Category identification and composition analysis of ancient glass products based on GA-BP neural network and factor analysis

Juele Xie #,*, Jiahui Ye#, Jun Zhou#

Hainan University, Hainan, Haikou, 570228 China

*Corresponding author: xjl020725@163.com

#These authors contributed equally.

Abstract: Ancient glass is a witness of early trade exchanges, and its study has profound historical significance. In this paper, we analyze the nature and chemical composition data of a batch of ancient glass products to investigate the compositional differences and identification characteristics of different types of glass products. In this paper, we use relevant information to achieve the effect of identifying cultural relics based on chemical composition data by mathematical modeling for a batch of ancient glass products in China (divided into high potassium glass and lead-barium glass), and analyze the correlation between the chemical composition of different categories of glass to compare the differences between classes.

Keywords: Genetic algorithm; GA-BP neural network; Spearman correlation analysis; Factor analysis.

1. Introduction

In ancient times, glass was a valuable physical evidence of early trade exchanges, and although our glass is similar in appearance to foreign but contains different chemical compositions, among which, lead-barium glass has a relatively high content of lead oxide and barium oxide, and high potassium glass is fired by grass ash as a flux, so its potassium content is high[1]. Ancient glass is susceptible to weathering by the environment, and the exchange of internal elements with environmental elements occurs, which in turn affects the judgment of the category.

For a batch of ancient glass products in China (divided into high potassium glass and lead-barium glass). A BP neural network with optimized weights and thresholds by genetic algorithm was used for the prediction of categorical variables. The prediction accuracy of the GA-BP neural network constructed by debugging was very high, and its prediction accuracy for the training set and the test set was as high as 98% and 100%, respectively, based on which the cultural relics of unknown categories were identified, after which the input samples were adjusted by a certain proportion for sensitivity analysis of this model, and finally the glass with artifact numbers A1, A6 and A7 belonged to high potassium glass, and the glass with artifact numbers A2, A3, A4, A5 and A8 belonged to lead-barium glass and the model was more stable. The Spearman correlation analysis was used to explore the correlation of chemical components within the class, and then factor analysis was introduced to condense the information, and the difference between chemical components was portrayed in the form of a factor weighted score formula and a comprehensive table of weighted components.

2. Model assumptions and notation

2.1 Assumptions

Assumption 1: The component test data in the annex are all true and reliable.

Assumption 2: Both sampling points of the weathered glass artifacts belong to the weathered area.

2.2 Notations

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

Table 1 Notations

Symbols	Interpretation and Explanation
x、x^*	Denote the sample data before and after normalization, respectively
m、n、l	The number of nodes in the input layer, output layer and hidden layer of the neural network
SD、σ	Standard deviation of the sample
D	Weathering variable, D=1 for weathering, D=0 for undifferentiated
$\rho_{X,Y}$	Spearman correlation coefficient between variables X and Y
F	Common factor in factor analysis

3. Model construction and solving

3.1 Classification prediction of glass artifacts based on unknown categories of GA-BP neural network algorithm

The advantage of using BP neural network algorithm[2-4], also known as forward feedback neural network, is that it can learn and store a large number of input-output pattern mapping relationships without revealing the mathematical equations describing such mapping relationships beforehand. The core idea of this algorithm is the "negative gradient descent" theory, which means that the error adjustment of the BP neural network always proceeds in the direction of the fastest decrease in error, and the network weights and thresholds are continuously adjusted by back propagation to minimize the sum of squared errors of the network.

Genetic algorithm[5] is an approach to search for the optimal solution by simulating the natural evolutionary process. Since the ordinary BP neural network is very sensitive to the initial weights, and the initial weights and thresholds are selected by random numbers, which may converge to a local minimal value, the genetic algorithm can be introduced to perform genetic operations on the initial weights and thresholds, and the optimal weights and thresholds are obtained by continuous elimination and selection, and then the neural network is operated, which not only can get the global optimal solution, but also can reduce the number of iterations to solve the shortcoming of slow convergence of neural networks[6].

Let W_1 be the input layer to hidden layer weights and B_1 be the input layer to hidden layer threshold; W_2 be the hidden layer to output layer weights and B_2 be the hidden layer to output layer threshold.

