Construction of glass composition analysis and identification model based on K-means algorithm optimized by genetic algorithm

Chenyu Zhang [1,*,#], Yunjian Wu [1,#], Mengna Zhou [2,#]

[1] DUniversity of Electronic Science and Technology of China, School of Electronic Science and Engineering, Chengdu, China

[2.]University of Electronic Science and Technology of China, School of Optoelectronic Science and Engineering,Chengdu, China

* Corresponding Author Email: 2927492849@qq.com

[#]These authors contributed equally.

Abstract. Ancient glass is highly susceptible to weathering by the environment in which it is buried. During the weathering process the proportion of its composition changes, thus affecting the correct judgment of its category. In this paper, the two categories of high potassium and lead-barium glass were subclassified separately, and firstly, the elbow method was used to combine the SC contour coefficients to obtain their optimal number of clusters as 3,7 respectively; after that, the initial point selection was set as the decision variable, the interclass distance and the minimum as the fitness function, and the initial population size was set as 50, the clustering distance as Euclidean distance and other parameters, and the genetic algorithm was used to optimize the K-means clustering to achieve the subclassification delineation. Taking lead barium as an example, after optimization, CHI is increased by 4, DBI is decreased by 0.3, contour coefficient is increased by 0.2, and the clustering effect becomes better. Finally, the Kappa consistency test was performed, and the Kappa coefficient was 0.822 indicating the reasonableness of the clustering results; by changing the selected feature vector dimensions for sensitivity analysis, the Kappa coefficients were all above 0.8, which were not sensitive to the dimensionality of the feature vectors. The model effectively implements the problem of subclassifying different classes of glass for subclassification.

Keywords: Glass; Subclass classification; Elbow method; Genetic algorithm; Kmeans clustering

1. Introduction

Glass is a scarce item in the cultural exchange between China and the West in ancient times. After absorbing the glass-making technology from the West, China began to make its own glass, which is somewhat similar in appearance, but somewhat different in chemical composition.

Ancient glass is highly susceptible to weathering by the environment in which it is buried. In the process of weathering its composition ratio changes, thus affecting the correct judgment of its category. A batch of data related to ancient glass products in China is available, and workers have classified them into two types: high potassium glass and lead-barium glass.

Analyze the classification laws of high potassium glass and lead-barium glass; select the appropriate chemical composition for subclassification of the two categories of glass respectively according to the relevant knowledge, specific division methods and results need to be given, and analyze the reasonableness and sensitivity of the results.

The mathematical model is based on the official form 1 of the 2022 Mathematical Modeling C problem. First visualize the Form 1 data combined with the statistical law above can be derived from the classification law; for the sub-classification problem, using the variance sorting chemical components, the greater the variance, the better the classification effect, select the top 6 as clustering features, and then determine the best k-value best clustering center after clustering respectively, the results of accuracy sensitivity verification, the selection of the top 6 rational verification.

2. Analysis of the classification law of high potassium glass and lead-barium glass

2.1. Analysis of the classification law of high potassium glass and lead-barium glass

In contrast to the above statistical patterns based on type analysis of the presence or absence of weathered chemical content on the surface of artifact samples, this section focuses on how to distinguish between high potassium glass and lead-barium glass.

Analysis of Form 1 in the Annex makes the following parallel coordinate plots as shown in Figure 1.

