Composition Analysis and Identification of Ancient Glass Products Based on Optimized $BP_Adaboost$ Strong Classifier

Yaodi Tan [1, *], Zhiyuye Jing [2], Gaofei Yang [3]

[1] School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Jiangxi, China

[2] School of Science, Jiangxi University of Science and Technology, Jiangxi, China

[3] School of Information Engineering, Jiangxi University of Science and Technology, Jiangxi, China

* Corresponding author: 210744358@qq.com

Abstract. Due to the different combustion agents used in the manufacture of glass in ancient China and abroad, there are apparent differences in the composition of ancient glass products in China and abroad. This paper presents a reduction analysis and identification of the composition of ancient glass products, which is of great significance to the study of the Silk Road and the history of the development of glass products. This paper analyzes and identifies the composition of ancient glass products based on the data of C questions of the CUMCM competition. The study first used a neural network optimized by a genetic algorithm to screen out the elements closely related to the weathering condition and used Logistic regression to reduce the composition of the weathered samples. The reduced samples were subclassified by systematic clustering according to their respective compositions, and high-potassium glass and lead-barium glass were classified into two subclasses each. The classification basis was reasonably clarified by factor analysis. Finally, the BP_Adaboost strong classifier was optimized to identify and classify the samples based on the unknown sample components. The correct identification rate reached 100%, which played a good role in distinguishing the weathered Chinese and foreign glass products.

Keywords: Ancient Glassware Analysis and Identification, Genetic Algorithm Optimization, Hierarchical Cluster Method, Factor Analysis, $BP_Adaboost$ Strong Classifier.

1. Introduction

The Silk Road was a channel for the exchange of Western culture in ancient China, in which glass was valuable evidence of our early trade, the primary material of glass is quartz sand (SiO_2), to reduce the melting point of quartz sand, China generally uses lead ore as an accelerant, whose main components are lead oxide (PbO) and barium oxide (BaO), while the main components of the accelerant for glass products in Southeast Asia and India The main component of the combustion aid for glass products in Southeast Asia and India is grass ash (K_2CO_3), and limestone ($CaCO_3$) is added as the corresponding stabilizer[1]. As ancient glass is highly susceptible to weathering by the burial environment, in the process of weathering, the internal elements of the glass samples are highly susceptible to a large number of exchanges with elements in the environment, resulting in changes in the composition ratio of the samples, which affects the correct judgment of the sample categories and has an impact on the determination of the origin of the samples [2]. In this study, the elements susceptible to weathering are screened by a neural network optimized by a genetic algorithm. Their pre-weathering chemical content is restored so that the glass samples can be sub-classified and analyzed according to their pre-weathering composition. Finally, a more accurate classification and origin identification of the samples based on their pre-weathering composition content is achieved.

2. Materials and Methods

2.1. Data sources and analysis

This paper used data on the composition of ancient glassware for question C of the 2022 CUMCM competition (https://www.shumo.com) for analysis and study. Descriptive statistics were performed on the obtained data.

