An Approach to The Classification of Ancient Glassware Based on K-Means Clustering Models

Jiebin Huang*, Jinghao Tang

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

*Corresponding author: ghooup123@126.com

Abstract: The study of ancient glassware can be very helpful by building a sound research model. In this paper, we first counted the data onto different glass categories for their chemical composition and quantified the patterns using the composition variability. Then, seven factor sets were selected based on data saturation and a K-Means clustering model were built to complete the classification of their sub-classifications. Finally, a test of the rationality and sensitivity of the model was completed by feeding the clustering model with noise-added glass component data.

Keywords: K-Means Clustering Model, Subclass Classification, Noise, Sensitivity

1. Introduction

In studies related to ancient glass, we learn how archaeologists classify glass according to its composition, and on the basis of this, work on subclassification of glass [1-3]. This requires us to consider many factors, such as, chemical composition, category, sensitivity, rationality, etc. The question focuses on mathematical modelling and statistical analysis to build a relevant model to complete the work of subclassification of samples and validation of the results.

In this paper, the data of high potassium glass and lead-barium glass are statistically analysed according to the basic classification information of glass artefacts and the corresponding chemical composition ratios, and the classification rules for them are derived, and on this basis, the relevant conditions are combined to build a model to select the appropriate factors to complete the sub-classification classification, and the analysis results are analysed by conducting sensitivity analysis.

2. Establishment and Solution of Model

2.1 Classification rule

In this paper, we first calculate the mean values of the components of the two types of glass by comparing the mean values and standard deviations of the components of the two types of glass, and we can see that the mean values of silica (SiO_2) components in high potassium glass are generally larger than those in lead-barium glass, and the lead oxide (PbO) and barium oxide (BaO) components in high potassium glass are much smaller than those in lead-barium glass. The PbO and BaO compositions of high potassium glass are much smaller than those of lead-barium glass. The differences are evident from the following graphical comparison.