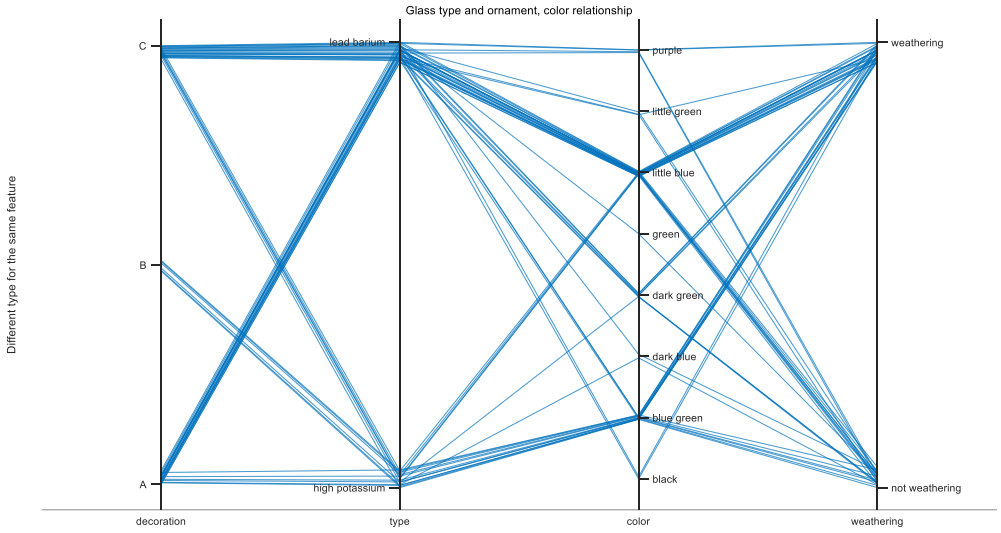


Figure 1. Relationship between glass type and ornamentation, color

Analyzing the above parallel plots, it is concluded that: the type of decoration B must belong to high potassium glass; the type of color is black, purple, green, light green must belong to lead barium glass.

Analyzing Form 2 in the Annex, the difference between the average values of the chemical composition of unweathered high potassium and unweathered lead-barium and the difference between the average values of the chemical composition of weathered high potassium and weathered lead-barium are shown in Figure 2.

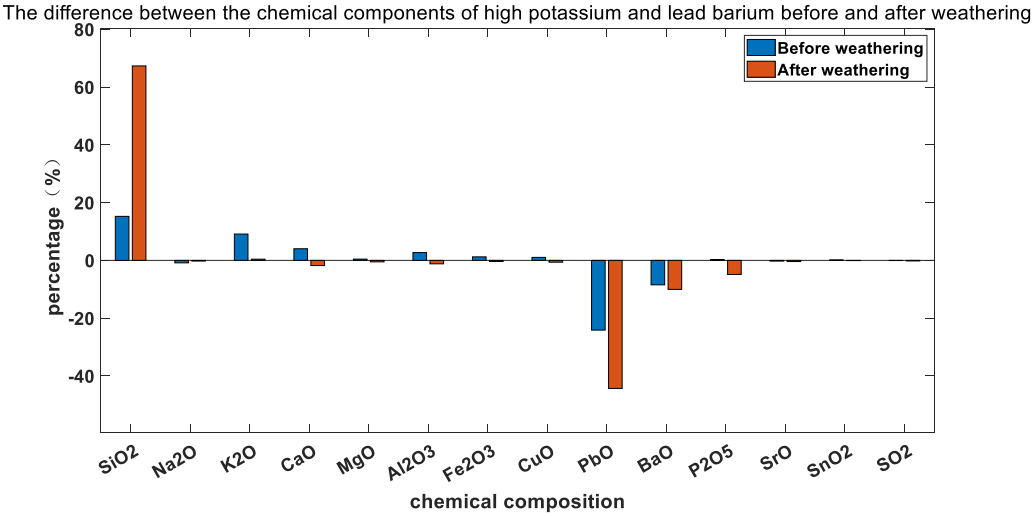


Figure 2. Difference between high potassium and lead-barium chemical composition before and after weathering

Analysis of Figure 2 combined with the radar chart made above, it is concluded that: SiO2 content above 90% must be high potassium glass, SiO2 content below 50% must be lead-barium glass; PbO content above 10% or about 10% must be lead-barium glass; BaO content above 5% or about 5% is generally lead-barium glass; SO2 content of about 15% must be lead-barium glass.

2.2. Selection of chemical composition

To classify the subcategories of high potassium glass and lead-barium glass separately, we first select some of the 14 chemical components as feature vectors.

Considering high potassium glass and lead-barium glass separately, the 14 chemical components were normalized by column to max-min, and the normalization formula was shown in Equation 1; then the variance was found for each column of the normalized data (Equation 1), and the different columns were arranged in descending order by variance; the first six indicators were selected.

The reason for the descending order of variance is that a larger variance represents a greater degree of variation in this chemical component; conversely, it represents a smaller degree of variation in this chemical component. It is better to select the index with the highest variance as the feature vector for clustering.

