Glass Classification and Identification based on K-means++ Clustering Analysis and BP Neural Network Method

Guanfu Cai *, Haojun Li, Jingming Huang

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, 510006, China

* Corresponding author email: 2632066077@qq.com

Abstract. The weathering caused by the burial environment will cause the ancient glass to exchange material with the environment, and the weathering layer produced by the material exchange will have an impact on the class judgment of ancient glass. In this paper, we analyze the classification rules of high potassium glass and lead-barium glass; select the appropriate chemical composition for subclass division and establish the classification model of glass. After this, a binary classification (high potassium/lead-barium) model is established by BP neural network training, and the subclassified main components of each type of glass derived from the classification model are used as adjustment variables to analyze the chemical composition of unknown categories of glass artifacts and identify the types to which they belong by controlling a single variable.

Keywords: K-means++ Clustering Analysis; BP Neural Network; Glass Artifacts.

1. Introduction

Glass, precious physical evidence of trade between ancient Chinese and Western cultures through the Silk Road. Quartz sand is the main raw material for glass making, the main component is silicon dioxide (SiO2) [1]. Pure quartz sand refining needs to add flux to reduce the melting temperature [2-3], the ancient often natural bubble soda, grass ash, saltpeter and lead ore as a flux, while the melting can be converted into calcium oxide (CaO) limestone as a stabilizer, the different fluxes also led to the different main chemical composition of glass. China's invention of glass is lead ore as a flux of lead barium glass, lead oxide (PbO), barium oxide (BaO) content is high [4-5], the glass of the Chu culture is a representative of it [6]. In contrast, potassium glass with high potassium content, which is popular in the Lingnan region, South Asia and Southeast Asia, is buried in an environment that causes weathering, and the weathering layer produced by the exchange of materials will have an impact on the classification of ancient glass [7-8]. In this paper, we use the data related to ancient glass products to establish a model for the identification and classification of glass [9-10].

2. Cluster Analysis Model

2.1 Glass Classification Modeling

In order to classify the subclasses among the existing categories of glass, the unweathered sampling points were selected for analysis and the data of the corresponding classified element oxides were excluded, i.e., the data of potassium oxide (K2O) content were removed when classifying the subclasses of high potassium glass; the data of lead oxide (PbO) and barium oxide (BaO) were removed when classifying the subclasses of lead-barium glass, and the content of other chemical components were subsequently normalized.

The K-means++ clustering analysis is optimized compared to the K-means algorithm in terms of the initialization of K cluster centers, and in terms of choosing the initial cluster centers, the K-means++ algorithm is based on the principle of ensuring that the initial cluster centers are as far away from each other as possible. In this point K-means++ can optimize the shortcomings of K-means algorithm which is sensitive to the initial value and more sensitive to the isolated point data, and has better statistical significance. the specific steps of K-means++ algorithm are as follows.

1) Randomly select a sample as the first cluster center.

2) Calculate the shortest Euclidean distance between each sample and the currently existing cluster center, the larger the value, the more likely it is to be selected as the cluster center; subsequently, select the next cluster center using the roulette wheel method (based on the size of the likelihood to draw).

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

3) Repeat the previous step until k cluster centers are selected, and after the initial points are selected, calculate the distances of the remaining data objects to the k initial cluster centers, and assign the data objects to the clusters where the nearest centers are located.

4) Adjust the new class and recalculate the new cluster centers.

5) Repeat step 3) and 4) to see if the cluster centers converge, if they converge or reach the number of iterations to stop the cycle and end the run.

We use the elbow method to select a suitable k value for the model. The sample partition of the model will become finer as the number of clusters k increases, and the degree of aggregation of each cluster will increase as the number of clusters k increases, while the SSE will gradually become smaller. When k is smaller than the true number of clusters, the decrease of SSE will be large because the increase of k will significantly increase the degree of aggregation of each cluster, and when k reaches the true number of clusters, the return of the degree of aggregation obtained by increasing k will become smaller rapidly, so the decrease of SSE will decrease abruptly and then level off as the value of k continues to increase, which leads to the graph of the relationship between SSE and k forming an elbow shape. The value of k-value at the elbow is the true number of clusters, and SSE is the core index of the elbow method.

$$SSE = \sum_{i=1}^{k}\sum_{p\in C_1}|p - m_i|^2 \quad (2)$$

Ci represents the i-th cluster, p represents the sample points in Ci, mi is the center of mass of Ci i.e., the mean of all samples in Ci, and SSE is the clustering error of all samples, which represents the clustering effect.

