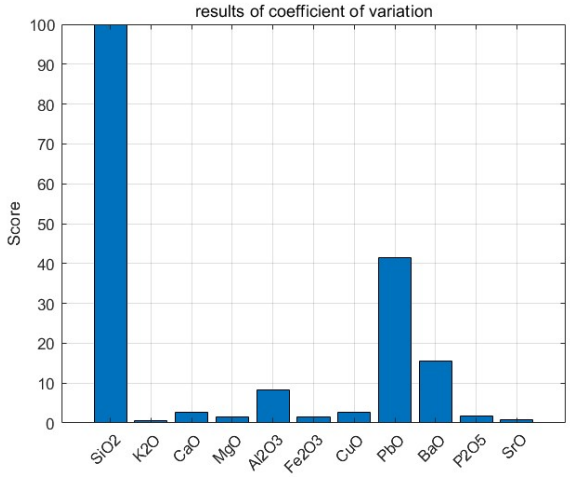

Figure 4. the scores of each chemical component of the lead-barium glass

After the selection of parameters is completed, some sample points have undetected chemical components, which may have a huge impact on the training of the network. Therefore, the method of hot deck imputation is used to complete the filling of missing values.

After finishing the sorting and pre-processing of the existing data, we begin to design the input data of the neural network, because the input data of the neural network needs to confirm the weathering type of the sampling point, but the weathering type of the sampling point of the weathered glass in the high-potassium glass and the lead-barium glass is unknown. Therefore, the average value of each chemical component of non-weathered glass is taken as the standard data of non-weathered sampling points, and the average value of each chemical component of weathered glass is taken as the standard data of weathering sampling points to verify the rationality of this analysis. Since it is known that the sampling points of non-weathered glass are all non-weathered, the standard data obtained by taking the average value is also non-weathered in probability. In the same way, even if there are non-weathered sampling points in weathered glass sampling points, taking their average value within the error range allows the interference of non-weathered sampling points and is conducive to improving the training accuracy of the neural network. By comparing the Euler distance of each set of weathered glass sampling points and the standard data, the qualitative judgment of weathering type is completed. The judgment here is only used for the training of the neural network and cannot be verified as the true value.

After the preparation of known sample points, we expand the number of sample points to improve the training accuracy of the neural network. The high-potassium glass and lead-barium glass are divided into weathering sampling points and non-weathering sampling points. For each chemical component in one of the two categories, the built-in function of Matlab is used to achieve the fitting of the probability density function of each chemical component, and the random number function is used to generate 1000 sample points between the minimum and maximum values. Then, the high-potassium glass weathering simulation data, high-potassium glass non-weathering simulation data, lead-barium glass weathering simulation data and lead-barium glass non-weathering simulation data are formed randomly. The histogram of known sample chemical compositions and fitted generated probability density function are shown in the figure below. Figure 5, figure 6 and figure 7 respectively show the fitting of three types of chemical components of high-potassium glass. Due to the limited number of samples and the error of interval setting, there is some error in fitting the probability density function.

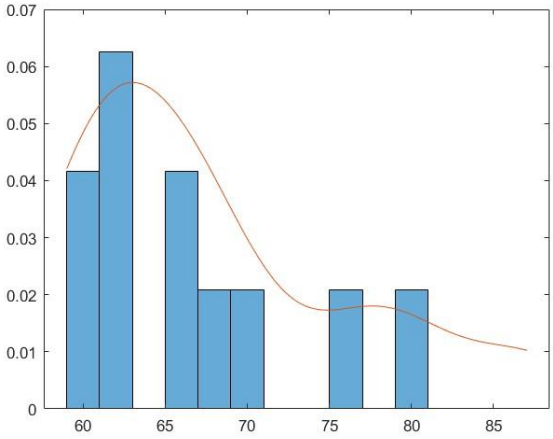

Figure 5. Fitting probability density function of chemical composition

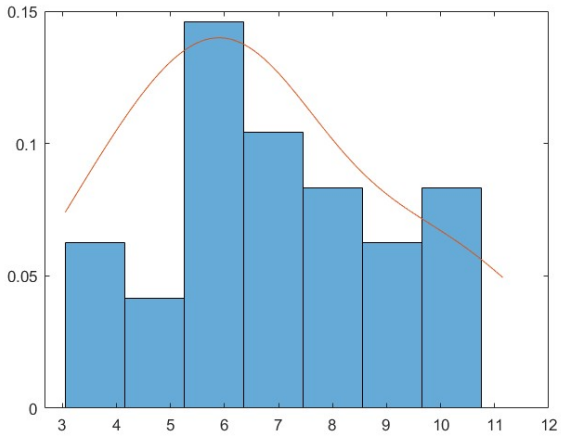

Figure 6. Fitting probability density function of chemical composition

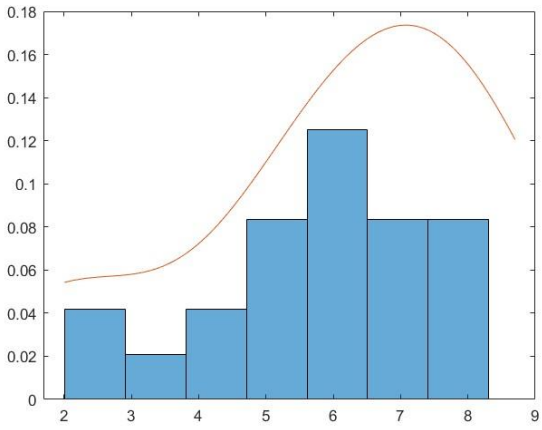

Figure 7. Fitting probability density function of chemical composition

3.2. Analysis of experimental results

Test Set: Our test set consists of a total of 38 real samples, including 12 samples of high-potassium glass and 26 samples of lead barium glass. These two sets of data are used as separate test sets for the high-potassium glass network and the lead-barium glass network, respectively, to evaluate the performance of the networks in terms of accuracy and generalization.