The clustering effect will be better if the indicators with the largest possible variance are selected as the feature vectors. The reasons for selecting the first six indicators are explained in section 2.6.2.

$$s^2 = \frac{1}{n} \times \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad (1)$$

n represents the number of rows in each column.

The top 6 chemical compositions selected for the sorted high-potassium glass and lead-barium glass are shown in Table 1.

Table 1. Chemical composition table of the first 6 positions of the two types of glass

Top 6 chemical composition (high potassium)	Top 6 chemical composition (lead barium)
Silicon Dioxide	Phosphorus pentoxide
Calcium oxide	Strontium oxide
Potassium oxide	Silicon Dioxide
Strontium oxide	Calcium oxide
Magnesium oxide	Lead oxide
Sulfur Dioxide	Copper oxide

2.3. Limitations of original K-means clustering

The flow chart of the K-means clustering algorithm is shown in Figure 3[4]; in the original K-means clustering, there are two major problems, one is the selection of the initial point, and the other is the number of classes of clusters, both of which are not properly selected will directly affect the clustering results, so this paper selects the elbow method combined with the SC contour coefficient using the original K-means clustering to determine the best number of clusters, and uses genetic algorithm to optimize the selection of the initial The location of the center point is optimized, and then the clustering is carried out for high potassium glass and lead-barium glass respectively[5]. The reason for choosing genetic algorithm to optimize k-means instead of k-means++ is that according to the literature [1] it is known that genetic algorithm optimizes k-means with about 10% higher accuracy than k-means++ [6].

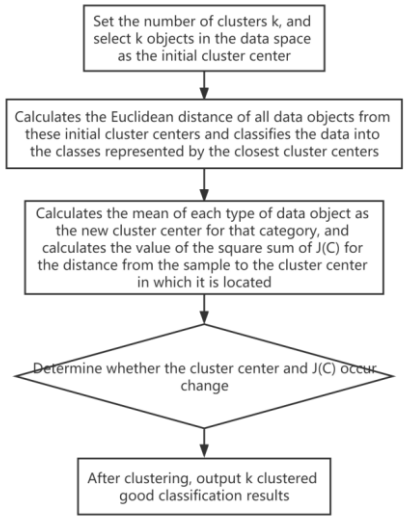


Figure 3. Flowchart of K-means clustering algorithm

2.4. Determining the optimal number of clusters using the elbow method combined with SC contour coefficients

2.4.1 Elbow method to determine the elbow point

The core indicator of the elbow method is the SSE [2].

$$SSE = \sum_{i=1}^{k} \sum_{p \in Ci} | p - m_i |^2 \tag{2}$$

where Ci is the ith cluster, p is the sample points in Ci, mi is the center of mass of Ci (the mean of all samples in Ci), and SSE is the clustering error of all samples, which represents how well the clustering works.

As the number of clusters k increases, the sample division will be finer, the degree of aggregation of each cluster will increase, and the error squared and SSE will become smaller. When k is less than the optimal number of clusters, the decrease in SSE will be large because the increase in k will substantially increase the degree of aggregation of each cluster, and when k is greater than or equal to the true number of clusters, the return on the degree of aggregation obtained by increasing k again will rapidly become smaller, so the decrease in SSE will plummet and level off as the value of k continues to increase The graph of the relationship between SSE and k is the elbow shape, and the value of k corresponding to the elbow is the optimal true clustering number.

The distance standard Euclidean distance was chosen and the initial points were randomly selected and repeated 100 times with a maximum number of generations of 50 for each iteration. the SSE values of high potassium glass and lead barium at a clustering number of 1:10 were calculated using the original K-means [3], as shown in Figures 4 and 5.