Step1: Determine the structure of BP neural network

There are three types of transfer functions commonly used in neural networks as follows.

$$\begin{cases} f(x)=\begin{cases} 1\ ,\ x>0 \\ -1\ ,\ x\le 0 \end{cases} \\ f(x)=x \\ f(x)=\dfrac{1}{1+e^{-x}}\ ,\ f(x)=\dfrac{1-e^{-x}}{1+e^{-x}} \end{cases} \tag{1}$$

The first threshold type is often used in simple classification models, the second linear type is generally used only in input neurons and output neurons, and the third S-type is often used in hidden layer neurons. In this paper, linear functions are used in the input and output layers, and the first S-type transfer function is used in the hidden layer.

Step2: Normalize the input and output samples to be trained

Since the input and output matrices have different data levels, in order to facilitate the subsequent data processing and speed up the convergence of the neural network, it is necessary to first normalize the data through the following formula to process the data set to the 0~1 interval.

$$x^*=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{2}$$

Step3: Selection of network learning parameters

The number of neurons in the hidden layer has a significant impact on the accuracy of the final prediction results of the BP neural network algorithm, so we refer to the pooling method to determine the optimal number of nodes in the hidden layer, which is calculated as

$$l = \sqrt{m+n} + a \ , \ a = 0,1,2,...,10 \tag{3}$$

Similarly, different values of network learning efficiency (speed) η are taken to observe how well the network output fits the actual output to determine the relative optimal learning efficiency.

Step4: Initialize network weights and thresholds

Generally, the initial W_1, B_1, W_2 and B_2 are generated by random numbers, so the initial values are often not the optimal solutions, and the initial neural network weights and thresholds (considered as the initial population) need to be encoded by genetic algorithms to perform genetic operations such as selection, crossover and variation to update the population, and the optimal weights and thresholds are output when the population meets the conditions, and the optimal values are substituted into the subsequent steps for The optimal value will be substituted into the subsequent steps for calculation.

Step5: GA-BP neural network training

For a given training pattern input, the output pattern of the network is calculated and compared with the desired output, if it meets the requirements, the result is output, if there is an error, the error is back-propagated to correct the weights and thresholds, and the new weights and thresholds are used to input again until the output.

Step6: Output data reduction (inverse normalization)

When the MSE of the sample is less than the set target, x^* is reduced to the original mathematically meaningful x by the transformation of the formula in Step 2, which means that the prediction accuracy of this neural network reaches the predetermined target and the training is completed and can be used for subsequent data prediction.

After continuous trial and contrast, and combined with previous experience, the two important parameters of the neural network (learning efficiency and training times) were taken several times and repeated experimental analysis, and the following Figure 1 were obtained.