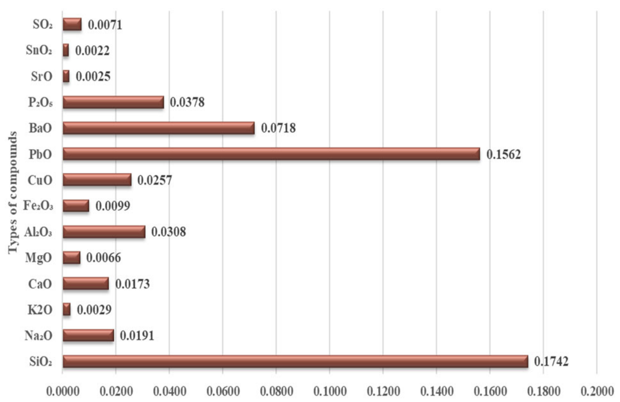


Figure 1 Mean value of change in chemical composition of lead barium glass

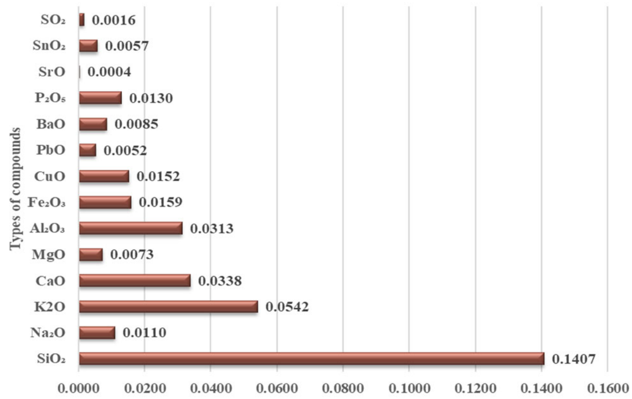


Figure 2 Mean value of change in chemical composition of high potassium glass

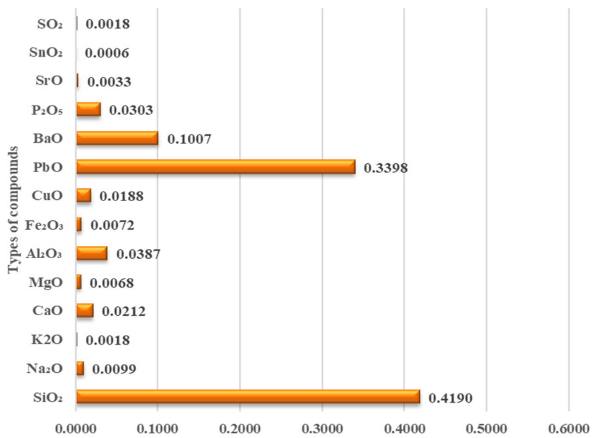


Figure 3 Standard deviation of the variation in the chemical composition of lead barium glass

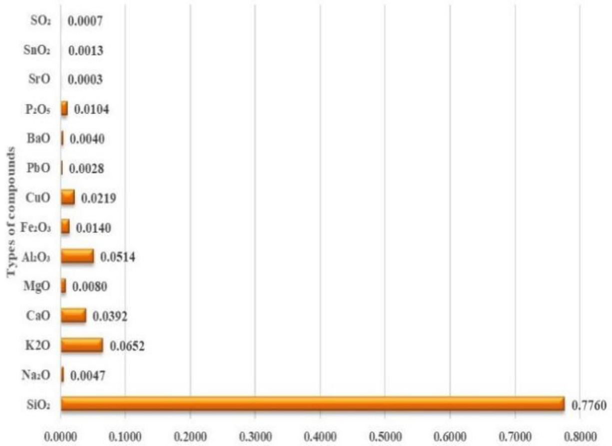


Figure 4 Standard deviation of chemical composition variation of high potassium glass

Using the obvious differences in these three components, we can further divide the two categories of glass.

2.2 Subclassification

2.2.1 Data processing

In order to further sub-classify high potassium glass and lead barium glass separately, the following table of seven components in each category was selected by excluding individual components with low amounts of data based on the saturation of the data, respectively.

Table 1 Types of the seven chemical components screened

Glass categories	Types of chemical components screened						
High potassium glass	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3
Lead barium glass	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3

Since these seven components contain too large a dimension, which is not conducive to the visual image display of the clustering division [4], this paper decided to further filter out the three components with the best aggregation effect from the seven components for K-means cluster analysis [5-9], and make a scatter plot of the components after cluster analysis to select the subcategories of each type of glass by cluster analysis.

2.2.2 Establish K-means clustering analysis model

The ultimate goal of the K-means method [10] is to converge the individual sample data into a given cluster of k classes. The algorithm steps are as follows.

Step1:

Initialize the k class cluster centers first using K-means.

$$\{Center_1, Center_2, Center_3, ..., Center_k\} \quad 1 < k < n \tag{1}$$

Step2:

Calculate the Euclidean distance to the center of the class cluster for each sample data.

$$dis(M_i, Center_j) = \sqrt{\sum_{t=1}^{n}(M_{it} - Center_{jt})^2} \tag{2}$$

where denotes the i-th sample data $Center_j$ denotes the j-th class cluster, M_{it} denotes the t-th dimensional attribute of the i-th sample data $Center_{jt}$ denotes the t-th dimensional attribute of the j-th class cluster.

Step3:

Modeling of class clusters by the following equation.

$$Center_t = \frac{\sum_{M_i \in S_t} M_i}{|S_t|} \tag{3}$$

$$1 \le i \le |S_t| \tag{4}$$

where $Center_t$ denotes the t-th class cluster, $|S_t|$ denotes the number of objects in the t-th class cluster, and M_i denotes the i-th sample data in the t-th class cluster.

Step4:

Assigning, by circular iteration, for each data sample, to the central class cluster with the smallest Euclidean distance.

Step5:

For the center of each class cluster, recalculate and update the center value of the class cluster $Center_j$.

After the clustering division of K-means, this paper names and classifies the chemical composition characteristics of the sample data in the clustered regions delineated in the scatter plot and analyses the sensitivity of the model. The specific flow chart is shown in Figure 5 below.