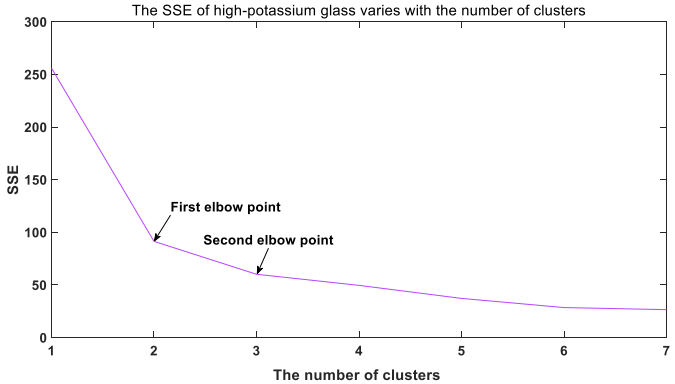


Figure 4. High Potassium Glass SSE Variation Chart

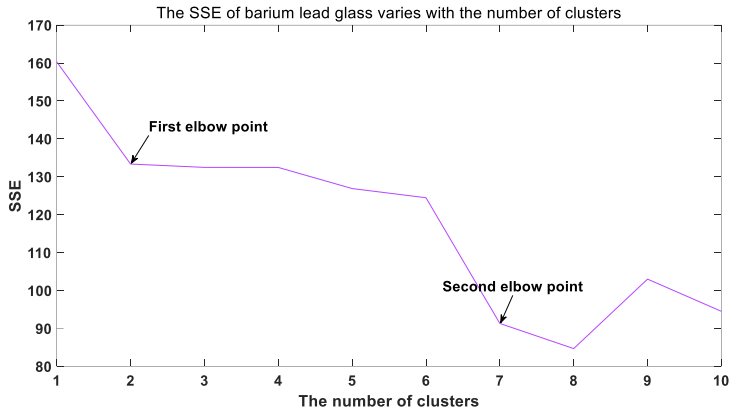


Figure 5. Lead barium glass SSE variation chart

2.4.2 SC profile factor

The contour coefficients take values in the range of [-1,1], and the closer the value is to 1, the better the clustering performance is, and conversely, the closer the value is to -1, the worse the clustering performance is. The contour coefficient s for a given sample is

$$s = \frac{b-a}{sum(a,b)} \tag{3}$$

a denotes the average distance of a sample from other samples in its cluster, b denotes the average distance of a sample from other cluster samples

The total contour coefficient SC of clustering is:

$$SC = \frac{\sum_{i=1}^{N} si}{N} \tag{4}$$

The distance standard Euclidean distance was chosen and the initial points were randomly selected and repeated 100 times with a maximum number of generations of 50 for each iteration. the SC profile coefficient values for high potassium glass and lead barium at a clustering number of 2:10 were calculated using the original K-means, as shown in Figure 6.

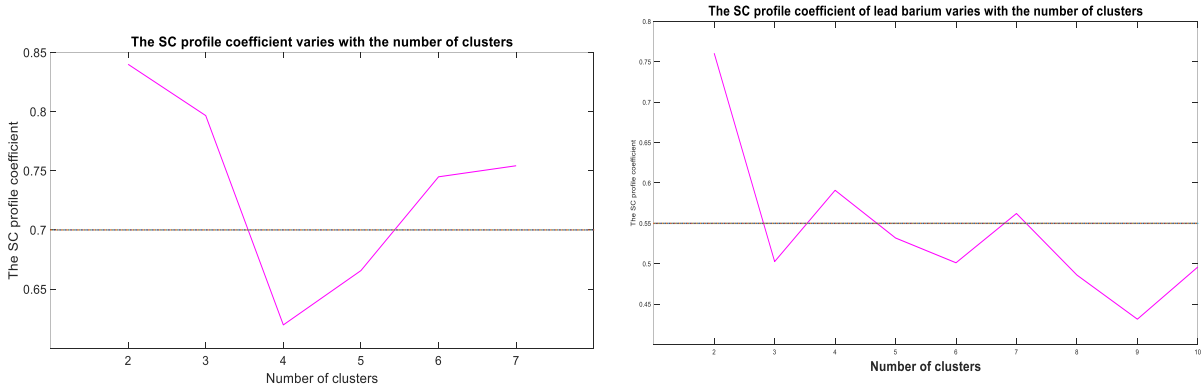


Figure 6. Plot of sc profile coefficients for two types of glass

2.4.3 Determining the optimal number of clusters for high potassium and lead barium

Combining the high potassium elbow diagram and SC contour coefficients, it is reasonable to classify high potassium into 2 or 3 categories, but considering that high potassium itself is weathered and unweathered, the optimal number of high potassium categories is 3.