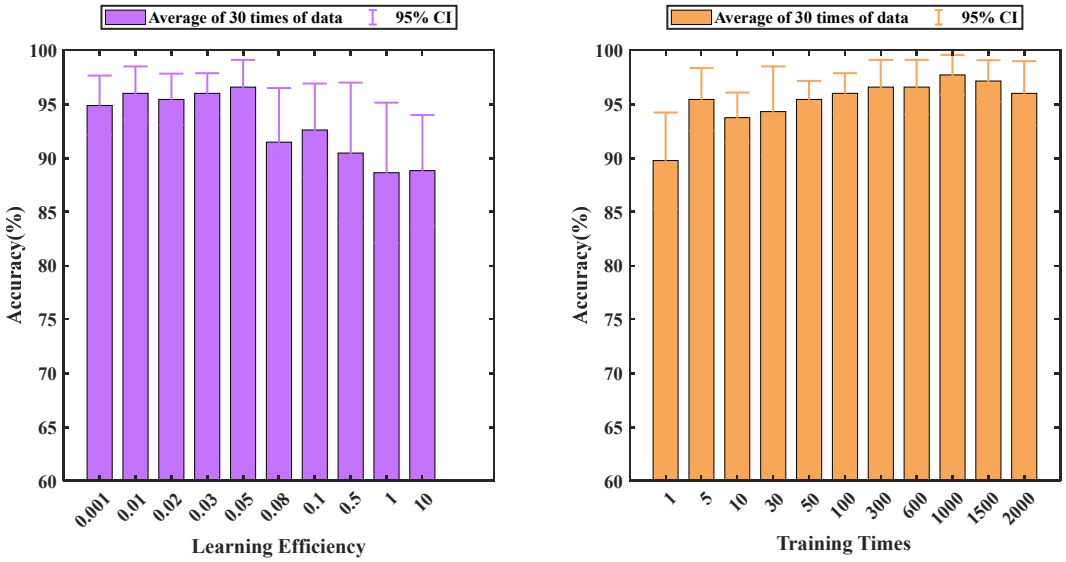


Figure 1 Relationship between training parameters and accuracy

Using the same method to determine other parameters, the optimal neural network learning parameters were finally determined one by one: the number of implicit layers was 1, the number of nodes in the implicit layer was 14, the learning efficiency was 0.05, the number of training was 1000, and the training target error MSE was set at 0.000001; and the optimal genetic algorithm parameters: the number of genetic generations was 50, the population size was 20, the parameter of the crossover probability was 0.8, and the parameter of the variation probability was 0.1.

Through these parameters and combined with MATLAB to decode the optimal individuals, the optimal weights and thresholds can be obtained, and the optimal values are used to train the BP neural

network. Meanwhile, to prevent overfitting, the samples are divided into training set (48 randomly selected from the samples) and test set (the rest of the samples) to train and test the neural network). Finally, obtained the comparison plot of the glass type prediction Figure 2 based on the GA-BP algorithm (1 indicates high potassium glass and 2 indicates lead-barium glass).

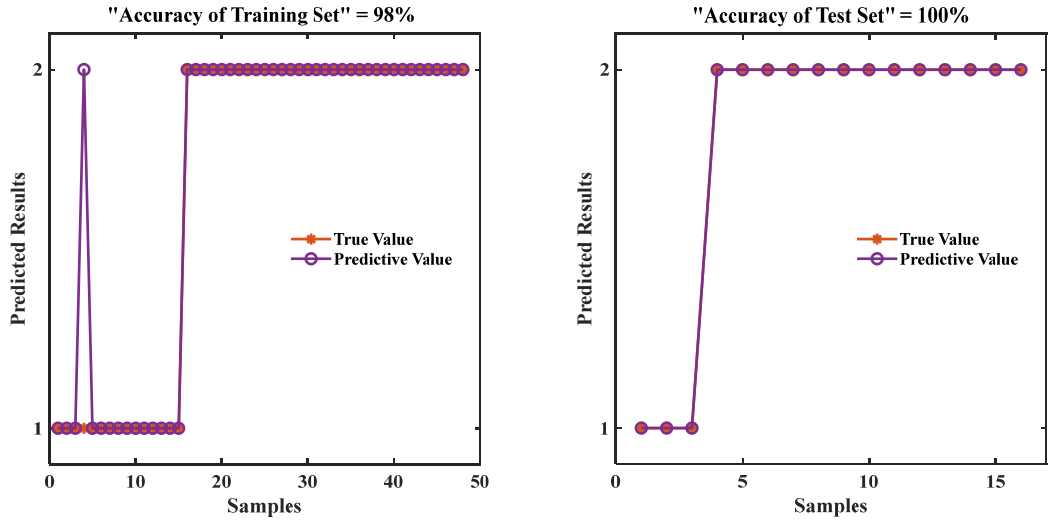


Figure 2 GA-BP neural network prediction results (left is the training set, right is the test set)

It can be seen that the BP neural network optimized based on GA has a good degree of overlap with the prediction results of the training set samples, and the prediction accuracy also reaches 98%, which indicates that the training effect is more satisfactory. Meanwhile, in order to prevent overfitting, the prediction accuracy of the trained neural network is tested using the test set, and at this time the accuracy even reaches 100%, which is very high, and the accuracy of the model is verified at the same time. It indicates that the model is suitable for the subsequent classification prediction of glass artifacts.

After the GA-BP neural network was trained, the weathering was first represented by the value of the weathering variable D, and then 15 data (14 chemical components and 1 weathering variable) of the unknown category artifacts were imported into MATLAB for prediction.