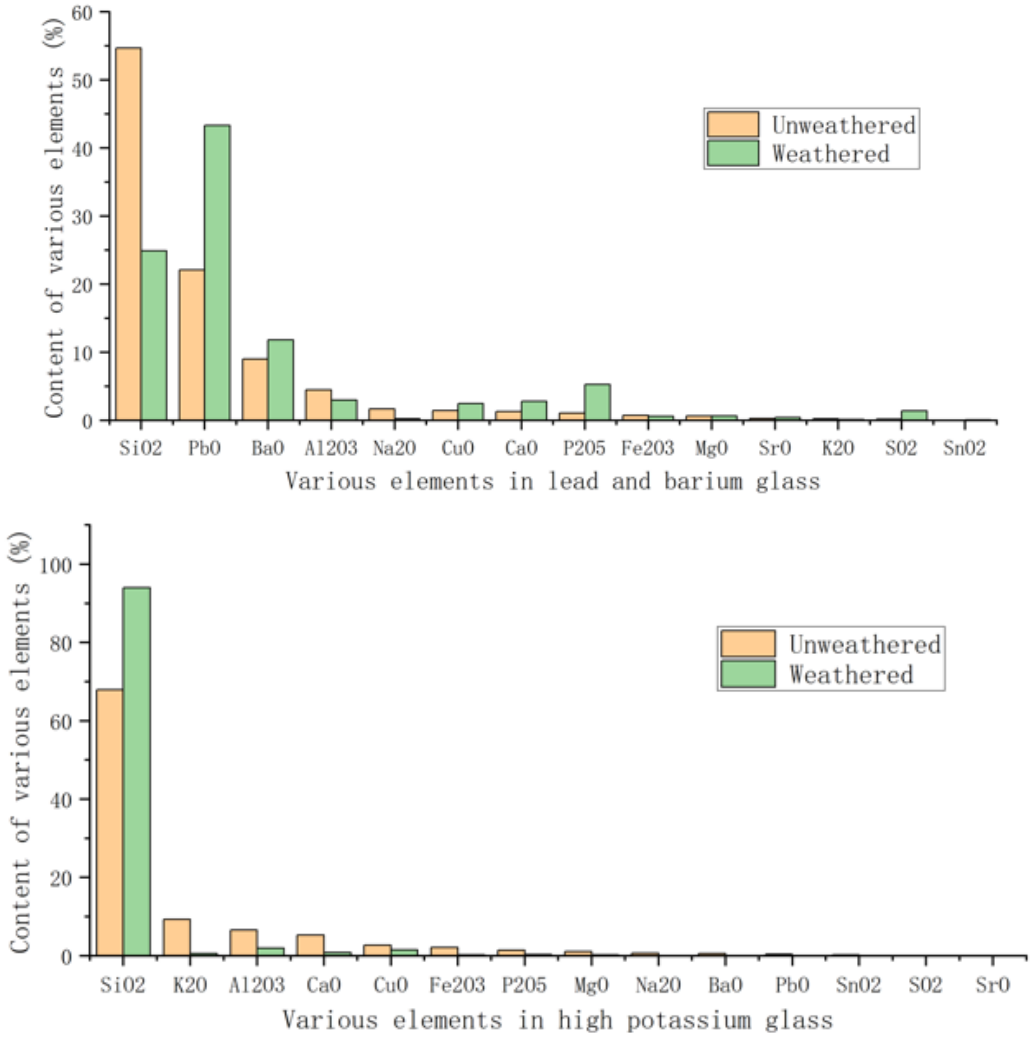


Figure 1. Chemical composition of two types of glass before and after weathering

By comparing the mean values of chemical composition content before and after weathering of the two types of glass, As shown in Figure 1, it can be found that lead-barium glass before and after weathering has a 29.75% decrease in silica content, a 21.23% increase in lead oxide content, and a slight increase in the percentage of phosphorus pentoxide content, but the rate of content after weathering has increased 4.03 times compared to that before weathering. It can be found that the percentage of silica content of high potassium glass increased by 25.98% before and after weathering, and the content of potassium oxide, calcium oxide, and alumina decreased significantly after weathering, even to 0. It can be inferred that for lead-barium glass type, silica content and alumina content have significant effects on weathering, and for high potassium glass type, silica content, potassium oxide content, calcium oxide content, and the content of silica, potassium oxide, calcium oxide, and alumina have significant effects on weathering.

2.2. Introduction to the method

After qualitatively examining the data, the article finds that not all elemental contents are strongly associated with weathering of glass samples. Moreover, the elements closely related to weathering are not the same for different glass types. To avoid the problem of overfitting, this paper first uses the BP neural network optimized by genetic algorithm to compress and downscale the independent variables to find out the elements that are closely related to the weathering of different types of glass and then uses the filtered variables to perform logistic regression to discriminate the weathering of different types of glass samples.

2.2.1. BP Neural Network for Dimensionality Reduction

To perform optimization computations using genetic algorithms, it is necessary to map the solution space to the coding space, where each code corresponds to a solution to the problem. In this paper, the length of the code is designed as 14, and each bit of the chromosome corresponds to an input independent variable. The value of each gene can only be "1" or "0" if the value of one bit of the chromosome is "1. Suppose the value of a chromosome is "1". In that case, it means that the input independent variable corresponding to that bit is involved in the final modelling, while the opposite is true if the input is "0". The inverse of the mean squared deviation of the test set data is selected as the genetic algorithm's fitness function so that the most representative independent variables are finally selected to participate in the modelling after continuous iterative evolution [3].

(1) Single BP model building: The BP model was first built with all 14 input independent variables to compare the prediction effects after optimization of the genetic algorithm.

(2) Initial population generation: N initial string structure data are randomly generated, i.e., one individual and N individuals become populations. The genetic algorithm starts iterating with these N string structures as the initial points.

(3) Adaptation function calculation: The function that measures an individual's fitness is called the fitness function. Here, the inverse of the sum of squared errors of the test set data is chosen as the fitness function:

$$f(x) = \frac{1}{SE} = \frac{1}{sse(\hat{T}-T)} = \frac{1}{\sum_{i=1}^{n}(\hat{t_i}-t_i)^2} \tag{1}$$

Where $\hat{T} = \{\hat{t_1}, \hat{t_2}, \cdots, \hat{t_n}\}$ is the predicted value of the test set $T = \{t_1, t_2, \cdots, t_n\}$ is the true value of the test set, and n is the number of samples in the test set.

To avoid the influence of the randomness of the initial weights and thresholds on the calculation of the fitness function, the weights and thresholds of the established BP neural network are optimized using a genetic algorithm for each individual to calculate the value of the fitness function.

(4) Select Operation: Calculate the sum of fitness of all individuals in the population:

$$F = \sum_{k=1}^{n_r} f(X_k) \tag{2}$$

The relative fitness of each individual in the population is calculated and used as the probability that the individual will be selected for an inheritance to the next generation:

$$P_k = \frac{f(X_k)}{F} \quad k = 1,2,3,\cdots,n_r \tag{3}$$

A simulated roulette operation is used to generate a random number between (0, 1) to determine the number of times each individual is selected. Individuals with greater fitness will have their genes expanded in the population.