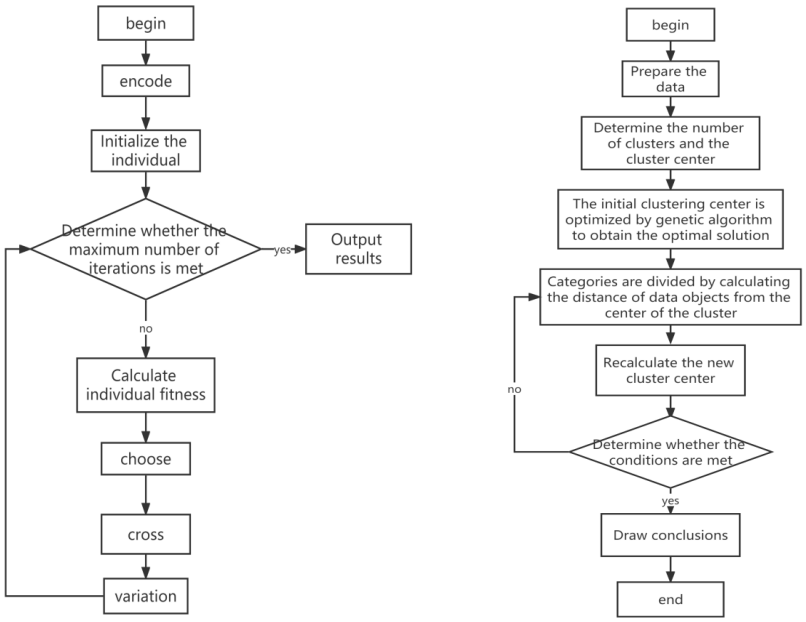
Combined with the elbow diagram of lead and barium and the SC contour coefficient, it is reasonable to classify lead and barium into 2 categories or 7 categories, but considering that lead and barium are weathered and unweathered as well as severely differentiated, the best classification number for high potassium is 7.

2.5. Optimal selection of initial centroids and classification of high potassium lead barium using genetic algorithm

2.5.1 Principle of genetic algorithm to optimize the initial centroid

The flow chart of the genetic algorithm algorithm and the flow of the genetic algorithm optimized K-means algorithm are shown in Figure 7; the general idea is that after determining the clustering categories, the different clustering centroids are used as the decision variables, the minimum sum of intra-class spacing after each clustering is used as the objective function, the fitness function[7], and the minimum value of the maximum value of each column variable of the data set is used as the upper and lower bound.

The initial population size was selected as 50, the crossover ratio was 0.8, the variation ratio was 0.1, the individuals with better fitness were selected, the maximum number of generations was set to 4000, and the genetic algorithm was used to optimize the solution [8].



(a) (b)

Figure 7. Genetic algorithm flow chart before and after improvement. (a) Original flow chart of genetic algorithm, (b) Genetic algorithm optimization K-means algorithm flow chart

2.5.2 Genetic algorithm optimization results and K-means clustering results after optimization

The variation of the fitness function with the number of iterations is shown in Figure 8: The analysis shows that the distinction between different categories is obvious, and the classification between the same categories is also better, which indicates good clustering [9].

