Classify Glassy Antiques Based on the Chemical Components

Shangwen Li *

Jinan University – University of Birmingham Joint Institute at Jinan University, Guangzhou, China

* Corresponding Author Email: sxl198@student.bham.ac.uk

Abstract. As one sort of cultural relic, glassy antiques can effectively convey the historical information of a certain era and reveal the cultural exchanges in different regions as a symbol of foreign trade. However, due to long-term weathering and corrosion when buried in soil, the shape, color, and chemical components of a glassy antique can change considerably, and hence identifying it and recognizing its category is particularly difficult. Clustering is a popular technique of data analysis and data mining. K-means is one of the most popular data mining algorithms, as it is simple, scalable, and easy to modify in different contexts and fields of application. This paper uses k-means to find the clustering center showing the characteristics of the chemical components of different categories of glassy antiques. Then the particle swarm optimization algorithm (PSO) that offers a globalized detailed search methodology is utilized to improve the k-means clustering. The new result compared to the previous one of traditional k-means clusterin shows better classification capacity. Finally, it compares the results of k-means with that of PSO-k-means and analysis the advantages and disadvantages of PSO-k-means.

Keywords: K-means clustering; particle swarm optimization algorithm; glassy antique classification.

1. Introduction

Through the combination of computer-aided technology and porcelain cultural relics protection technology, cultural relics restoration will achieve a leap from digitalization to intelligence and high-speed. Computer aided technology provides a new solution for the efficient and accurate restoration of cultural relics. With the high speed and parallelization of computer computing power, the development of deep convolution neural network is further accelerated. Computer vision tasks such as image classification and image semantic segmentation based on the powerful automatic feature extraction ability of deep convolution neural network have achieved significant performance improvement. In image classification, the convolution neural network avoids the complex feature extraction process by taking the original image as the input, and then effectively learning the corresponding features from a large number of samples, and can extract more abstract and refined features, thus achieving better image classification performance. Therefore, no matter how to effectively classify ceramic cultural relics fragments or cultural relics fragments, convolution neural network can avoid the shortcomings such as the loss of feature information easily caused by manual feature extraction and the poor classification effect, and help cultural relics restoration workers to improve the accuracy and efficiency of the restoration of Qin Terracotta and ceramic cultural relics fragments, which has become a feasible and worthy of in-depth research. With the rapid development of virtual reality, augmented reality and other technologies, virtual digital museum is an application scenario with great development potential.

2. Classification of Ancient Glassy antiques by PSO-k-means

2.1. Processing the Data

Before any analysis, the original data need some preprocessing. Following is raw data supplied by the China Society for Industrial and Applied mathematics. As shown in Table 1. It includes 77 sampling points from 66 pieces of glassy antiques with respective percentages of 14 sorts of chemical components contained as well as the category and the weathering status. Glassy antiques here are divided into two categories: high-potassium glass & lead-barium glass. On account of the small

volume of data, here consider the sampling points on the same glassy antique but in different positions as taken from different antiques, as to richen training sample. For the same reason, 10% of the data is used as the test set, and the remaining data is used as the training set. Here the last 8 samples are selected as the test set. In the following part of this paper, they will be used to test the clustering result. And the remaining 69 pieces of data will be used as the training set of the model.

Table 1. Raw data parameters

K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	PbO
BaO	P2O5	SrO	SnO2	SO2	category	weathering

Before applying the clustering method, the raw sample data of training set needs to be properly preprocessed. ①As the China Society for Industrial and Applied mathematics, the data supplier suggests, delete samples whose total chemical components proportion is lower than 85% and higher than 105%. ② Delete variables with massively missing data: SnO_2 (89.86% data missed) and SO_2 (88.41% data missed). ③ Fill in 0 for the remaining missing data.There are still 67 samples left in total, including 18 high-potassium glassy samples, of which 12 are weathered and 6 are unweathered, and 49 lead-barium glassy samples, of which 13 are weathered and 36 are unweathered.

The data of the percentage of chemical components hold the constant-sum property that for a single sampling point, all the numbers of each component should sum up to 100%. Hence there can be multi-colinearity between different components thus increasing the cost of explaining the model results. Additionally, limitations of the production formula can involve some correlations between components [1]. For example, to control the pH value during the whole production process, KOH and NaOH are widely used which have different influences on the production formula of different sorts of glassy antiques [2]. Hence PCA is used here to cancel the correlation of components. The result is as follows. As shown in Table 2.