(5) Crossover operation: For the compressed dimensionality reduction of the input independent variables, the crossover operation uses the simplest single-point crossover operator.

(6) Mutation operation: The simplest single point variation operator is used for the compressed dimensionality reduction variation operation of the independent variables.

(7) Output optimization results and build optimized BP model: When the iteration termination condition is satisfied, the output last generation population is the optimal or near-optimal solution, and the most representative combination of output independent variables is selected. The results are analyzed by rebuilding the model simulation test using BP neural network.

2.2.2. $Logistic$ regression to quantify the relationship between weathering and chemical composition content

In this problem, the weathering of the glass is the dependent variable, the content of various chemical components is the independent variable, and whether the sample is weathered or not is a dichotomous variable, let y_i denote the weathering of class i glass, y_i can only take 0 or 1, and the values and meanings are as follows [4].

$$y_i = \begin{cases} 1, \ Weathering \\ 0, \ Unweathered \end{cases} \tag{4}$$

For the case where the dependent variable is category, we can use $Logistic$ regression to construct a functional relationship between chemical composition content and weathering. Consider Yas the probability of event occurrence, $Y \geq 0.5$ means occurrence; $Y \leq 0.5$ means no occurrence. Let compound j in class i glass be the independent variable denoted as x_{ij}, $i = 0$for lead-barium glass, $i = 1$ for high-potassium glass

The linear probability model is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \mu_i \tag{5}$$

If the model is used directly, the predicted value$y_i$may appear as unrealistic as $y_i > 1$ and $y_i < 0$ To avoid this situation, we use a connection function $F(X_i, \beta)$ with a value domain of [0,1] for X_i. We use the $Sigmod$ function here as the connection function then we have the following equation:

$$F(X_i, \beta) = \frac{e^{x_i}}{1+e^{x_i}} \tag{6}$$

For the case given X_i, considering the two-point distribution of y_i, we interpret $Y_i = F(X_i, \beta)$ as the probability of y_i=1. If $Y_i \geq 0.5$, then we consider its prediction of y_i=1.

In summary, the $Logistic$ regression model can be expressed as:

$$Y_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\mu_i}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\mu_i}} \tag{7}$$

2.2.3. Subclassification of different species of glass using systematic clustering

The basic idea of systematic cluster analysis is to cluster variables that are close together into classes first and those that are farther apart into classes later, in order until each variable is grouped into the appropriate class [5]. This method is well suited for classifying ancient glass samples because early ancient glasses were made from different recipes, so their chemical compositions differed considerably. Therefore, it is reasonable to treat the chemical composition as a variable and cluster the samples step by step, which is used to analyze the classification pattern of high potassium glass and lead-barium glass.

2.2.4. $BP_Adaboost$ Strong classifier classification process

(1) Data selection network initialization. The sample space is randomly selected with m sets of training data, initialized with test data distribution weights $D_i(i) = \frac{1}{m}$, determined the neural network structure according to the sample input and output dimensions and initialized the weights and thresholds of the BP network.

(2) Weak classifier prediction. When the $NO.t$ weak classifier is trained, the BP network is trained with the training data, and the output is predicted to obtain the prediction sequence $g(t)$ east prediction error and e_i.

$$e_i = \sum_i D_{i(i)} \quad i = 1,2,\cdots,m(g(t) \neq y) \tag{8}$$

Where $g(t)$ is the predicted classification result y is the desired classification result.

(3) Calculate the predicted sequence weights.

$$a_i = \frac{1}{2} ln\left(\frac{1-e_i}{e_i}\right) \tag{9}$$

(4) Test data weight adjustment. Adjust the next round of training sample weights according to the predicted sequence weights a_i.

$$D_{i+1}(i) = \frac{D_{i(i)}}{B_i} \times e^{[-a_i y_i g_i(x_i)]} \quad i = 1,2,\cdots,m \tag{10}$$

(5) Strong classification function. The strong classification function $h(x)$ is obtained by combining the T groups' weak classification function $f(g_t, a_t)$ after training T rounds [6].

$$h(x) = sign[\sum_{i=1}^{T} a_i \cdot f(g_i, a_i)] \tag{11}$$

2.3. Evaluation of the model

2.3.1. Factor analysis to justify different classes of glass subcategories

Factor analysis is a technique for data simplification. It explores the underlying structure in the observed data by examining the internal dependencies among many variables and representing their underlying data structure with a few dummy variables. These few dummy variables can reflect the main information of the numerous original variables. The original variables are explicit observable variables, while the dummy variables are unobservable latent variables called factors. And the characterization of factors can be relatively reasonable to indicate the feasibility of subclassification.

Assume that the p-dimensional random vector $x = (x_1, x_2, \cdots, x_p)'$ has mean $u = (u_1, u_2, \cdots, u_p)'$ and covariance matrix $u = (u_1, u_2, \cdots, u_p)'$.