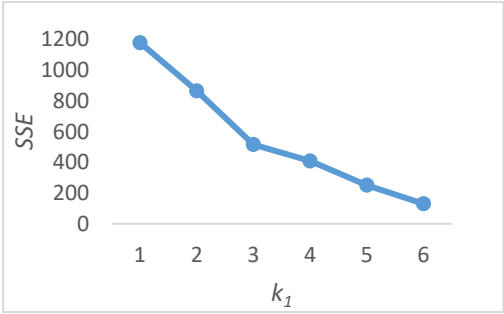

Figure 1. Relationship between k1 value and SSE for high potassium

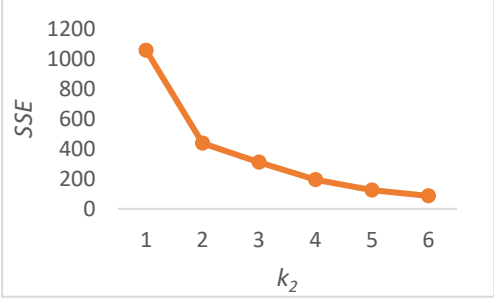

Figure 2. Relationship between k2 value and SSE for high potassium glass

We set the range of values of k1 and k2 for both high potassium glass and lead-barium glass as [1,6], and plotted the SSE versus k values for the two types of glass using SPSS software using the elbow method, respectively. The results are shown in the figure1 and 2.

From the figure, it can be concluded that the true clustering number of high potassium glass k1 =3, then the subclasses corresponding to high potassium glass should be divided into three categories; the true clustering number of lead-barium glass k2 =2, then the subclasses corresponding to lead-barium glass should be divided into two categories.

The clustering is as follows table 1.

Table 1. Number of cases in clusters

High Potassium Glass			**Lead barium glass**		
Clusters	1	7	Clusters	1	9
	2	2			
	3	3		2	4
Valid		12	Valid		13
Missing		0	Missing		0

The final clustering centers is as table 2.

Table 2. Final clustering centers

High Potassium Glass				**Barium lead glass**		
	Clustering			Clustering		
	1	2	3		1	2
SiO2	75.18	91.14	69.48	SiO2	96.25	73.41
CaO	7.85	1.06	5.37	CuO	0.23	19.44
Al2O3	7.34	3.88	10.13			

From the above, it can be seen that the high potassium glass classification category is the best for three-time clustering, and the lead-barium glass classification category is the best for three-time clustering. The differences in subcategories of high potassium glass are reflected in silica content, calcium oxide content, and alumina content, and the differences in subcategories of lead barium glass are reflected in silica content and copper oxide content, and the different categories of the two categories of glass clustering are now analyzed.

Analysis of the different categories of high-potassium glass.

(1) the first category of high potassium glass silica content is much lower than the second category, similar to the third category; calcium content is the highest; aluminum content is also the highest. Therefore, this type of high potassium glass can be classified as low silica, high calcium and high aluminum subcategory.

(2) the second type of high potassium glass silica content is much higher than the first and third type, the highest; calcium content is much lower than the rest of the class for the lowest; aluminum content is also much lower than the rest of the type for the lowest. Therefore, this type of high potassium glass can be classified as high silica low calcium low aluminum subclass.

(3) The silica content of the first type of high potassium glass is much lower than the second type, which is similar to the first type; the calcium content is close to the average; the aluminum content is also close to the average. Therefore, this type of high potassium glass can be classified as low silicon in calcium in aluminum subcategory.

Analysis of the different categories of lead barium glass.

(1) the first category of lead barium glass silica content is much higher than the second category as the highest; copper oxide content is much lower than the second category as the lowest. Therefore, this type of lead barium glass can be classified as high silica and low copper subcategory.

(2) The second type of lead barium glass has the lowest silica content compared with the first type, and the highest copper oxide content compared with the second type. Therefore, this type of lead barium glass can be classified as low silicon and high copper subcategory.