Among them, the cumulative contribution rate of the first seven principal components reaches 89.63%. Hence in the following part of this paper, the original variables will be replaced with these seven variables. After conversion, the data is as follows. As shown in Table 3.

Table 2. Partial PCA processing

principal component	eigenvalue	contribution	Accumulated contribution
r1	3.8625	32.1876	32.1876
r2	2.4083	20.0691	52.2567

Table 3. Partial after conversion

Samples	r1	r2	r3	r4	r5	r6	r7
01	33.35463	-16.7142	2.081832	-8.87109	7.470024	10.93096	8.450387
02	-2.55222	2.321177	-14.5419	-3.36204	-6.1809	-4.69827	4.261953

2.2. Use the k-means to find clustering centers

In this part, k-means will be used to find the clustering center [3,4]. Then the result will be compared with the real data and the clustering center will be used to predict the test set to see if it works well.

2.2.1 Apply k-means and compare the result with real data

First, the elbow method is utilized to find an appropriate k to use. Here input the data expressed by the principle components obtained from the previous part, and the output is as the following Fig.1. It can be seen that a clear change of gradient appears when k=4. Therefore, k=4 will be selected as the number of clusters to apply the k-means. And the result is as follows Table 4. Compared with the real data, it can be found that this result doesn't tell the status of weathering well that in each class

there is a mixture of weathered samples and unweathered samples. However, it classifies the category better that class 1 & class 3 consist completely of lead-barium samples and class 4 only contains high-potassium samples. But in class 2, these two kinds of samples account for almost half each. To an extent, it reflects that weathering impacts the proportion of chemical components as materially so that some characteristics of the category are obscured.

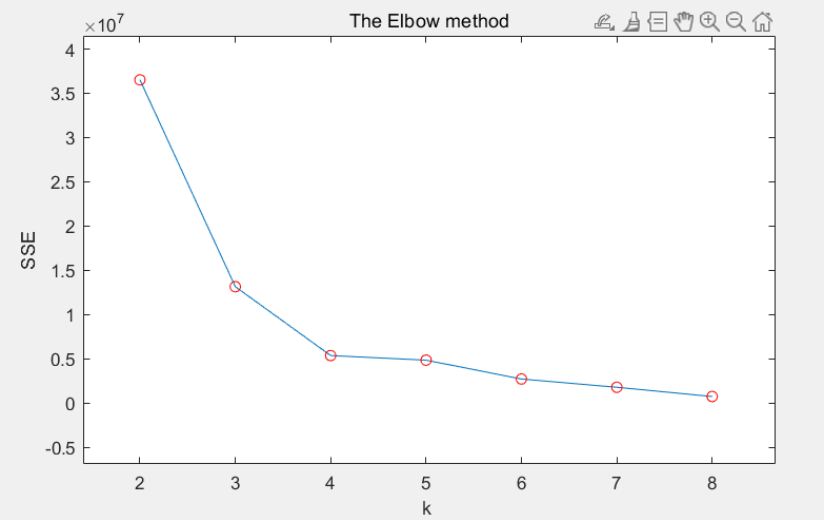

Fig 1. SSE

Table 4. k-means result

class 1	2	10	11	20	27	28	32 60	37 62	41 63	42 65	43 66	44 67	47	48	54	56	58	59
class 2	1	4	5	6	7	8	16	17	18 53	30 61	31	34	35	36	38	40	49	50
class 3			14	21	24	25	26	33	39	45	46	51	52	55	57	64		
class 4						3	9	12	13	15	19	22	23	29				

2.2.2 Predict the test set and compare the real data

From the previous part, 4 clustering centers are obtained, which will be used to classify the test set. First, express the test set by the principal components. Next, match each sample to the nearest clustering center and the following result can be obtained. It implies that according to the clustering result, all the samples in the test set are lead-barium glassy antiques. Except for the 5th one, the rest are predicted correctly. Although this clustering result doesn't classify the training set well, it has good prediction ability.

2.3. Use the PSO-k-means to find clustering centers

In this part, PSO is introduced to improve the clustering accuracy of k-means [5,6].