Training Set: We generated simulated samples for the training set using probability density functions. The training set comprises 1000 weathered samples and 1000 non-weathered samples, resulting in a total of 2000 samples available for training in each category: high-potassium glass and lead barium glass.

Figure 8 shows that the training results of the high-potassium glass network include the values of the training loss, training set accuracy, and test set accuracy. Figure 9 shows that the training results of the lead-barium glass network include the values of the training loss, training set accuracy, and test set accuracy.

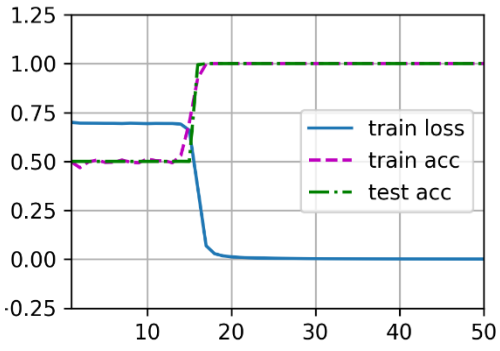

Figure 8. High-potassium glass

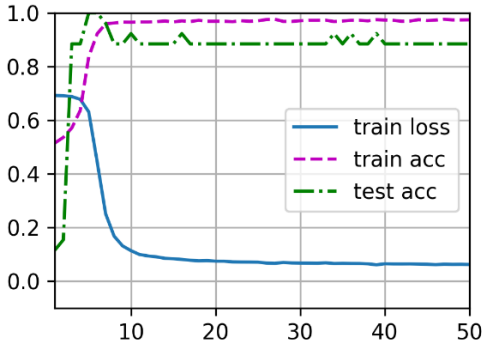

Figure 9. Lead-barium glass

Table 1. Neural network prediction results

	High Postassium Glass	Lead Barium Glass
Final Training Loss	1.65e-03	6.29e-02
Training Accuracy	100%	97.3%
Test Accuracy	100%	88.47%

Table 1 shows the prediction results of neural network. According to the table of the results obtained, it can be clearly observed that both the high potassium glass and the lead barium glass show very low loss values and have significantly high training and test accuracy. These findings strongly support the conclusion that, despite the limited size of the dataset, neural networks are successfully modeled by applying precise and efficient data preprocessing techniques such as reducing invalid variables, highlighting key variables, and generating simulation data using fitted probability distribution functions. Thus, the obtained results indicate the achievement of a significant level of accuracy.

4. Conclusions

The chemical composition analysis method provided in this paper can effectively solve the problem that the sample is small and the chemical composition is relatively complex, and through selecting a reasonable algorithm to expand samples and design a training set that meets the requirements of the neural network. Finally, the weathering type analysis of the sampling point can be completed with high efficiency and high accuracy through the neural network.

If the method designed in this paper is applied to industrial production, we need to further consider the correlation of each chemical component and adjust the structure of the neural network reasonably.

References

[1] Zhou X, Lv H, Cui J,et al.Fluorite used in ancient Chinese glassmaking during the 10th to 12th centuries: Evidence from glass products excavated in the capital city site of the Liao dynasty [J].Archaeometry, 2022.

[2] Bao C, Jia S, Liu A, et al.Weathering corrosion of underwater stone relics [J].Corrosion and Protection, 2018, 39(3):218-221.

[3] Heimann R B. Weathering of ancient and medieval glasses—potential proxy for nuclear fuel waste glasses. A perennial challenge revisited [J]. International Journal of Applied Glass Science, 2018, 9(1).

[4] Yuan G X, Bai Y L, Yan C H, et al.Surface Geotechnical Weathering and denudation in the Northeast Margin of Tarim [C]//2019:032014.

[5] Juhasz, E., Haubner, R., Toth, M., & Szabo, G. (2014). Assessment of weathering degree of architectural glasses using hardness measurement. International Journal of Architectural Heritage, 8(2), 233-240.

[6] Wu K, Peng B, Xie H, et al.A Coefficient of Variation Method to Measure the Extents of Decentralization for Bitcoin and Ethereum Networks[J].International Journal of Network Security, 2020, 22(2):191-200.

[7] Eldridge S M, Deborah A, Sally K. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method[J]. International Journal of Epidemiology, 2006(5):1292.

[8] Xu H, Zhang Y, Liu J, et al.Feature Selection Using Maximum Feature Tree Embedded with Mutual Information and Coefficient of Variation for Bird Sound Classification[J].Mathematical Problems in Engineering, 2021, 2021.

[9] Qiu B, Luo D. A Grey Multi-Level Evaluation of Industrial Park Ecology Based on a Coefficient of Variation-Attribute Hierarchy Model [J]. Sustainability, 2021, 13.

[10] Rebecca A, Laura B, Jenny T K. Finding a Flexible Hot-Deck Imputation Method for Multinomial Data [J]. Journal of Survey Statistics and Methodology, 2020(4):4.

[11] Rumelhart D E, Hinton G E, Williams R J. Learning Representations by Back Propagating Errors [J]. Nature, 1986, 323(6088):533-536.

[12] Andreieva V, Shvai N. Generalization of Cross-Entropy Loss Function for Image Classification [J]. 2021.

[13] Szegedy C, Liu W, Jia Y, et al.Going Deeper with Convolutions [J].IEEE Computer Society, 2014.