applied sciences

MDPI

Article

K-Means++ Clustering Algorithm in Categorization of Glass Cultural Relics

Jie Meng [1,*], Ziyang Yu [1], Yuxin Cai [2] and Xiuling Wang [3]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100083, China
[2] Weiyang College, Tsinghua University, Beijing 100083, China
[3] College of Information Engineering, Inner Mongolia University of Technology, Hohhot 010080, China; wxltt3756@imut.edu.cn
* Correspondence: meng-j20@mails.tsinghua.edu.cn

Abstract: We used statistical methods to study the classification of high-potassium glass and lead–barium glass and analyzed the correlation between the chemical composition of different types of glass samples. We investigated the categorization methodology of glass cultural relics, conducted a principal component analysis on the chemical composition data of the glass, and developed a case-specific clustering algorithm (K-Means++) to further categorize the glass cultural relics. K-Means++ was developed to reduce the sensitivity of a traditional K-Means clustering algorithm, by choosing the next clustering center with probability inversely proportional to the distance from the current clustering center. Then we verified the validity of the six subcategories we defined by inertia and silhouette score and evaluated the sensitivity of the clustering algorithm. We obtained a robustness ratio that maintained over 0.9 in the random noise test and a silhouette score of 0.525 in the clustering, which illustrated significant divergence among different clusters and showed the result is reasonable. With our proposed algorithm and classification result, a more comprehensive understanding of glass relics can be gained.

Keywords: K-Means++ clustering; principal component analysis; machine-learning classification

Citation: Meng, J.; Yu, Z.; Cai, Y.; Wang, X. K-Means++ Clustering Algorithm in Categorization of Glass Cultural Relics. *Appl. Sci.* **2023**, *13*, 4736. https://doi.org/10.3390/app13084736

Academic Editors: Vincent A. Cicirello, Wei Liu and Chia-Huei Wu

Received: 7 March 2023
Revised: 31 March 2023
Accepted: 5 April 2023
Published: 9 April 2023

1. Introduction

Glass has long been recorded among Chinese historical materials, but research on ancient Chinese glass started late. There is a lack of research on the weathering and composition of ancient silicate glass, and most of it is from the perspective of dynasty replacement. The cultural and artistic forms of glass and the laws of its own operation and development are studied in terms of cultural exchange and chemical analysis. Few scholars have systematically established mathematical models and used intelligent algorithms to qualitatively and quantitatively predict the original composition and subclassification methods of weathered silicate glass.

Machine-learning algorithms play an important role in the exploration of ancient cultures nowadays, helping in the search for statistical insight and classification. Data measurements and statistical analyses of the chemical compositions of ancient cultural relics, looking for statistical rules and classifying the types of cultural relics, can provide a reliable basis for identifying the types of ancient cultural relics and tracing the history of ancient cultural relics from a statistical perspective. In the most basic way, ancient glass cultural relics discovered in China are mainly divided into two categories: high-potassium and lead–barium glass. This is because in the process of making these glasses, people needed to add many kinds of auxiliary solvents into the main part, SiO_2. Commonly seen in the southeast of China, the high-potassium glass usually has plant ash as its auxiliary solvent, which is rich in potassium (in the form of K_2O), while the lead–barium glass is located elsewhere in China, rich in PbO and BaO [1,2]. As time went by, some of the glass relics were weathered and their chemical compositions changed accordingly. Wang Chengyu's [3]

in-depth study on the mechanism of weathering has certain reference significance for component prediction. Zhao Fengyan, et al. [4] classified the chemical composition of glassware by nondestructive analysis of pXRF. However, the existing chemical research methods cannot accurately and reasonably classify according to the composition of glass. Therefore, we consider introducing machine learning to solve practical problems by using classification models and intelligent algorithms. Intelligent algorithms have been widely used in scientific research in the field of materials in recent years, such as for the construction and application of the database of ancient Chinese unearthed glass beads studied by Feng Bailing [5] and the summary by Zhang Liyan [6] of the main theoretical basis, simulation process, and application status of each simulation method by using seven simulation methods of glass composition properties, but there are few methods that shed light on the specific field of glass relics. Li Jiangang [7] conducted glass defect detection based on deep learning, but there is still a gap in the use of machine learning to study the weathering and subclassification of ancient glass at home and abroad [8].

In this paper, we first study the classification rules of high-potassium glass and lead–barium glass by statistical methods and try to recover the original proportion of chemical elements of weathered points, i.e., their chemical composition before being weathered, by statistical methods. After that, we conduct the subcategorization based on dimensionality reduction by PCA and make an evaluation. The K-Means++ clustering model is established using three principal components, and the rationality and sensitivity of the model are tested.

The rest of this paper is organized as follows: In Section 2, the classification law of high-potassium glass and lead–barium glass are studied based on sample data. Section 3 proposes the K-Means++ clustering algorithm and establishes the classification model. In Section 4, the validity of the model is analyzed. Section 5 analyzes the model sensitivity. Section 6 discusses limitations of the study and future work. The conclusions are provided in Section 7.

2. Classification Law of High-Potassium Glass and Lead–Barium Glass

2.1. Sample Data Acquisition and Exploratory Spatial Data Analysis

2.1.1. Data Overview

The data were acquired from a private archaeology database, which contains the results of chemical composition detection of 58 glass cultural relics in total. Because two parts of one glass relic (for example, the inside and outside of a glass vase) may have significant differences in chemical composition, 11 glass relics underwent detection on two different points. Therefore, the data mainly consisted of 69 data points collected from 58 glass relics.

For each glass relic, its ornamentation, color, and type (high-potassium/lead–barium) were recorded, and for each detection point, its weathering degree (unweathered, weathered, severely weathered) and 14 chemical compositions were recorded. The 14 chemical compositions were SiO_2, Na_2O, K_2O, CaO, MgO, Al_2O_3, Fe_2O_3, CuO, PbO, BaO, P_2O_5, SrO, SnO, and SO_2. The first three lines of the data are shown in Table 1.