2.3.1 Improve k-means with PSO

The first thing to do is the transformation of the problem type. The clustering problem can be transformed into an optimization problem: find k center points to minimize the Euclidean distance from each sampling point to the center point in each cluster range:

$$\min \quad F(D^K, \mu) = \sum_{k=1}^{K} \sum_{i=1}^{nk} \text{dist}\left(d_i^{(k)}, \mu_k\right) \tag{1}$$

Where $D^K = \{d^{k1}, d^{k2}\ldots, d^{kn}\}$: It represents the data set to be divided into k classes. $C = \{c_1\ldots, c_k\}$: It represents the center set of clustering. Dist(a, b): It calculates the Euclidean distance between a& b. Specific steps of PSO-k-means: 1) Initialize a group of particles, including random position and velocity; 2) Evaluate the fitness of each particle, where fitness here is how well the clustering center

minimizes the sum of Euclidean distance from itself to the points in its class; 3) For each particle, compare its fitness value with the best position it has been in, and if it is better, take it as the current best position; 4) For each particle, compare its fitness value with the best position record the group hold, and if it is better, take it as the current best position; 5) Move particles and update the velocity; 6) Determine whether the stop signal is reached (reach the optimization threshold, or reach the upper limit of iterations). Particle position update:

$$\begin{cases} v_{id} = w \cdot v_{id} + c_1 \cdot r_1 \cdot (p_{id} - x_{id}) + c_2 \cdot r_2 \cdot (p_{gd} - x_{id}) \\ x_{id} = x_{id} + v_{id} \end{cases} \tag{2}$$

$$\begin{cases} x_i = [x_{i1}, x_{i2} \dots x_{id}] \\ v_i = [v_{i1}, v_{i2} \dots v_{id}] \end{cases} \tag{3}$$

Where v_{id}: The velocity of a particle, x_{id} is the position of this particle. w: The inertia controls the residual proportion of the velocity in the last movement. r1&r2: A random number between (0,1). p_{id}: The best position this particle has been in. p_{gd}: The best position of the group this particle belongs to. c_1&c_2: Controls to what extent the new velocity learns from the best position record of this particle& group's best position record. Specific steps of particle position update: Step 1: Initialize two parameters: position and speed. They determine the motion state of a particle; Step 2: The result (Euclidean distance) of each search is the particle fitness, and then record the individual historical optimal position and the group's historical optimal position; Step 3: The historical optimal positions of an individual and the group are equivalent to two forces, which jointly affect the motion state of a particle in combination with the inertia of the particle itself. The particle swarm optimization algorithm and k-means algorithm are combined to find better clustering centers that minimize the Euclidean distance as much as possible [6,7].

2.3.2 Apply PSO-k-means and predict the test set

After improvement, PSO-k-means is applied to the data again to find better clustering centers. The stop threshold is set at 10e-3. After many adjustments of parameters to fit in the volume characteristics of sample data, a version that can output a stable result is found and the result is as follows Fig 2.

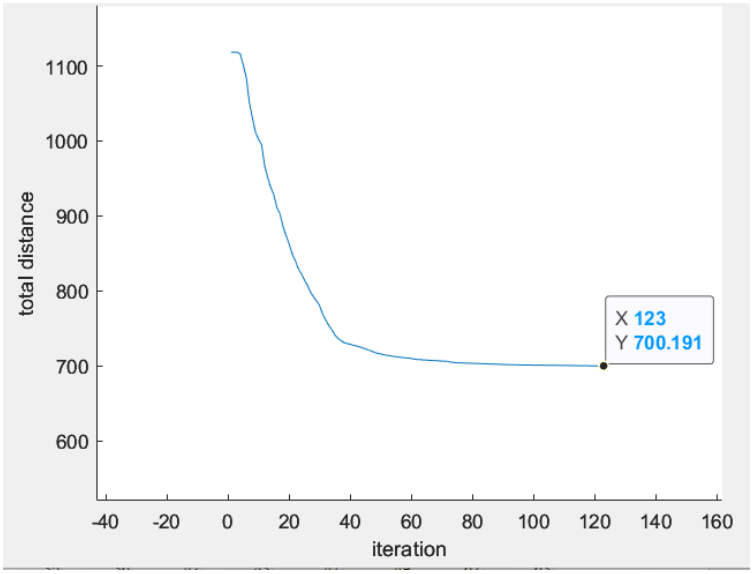

Fig 2. Output a stable result

It is a considerably beautiful result because class 2 is completely the same as the sample group of high-potassium glassy antiques and the rest classes contain all samples in the sample group of lead-barium glassy antiques, which implies that this clustering result recognizes the category of samples in the training set completely correctly. And the clustering result also implicates the possibility of sub-classification within lead-barium glassy antiques [8]. Then when using this result of clustering centers to predict the test set. All samples are correctly predicted.