The general model for factor analysis is:

$$\begin{cases} x_1 = u_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_2 = u_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \vdots \\ x_p = u_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases} \tag{12}$$

Where $f_1, f_2, \cdots, f_m$ are called common factors and ε_i $(1,2,\cdots p)$ are special factors, and they are all unobservable random variables.

The common factors $f_1, f_2, \cdots f_m$ appear in the expression of each original variable x_i $(1,2,\cdots,p)$ and can be interpreted as certain characteristics shared by the original variables (with common

influences): each special factor ε_i $(1,2,\cdots,p)$ appears only in the expression of the $NO.i$ original variable x_i corresponding to it, and it acts only on this original variable. The above equation is expressed in matrix form as $x = u + AF + \varepsilon$. where $f = (f_1, f_2, \cdots, f_m)'(m \leq p)$ is the common factor vector, $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m)'$ is the special factor vector, and $A_{p\times m} = (a_{ij})$ is called the factor loading matrix, and the rank of A is assumed to be m [7]. We can use the features represented by the factors to explain the rationality of subclassification.

2.3.2. Verifying the stability of the prediction model using Monte Carlo simulation

Monte Carlo methods, also known as computerized stochastic simulation methods, are based on the statistical results of many events to calculate some deterministic problems [8]. In this paper, the stability and accuracy of the established classification model for sample classification are observed by computer perturbation of the component content of the glass in a random simulation.

3. Model Building and Solution

3.1. Regression equation of whether the sample is weathered or not using the variables with reduced dimensionality

3.1.1. Genetic Algorithm Optimized BP Neural Network for Independent Variable Compression and Dimensionality Reduction

Considering that the elements closely related to the weathering degree of high-potassium glass and lead-barium glass are not the same, the article carried out the dimensionality reduction of the independent variables for different types of glass samples separately to obtain the respective elements closely related to the weathering degree.

The genetic algorithm adaptation curve shown in Figure 2 below shows that the value of the adaptation function increases with the increase of the number of iterations. And in this paper, a 10-fold cross-test was conducted on the trained neural network using known samples, and its result is shown in Figure 3 below. And the correct prediction rate of the test set reached 95.65%. It can be seen that the independent variables obtained from the screening have a high correlation with the degree of weathering.

By screening the variables through this neural network, the chemical components in high potassium glass that have a greater influence on weathering are silicon dioxide (SiO_2), lead oxide (PbO), phosphorus pentoxide (P_2O_5), tin oxide (SnO_2), and potassium oxide (K_2O); the chemical components in lead-barium glass that have a greater influence on weathering are silicon dioxide (SiO_2), aluminium oxide (Al_2O_3), lead oxide (PbO), tin oxide (SnO_2), calcium oxide (CaO).