According to the results of the analysis, the subclasses of high potassium glass and lead-barium glass can be divided into the following figure 3:

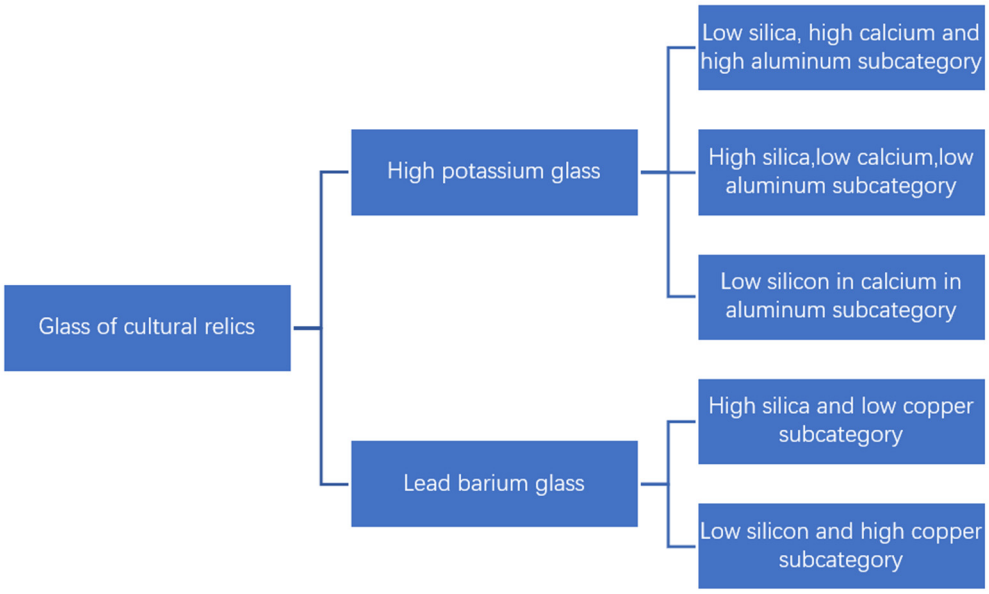

Figure 3. Sub-class division of glass artifacts

2.2 Glass Classification Model Stability Check

In the classification of the subclasses of the two types of glass, the final clustering centers of the two types of glass were obtained using SPSS software, and the silica content, calcium oxide content, and alumina content in the clustering center of the second type of high potassium glass were selected, and the silica content and copper oxide content in the clustering center of the first type of lead-barium glass introduced errors, and the chemical components that did not introduce errors were normalized using SPSS software to obtain the true number of clusters of the two types of glass using the elbow method (k1, k2) are listed in the following table 3.

Table 3. Introduction of error values

errors	k1	k2
15%	3	2
10%	3	2
5%	3	2
-5%	3	2
-10%	3	2
-15%	3	2

From Table 3, it can be seen that after the introduction of -15%-15% error, the true clustering numbers (k1, k2) of both types of glass do not deviate, and the change rate of each numbered artifact type is 0%, so the stability of the model can be proved.

3. Glass Identification Model Building

3.1 BP Neural Network Model

BP (Back Propagation) neural network is a multilayer feedforward network trained by error back propagation algorithm and is one of the most widely used neural network models. BP network can learn and store a large number of input-output pattern mapping relations without revealing the mathematical equations describing such mapping relations beforehand. Its learning rule is to use the fastest descent method to continuously adjust the weights and thresholds of the network by back propagation to minimize the sum of squared errors of the network. The BP neural network model topology consists of an input layer, hide layer and output layer, As shown in the figure 4.

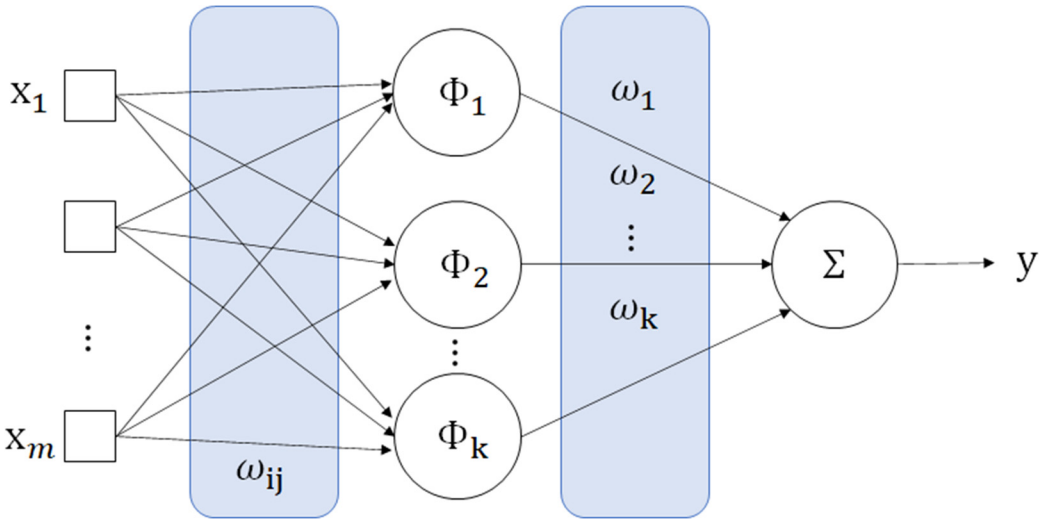

Figure 4. BP neural network topology

The algorithm of BP neural network has good ability of adaptive and classification recognition, etc. The neural network can be viewed as a function mapping for which there is a clear correspondence between input and output. The type prediction for the unknown category of glass artifacts in Annex Form 3 can be viewed as a complex function mapping problem with each chemical composition as input and artifact type as output.