Table 1. The first three lines of the data.

Relic No.	Ornamentation	Type	Color	Weathering	Detection Point	SiO_2	CaO	MgO	Fe_2O_3	PbO	...	Sum
01	C	high-potassium	blue–green	unweathered	01	69.33	6.32	0.87	1.74			97.61
02	A	lead-barium	light green	weathered	02	36.28	2.34	1.18	1.86	47.43		99.89
03	A	high-potassium	blue–green	unweathered	03position1	87.05	2.01			0.25		100
					03position2	61.71	5.87	1.11	2.16	1.41		98.88

Because the chemical detection method in archaeology may produce a slight error, which makes the sum of the 14 compositions deviate from 100%, we regard a sum between 85% and 105% as reasonable. Based on this, we dropped two data points which have the sum 79.47% and 71.89%, respectively, and 67 data points remained for further processing.

2.1.2. Exploratory Data Analysis

Based on the chemical composition of the artifact samples and other testing means, the high-potassium glass and lead–barium glass were classified from a total of 67 data samples (including ornamentation, color, weathering degree, the proportion of the main components, etc.). The data distribution of sample color and the data distribution of color based on type are shown in Figure 1.

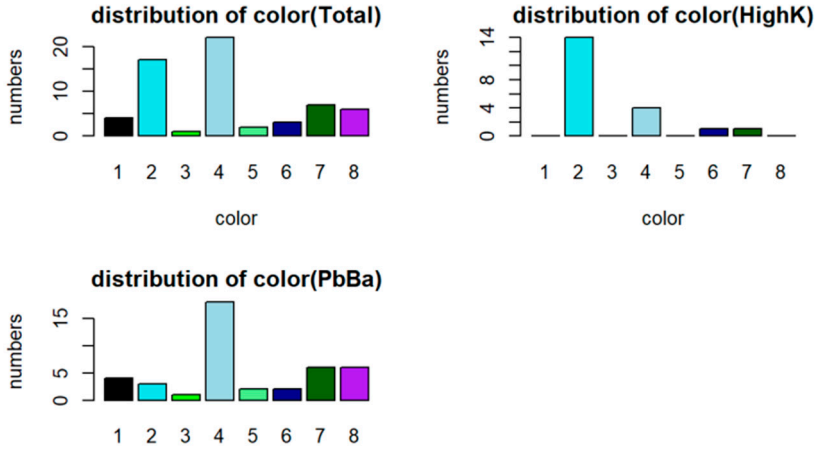

Figure 1. Samples color (type-based) data distribution.

The samples are marked 01–67, including 18 high-potassium and 49 lead–barium samples. To see how much information the basic data provided, a Kendall correlation coefficient matrix was established to initially observe the correlations among the four qualitative variables, and a correlation diagram was drawn, as shown in Figure 2. Additionally, because in the later section we recover the original chemical compositions of weathered points, we took a brief look at the linear relationship between weathering degree and other variables (including 3 qualitative variables and 14 continuous variables) by linear regression. This aimed to provide insight into how well we could do to discover the nonlinear clusters and distinguish weathered data from unweathered data using other variables. In the regression result shown in Figure 3, we obtained adjusted an R-squared value of 0.75. In Figure 4, we can see the positive/negative correlations between weathering degree and certain chemical compositions. From the model diagnostics graph shown in Figure 5, we conclude that there is a nonlinear relationship (from Residuals vs. Fitted), the residual is statistically normally distributed (from Normal Q-Q), heterogeneity appears in the model (from Scale–Location), and there are no global outliers in the data (from Residuals vs. Leverage).

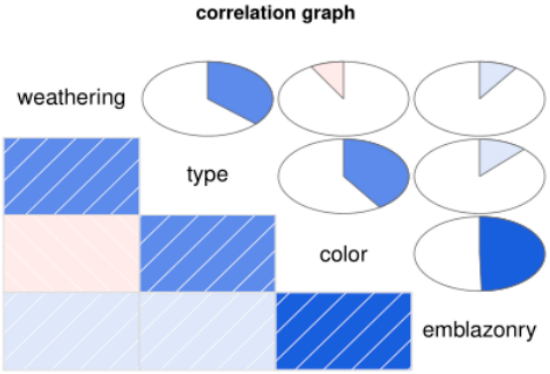

Figure 2. Correlation diagram among columns.