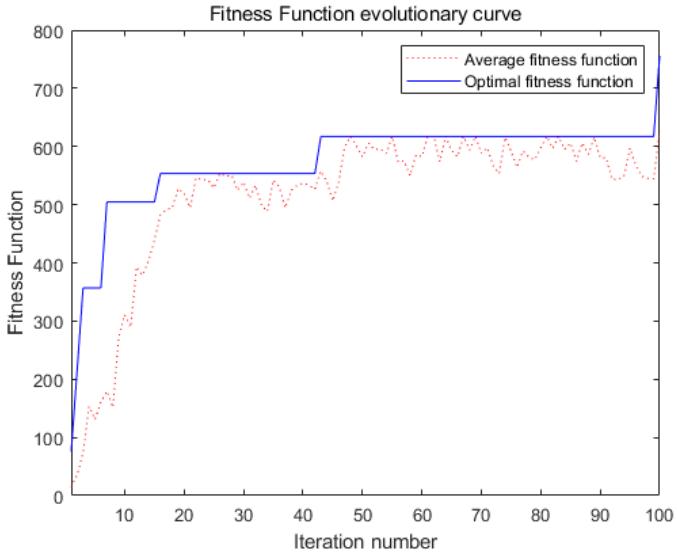


Figure 2. The fitness function curve of the legacy algorithm

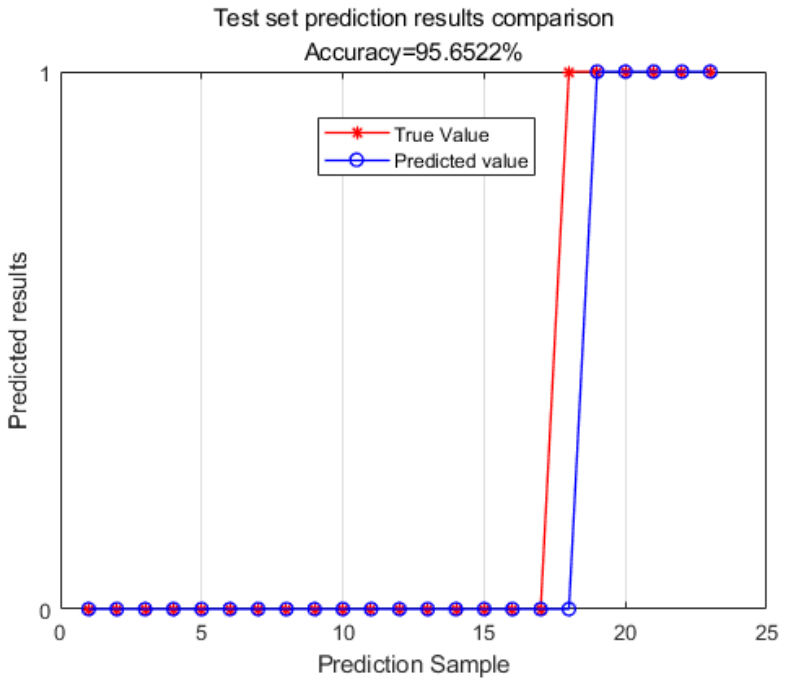


Figure 3. Predicted results of weathering conditions

3.3.2. Quantitative solution of the relationship between chemical composition content and weathering condition using $logistic$ regression

To reduce the multicollinearity among the independent variables, the regressions were conducted separately for different types of samples using the independent variables derived from the previous dimensionality reduction, considering that not all chemical components were closely related to whether the samples were weathered or not and the chemical components closely related to the weathering status of different types of glass samples were not the same. The result is shown in Table 1 and Table 2.

Table 1. Lead-barium glass $Logistic$ regression results

Variables	B	Standard deviation	Wald	Freedom	Significance	Exp(B)
SiO_2	-1.854	270.602	0	1	0.995	0.157
CaO	22.222	2633.18	0	1	0.993	4.48E+09
Al_2O_3	-15.776	2925.555	0	1	0.996	0
PbO	3.909	408.477	0	1	0.992	49.867
SnO_2	204.067	35614.49	0	1	0.995	4.22E+88
Constant	-56.435	14436.45	0	1	0.997	0

Table 2. High-potassium glass $Logistic$ regression results

Variables	B	Standard deviation	Wald	Freedom	Significance	Exp(B)
SiO_2	-0.251	2.449	0.01	1	0.918	0.778
PbO	-15.539	58.963	0.069	1	0.792	0
P_2O_5	-0.566	50.779	0	1	0.991	0.568
Al_2O_3	-0.718	10.966	0.004	1	0.948	0.488
K_2O	-1.948	5.547	0.123	1	0.725	0.143
Constant	33.428	251.328	0.018	1	0.894	3.29E+14

The final $logistic$ regression equation is obtained from the above table as:

$$y_0 = \frac{e^{-1.854x_1+22.222x_2-15.776x_3+3.909x_4+204.067x_5-56.435}}{1+e^{-1.854x_1+22.222x_2-15.776x_3+3.909x_4+204.067x_5-56.435}} \quad (13)$$

$$y_1 = \frac{e^{-0.251x_1-15.539x_2-0.566x_3-0.781x_4-1.948x_5+33.428}}{1+e^{-0.251x_1-15.539x_2-0.566x_3-0.781x_4-1.948x_5+33.428}} \quad (14)$$

For lead-barium glass, ten samples were randomly selected as the test set, and 39 samples as the training set several times, and the average correct prediction rate was 95.63%.

3.2. Subclassification of glass samples using systematic clustering and tested by factor analysis

The relationship between weathering conditions and chemical composition content has been quantitatively analyzed above. We reduce the chemical composition content of the samples after weathering to the content before weathering for the following analysis based on the changing pattern of chemical composition content before and after weathering for different classes of samples. In this paper, based on the two major classes of lead-barium glass and high-potassium glass, the glass samples were subclassified by systematic clustering. Each of them was divided into two subclasses. The lead-barium glasses are classified into stable and colour-rich classes, while the high-potassium glasses are classified into high-silica and low-silica classes.

Referring to the relevant literature[9], it is concluded that CuO and BaO are both coloring oxides in ancient glasses, while $Ca^{2+}, Fe^{3+}$and Al^{3+}can replace Si^{4+} in the $[SiO_4]$ tetrahedron to maintain the stability of the glass and are stable oxides in ancient glasses[10]. We choose these five elements as independent variables, and the classification of lead-barium glass is verified by factor analysis. According to the gravel test results, two factors are selected for analysis in this paper. The obtained rotated factor matrix is shown in Table 3 below.

Table 3. Rotation factor matrix of main components of lead barium glass

Oxide variables	F1	F2
CaO	-0.057	0.87
Al_2O_3	-0.482	0.413
Fe_2O_3	-0.165	0.8
CuO	0.888	-0.06
BaO	0.904	-0.136

The larger the absolute value of the correlation coefficient between the oxide variable and a factor, the closer the factor is to the variable and the greater the contribution of the factor embodying that variable. Then it is obvious that F1 can be regarded as a colourant factor and F2 is a stability factor. And concerning the classification results of systematic clustering, the mean values of the five oxide components of the two subclasses are shown in Figure 4 below, which can be interpreted by factor analysis as the classification principle of the two subclasses as the property stability class and the colour-rich class.