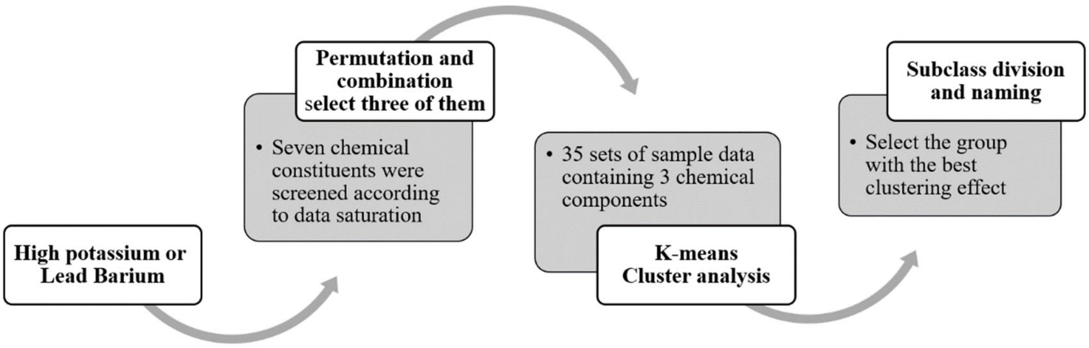


Figure 5 Sensitivity analysis flow chart

2.2.3 Solution of K-means Clustering Analysis Model

The two classes of glass were processed separately: in this paper, the seven chemical composition factors that were initially screened were arranged and combined to obtain a total of seven combinations, and three chemical compositions from each combination were used as input to the K-means model to obtain a total of 35 as output. where is the sum of the distances from all points between classes to the center of the class cluster， the formula is as follows.

$$ds = \sum_{i=1}^{k} d_i \tag{5}$$

where d_i denotes the sum of the distances from the point between the i-th class to the center of the i-th class.

This metric provides a good quantitative representation of the effect of clustering. The results of ranking the size of the clusters are shown in the table below

Table 2 Ranking of the effect of clustering of quantitative indicators (partial data)

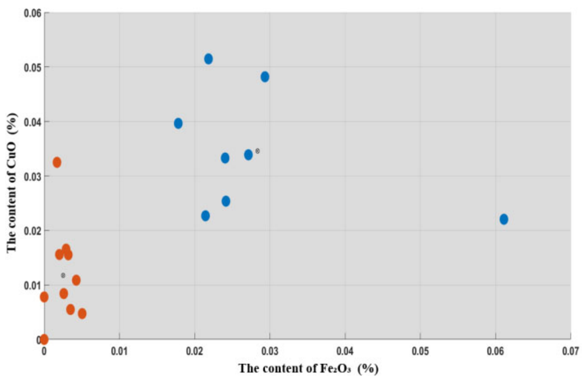
Group	1	2	3	4	5	6	7	8	9	10
Clustering effec	0.0677	0.0669	0.0748	0.0633	0.0626	0.0563	0.0531	0.0536	0.0568	0.0522
Ranking	34	33	35	32	31	29	26	27	30	23

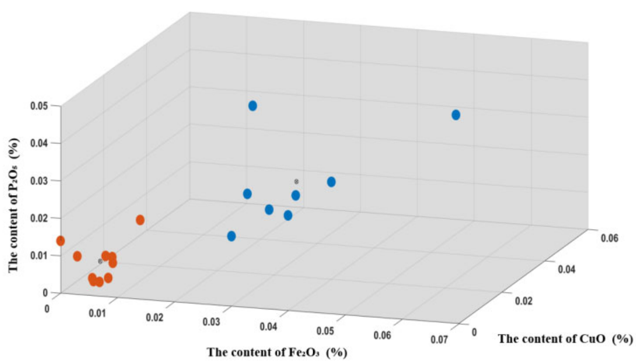
In accordance with the ranking in the table, the combinations with the best clustering results were selected for this paper as shown in the following table.

Table 3 Optimal combination of scientific components for clustering effects

Optimum combination table	Clustering effects optimize the combination of chemical component components		
High potassium glass	Fe_2O_3	CuO	P_2O_5
Lead barium glass	CuO	CaO	Al_2O_3

In the next step the three chemical components in the optimal combination were used as model inputs to obtain the scatter plot of the results after clustering division as shown below.