By pre-processing the data, the training set of BP neural network is expanded to 67 groups, because the training set data is relatively small, and the repeated use of training set data for validation easily leads to overfitting of the validation results, therefore, in order to expand the validation, set data, the validation set is validated using the chemical composition of weathered artifacts before weathering predicted by the model establishment of problem 1.

Vanilla gradient descent, also known as batch gradient descent, computes the gradient of the loss function with respect to the parameter θ over the entire training data set. With less training data, batch gradient descent uses all the training data in each update, ensuring that each training set data has a weighting effect on the final training result and making the training move iteratively toward the minimum.

The three-layer BP network has m input nodes (the total number of chemical compositions of the given artifacts), c output nodes (high potassium/lead-barium types), and r hidden nodes

Since the chemical composition content of unweathered culture is more stable, which can make the neural network reach convergence faster and therefore more meaningful for classification, the activation function from the input layer to the hidden layer uses a modified Sigmoid function for artifacts with different weathering degrees, and introduces a weight α to map the different inputs to [0,1]. The improved Sigmoid function is as follows.

$$\hat{v}_j = \frac{1}{1+e^{-\alpha \cdot v_j}} (j = 1,2,\dots,r\,; i = 1,2,\dots,m) \tag{3}$$

When the input artifacts are unweathered, α=1, when the input artifacts are weathered, α=0.6, when the input artifacts are severely weathered, α=0.2, where v_j is the output of the implied layer and v^_j is the actual implied layer output after activation by the activation function.

The activation function from the implied layer to the output layer uses the linear function Purelin function, and the relationship between the network input and output can be obtained as follows:

$$\hat{y}_k = \sum_{j=1}^{r} v_j \cdot f[\sum_{i=1}^{m} w_{ij} \cdot P_i + \theta_j] \tag{4}$$

Among them $k = 1,2,3,\ldots N$, w_{ij} is the connection weight, θ_j is the threshold value, y_k is the desired output, $\hat{y}_k$ is the actual output of the network, $m = 14$(Total chemical composition of the given artifacts), $c = 2$(high potassium/lead-barium type), $r = 10\,(\lfloor (m+c) \cdot 2/3 \rfloor)$

The parameters in the BP neural network prediction model are listed in table 4:

Table 4. Parameters in the BP neural network prediction model

Regularization term weight	Motivator factor	Optimizer type	Number of hidden layer layers
0.00001	0.8	lbfgs	10

The predicted and actual values of the BP neural network are shown in the following figure:

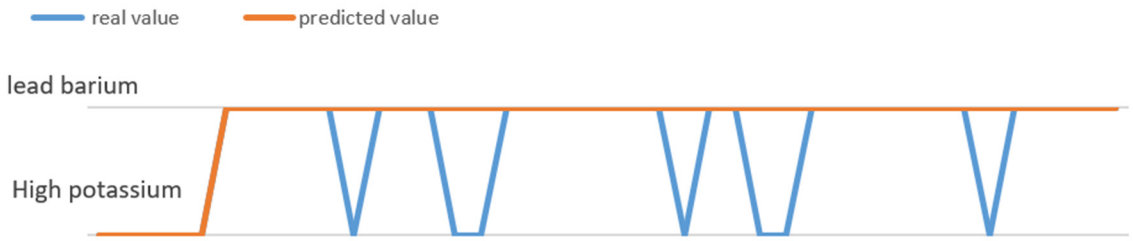

Figure 5. BP neural network real value vs. predicted value line graph

From Figure 5, it can be obtained that the prediction of the network has a good fit and the accuracy of the prediction can reach 85.71%. The higher accuracy also confirms the high reliability and reasonableness of the chemical composition content of weathered artifacts before weathering inferred from the model of problem one.