And the subclassification of high potassium glass was divided into high silica and low silica. Based on the results of systematic clustering, descriptive statistical analysis was performed on the content of high potassium glass in different classes. The elemental silica content was found to be fully consistent with the division of subclasses. The results are shown in Figure 5.

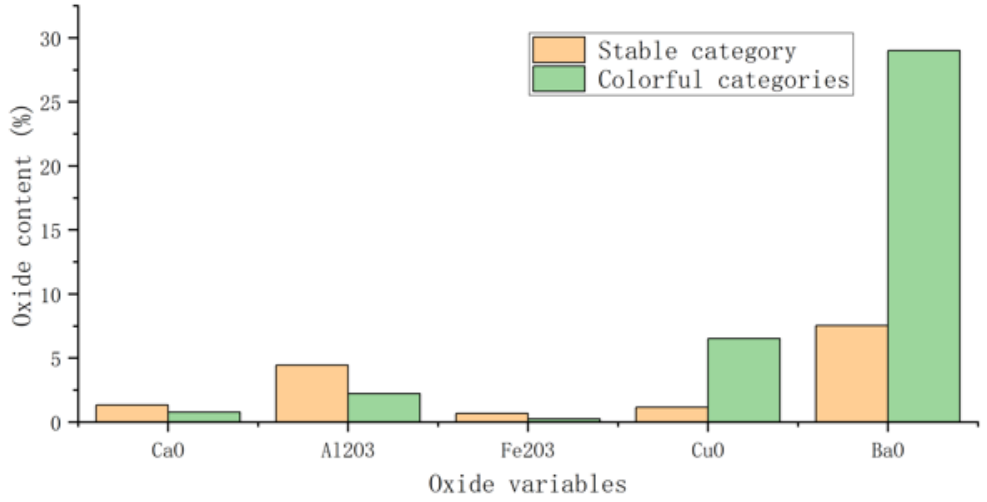


Figure 4. Lead-barium glass subclasses oxide content

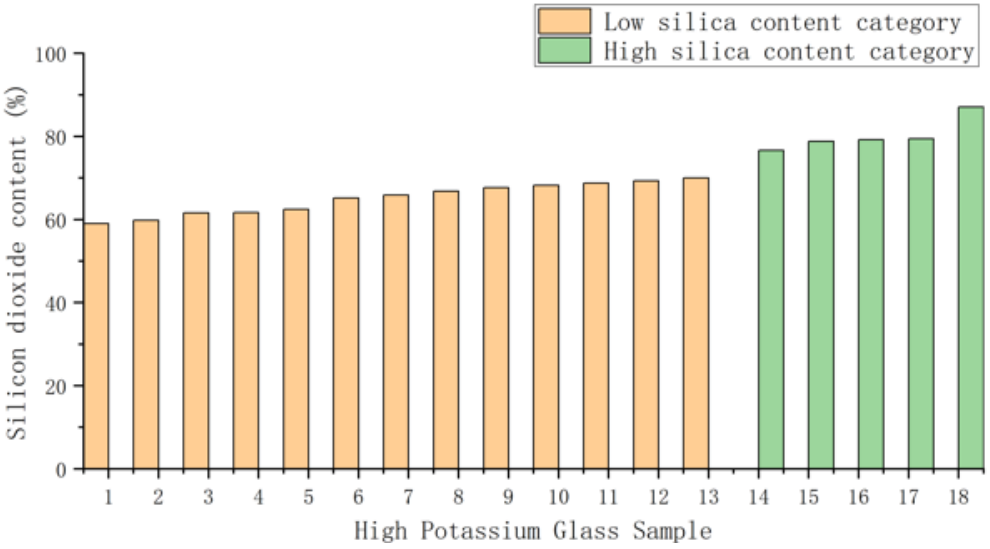


Figure 5. High-potassium glass subclasses oxide content

3.2.2. Verification of the stability of the classification results using Monte Carlo simulation

The above process has subdivided the glass types into four categories. The sensitivity of the subclassification results for lead-barium and high-potassium glass will be discussed separately below.

In this paper, the classification results are first used as training data for the *BP* neural network. Then a random perturbation is simulated using Monte Carlo to gradually increase the chemical composition change rate from 1% to 100%. If the results are consistent, then the sample classification is consistent with the original category. If the results are consistent, the classification model is stable.

The following graphs show the accuracy of category prediction of chemical composition content of ancient glass samples after perturbation. As shown in Figure 6 and Figure 7.