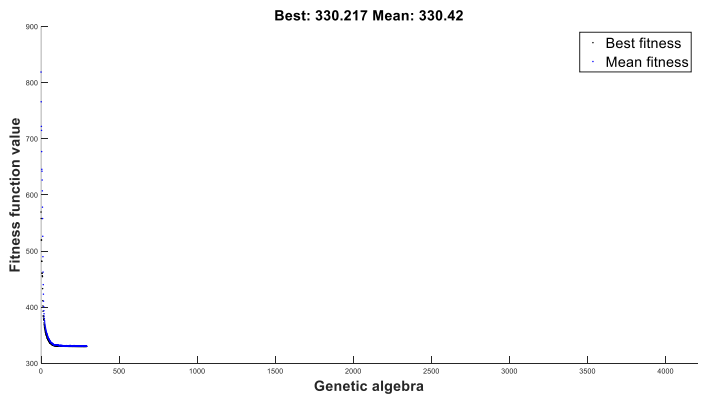


Figure 8. Adaptation function change

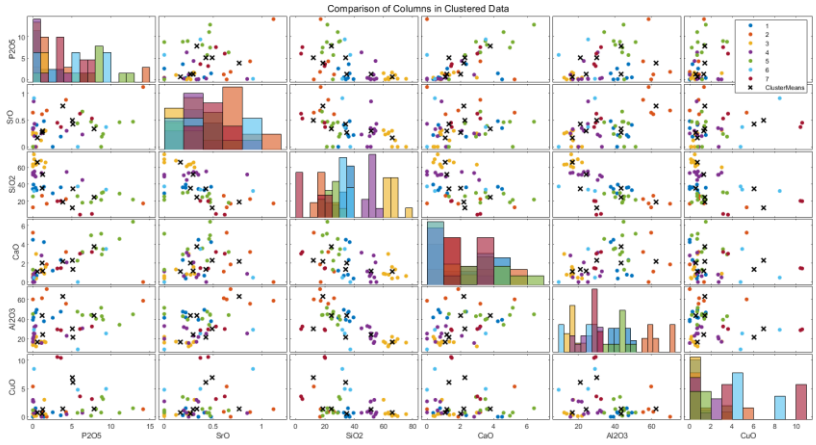


Figure 9. Scatter matrix after genetic algorithm optimization

The scatter matrix after optimization of the genetic algorithm is shown in Figure 9: where the distribution of the same chemical substance on the diagonal are relatively uniform, indicating a better clustering effect.

The results of the genetic algorithm after classification are shown in Figure 10.

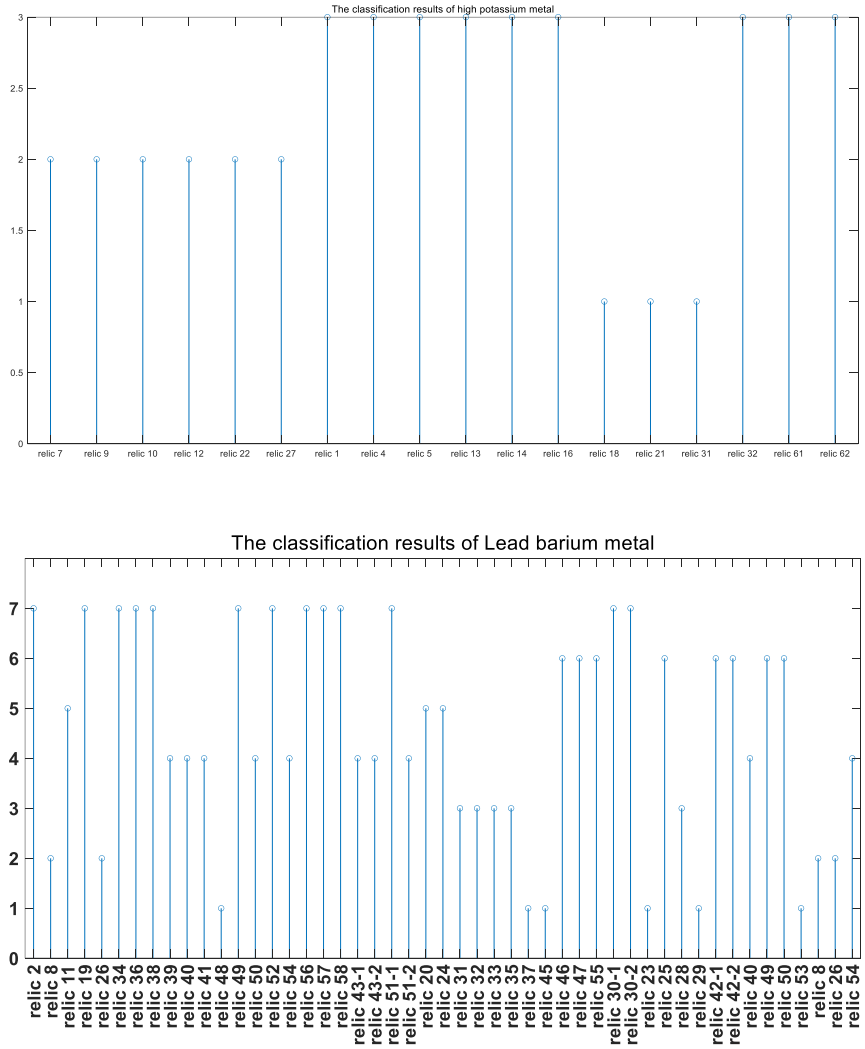


Figure 10. Graph of classification results of two glass genetic algorithms

2.5.3 Validation of pre-optimization and post-optimization effects

After clustering for K-means, the larger CHI index is better, the smaller DBI index is better, and the larger contour coefficient is better.