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4472 -0.1217  0.0225  0.0934  0.5333
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.243e-01  1.044e+00   0.119 0.905851
## SnO2         5.393e-01  1.966e-01   2.743 0.009340 **
## SO2          5.683e-02  2.409e-02   2.359 0.023701 *
## SiO2        -1.153e-05  1.224e-02  -0.001 0.999253
## Na2O         8.776e-02  2.902e-02   3.024 0.004519 **
## K2O         -2.124e-02  2.526e-02  -0.841 0.405795
## CaO         -3.321e-03  2.460e-02  -0.135 0.893350
## MgO         -4.522e-02  9.241e-02  -0.489 0.627473
## Al2O3        5.774e-02  1.927e-02   2.997 0.004850 **
## Fe2O3       -4.136e-02  4.462e-02  -0.927 0.359924
## CuO          7.001e-02  3.141e-02   2.229 0.031964 *
## PbO          2.670e-02  1.303e-02   2.049 0.047621 *
## BaO         -1.537e-02  2.046e-02  -0.751 0.457144
## P2O5        -1.803e-03  2.397e-02  -0.075 0.940430
## SrO         -6.486e-01  2.032e-01  -3.193 0.002875 **
## color2      -3.800e-01  2.220e-01  -1.712 0.095356 .
## color3      -1.537e+00  3.518e-01  -4.368 9.74e-05 ***
## color4      -4.889e-01  1.879e-01  -2.602 0.013267 *
## color5      -9.700e-01  3.120e-01  -3.108 0.003607 **
## color6      -1.222e+00  3.072e-01  -3.977 0.000312 ***
## color7      -7.160e-01  2.665e-01  -2.687 0.010746 *
## color8      -5.223e-01  3.493e-01  -1.495 0.143268
## emblazonry2  1.071e+00  2.488e-01   4.304 0.000118 ***
## emblazonry3  9.602e-02  1.456e-01   0.660 0.513570
## type2        1.826e-01  3.001e-01   0.609 0.546564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2498 on 37 degrees of freedom
## Multiple R-squared:  0.8485, Adjusted R-squared:  0.7502
## F-statistic: 8.633 on 24 and 37 DF,  p-value: 5.475e-09
```

Figure 3. Linear regression result.

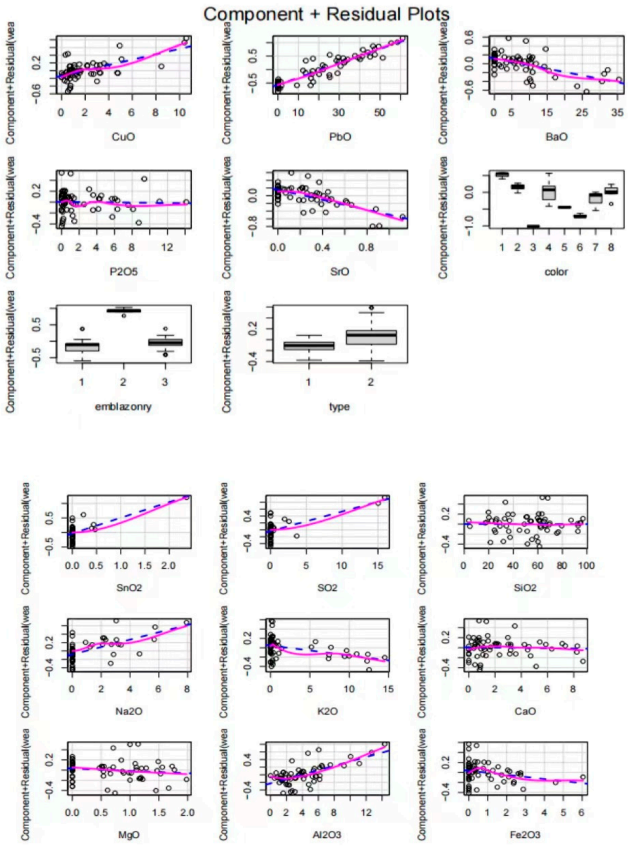

Figure 4. Component + Residual Plots.

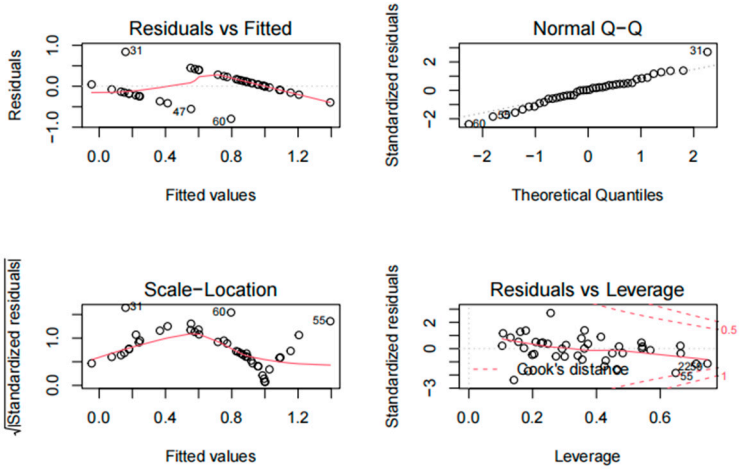

Figure 5. Model diagnostics graphs.

2.2. Analysis of Classification Laws Based on Sample Data

It is necessary to explore the classification pattern of high-potassium and lead–barium glass, thus enabling the further classification of subcategories in terms of chemical composition content. In order to explore the relationship between the proportion of chemical components and the division of main categories (high potassium, lead and barium), we used the supervised learning algorithm KNN to construct a K-nearest neighbor graph between sample points, which proved well that K_2O content and (PbO + BaO) content were the key indicators for distinguishing the main categories.

Semantically, "high-potassium glass" should be glass with a high potassium oxide content (noted as K_2O below), while "lead–barium glass" is glass with a high content of lead oxide (PbO) and barium oxide (BaO). It can be assumed that if the K_2O content and the (PbO + BaO) content of the glass cultural relics are taken as two dimensions, a clear demarcation line can be drawn under the plane right-angle coordinate system. We validated this idea based on 67 data samples by the KNN (K-Nearest Neighbor) algorithm of supervised learning.

The KNN algorithm is a method to classify each record in a dataset, which is a typical supervised learning algorithm. The process of a KNN algorithm classifying one new point is as follows: the distances between this point and all marked points are calculated, from which n_neighbors points with the closest distance are selected. The category with the largest proportion of these n_neighbors points is the classification of the new point [9–12].

First, the values of the two independent input variables K_2O and (PbO + BaO) were calculated for the 67 samples. Among the 67 samples, 70% were randomly selected as the training set and 30% as the test set, taking n_neighbors = 5 to obtain a 100% correct classification rate, as shown in Figure 6 (both training and testing data are plotted in the figure).

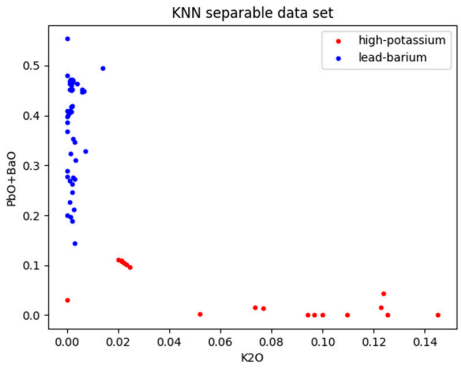

Figure 6. KNN classification results.

To further determine the correctness of this classification criterion, the proportion of the training set was reduced to 50% and the other 50% selected as the test set, taking n_neighbors = 5, and we still obtained a 100% correct classification rate.

From this, we can infer statistically that the classification of high-potassium glass and lead–barium glass is precisely binary clustering according to the K_2O, (PbO + BaO) content.

3. K-Means++ Subclass Classification

3.1. Prediction of the Chemical Composition Content before Weathering on Weathered Samples

The weathering of glass products will lead to significant changes in their composition, so it is necessary to predict the composition content before weathering based on the weathering point detection data. We chose to use the similarity between glass products to predict the chemical composition of specific glass products when they are not weathered.

Because the present compositions of weathered data deviate from their original compositions, we cannot compare the similarities using given chemical compositions directly. As an alternative, we chose to conduct a principal component analysis first, using only unweathered data so that the unimportant compositions would be eliminated. Second, we applied PCA to weathered data, and the principal components (PCs) obtained in this way would be irrelevant to the weathering degree.