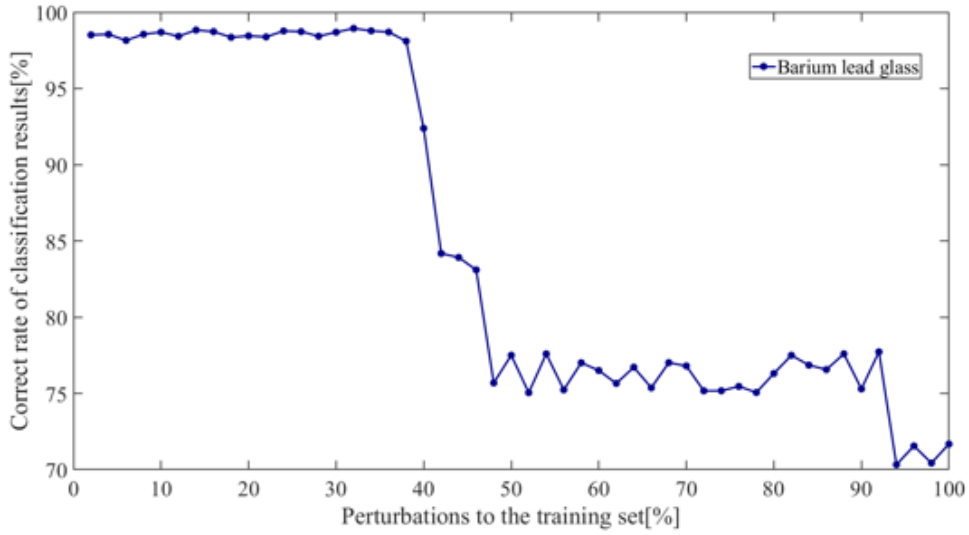


Figure 6. Classification prediction accuracy of lead-barium glass after perturbation

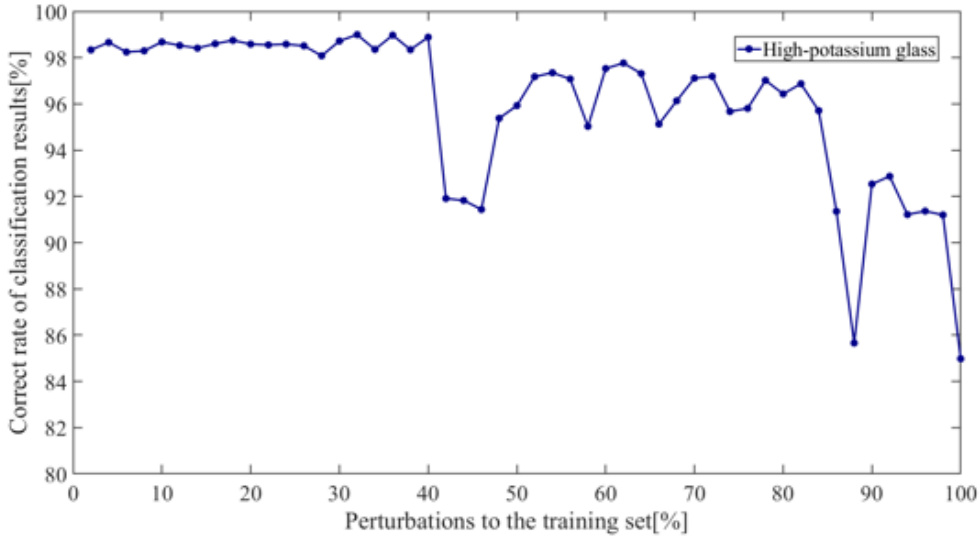


Figure 7. Classification prediction accuracy of high-potassium glass after perturbation

With the change in the amount of perturbation to the chemical composition, there will be a certain fluctuation of the correct test rate, which is shown by the gradual decrease of the accuracy of the added test with the increase of the disturbance term, and the prediction accuracy of the test set is still above 70% when the perturbation term is 100%, indicating that the subclassification stability of the glass samples is good.

3.3. Classification of unknown samples using an optimized *BP_Adaboost* strong classifier

Based on the results already obtained above. This article can use the content of the pre-weathering samples and the final sub-classification results as the training set to train the neural network model. Then this article uses the final derived model to perform the weathering detection for samples with known content, unknown state and type. Firstly, to detect whether they are weathered or not, to restore the weathered samples, and finally, to classify the restored samples, to predict the class to which the samples belong.

In this paper, 80% of all glass samples are randomly selected as the training set, and the remaining 20% of samples are used as the test set to build a perfect detection model. We used two different network models, BP neural network and optimized *BP_Adaboost* strong classification, and compared the final recognition correct rate. The final detection results.

In terms of the results, as shown in Figure 8 and Figure 9 below, for multiple randomly selected unknown samples for component reduction and category determination, the experiments yielded an

accuracy of 92.86% for the classification model built based on BP neural network, and up to 100% for the classification model built based on $BP_Adaboost$ strong classifier. Thus, it is concluded that the classification model built based on $BP_Adaboost$ strong classifier is more suitable for judging the category of glass samples.