Post-optimization clustering CHI index: 73.1331 Pre-optimization: 69.8062

Post-optimization clustering DBI index: 0.85063 Pre-optimization: 0.879

Post-optimization clustering contour coefficient: 0.57498 Pre-optimization: 0.56

Accordingly, it can be concluded that the overall classification effect becomes better after optimizing the initial centroid of K-means using genetic algorithm.

2.6. Kappa consistency validates classification rationality, sensitivity

Kappa coefficient, which can be used to measure the effectiveness of the classification. The kappa coefficient based on the confusion matrix is calculated as follows.

$$kappa = \frac{p0-pe}{1-pe} \tag{5}$$

$$p_0 = (TP + TN)/N \tag{6}$$

$$p_e = \frac{(TP+FP)(TP+FN)+(FP+FN)(FP+TN)}{N^2} \tag{7}$$

where FP: false positive cases, FN: false negative cases, TP: true cases, TN: true negative cases, and N is the number of samples.

2.6.1 Kappa Consistency Validates Classification Rationality

The confusion matrix is established for lead barium glass as an example. Table 2 shows the statistics based on the frequency of different subclasses belonging to unweathered, weathered or severely weathered, and then the clustering results are used to determine whether different test points belong to the weathering category; since the specific test points belong to unweathered, weathered or severely weathered are known, Table 3 establishes the confusion matrix based on the judged weathering category and the real weathering category. The Kappa coefficients of the above clustering results were calculated to be 0.822 for lead barium and Kappa ≥ 0.75, indicating that a fairly satisfactory degree of agreement has been achieved. It indicates that the classification is reasonable.

Table 2. Confusion Matrix 1

Subclass	Weathering	Unweathered	Severe weathering	
1	1	5	0	Unweathered
2	2	0	2	Severe weathering
3	0	5	0	Unweathered
4	9	1	0	Weathering
5	1	2	0	Weathering
6	0	8	0	Unweathered
7	11	2	0	Weathering

Table 3. Confusion Matrix 2

Detection points	Weathering	Unweathered	Severe weathering
Weathering	22	3	2
Unweathered	2	20	0
Severe weathering	2	0	2

2.6.2 Kappa Consistency Verification Classification Sensitivity

In this paper, the top six chemical indicators are selected as the feature vectors for classification, and the top 7,8,9,10,11,12 chemical indicators are selected to calculate the Kappa as shown in Figure 11: As can be seen from the figure, the change of Kappa coefficient is above 0.8 when different feature vector dimensions are selected, which shows that the results are not sensitive to the dimensionality of the feature vectors.

However, the efficiency of the algorithm with different feature vector dimensions (the running time will be different) is shown in Figure 12[10]: As the figure shows, the running time increases sharply with the increase of the feature vector dimensions, but the classification accuracy and the classification results do not change much, so the top 6 chemical components are selected as the feature vectors to ensure the running efficiency of the algorithm and the accuracy of the results at the same time.