First, the original data were tested for validity, and only unweathered samples were selected, from which we had 35 data points [13–15]. The 14 chemical content compositions of these data were taken into PCA. The percentage of variance explained by each orthogonal index was analyzed, and the first four items were [0.72766864 0.15677105 0.05843352 0.02400202], and the first three items were taken as new PCs (for explained variance greater than 0.05), which were noted as PC_1, PC_2, and PC_3.

PCA process:

- Use 35 unweathered samples, with 14 chemical compositions, as matrix X (35 × 14).
- Calculate the eigenvalue decomposition: $X^TX = Q\Sigma Q^T$.
- Choose the largest 3 eigenvalues and their eigenvectors, form matrix P (14 × 3).
- Calculate PCs for unweathered data: $Y = XP$.
- Use 32 unweathered samples, with 14 chemical compositions, as matrix X1 (32 × 14).
- Calculate PCs for weathered data: $Y_1 = X_1P$.

After that, this PCA model was applied to calculate the PCs of the weathered samples. The PC matrix obtained in this way is significantly less related to weathering because when constructing the model, no factors related to weathering are introduced, and when using the model to calculate the factor values of weathered samples, it is equivalent to making it as reductive as possible for the unweathered samples, in line with the prediction requirements.

For the definition of similarity, we used the Euclidean distance to measure the similarity. Due to the large number of chemical components and the fact that some components are impurities, we first used the principal component analysis to perform the dimensionality reduction of the chemical content, which also ensured that some impurities in the glass products (such as strontium oxide, which is less than 0.01 in all glasses) can be ignored sufficiently as to not affect the subsequent prediction results [16–18].

Based on the above definition of "relevance" and comparative analysis, the three PCs were used as indicators of similarity for prediction, as follows: similarity of glass products a and b,

$$\text{Similarity}_{a,b} = \sqrt{(PC_{1a} - PC_{1b})^2 + (PC_{2a} - PC_{2b})^2 + (PC_{3a} - PC_{3b})^2}, \quad (1)$$

For all weathered points, the n points that were most similar to them were taken, and the content prediction of this point was obtained by taking the similarity between the n points and this point as the weight and weighting the average of the chemical composition content of the n points. For the selection of n, we tried = 3, 4, 5, and found that the difference between the predicted values was less than 5%. Thus, the three cases could be regarded as

equivalent within the error tolerance, and the prediction results were obtained by taking n = 4.

3.2. K-Means++ Algorithm

Next, we performed the subclass classification based on high-potassium glass and lead–barium glass. In the absence of a priori knowledge of the subclass classification criteria and no real labels for reference, we first chose to use the unsupervised clustering algorithm K-Means for modeling and analysis, using the three PCs obtained in the PCA section as three feature variables in the clustering.

The K-Means algorithm at random uniformly selects K points as the center of mass at initialization, and in each iteration, calculates the distance from each point to the K centers of mass, divides the samples into the clusters corresponding to the closest center of mass, and at the same time, calculates the mean value of all samples within each cluster and updates the center of mass of the cluster using this mean value, until the position change of the center of mass is less than the specified threshold (default 0.0001) or the maximum number of iterations is reached [19–22]. Because the classification of glass products is not related to weathering, it was necessary to eliminate the influence of weathering on chemical composition and classification and to select the PCA obtained above. The three-factor index was used as the factor parameter of the sample.

The weakness of K-Means lies in its initialization. Because the probability of choosing any data point as the center of mass is equal, there is a significant chance that several close points are simultaneously selected as centers, which means that in the later process of iteration, these points are highly likely to be divided into different clusters. However, if two points are close to each other in the context of feature variables, they should be in the same cluster for unsupervised learning. Therefore, we considered improving the initialization strategy and developing K-Means++.

Since K-Means is more sensitive to initial values, the algorithm was improved. After setting the probability of each sample point becoming a cluster center to be inversely proportional to the distance from the current cluster center, the more distant the sample point is from the existing cluster center, the more likely it is to be selected as the next cluster center. In the mathematical formulation, we denoted the set of existing cluster centers as S, and the probability of choosing x as the next cluster center is

$$p(x) = \frac{\min_{a \in S} D(x, a)}{\sum_x \min_{a \in S} D(x, a)}$$

$$\text{in which } D(x, a) = \text{Similarity}_{x,a}{}^2.$$

In other words, K-Means++ is most likely to select the point which is far from all the existing cluster centers. In this way, the initialization will guarantee enough divergence (in the sense of distance) among different clusters, which improves the effectiveness of clustering and the convergence rate.

The algorithm was tested and found to be significantly less sensitive, so the improved algorithm was named K-Means++. All 67 samples were classified into six classes using the K-Means++ algorithm, and the classification results are shown in Figure 7 and Table 2. By contrast, the classification result of K-Means is shown in Figure 8.

