Research on an Integrated Intelligent Classification Algorithm Based on K-Means PCA-RF Machine Learning

Bingkun Song [1,2,*,#], Wenxuan Fan [1,2,#], Shuo Zhang [1,#]

[1] Chang'an University, Xian, China

[2] University College Dublin, Dublin, Ireland

* Corresponding Author Email: 2021901434@chd.edu.cn

[#] These authors contributed equally

Abstract. With the rapid development of machine learning and artificial intelligence, research on classification models is gradually becoming popular. This article aims to propose a general classification model and classify indicator features. First, this paper constructs the data preprocessing based on K-Means, and data dimensionality reduction based on PCA algorithm. Finally, random forest algorithm (RF) is used for feature classification, and 325 groups of data are used for training. The results show that: (1) The K-Means PCA-RF algorithm constructed in this paper has good robustness and classification performance. (2) K-Means PCA-RF can effectively classify features and perform sensitivity analysis.

Keywords: Intelligent classification, K-means, machine learnings, random forest.

1. Introduction

In recent years, with the widespread application of glass materials in various industries, classifying them has become an important research direction. However, traditional classification methods require a lot of manual experience and time, resulting in low efficiency and inconsistent results, making it difficult to meet practical needs. Therefore, machine learning has become an effective classification method for glass materials. Conducting machine learning research on glass material classification can improve classification accuracy, improve classification efficiency, promote progress in glass material research, and enhance the safety of glass products. In summary, the importance of machine learning in glass material classification research is self-evident. With the development of technology and artificial intelligence, machine learning will become the mainstream technology for glass material classification, bringing more refined and safe glass products to various industries [1-2]. In addition, the use of grass ash as a flux to fire the resulting glass products, also known as high potassium glass, widely popular in the Lingnan region of China. Most of the existing scientific studies on ancient glass use optical and spectroscopic techniques to analyze the physical and chemical properties of the artifacts, such as chemical composition, physical phase, microscopy, cross-sectional structure, and other physical. The study can yield information on the chemical composition of ancient glass, flux composition, production processes, and other basic research. However, relatively few studies have been conducted on the chemical composition of glass products using mathematical methods. Therefore, this paper focuses on the analysis of chemical composition of these two main ancient glass products. Based on a batch of data of chemical composition content of Chinese ancient glass samples, a systematic study of the relationship between classification and composition of these two main ancient Chinese glasses was conducted using principal component analysis based on K-means clustering, and the relationship between composition content and chemical composition was obtained, and moreover, an innovative subclass classification and a test of the rationality and sensitivity of the classification were conducted under the two main classes of glasses again based on the composition data, which is of great practical significance.

2. Analysis of the classification problems of the main subclasses of ancient glass

In this paper, we search for the classification basis of high potassium glass and lead-barium glass; set up the law of subclass division under the two categories and give the results and test the rationality and sensitivity of the classification law.

For the principal class analysis, we used the inverse method, using principal component analysis based on K-mean clustering, to divide the data into two classes and compare the classification results obtained with the classification data given in the question, and if the matching rate is high, we can assume that the expression between the principal component and the chemical component content involved in the principal component analysis can be used as a law for the classification of ancient glass products. For the subclasses, we excluded the chemical component species that were not relevant or of minimal relevance to the subclass classification, and again used the principal component analysis based on K-means clustering to process the remaining data to obtain the subclass classification laws and their results.

Finally, the subclass classification results are set as the real classification results, and the results are checked by using other classifications or getting the results, and the match between the obtained results and the real results is checked to get the reasonableness; the original data are classified again by adding the deflection, and the match between the obtained results and the real results is checked to get the sensitivity.

3. Principal component analysis model based on K-means clustering

3.1. Model Preparation

According to the literature, it is known that the content of oxygen PbO,BaO is high due to the addition of lead ore in the firing process of lead-barium glass, high potassium glass is fired with grasswood ash as a flux, which is rich in potassium; Fe and Cu elements in the glass only play a role in color development [3-4]. Therefore, Na_2O, K_2O, PbO, and BaO were initially selected as the targets of the study to influence the classification of glass types.

3.2. Principal component analysis model building based on K-means clustering

(1) Principal component analysis

The idea of the principal component analysis method is that multiple complex variables in the problem are considered as the original coordinate system, and after the rotation translation transformation of the coordinates, the center of gravity g of the data points in the new coordinate system is made consistent, thus reducing the data to a relatively small number of new comprehensive generalized indicators, and using the resulting comprehensive generalized indicators to reflect most of the information of the original variables. While ensuring less loss of the original information, the data and workload are simplified, some of the interfering terms are excluded, and the intrinsic laws are essentially revealed [5-7].