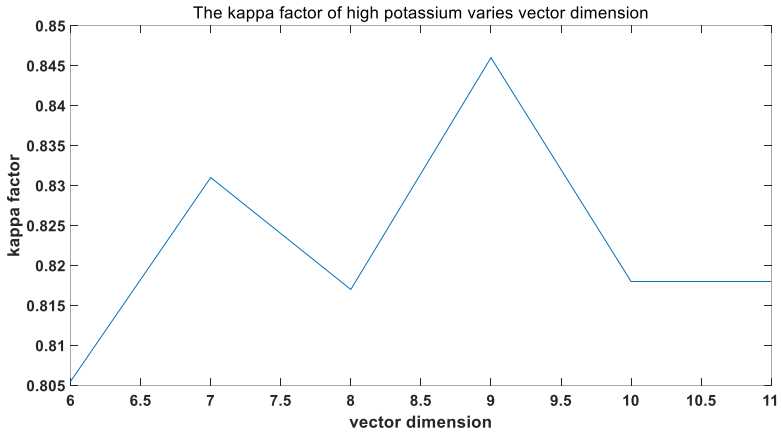


Figure 11. Eigenvectors of the first 6 indicators of lead-barium glass

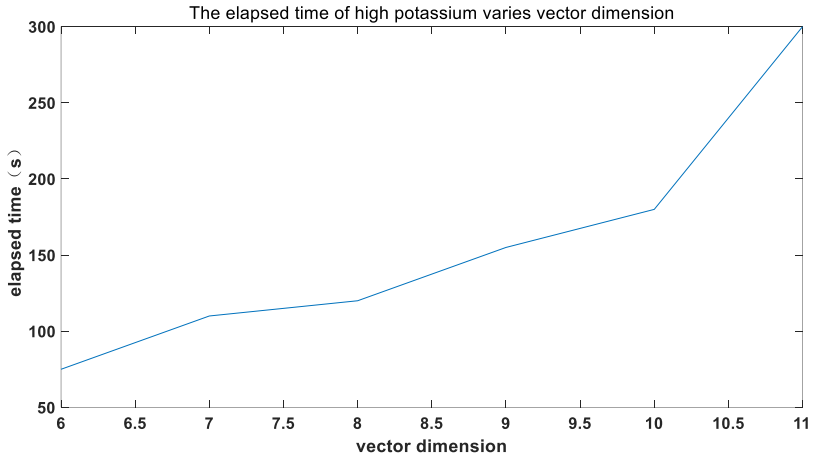


Figure 12. Efficiency of the algorithm with different eigenvector dimensions

3. Conclusions

In this study, the kmeans clustering k values are cross-validated using elbow method and contour coefficient method. Genetic algorithm is used to overcome the effect of instability of kmeans clustering cluster centers. However, there are also some shortcomings, such as the dependence of the model on the data, and the machine learning model has a strong dependence on the data to ensure the absolute accuracy of the data.

In future research, the kmeans clustering k-values and centroids will be optimized, which can significantly improve the clustering effect. Based on the improved kmeans can be applied to many clustering scenarios such as classifying risk levels and so on. Meanwhile the index isotropy optimization gray correlation analysis model can be applied to a wider range of data.

References

[1] SUN Hongyan. K-means algorithm for optimizing initial clustering centers by genetic algorithm. Audio Engineering,2019,43(11):32-33+47. DOI: 10.16311/j.audioe.2019.11.011.

[2] WANG Yingji, DONG Hongbin. Determination of the number of classes based on density peak and elbow method [J]. Applied Science and Technology, 2021,48(02):74-79.

[3] Sammouda Rachid, El Zaart Ali. An Optimized Approach for Prostate Image Segmentation Using K-Means Clustering Algorithm with Elbow Method [J]. Computational Intelligence and Neuroscience, 2021,2021.

[4] Yan Li, Lin Cheng, Chenhao Zhang, Hanqing Zhao. Analysis of Spatiotemporal Characteristics of Online Car Hailing Based on k-means Clustering[C]//. Proceedings of the World Transport Conference 2022 (WTC2022) (Traffic Engineering and Air Transport Chapter), 2022: 14-23. DOI: 10.26914/c.cnkihy.2022.019760..

[5] Zhu Donglin, Xie Linpeng, Zhou Changjun. K-Means Segmentation of Underwater Image Based on Improved Manta Ray Algorithm[J]. Computational Intelligence and Neuroscience,2022,2022.

[6] Chan Jan Y. K., Leung Alex Po, Xie Yunbo. Efficient High-Dimensional Kernel k-Means++ with Random Projection[J]. Applied Sciences,2021,11(15).

[7] Hongjiang Ma, Xiangbing Zhou. A GPS location data clustering approach based on a niche genetic algorithm and hybrid K-means[J]. Intelligent Data Analysis,2019,23(S1).

[8] Ankita Sinha, Prasanta K. Jana. A hybrid MapReduce-based k-means clustering using genetic algorithm for distributed datasets[J]. The Journal of Supercomputing,2018,74(4).

[9] Tessy Badriyah. Hybrid Modeling KMeans – Genetic Algorithms in the Health Care Data[J]. Emitter: International Journal of Engineering Technology,2013,2(1).

[10] David K. Hale, Byungkyu Brian Park, Aleksandar Stevanovic, Peng Su, Jiaqi Ma. Optimality versus run time for isolated signalized intersections[J]. Transportation Research Part C,2015,55.