In order to see the improvement in the sensitivity of K-Means++, we compared K-Means++ with K-Means in the classification results, i.e., we ran both algorithms 1000 times in clustering the 67 samples. Initially, we obtained a result for both algorithms, as shown in Figures 7 and 8. Every time, we got a classification result with six cluster centers, and we recorded the accumulated difference among these six centers from the initial six centers (represented by the absolute difference of center distance). The larger the accumulated difference is, the more volatile the classification is and the more sensitive the algorithm is. From Figure 9, we can see that the accumulated difference of K-Means++ increases more slowly than that of K-Means, which verifies the sensitivity improvement.

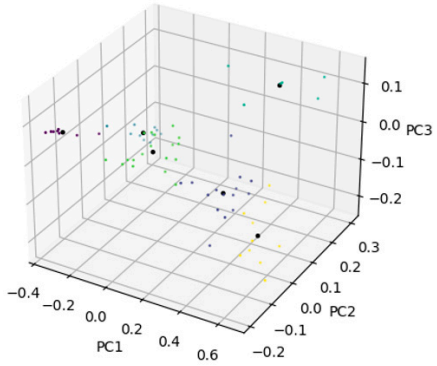

Figure 7. Three-dimensional K-Means++ clustering visualization.

Table 2. Results of K-Means++ clustering for all samples.

Clustering Name	Sample
0	08 *, 08 Severe weathering point, 11, 20, 24, 26, 26 Severe weathering point.
1	01, 03 Part 2, 04, 05, 06 Part 1, 06 Part 2, 13, 14, 16
2	02, 19, 30 Part 1, 30 Part 2, 34, 36, 38, 41, 43 Part 2, 49, 50, 51 Part 1, 52, 56, 57, 58
3	03 Part 1, 07, 09, 10, 12, 18, 21, 22, 27
4	23 unweathered point, 25 unweathered point, 28 unweathered point, 29 unweathered point, 31, 32, 33, 35, 37, 42 unweathered point 1, 42 unweathered point 2, 44 unweathered point, 45, 46, 47, 48, 49 unweathered point, 50 unweathered point, 53 unweathered point, 55
5	39, 40, 43 Part 1, 51 Part 2, 54, 54 Severe weathering point

* The number (for example, 08) represents the glass relic sample. If two different points are detected in one glass relic, the description following the number distinguishes them (for example, 08 and 08 Severe weathering point; 03 Part 1 and 03 Part 2).

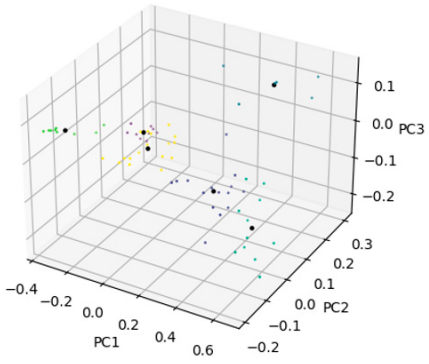

Figure 8. Three-dimensional K-Means clustering visualization.

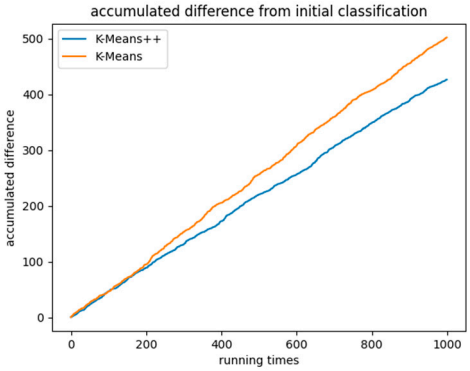

Figure 9. Accumulated difference of two clustering algorithms.

K-Means++:

Initialization: Select 6 cluster centers $\mu_1^{(0)}, \mu_2^{(0)}, \cdots \mu_6^{(0)}$ one by one, using $p(x) = \frac{\min_{a \in S} D(x, a)}{\sum_x \min_{a \in S} D(x, a)}$.

Define error function $J(c, \mu) = \sum_{i=1}^{n} ||x_i - \mu_{c_i}||^2$, where c_i is the cluster which x_i belongs to.

Iterate until $J(c, \mu)$ converges (or reaches the iteration time limit):

For i = 1, 2, ..., n:

$$c_i = \underset{k}{\operatorname{argmin}} ||x_i - \mu_k||^2$$

For k = 1, 2, ..., 6:

$$\mu_k = \underset{c_i = \mu_k}{\operatorname{mean}} x_i$$

4. Model Validity Analysis

Whether the division into parent classes is satisfied is an intuitive indicator of model validity. The subclass classification relies on the chemical composition content, which makes the samples in the same subclass highly similar with respect to some chemical compositions. In order to better reflect this similarity, linear regression can be used to explore the composition characteristics of each subclass as another criterion for subclass classification. The contour coefficient of the clustering results is a measure of whether the cluster is reasonable and valid [23]. In this paper, we mainly analyzed the reasonableness of the K-Means++ clustering model from the above three aspects.

4.1. Comparison Verification with Parent Classes

A more intuitive indicator is to verify that the division of subclasses satisfies the division of the parent classes (i.e., high-potassium vs. lead–barium). The statistical results of the K-Means++ clustering algorithm are shown in Table 3.

Table 3. Results of clusters by parent class obtained by K-Means++ clustering algorithm.

Clustering Name	High-Potassium	Lead–Barium
0	0	7
1	9	0
2	0	16
3	9	0
4	0	20
5	0	6

It can be found that although the six subclasses were divided without introducing any parent variables and clustering was performed only by relying on the dimensionality reduction factor of chemical content, the samples of the same subclass all belonged to the same parent class. This indicates that the clustering results obtained by this method have a certain degree of validity.

4.2. Significance Test of Linear Regression

Subclass classification relies on chemical composition content, which makes samples within the same subclass highly similar with respect to some chemical compositions. To better represent this similarity, linear regression can be used to explore the compositional characteristics of each subclass, using different subclasses within the same parent class as dependent variables and the chemical composition content of each sample before weathering (including predicted values for weathered samples) as independent variables [24].

Before the component regression, the subclass numbers need to be adjusted so that the subclasses belonging to the same parent class have consecutive serial numbers. This is because there is a difference in component content between the parents themselves, and

we need to explore the subclasses based on the difference in parent content, rather than regressing the subclasses directly without a priori knowledge.

For the six clusters of the K-Means++, the component regression was performed, and the results are shown in Figure 10. It can be seen that there are many chemical components with significant results, such as SiO_2, Na_2O, CaO, etc., which means that we can use the chemical component as one criterion for subclass classification.