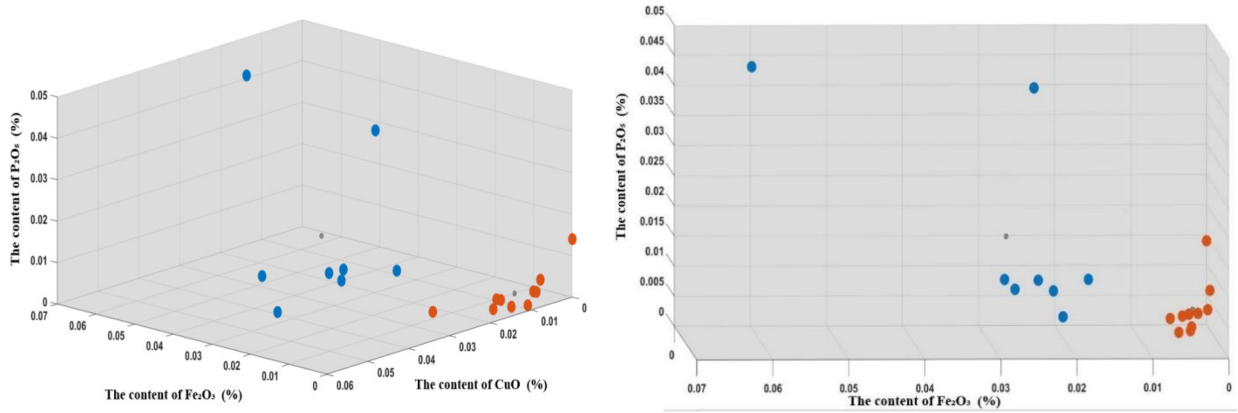


Figure 6 Scatter diagram of subdivision of high potassium glass

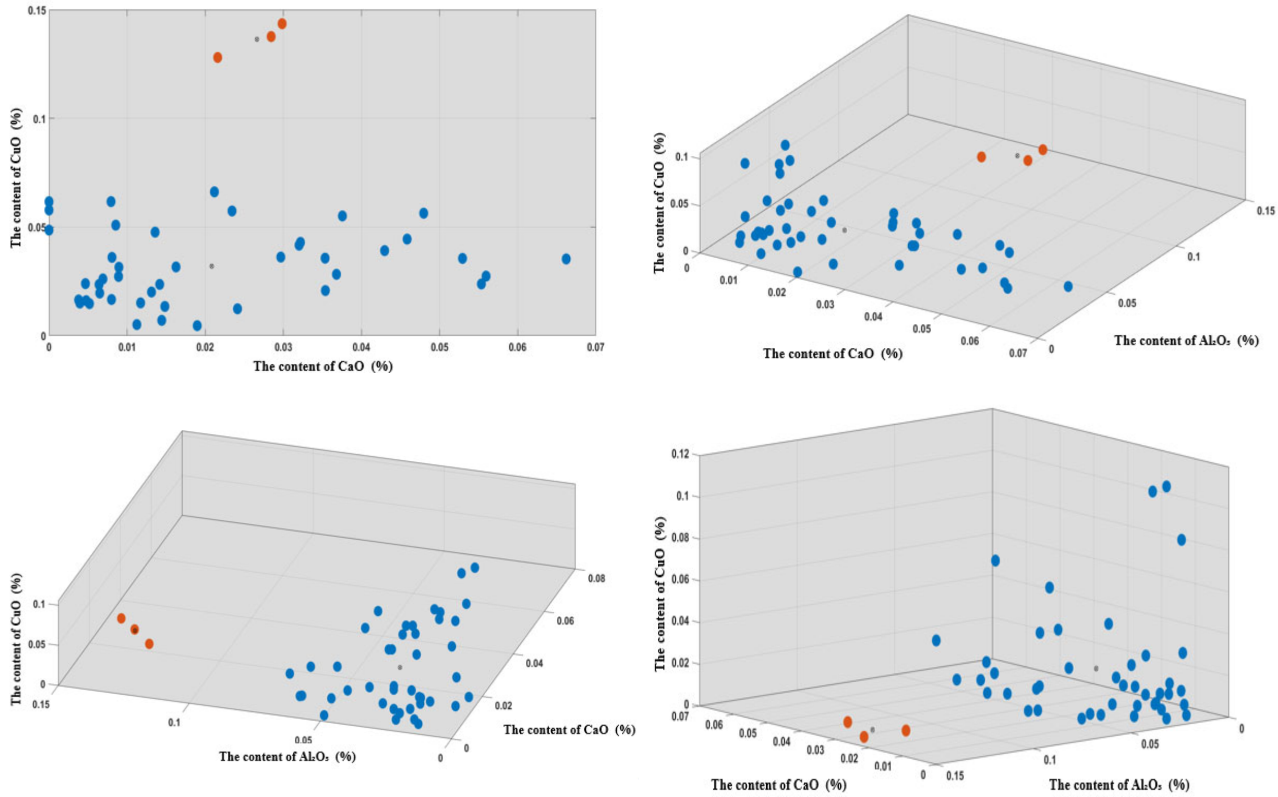


Figure 7 Scatter diagram of subdivision of lead barium glass

And output the corresponding observation tree as follows.

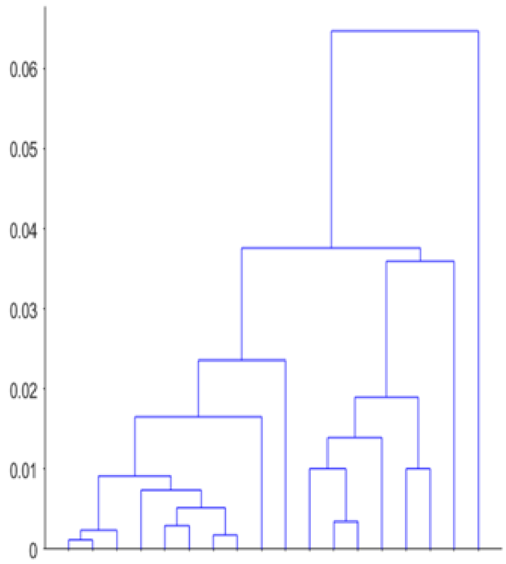


Figure 8 Observation tree for high potassium glass

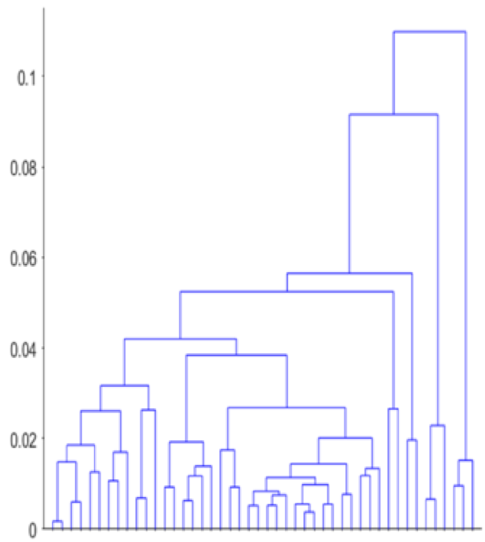


Figure 9 Observation tree for lead barium glass

The scatter plot of the high potassium shows that the glassware samples, shown by the orange dots, have very small amounts of copper oxide and iron oxide. A review of the relevant literature shows that the color of this type of glassware should not appear pale blue and blue-green. Due to the low content of iron and copper, it is not susceptible to a range of electrochemical reactions as well as oxidation reactions. It has been decided to sub-classify this glass and name it "high potassium antioxidant glass".

The scatter diagram of lead and barium shows that the glass samples, shown by the orange dots, have a very high content of alumina. Based on the chemical reaction equation for alumina:

$$4Al + 3O_2 = 2Al_2O_3 \tag{6}$$

we can tentatively conclude that a dense film of alumina forms on the surface of the glass, which is less susceptible to weathering due to the chemical stability of alumina. In this paper, we have decided to sub-classify the glass and name it "lead-barium anti-corrosion glass".