Let the number of samples examined be n , where there are m = 14 random variables $\xi_1,\xi_2,...,\xi_m$ form a matrix of size n*m .

$$Y = XU^{\sim} \tag{1}$$

Using principal component analysis, construct new variables $\eta_1,\eta_2,...,\eta_p$, where p < m and $\eta_1,\eta_2,...,\eta_p$ are independent of each other. The expressions are.

$$\begin{cases} \eta_1 = u_{11}\xi_1 + u_{21}\xi_2 + \cdots + u_{m1}\xi_m \\ \eta_2 = u_{12}\xi_1 + u_{22}\xi_2 + \cdots + u_{m2}\xi_m \\ \vdots \\ \eta_p = u_{1p}\xi_1 + u_{2p}\xi_2 + \cdots + u_{mp}\xi_m \end{cases} \tag{2}$$

The sample means with the following variances.

$$x_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)}(j = 1, 2, \cdots, m) \tag{3}$$

By normalizing the data, the normalized matrix is obtained as X˜, where

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots m) \tag{4}$$

The sample correlation matrix R is

$$R = (r_{ij})_{mm} = \frac{1}{n-1} X^T X \tag{5}$$

The diagonal array Γ consisting of the eigenvalues of R_η and the original correlation matrix R from the eigenvalue decomposition of the correlation matrix is related as follows.

$$R_\eta = \frac{1}{n-1} Y^T Y = \Gamma = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix} \tag{6}$$

The eigenvectors of the standard orthogonalization of the correlation matrix R, after the side-by-side columnwise formation of the principal component loading matrix U, combined with equation (7), can be obtained

$$\frac{1}{n-1} Y^T Y = \frac{1}{n-1} U^T \tilde{X}^T \tilde{X} U = U^T R U = U^T U \Gamma U^T U = \Gamma \tag{7}$$

In turn, we can find η1, η2, ...,, ηm, and use its eigenvalues λ1, λ2, ... , λm to calculate the variance contribution a_i:

$$a_i = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_m} \tag{8}$$

Since the variables η1, η2, ..., ηm are independent of each other and the eigenvalues of the variables corresponding to their contribution sizes decrease in order, the cumulative variance contribution of the first j principal components can be set according to their eigenvalue sizes as

$$G(j) = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_m} \tag{9}$$

With this equation, the weight of the total sample information contained in the first j principal components selected is given. We select two principal components with eigenvalues greater than or equal to 0.8, which are considered to contain a large amount of information about the indicators of the original sample variables. It is tested that the cumulative variance contribution values of the selected principal components are all located above 70%, which reduces the workload and improves the efficiency of analyzing the actual problem at the same time.

The importance of each principal component is reflected by using a gravel plot of the chemical composition variables of glass products (the principal components are listed in descending order by the size of the corresponding eigenvalues).

As shown in Fig. 1, the trend of change decreases sharply between point 1 and point 2, and then the trend decreases rapidly, so the extracted principal components are reasonable.

The coefficients of the principal component scores were calculated by regression method, and the variables were recorded as x_1, x_2, x_3, x_4 in the order of Na_2O, K_2O, PbO and BaO, and the principal components were recorded as F_1F_2. The relationship is obtained as shown in equation (10).

$$\begin{cases} F_1 = -0.1048x_1 - 0.6132x_2 + 0.5999x_3 + 0.5031x_4 \\ F_2 = 0.9604x_1 - 0.0601x_2 - 0.1044x_3 + 0.2513x_4 \end{cases} \quad (10)$$

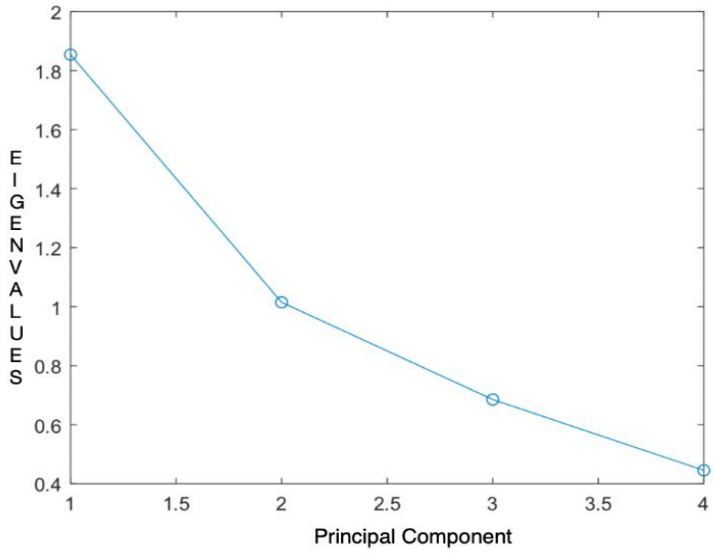

Figure 1. Gravel map

(2) K-means clustering

The standardized data of chemical composition of each glass product were brought into the principal component expression, and the principal component scores of each sample were calculated as shown in Table I.