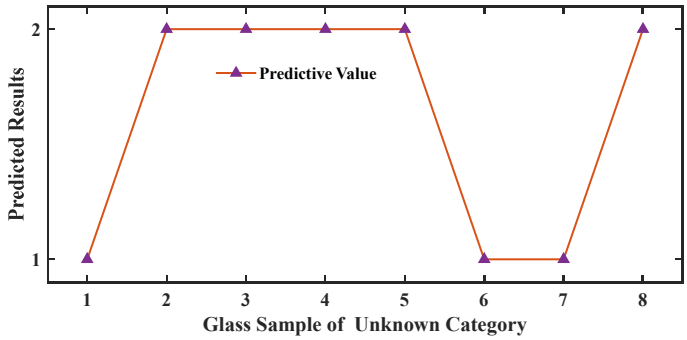


Figure 3 Glass artifact identification results

Based on the output of Figure 3 above, I can tell that the artifacts numbered A1, A6, and A7 belong to high potassium glass, and the artifacts numbered A2, A3, A4, A5, and A8 belong to lead-barium glass.

For the sample input data (except for the weathering variable D) were fluctuated up and down by 10%, the coefficient of variation was increased from 0.9 in the order of 0.02 to 1.1, and the experiment was repeated several times to obtain the following results.

The change of the average prediction accuracy of the GA-BP model with the change of the sample data can be observed in Figure 4, and it is found that the accuracy of the model decreases to different degrees regardless of the positive (increasing) or negative (decreasing) change of the data, and the more the deviation from the original value of the sample, the lower the average accuracy, but the overall can maintain the level of more than 90%, i.e., regardless of the fluctuation of the sample, the

model can still The overall accuracy of the model is above 90%, that is, regardless of the fluctuation of the sample, the model is still at a high level of prediction accuracy. In summary, it shows that the model is less sensitive to sample changes and has good stability.

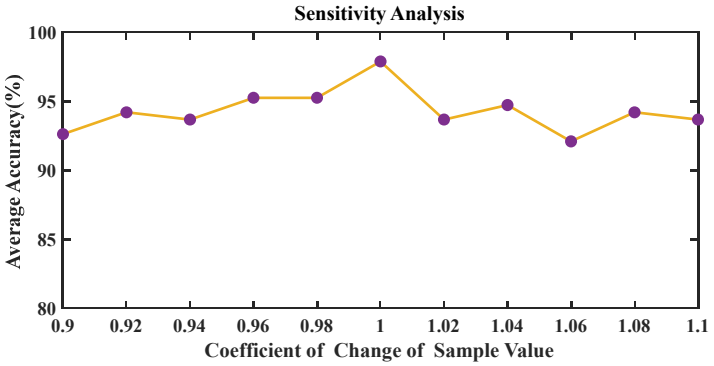


Figure 4 Sensitivity Analysis Chart

3.2 Chemical composition analysis of glass

In the field of natural science, Spearman's correlation coefficient is widely used to measure the degree of correlation between two variables with non-normal distribution, while the data of each component of the glass samples were tested to be not normally distributed (as can be seen visually in the figure below), which is obviously not suitable for Pearson's correlation analysis.

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}\left(p_i - \overline{p}\right)\left(q_i - \overline{q}\right)}{\sqrt{\sum_{i=1}^{n}\left(p_i - \overline{p}\right)^2}\sqrt{\sum_{i=1}^{n}\left(q_i - \overline{q}\right)^2}} \tag{4}$$

The Spearman correlation coefficients between two sets of random variables $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ can be obtained by the above calculation, where, $\rho_{X,Y}$ is the Spearman correlation coefficient, p_i and q_i respectively are the ranking of x_i and y_i, if the values appear equal in the variables, the ranking corresponding to that value is the average of the rankings corresponding to those values[7]. Visualization of data results was performed through Rstudio.