2.3 Models checkout

(1) Reasonableness of the model:

Based on the results of the sub-classification of the various types of glass obtained from the above model, the results of the model are justified on the basis of relevant scientific evidence and literature, and the results of the naming of "high potassium antioxidant glass" and "lead-barium anti-corrosion glass" are derived. The derivation process is scientific and consistent with common sense, and the model is judged to be reasonable.

(2) Sensitivity of the model:

In order to quantify the sensitivity of the model, random noise was added to the sample data as data perturbation, and the noise added sample was used as input to the model, and the volatility of the clustering effect corresponding to different noisy data was calculated by the model with the following formula.

$$\sigma = \frac{|ds_i - ds_0|}{ds_0} \tag{7}$$

Where, ds_i indicates the clustering effect with the i-th noise added and ds_0 indicates the clustering effect without the noise added.

The following graphs are drawn from the clustering effect volatility versus different noise.

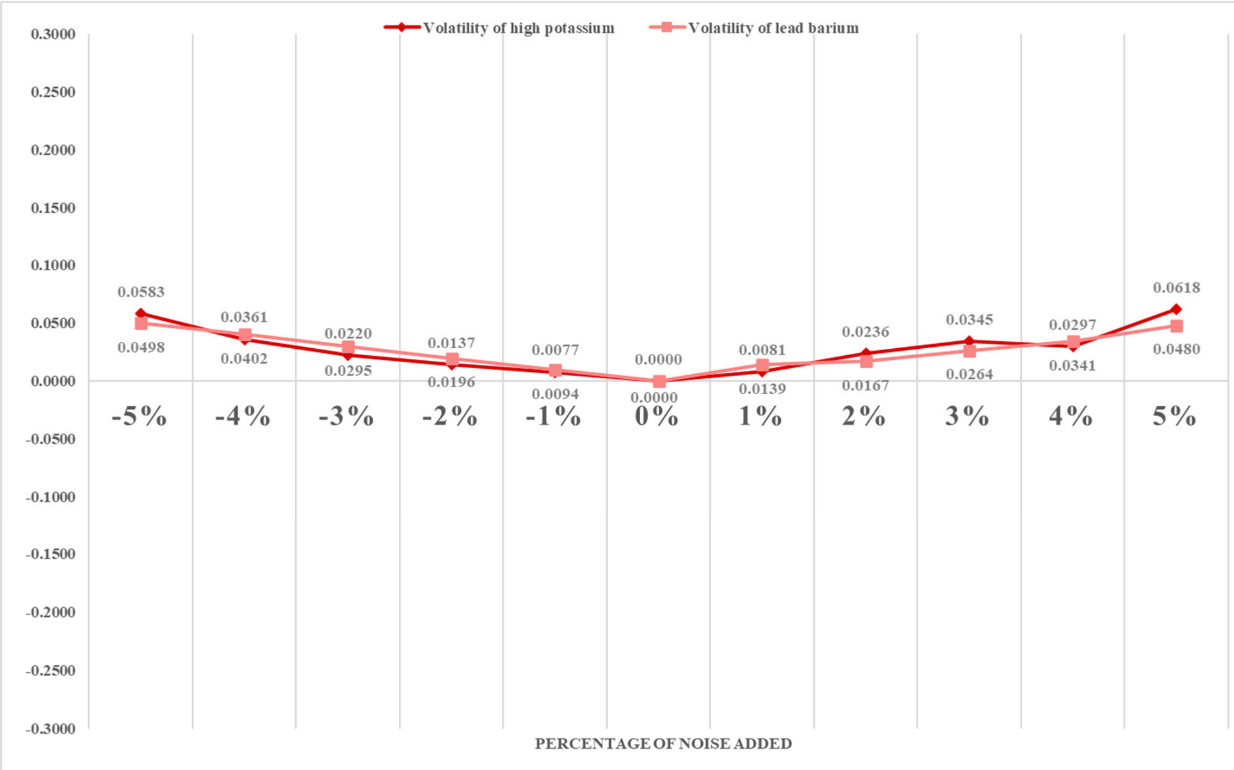


Figure 10 Volatility versus different noise

The graph above shows that the volatility of the clustering effect tends to increase slightly exponentially as the noise range increases, but is generally based on a stable level. To measure the sensitivity of the model, this paper calculates the variance of the clustering effect volatility to obtain the sensitivity ρ.