```
. reg change0 sio2 na2o k2o cao mgo al2o3 fe2o3 cuo pbo bao p2o5 sro sno2 so2, robust

Linear regression                               Number of obs     =         67
                                                F(14, 52)         =      25.18
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7295
                                                Root MSE          =    0.93125
```

change0	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
sio2	29.45664	10.51482	2.80	0.007	8.357103	50.55617
na2o	23.19608	10.89242	2.13	0.038	1.338838	45.05332
k2o	23.04289	13.75503	1.68	0.100	-4.5586	50.64438
cao	58.18837	15.58324	3.73	0.000	26.9183	89.45844
mgo	51.04194	30.33607	1.68	0.098	-9.83183	111.9157
al2o3	12.52075	9.973145	1.26	0.215	-7.491829	32.53333
fe2o3	7.955452	17.92288	0.44	0.659	-28.00943	43.92034
cuo	53.44195	13.32424	4.01	0.000	26.70491	80.17898
pbo	24.80044	10.55115	2.35	0.023	3.628006	45.97287
bao	20.05019	14.40876	1.39	0.170	-8.863109	48.96349
p2o5	53.63678	19.19719	2.79	0.007	15.11479	92.15876
sro	-63.96303	96.24492	-0.66	0.509	-257.0926	129.1665
sno2	166.6073	52.18478	3.19	0.002	61.89093	271.3238
so2	-23.30729	9.583671	-2.43	0.018	-42.53833	-4.076249
_cons	-25.33151	10.22957	-2.48	0.017	-45.85865	-4.804365

Figure 10. Components regression results of K-Means++ clustering.

Analyzing the composition table, we can clearly see that the two subclasses of the "high-potassium" parent class 1.3 differ significantly in Ca_2O content, so the subclasses are named "high-potassium high-calcium" and "high-potassium low-calcium" classes. At the same time, we can clearly see that the two subclasses 0.2 of the parent class "Pb-Ba" differ significantly in the content of CuO. The value for subclass 0 is significantly higher than the mean value; subclass 2 has a value significantly lower than the mean value, and that of subclass 4, 5 is close to the mean value. The difference in the content of subclass 4 and 5 is reflected in Ca_2O. Therefore, the subclasses were named "lead–barium high-copper", "lead–barium low-copper", "lead–barium medium-copper low-calcium", and "lead–barium medium-copper high-calcium". The details of the subclasses are shown in Table 4.

Table 4. Subclass results obtained by the K-Means++ clustering algorithm.

Subclass Name	Subclass Number
Lead–barium high-copper	0
High-potassium and calcium	1
Lead–barium low-copper	2
High-potassium and low-calcium	3
Lead–barium mid-copper low-calcium	4
Lead–barium mid-copper high-calcium	5

4.3. Silhouette Coefficient and Square Sum of Error

The contour coefficient refers to a method that reflects the consistency of the data clustering results and can be used to assess the degree of dispersion among clusters after clustering. For a sample u belonging to cluster C_i, we denote d (u, v) as the distance between u and v, defined as

$$a(u) = \frac{1}{|C_i| - 1} \sum_{v \in C_i, u \neq v} d(u, v),\ b(u) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{v \in C_k} d(u, v). \quad (2)$$

We define the contour coefficient of u as $s(u) = \frac{b(u)-a(u)}{\max\{a(u),b(u)\}}$, lying between -1 and 1. If $s(u)$ of a sample is close to 1, it means the sample is reasonably clustered; if it is close to -1, it means it should be classified into other clusters. If the silhouette is close to 0, it means the sample is on the boundary of two clusters. The mean value of all sample contours is called the silhouette coefficient, which is a measure of whether the clustering is reasonable and valid.

The square sum of errors inertia is defined as the sum of squares of the distances between all samples and the center of mass of the cluster to which they belong, and the optimal number of classifiers should be taken at the point where the deformation of inertia is most intense. For the optimal number of classifications for K-Means++ clustering, two evaluation metrics (inertia and silhouette coefficient) are used. The traversal is performed for the possible number of classifications n = 2, 3, ..., 19, 20, varying the number of clusters k, using the silhouette_score function implemented in the python sklearn library for validation and plotting the curve of inertia and silhouette coefficient, as shown in Figures 11 and 12. From the square sum of errors image and the silhouette_score image, it is best to be divide the data into six categories.

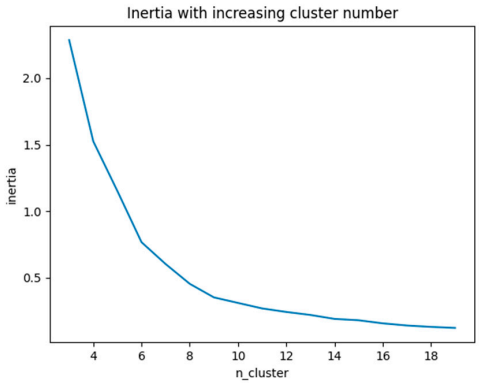

Figure 11. Variation in square sum of errors with the number of clusters for the K-Means++ clustering algorithm.

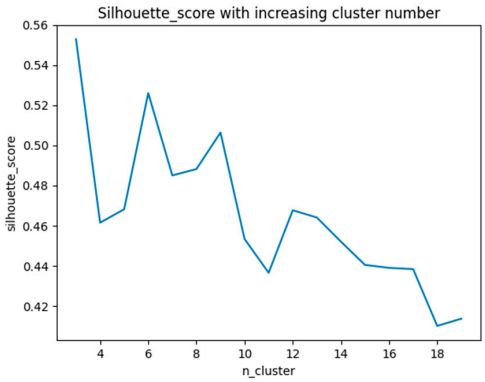

Figure 12. Variation in silhouette_score with the number of clusters for the K-Means++ clustering algorithm.

5. Model Sensitivity Analysis

For a certain type of glass artifact, the randomness associated with the production process or prolonged exposure to the environment may cause some change in the proportions of the various chemical components, but the change is relatively minor to the extent that it should not affect our determination of that glass category. We consider such perturbations as noise added to the sample data [25]. Since the sample of data in this study is small and estimates of noise are highly susceptible to overfitting, it is useful to assume that the prior probability distribution of the noise is F = N (0, 0.0001), i.e., a Gaussian distribution with mean 0 and standard deviation 0.01. For any $x \in F$,