2.4. Compare the results of k-means and PSO-k-means

From the clustering results, the clustering results of PSO-k-means are significantly better than that of k-means. PSO-k-means correctly recognizes the category of all samples in the training set, while k-means creates a bad class in which samples of two categories almost account for half. From the perspective of prediction ability, PSO-k-means and k-means do not show obvious differences. The results of both are almost completely correct, except that one of the eight test samples cannot be accurately recognized due to the poor performance of class 2 of k-means. However, it can also be the result of the small volume of the test set and if the volume of the test set increases, some differences in prediction results may appear [9,10]. In addition, although PSO-k-means shows stronger clustering ability, as an improved k-means algorithm based on an optimization algorithm, it may still fall into the problem of local optimization that requires to take measures to prevent. Its parameters need to be adjusted according to the volume characteristics of the input sample and the result is not very stable.

3. Conclusion

This paper focuses on the classification of glassy antiques by the composition ratio of chemical components, Firstly, it uses principal component analysis to cancel the constant-sum property of this kind of data that shows the proportion of each component and break the multi-colinearity that may exist between different components to decrease the cost of explaining the results by subsequent models. Then the elbow method helps to find out that k=4 should be used in the next part of k-means. But the clustering result of k-means is not good that it classifies all samples into 4 groups, among which the second one is a serious mixture of different categories. It also affects the result of subsequent category prediction of the test set that the category of one sample cannot be determined directly. Hence particle swarm optimization is utilized to improve k-means. And the result of PSO-k-means is much better than k-means. It clusters the data clearly that 4 groups are all free of the mixture and one group only contains one category. As a consequence, the prediction result is also good that all samples in the test set are correctly predicted. However, the clustering result also implicates the possibility of sub-classification. It can be seen that in the final result, high-potassium samples all stay in one group which means that the chemical composition of high-potassium samples are all around the same center, while lead-barium samples are divided into 3 different groups, which implies that these 3 groups of lead-barium samples have a characteristic each in chemical composition. Therefore, this can be used as a follow-up research direction and there is something about different manufacturing processes and the use of materials in different regions and countries.

References

[1] Gan, FX. The Road of Glass and Jade -- Also on the cultural and technical exchanges between China and foreign countries on the silicate cultural relics of the pre-Qin Dynasty (English) [J]. Journal of Silicate, 2013,41 (04): 458-466

[2] Li, M. Preparation of Feons products and research on lead barium glass during the Warring States, Qin and Han Dynasties [D]. Beijing University of Chemical Technology, 2014

[3] Liu, Z., Di, X., Wang, Q., & Wang, L. (2022). Composition analysis of glass relics based on a clustering model. Highlights in Science, Engineering and Technology, 21, 246–253.

[4] Yin, H. (2022). Identification model of surface composition of glass relics based on K-means clustering and boosting learning strategy. Highlights in Science, Engineering and Technology, 22, 340–346.

[5] Fan Yixuan, Kan Xiu, Cao Le, Shen Jie. Application of improved PSO-K-means algorithm in vehicle driving cycle estimation [J]. Intelligent Computer and Application, 2021,11 (07): 80-85+90

[6] Jin, J., Zhong, M., Lin, X., Tian, CH. A New Cluster Analysis Based on Combinatorial Particle Swarm Optimization Algorithm. Proceedings of the 2016 International Conference on Education, Management, Computer and Society. Advances in Computer Science Research, 2016, 476-479

[7] Taher Niknam, Babak Amiri, an efficient hybrid approach based on PSO, ACO and k-means for cluster analysis, Applied Soft Computing, Volume 10, Issue 1,2010, Pages 183-197

[8] Liu Song, Lv Liangbo, Li Qinghui, Xiong Zhaoming. Glass beads unearthed from Lingnan Han Tombs and Sino-foreign exchanges on the Maritime Silk Road of the Han Dynasty [J]. Cultural Relics Protection and Archaeological Sciences, 2019, 31 (04): 18-29.

[9] Shakhatreh M, Shakhatreh H, Ababneh A. Efficient 3D Positioning of UAVs and User Association Based on Hybrid PSO-K-Means Clustering Algorithm in Future Wireless Networks[J]. Mobile Information Systems, 2023, 2023.

[10] Sadeghi M, Dehkordi M N, Barekatain B, et al. Correction to: Improve customer churn prediction through the proposed PCA-PSO-K means algorithm in the communication industry[J]. The Journal of Supercomputing, 2023: 1-1.