3.2 Model Solution

The data were used as input to the BP neural network to derive the predicted classification probability results for each unknown artifact glass as follows Table 5:

Table 5. Results of glass classification of unknown artifacts

Artifact number	Prediction type
A1	High potassium
A2	lead barium
A3	lead barium
A4	lead barium
A5	Lead barium
A6	High Potassium
A7	high potassium
A8	lead barium

The artifacts with high potassium and lead-barium were subclassified in the previous subsection to derive the main chemical composition content of each type of artifact, and the predictive analysis was repeated using BP neural network in turn by using the single control variable method by adjusting the percentage of the main chemical composition of the artifacts to be tested (downward by 20%), after re-normalizing all the chemical compositions, and the results are shown in the following table:

Table 6. Single control variable method classification prediction table

Artifact Number	Downward adjustment of elements			
	Silicon dioxide	Aluminum oxide	Calcium oxide	Copper oxide
	Prediction Type			
A1	High potassium	High potassium	High potassium	High potassium
A2	lead barium	lead barium	lead barium	lead barium
A3	lead barium	lead barium	lead barium	lead barium
A4	lead barium	lead barium	lead barium	lead barium
A5	Lead barium	Lead barium	Lead barium	Lead barium
A6	High Potassium	High Potassium	High Potassium	High Potassium
A7	high potassium	high potassium	high potassium	high potassium
A8	lead barium	lead barium	lead barium	lead barium

From the data in Table 6, it can be obtained that the original prediction results did not change after the percentage adjustment of the main chemical components, i.e., the BP neural network model can pass the sensitivity test.

4. Conclusion

In this paper, we subclassify two types of artifact glass, high potassium and lead-barium, to establish a model that portrays all glass artifacts, and we subclassify artifact glass based on the idea that "things cluster together" and the composition of each glass artifact. The data from the unweathered sampling sites were selected, and the data corresponding to the content of oxides were excluded, followed by the normalization of the content of other chemical components. K-means++ cluster analysis using SPSS software was used to produce line graphs of the relationship between k-values and SSE for the two types of glass artifacts, and the subclasses were divided according to the principle of elbow method setting k1=3 for high potassium glass and k2=2 for lead-barium glass with the content of significant chemical components to obtain five subclasses: high potassium-low silica-high calcium-high aluminum, high potassium-high silica-low calcium-low aluminum, high potassium-low silica-medium calcium-medium aluminum, lead-barium- High silicon and low copper, lead and barium - low silicon and high copper. Finally, an error of -15%~15% was introduced for the content of significant chemical components to see if the subclass division changed, and it was calculated that the subclass division did not change and the change rate of each sampling type was 0%, which shows the stability of the model. After the establishment of the classification model, this paper identifies the unknown glass artifacts, and uses BP neural network training to establish a dichotomous classification (high potassium/lead-barium) model, and uses the predicted chemical composition content of glass artifacts before weathering to fit the neural network for validation and expand the neural network subject to training data. An error of -20% was introduced for the chemical composition of the unknown category of glass artifacts, and the prediction was repeated using the network model, with 0% change in the predicted results, proving that the model has stability.

References

[1] Wu B., Yu D. Deep convolutional neural network-based classification and detection of cell phone glass cover surface defects[J]. Software Engineering,2021,24(12):6-10. DOI:10.19644/j.cnki.issn2096-1472. 2021. 012.002.

[2] Ren Ru. A study of dictionary learning in classification of defects in glass fiber cloth [D]. Xi'an University of Engineering, 2019. DOI: 10.27390/d.cnki.gxbfc.2019.000047.

[3] Lu Yue. Deep learning-based classification and detection of cell phone glass defects[D]. Zhengzhou University,2019.

[4] Xue Yuan. Research on glass defect classification and recognition based on machine vision[D]. Hefei University of Technology,2018.

[5] Li Q., Zhou B. Research on neural network classification model of substrate glass defects based on cloud computing[J]. Computer and Digital Engineering,2017,45(07):1373-1376.

[6] Dai Qun. Research on hybrid neural network based on ICBP model with diverse integration methods[D]. Nanjing University of Aeronautics and Astronautics, 2009.

[7] Chen, Tianbao. Research on the classification algorithm of float glass defects based on BP network[D]. Huazhong University of Science and Technology, 2007.

[8] Ai Jieyan. Research on color classification and defect detection of wall and floor tiles based on computer vision [D]. South China University of Technology, 2003.

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

[10] Liu C. X., Deng P. F., Pan S. H., Luo W. D., Zhang S. L., Li T. M. Classification and research status of fireproof glass [J]. Guangzhou Chemical Industry,2021,49(15):16-18.