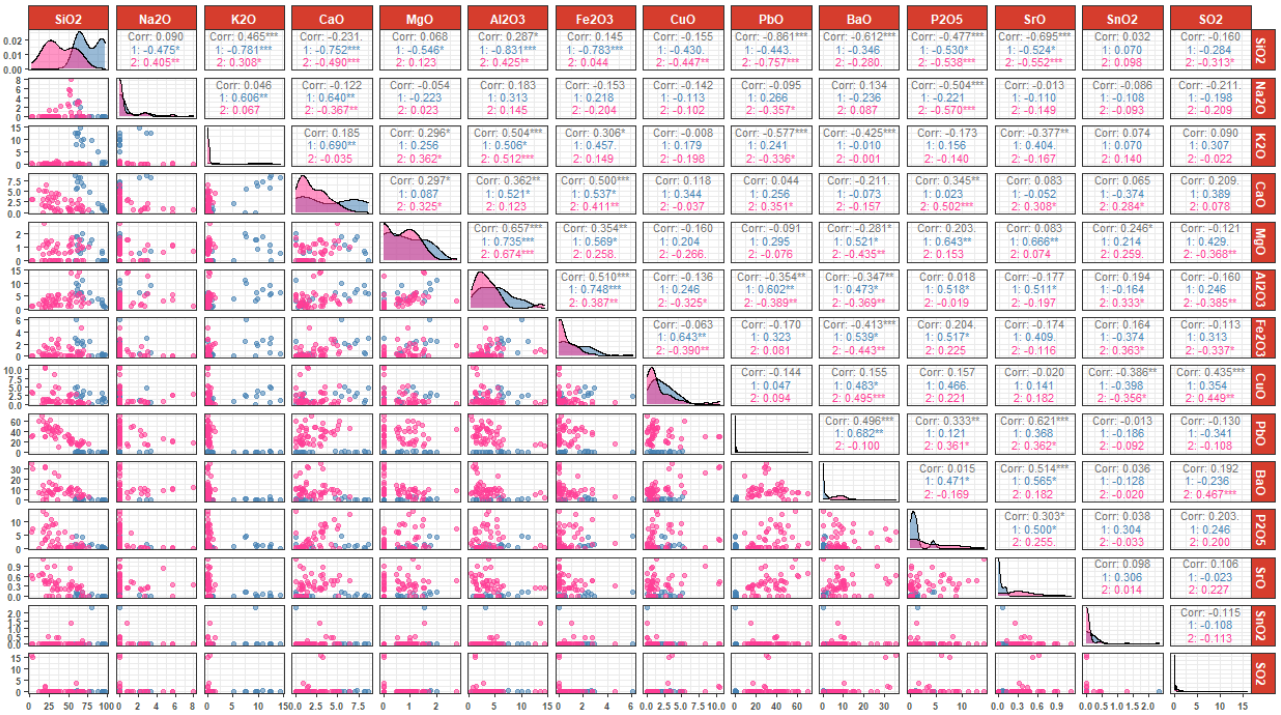


Figure 5 Spearman correlation analysis graph

The lower triangular part of the figure presents the scatter plot between each two variables, the upper triangular part shows the Spearman correlation coefficient between each two variables, *

represents significance, (blue indicates the data results of high potassium glass, red indicates the data results of lead-barium glass, and black indicates the data results of all samples), and the diagonal part shows the density distribution of the data of each variable.

For the data without *, it can be assumed that there is no significant correlation between the two variables. Preliminary analysis shows that there are significant differences in the distribution of SiO_2, CaO, BaO, P_2O_5 and SrO between lead-barium glass and high potassium glass, which indicates that there is a certain correlation between these elements and glass type, while the rest of the components are weakly influenced by the type.

At the same time, some basic correlations can be analyzed according to Figure 5: for high potassium glass, there is a strong negative correlation between SiO_2 and PbO and BaO; there is a strong positive correlation between K_2O, CaO and Al_2O_3, while all three show a strong negative correlation with PbO and BaO; while for lead-barium glass, there is still a strong negative correlation between SiO_2 and PbO and BaO; the correlation between K_2O, CaO and Al_2O_3 has been very weak; it can also be considered that the three are not correlated with PbO and BaO.

However, at this time, the data dimension is still high, and there are 14 metal elements, which make it difficult to read the information. Based on the above analysis, the data were condensed for further analysis by factor analysis[8-11]. The data were divided into four categories: high potassium-weathered, high potassium-unweathered, lead-barium-weathered, and lead-barium-unweathered. Meanwhile, in order to prevent the influence of poor quality data on the analysis results, the sampling point data with the total chemical content less than 90% were eliminated, and the data sets of each category were tested before factor analysis separately, and the results of the following Table 2 were obtained after eliminating the chemical composition data with low absolute value of factor load coefficient one by one:

Table 2 KMO and Bartlett's test

Indicators	Type			
	Lead-Barium - Weathering	Lead-Barium - Unweathered	High Potassium - Weathering	High Potassium - Unweathered
KMO value	0.660	0.524	0.615	0.570
Bartlett's spherical test P-value	0.000	0.000	0.000	0.031

The above table shows that the final KMO values of the lead-barium-weathering, lead-barium-unweathering, high potassium-weathering and high potassium-unweathering models are 0.660, 0.524, 0.615, and 0.570 respectively, which are all greater than 0.5, and are considered to be factor testable with the model, and the Bartlett's spherical test p-value is less than the significant level of 0.05, which is considered to be a significant correlation between the variables. The premise of factor analysis was satisfied.