$$P(-\sigma < X < \sigma) = \int_{-\sigma}^{\sigma} f(t)dt = 0.682, \tag{3}$$

where f(t) is the probability density function of the distribution F. There is about a 68.2% probability that the noise lies in the thousandths interval and has a small effect on the weight of the chemical composition.

The process of a random noise test is as follows: The number of random tests is initially set T = 100. In the i'th ($1 \leq i \leq T$) test, all original unweathered samples are added with noise sampled by the above prior distribution to obtain the noise-containing sample data, and the prediction labels are obtained by the clustering algorithm. The number of samples for which the predicted labels are the same as the original predicted labels after adding noise is counted and is denoted as t_i. At the end of T times of testing, it is calculated that (N is the number of samples)

$$\text{Random noise test ratio} = \frac{\sum_{i=1}^{T} t_i}{N \times T}. \tag{4}$$

As a measure, its value ranges from [0, 1], and the closer to 1, the better the noise immunity of the model.

The K-Means++ clustering algorithm was tested 10 times for random noise and averaged to obtain a random noise test ratio = 0.980, indicating that the model is insensitive to noise under the current noise prior distribution.

To further investigate the relationship between model sensitivity and noise standard deviation, we used the following equation:

$$\sigma = 0.01 + 0.00474t, \ 0 \leq t < 20. \tag{5}$$

The result of increasing the standard deviation in the above random noise test for the K-Means++ model reveals the robustness of K-Means++. Compared with the K-Means algorithm, when σ increases from 0.01 to 0.1, the clustering accuracy of K-Means++ remains significantly higher and shows a smaller variance, which illustrates the performance improvement of K-Means++. The random noise test ratio with σ is shown in Figure 13.

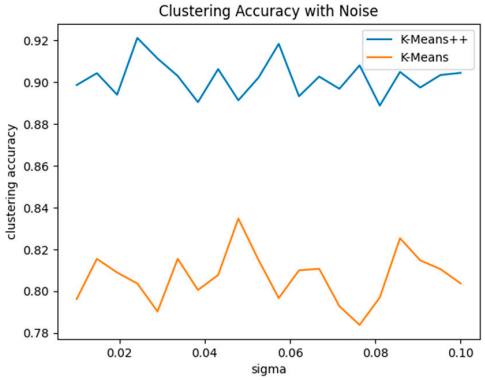

Figure 13. K-Means++ and K-Means clustering model sensitivity with noise standard deviation.

6. Limitations and Future Work

6.1. Size of Glass Relics Data

The data sample contained 67 effective samples, which is relatively small compared to typical machine-learning tasks in big data. However, it should be noted that unlike general regression tasks, a small dataset works for a clustering algorithm because we divided the 67 samples into only six subcategories, which is acceptable and free from the overfitting problem. In the K-Means++ algorithm, because there are no parameters to be learned and the first cluster center is chosen uniformly at random, we can safely say that K-Means++ performs equally well on each size of data, as long as the number of clusters is significantly smaller than the size of the data.

6.2. Potential Improvement of K-Means++

Since K-Means++ has a more complicated process during initialization, it on average requires a significant amount of time for the first step, especially when the dataset is large. The time complexity of initialization is O(mn), where m is the number of cluster centers and n is the number of total data points. Also, it certainly needs more space to store the probability vector in every iteration, with space complexity O(n). These are the extra costs using K-Means++.

However, when it comes to optimization problems, we know that a good initialization significantly contributes to a fast convergence. The same logic applies to K-Means++. Although K-Means++ costs more time during initialization, it has a theoretically faster convergence rate. Related experiments may be conducted using a large dataset to highlight the convergence improvement in a future study.

7. Conclusions

Based on the KNN, we constructed the K-nearest neighbor graph between the sample points. According to the K_2O content and (PbO +BaO) content, the main class of the sample was divided, and the principal component analysis was used to find the weathering-independent principal components to establish the relationship between the weathered sample and the unweathered sample, based on the improved K-Means++ algorithm model. The samples were then divided into subclasses. Before modeling, we pre-processed the color visualization in data processing, which highlighted the obvious effect and improved the efficiency of the data analysis. Despite the small number of samples, we still obtained ideal classification results. K-Means++ clustering algorithm has a good performance regarding the subclassification of glass cultural relics based on chemical contents, and the model evaluation verified this according to the comparison with parent classes, validation metrics, and noise test. We obtained a robustness ratio which maintained over 0.9 in the random noise test and a silhouette score of 0.525 in the clustering, which illustrated a significant divergence among different clusters and showed the result is reasonable.