$$\rho = \sum_{i=1}^{n} \frac{(ds_i - \overline{ds})}{n} \tag{8}$$

where ds_i denotes the clustering effect with the i-th noise added and $\overline{ds}$ denotes the mean of the clustering effect.

After calculation, we obtained the sensitivity of the glass data for both categories as shown below.

Table 4 Sensitivity of data for two categories of glass

Category	Sensitivity
High potassium glass	0.000364121
Lead barium glass	0.000232067

From the table, it can be found that the sensitivity obtained by substituting the high potassium data into the model is greater than that of the lead-barium data, i.e.: noise has a greater effect on the model for high potassium clustering, while the clustering model for lead-barium works better.

3. Model evaluation

The K-means algorithm in this paper is a hard clustering algorithm, which is representative of the prototype-based objective function clustering method, using the error sum-of-squares criterion function as the clustering criterion function. The advantage of using the K-means Euclidean cluster analysis model to solve the problem is that by filtering the data in a step-by-step manner it makes full use of the valid data of the problem to classify the glassware, simplifies the operations of the model and makes its conclusions reliable, which can be extended for the classification of more subclasses of other types of glass, and subsequently the model used in this paper can also be put into prediction or classification systems in everyday life The model used in this paper can be subsequently applied to prediction or classification systems in everyday life. However, the disadvantage of the model is that it does not discuss weathered or untethered glassware separately, which may have an impact on the actual results.

Also, in the testing of the model results, a perturbation method is adopted by introducing noise to the input data, which is substituted into the clustering model to obtain the volatility of the clustering effect under different noises, and then the volatility is calculated by the variance principle to produce a quantitative representation of the sensitivity. This helps to reduce overfitting and improve generalization performance, and can be widely used in deep neural network models as a way of data augmentation techniques that help to overcome the problem of having less data in a particular category.

References

[1] Musteline Saul Martello et al. Machine learning unveils composition-property relationships in chalcogenide glasses[J]. Acta Materialman, 2022, 240

[2] Rossana S. et al. Glass colorations caused by Mn-Fe redox pair : Application to ancient glass technology[J]. Journal of Non-Crystalline Solids, 2022, 594

[3] Liu Lyngen et al. Composition optimization of Zr55Cu30Al10Ni5 bulk metallic glass using cluster formula approach[J]. Journal of Non-Crystalline Solids, 2022, 595

[4] Shravan Nawaf H. M. M. et al. Microwave Nondestructive Testing for Defect Detection in Composites Based on K-Means Clustering Algorithm[J]. IEEE ACCESS, 2021, 9 : 4820-4828

[5] Caching, Hu et al. Real-time Flood Classification Forecasting Based on k-means++ Clustering and Neural Network[J]. Water Resources Management, 2021, : 1-15.

[6] Yong-Lang Liu and Zhong Cai and Ji-Tao Zhang. An improved image classification based on K-means clustering and Bow model[J]. Int. J. of Grid and Utility Computing, 2018, 9(1) : 37-42.

[7] Hong Ling et al. Transcriptome sequencing and annotation of the halophytic microalga Daniella salina.[J]. Journal of Zhejiang University. Science. B,2017,18(10).

[8] Zhang Zhen .et al .A local search algorithm for k-means with outliers[J]. Neurocomputing,2021,450.

[9] Al Guliyev Rasmi et al Parallel batch k-means for Big data clustering[J]. Computers & Industrial Engineering,2021,152.

[10] Daniel Ranti et al. K-Means Clustering to Elucidate Vulnerable Subpopulations Among Medicare Patients Undergoing Total Joint Arthroplasty[J]. The Journal of Arthroplasty, 2020,