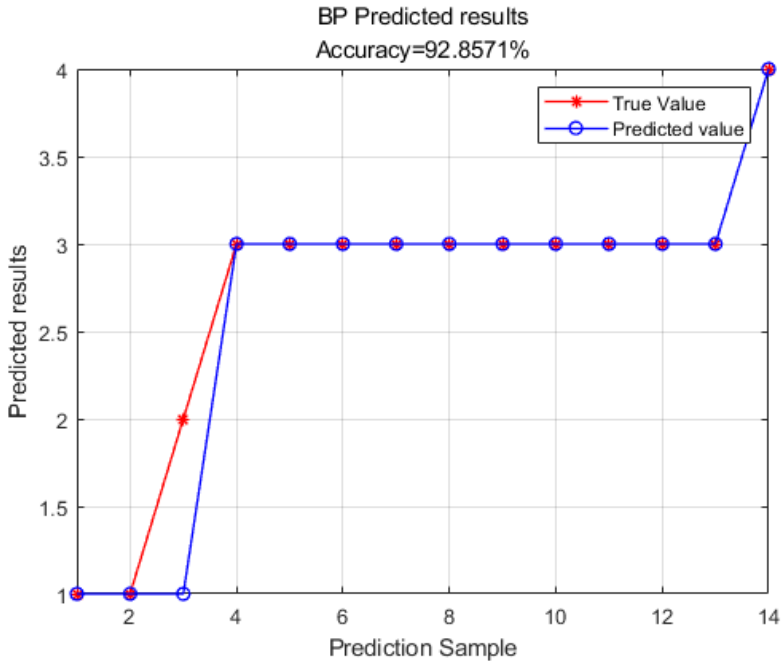


Figure 8. Prediction results of BP Neural Network model

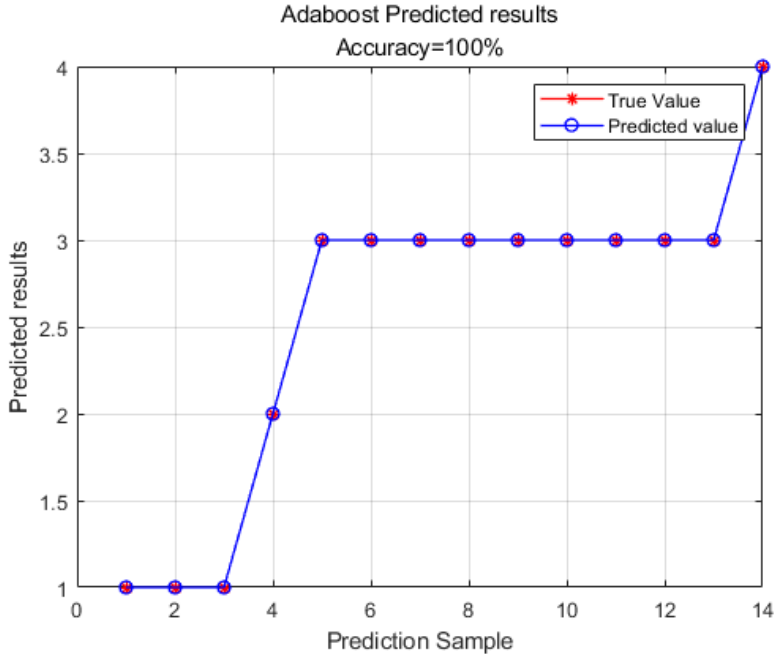


Figure 9. Prediction results of $BP_Adboost$ Neural Network model

4. Conclusions

Based on the strong class classification ability of the optimized $BP_Adaboost$ strong classifier, this paper proposes a completer and more reliable model for judging the class of ancient glass samples. It realizes the discrimination of ancient glass sample categories that only know the chemical composition content. The correct rate of judgment is close to 100% during the experiment, with more

accurate judgment ability, and then realizes the tracing of unknown ancient glass sample origins, which plays a role in the research of ancient glass and the history of the Silk Road.

References

[1] Brill R H, Hood H P. A new method for dating ancient glass[J]. Nature, 1961, 189 (4758): 12 - 14.

[2] Dal Bianco B, Bertoncello R. The development of growth rings on ancient glass surfaces: Description and simulation of the weathering [J]. Journal of non-crystalline solids, 2008, 354 (2-9): 773 - 779.

[3] Mirjalili S. Genetic algorithm [M]//Evolutionary algorithms and neural networks. Springer, Cham, 2019: 43 - 55.

[4] LaValley M P. Logistic regression [J]. Circulation, 2008, 117 (18): 2395 - 2399.

[5] Ezugwu A E, Shukla A K, Agbaje M B, et al. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature [J]. Neural Computing and Applications, 2021, 33 (11): 6247 - 6306.

[6] Li H, Liu S, Hassan M M, et al. Rapid quantitative analysis of Hg2+ residue in dairy products using SERS coupled with ACO-BP-AdaBoost algorithm [J]. Spectrochemical Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 223: 117281.

[7] McDonald R P. Factor analysis and related methods [M]. Psychology Press, 2014.

[8] Harrison R L. Introduction to monte carlo simulation [C]//AIP conference proceedings. American Institute of Physics, 2010, 1204 (1): 17 - 21.

[9] Rehren T, Freestone I C. Ancient glass: from kaleidoscope to crystal ball [J]. Journal of Archaeological Science, 2015, 56: 233 - 241.

[10] Fu Xiu-Feng, Gan Fu-Xi. A study on the composition of a group of ancient glass from southern and southwestern China based on multivariate statistical analysis method [J]. Conservation and Archaeological Science, 2006 (04): 6 - 13. DOI: 10.16334/j.cnki.cn31 - 1652/k.2006.04.002.