Author Contributions: Conceptualization and methodology, J.M., Z.Y. and Y.C.; software and validation, J.M. and Z.Y.; investigation, J.M. and Y.C.; writing—original draft preparation, J.M.; writing—review and editing, J.M.; project administration and funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the central government guides' local science and technology development fund projects (JY20210061).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used is from a private database.

Acknowledgments: The authors gratefully acknowledge the administrative and technical support of Tsinghua University. This work has been funded by Inner Mongolia University of Technology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yin, Y. Component analysis of ancient glass products by association prediction. *Contemp. Chem. Res.* **2003**, *1*, 122–126.
2. Cao, C.; Guo, H. Using fuzzy mathematics knowledge to classify the disease types of stone, ceramic and glass relics in the collection. *J. Beijing Union Univ.* **2009**, *4*, 58–60.
3. Wang, C.; Tao, Y.; Chen, M.; Huang, M. Weathering of sodium-calcium-aluminum-magnesium silicate glass and alkali-lead silicate glass. *Silic. Bull.* **1989**, *6*, 1–9.
4. Zhao, F.; Chen, B.; Chai, Y.; Dong, J.; Li, Q. The PXRF analysis and related issues of some glassware unearthed in Xi′an. *Archaeol. Cult. Relics* **2015**, *4*, 111–119.
5. Feng, B. Construction and application of ancient glass beads database unearthed in China. Master's Thesis, Northwest University, Xi'an, China, 2021.
6. Zhang, L.; Li, H.; Chen, S.; Li, Z.; Ruan, M. Simulation method of glass composition and properties. *J. Silic.* **2022**, *50*, 2338–2350.
7. Li, J. Glass defect detection based on deep learning. Master's Thesis, Fujian University of Engineering, Fuzhou, China, 2021.
8. Wang, Z.; Zhao, X.; Li, Z.; Guo, M.; Xiao, W.; Liu, Z. Original composition prediction and sub-classification method of weathered silicate glass based on machine learning. *J. Silic.* **2023**, *51*, 416–426.
9. Yigit, H. A weighting approach for KNN classifier. In Proceedings of the 2013 International Conference on Electronics, Computer and Computation (ICECCO), Ankara, Turkey, 7–9 November 2013; pp. 228–231.
10. Taneja, S.; Gupta, C.; Aggarwal, S.; Jindal, V. MFZ-KNN—A modified fuzzy based K nearest neighbor algorithm. In Proceedings of the 2015 International Conference on Cognitive Computing and Information Processing(CCIP), Noida, India, 3–4 March 2015; pp. 1–5.
11. Kuang, Z. Optimization of KNN classification algorithm in high-dimensional data. Master's Thesis, Guangdong University of Technology, Guangzhou, China, 2019.
12. Angün, E.; Altınoy, A. A New Mixed-Integer Linear Programming Formulation for Multiple Responses Regression Clustering. In Proceedings of the 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), Paris, France, 23–26 April 2019; pp. 1634–1639.
13. Rani, A.J.M.; Parthipan, L. Clustering analysis by Improved Particle Swarm Optimization and K-means algorithm. In Proceedings of the IET Chennai 3rd International on Sustainable Energy and Intelligent Systems (SEISCON 2012), Tiruchengode, India, 27–29 December 2012; 2012; pp. 1–6.
14. Dehariya, V.K.; Shrivastava, S.K.; Jain, R.C. Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms. In Proceedings of the 2010 International Conference on Computational Intelligence and Communication Networks, Bhopal, India, 26–28 November 2010; pp. 386–391.
15. Li, B.; Guan, Y.; Gong, W.; Wei, X.; Xue, D. Density-based K-means initial clustering center selection algorithm. *J. Suihua Univ.* **2022**, *42*, 148–151.
16. Zhang, S.; Chen, X. Dimensionality reduction of principal component analysis data based on mutual information credibility. *J. Hubei Univ. Natl. (Nat. Sci. Ed.)* **2019**, *4*, 425–430.
17. Zhao, B. Discussion on the application of machine learning classification based on data dimensionality reduction. *Mod. Inf. Technol.* **2018**, *2*, 144–145.
18. Dong, H. Comparative study of principal component analysis and linear discriminant analysis. *Mod. Comput. (Prof. Ed.)* **2016**, *29*, 36–40.
19. Liu, Z.; Zhou, B.; Yu, L.; Niu, C.; Xu, X. K-means algorithm combining Pearson similarity and minimum spanning tree. *J. Nanchang Inst. Technol.* **2022**, *41*, 91–96.
20. Zheng, R.; Zheng, X.; Huang, W. Fast sorting method for retired power batteries using improved K-means algorithm. *J. Xiamen Univ. Technol.* **2022**, *30*, 74–81.
21. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* **2012**, *2.1*, 86–97. [CrossRef]
22. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
23. Yin, S.; Wang, T.; Xie, F.; Liu, L.; Qu, Z.; Zhang, B. Evaluation method of clustering results based on mutual information and contour coefficient. *J. Ordnance Equip. Eng.* **2020**, *8*, 207–213.
24. Li, C.; Liu, Z.; Liu, S. Improved KNN data classification model based on linear regression method. *J. Lanzhou Petrochem. Vocat. Tech. Coll.* **2021**, *3*, 20–23.
25. Sun, L.; Liu, M.; Xu, J. K-means clustering algorithm based on optimizing initial clustering center and contour coefficient. *Fuzzy Syst. Math.* **2022**, *1*, 47–65.