By solving the factor analysis model through SPSSAU[12], the total variance interpretation and the rotated component matrix of the four classification models can be obtained separately. (Due to space limitation, this paper only presents the analysis process for the lead-barium-weathering glass, and the other three models are analyzed in exactly the same way.)

Table 3 Explanation of Variance

Factor Number	Eigen Value			Explanation of Variance Before Rotation			Explanation of Variance After Rotation		
	Eigen Value	Explanation of Variance	Accumulation	Eigen Value	Explanation of Variance	Accumulation	Eigen Value	Explanation of Variance	Accumulation
1	2.757	45.951	45.951	2.757	45.951	45.951	2.625	43.749	43.749
2	1.811	30.18	76.131	1.811	30.18	76.131	1.943	32.382	76.131
3	0.872	14.54	90.671	----	----	----	----	----	----
4	0.268	4.468	95.139	----	----	----	----	----	----
5	0.153	2.555	97.694	----	----	----	----	----	----
6	0.138	2.306	100	----	----	----	----	----	----

Table 3 above shows the variance contribution and cumulative contribution of the rotated indicators of chemical components of lead-barium-weathering glass. The factors with eigenroots greater than 1 were extracted as principal components, and the variance of the 1st principal component accounted for 43.749% of all principal components, while the cumulative variance contribution of the first 2 principal components reached 76.131%, so the use of the first 2 principal components was sufficient to represent the role of the main chemical components in lead-barium-weathering glass.

Table 4 Component score coefficient matrix

Name	Ingredients	
	Ingredients 1	Ingredients 2
SO_2	0.364	0.028
BaO	0.336	-0.060
P_2O_5	0.083	0.479
PbO	-0.010	0.240
CuO	0.369	0.075
CaO	0.023	0.478

The expressions of the chemical components corresponding to each common factor can be obtained from Table 4 above. Further, by normalizing the matrix of coefficients of the common factor scores, denoted as *F*, we can obtain the composite scores calculated with the proportion of the variance contribution of each common factor corresponding to the weights, and the final results are shown in Table 5 below.

Table 5 Factor score formula

Lead-barium-weathering glass
$F_1 = 0.364x_{SO_2} + 0.336x_{BaO} + 0.083x_{P_2O_5} - 0.010x_{PbO} + 0.369x_{CuO} + 0.023x_{CaO}$
$F_2 = 0.028x_{SO_2} - 0.060x_{BaO} + 0.479x_{P_2O_5} + 0.240x_{PbO} + 0.075x_{CuO} + 0.478x_{CaO}$
$F = (43.749/76.131) \times F_1 + (32.382/76.131) \times F_2$

Similarly, the composition score coefficient matrix and factor score equations for lead-barium-unweathered, high potassium-weathered, and high potassium-unweathered glasses can be obtained, and by further collating the above results we can obtain the composition weight tables for different chemical components for different types of glasses, and the results are shown in Table 5. The principal components extracted from the factor analysis can explain the vast majority of the variance of the variables, and the weights of each element in the principal components can be calculated based on the weighted scores, so the influence of each major element on the four models and the differences in different models can be observed visually.

Table 6 Composition-weighting table for different types of glass

Type	Principal Components	Weights	Chemical Composition
Lead-Barium-Weathered	Principal Component 1	57.47%	SO_2、BaO、CuO
	Principal Component 2	42.53%	P_2O_5、PbO、CaO
Lead-Barium-Unweathered	Principal Component 1	51.83%	SiO_2、PbO、CaO
	Principal Component 2	48.17%	BaO、CuO
High Potassium-Weathered	Principal Component 1	58.33%	K_2O、CuO、P_2O_5
	Principal Component 2	41.67%	SiO_2、CaO、Fe_2O_3
High Potassium-Unweathered	Principal Component 1	58.95%	Na_2O、K_2O、CaO
	Principal Component 2	41.05%	MgO、Fe_2O_3

Through the comprehensive analysis of Table 4, Table 5 and Table 6, the following information can be obtained: for lead-barium glass, PbO and BaO are often combined as the explanatory factors for the main components of lead-barium glass, because lead ore is added as the flux during the firing process of lead-barium glass; For high potassium glass, K_2O and CaO often jointly explain the composition content of high potassium glass. The reason may be that the ancient glass in some regions uses plant ash as the flux, while the plant ash contains potassium, calcium and other alkaline elements, and limestone is added as the stabilizer during the firing process. After calcination, the limestone is converted into CaO, so this is fully explained; Fe_2O_3 is only extracted as the main component of the high potassium glass, which indicates that Fe_2O_3 should play an independent role in explaining the high potassium glass. Through consulting the relevant literature, it is inferred that it is likely to be used as the colorant of the high potassium glass, which also conforms to the actual situation in ancient times[13]; At the same time, we can observe that although SiO_2 is the main component of glass, it only appears in the main components of weathered high potassium glass and unweathered lead-barium glass. This is because weathering will reduce the proportion of SiO_2 in lead-barium glass and increase the proportion of SiO_2 in high potassium glass. When the content of SiO_2 changes significantly due to weathering factors, the proportion of other components will change significantly in the opposite direction, which leads to the above situation, Therefore, few components can be explained jointly with them in the two types of glasses.

4. Conclusion

In this paper, for the identification of unknown categories of glass artifacts, the GA-BP neural network algorithm is used. Compared with the ordinary BP neural network, the optimized algorithm has a very high prediction accuracy, and also circumvents the complex mathematical equation prediction, which is highly operable and generalizable. At the same time, Spearman correlation analysis and factor analysis are used to explore the relationship between the chemical components in glass, and the role of the main chemical components is reasonably explained in combination with the actual situation. This paper has a certain guiding significance for the composition analysis and identification of ancient glass relics.

References

[1] Duan Pengfei. Roughness prediction of robot polished surface based on GA-BP neural network [J]. Journal of Jiamusi University (Natural Science Edition), 2022,40(4):76-79

[2] Zhu Ran, Ren Xuanyu, Yao Qi, et al. Alloy yield prediction based on BP neural network [J]. Computer Programming Skills and Maintenance, 2020(4):7-9

[3] Li TG, Chen Y, Zheng W, et al. Alloy yield prediction model based on BP neural network [J]. Shanxi Metallurgy, 2019, 42(3):15-16.

[4] Ai Xiaoli, Wang Yusheng, Tang Wenming. Prediction of blown oxygen quantity in converter steelmaking based on BP neural network [J]. Steelmaking, 2013,29(2):34-37.

[5] Liu J, Li FH, Liu Xiu X. Optimization of GA-BP neural network model and foundation pit deformation prediction [J]. Tunnel Construction (in English and Chinese) , 2021, 41(10):1-7.

[6] He Xiaoyu, Wang Min, Wang Zhongliang et al. Prediction of Mn element alloying yield in converter steel discharge process based on GA-BP neural network[J]. Steelmaking, 2022, 38(4): 14-20.

[7] Worksop, Zhang, Jinqiannan, Li, Shanshan, et al. Exploration of expert review behavior analysis and evaluation method based on Spearman coefficient and multilayer perceptron[J]. Science Bulletin, 2022, 38(5):107-112.

[8] He Xiaoqun. Modern statistical analysis methods and applications. 3rd edition [M]. People's University of China Press, 2012.

[9] Zhang Wentong. Advanced tutorial of SPSS statistical analysis [M]. Beijing: Higher Education Press, 2018.

[10] Cao Xudong. Principles and Methods of Mathematical Modeling [M]. Beijing: Higher Education Press, 2014.

[11] Shi Shoukui. Mathematical Modeling Algorithms and Applications [M]. Beijing: National Defense Industry Press, 2011.

[12] The SPSSAU project (2022). SPSSAU. (Version 22.0) [Online Application Software]. Retrieved from https://www.spssau.com.

[13] Gan Fuxi. The Origin of Ancient Chinese Glass - A Study of the Earliest Ancient Glass in China [J]. Science of China, 2007, 37(3): 382-391