Table 1. Principal component scores of the sample (partial)

Sample	Components 1	Components 2
1	-2.45405	-0.6839
2	0.417214	-0.79913
3	-1.6876	-0.61082
4	-2.61613	-0.64302
5	-2.40347	-0.67894
6	-2.60581	-0.69878
7	-1.95133	-0.60318
8	-2.02021	-0.62102
9	-0.87487	-0.52903
...	...	...
59	0.703393	-0.80377
60	1.021522	0.183546
61	-0.12819	1.430034
62	1.199901	-0.62077
63	0.922271	-0.84188
64	0.440086	1.105976
65	1.315909	-0.28888
66	1.544747	-0.25429
67	0.738525	-0.51637

K-mean clustering algorithm (fast clustering method), which is based on the idea that K clustering centers are randomly selected, the sample is divided into K clusters by the principle of minimum Euclidean distance, and the centroids of each cluster are iteratively updated until the result is satisfactory, using the global error function as the standard function (as shown in equation (11))[8-9].

$$E = \sum_{i=1}^{k} \sum_{x \in D_i} (x - c_i)^2 \tag{11}$$

Where: E is the error value, k is the number of clusters, each class of clusters is Di, ci is its centroid, and the data set is x.

Using the principal component score, instead of the initial data, the data to be classified were clustered into 2 clusters by the K-means clustering method, and the matching rate was up to 100% when compared with the original glass classification case. Thus, the expression (10) obtained by using principal component analysis can be used to represent the classification pattern of both glasses.

4. Subclass classification model building

4.1. Model Preparation

According to the previous questions: from the gray prediction of the first question, we know that SiO_2 and CaO are the substances that mainly distinguish the weathering of glass; CuO and Fe_2O_3 are the substances that show color; from the first question of the second question, we know that Na_2O, K_2O, BaO and PbO are the main components that distinguish high potassium glass from lead-barium glass. Besides, SO_2 and SnO_2 are too few samples and too little content. Since the question requires subclasses based on chemical composition, the chemical components related to weathering degree, color development, main class division and low amount are not considered here.

The remaining four chemical classes: MgO, P_2O_5, Al_2O_3 and SrO, combined with the relevant thesis materials and chemical knowledge, MgO and SrO are easily formed under alkaline conditions, P_2O_5 is easily formed under acidic conditions, and Al_2O_3 may be generated in both environments, so we tentatively believe that the subclass division is related to acidity and alkalinity.

4.2. Model Building

The K-mean clustering-based principal component analysis was used again for the remaining four categories of high potassium glass and barium lead glass, and the K-value was set to 3 (initially conceived as "alkaline, neutral, and acidic"). The cumulative variance contribution of G(j) was kept above 70%, and one principal component was calculated for high potassium glass (see equation (12) for the relationship with the original variables), and two for lead-barium glass (see equation (13) for the relationship with the original variables).

$$F = 0.5042x_1 + 0.4874x_2 + 0.5056x_3 + 0.5026x_4 \tag{12}$$

$$\begin{cases} F_1 = -0.1048x_1 - 0.6132x_2 + 0.5999x_3 + 0.5031x_4 \\ F_2 = 0.9604x_1 - 0.0601x_2 - 0.1044x_3 + 0.2513x_4 \end{cases} \tag{13}$$

As a schematic diagram of the explanation coefficients of the chemical components to the principal components is shown in Fig. 2 and Fig. 3, it can be seen that the sum of the contributions of the alkaline substances MgO and SrO is much larger than that of the other properties, so the average content of the alkaline substances is used to define the three categories of the well-sorted substances. The classification was obtained using K-means clustering processing as shown in Table II.

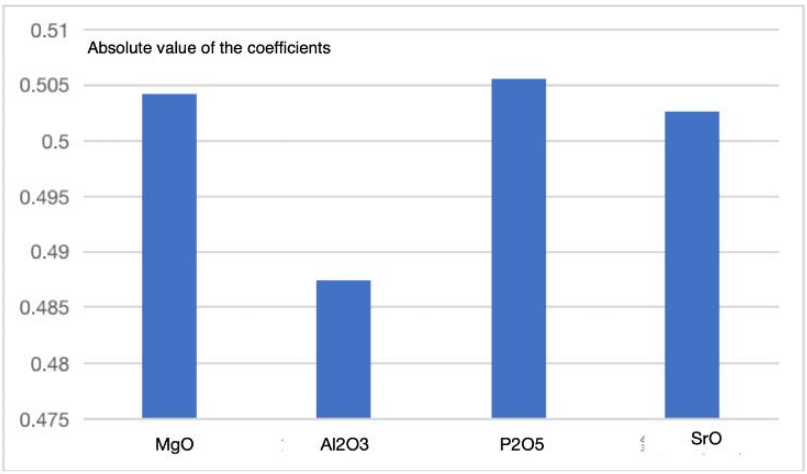

Figure 2. Diagram of high potassium coefficient

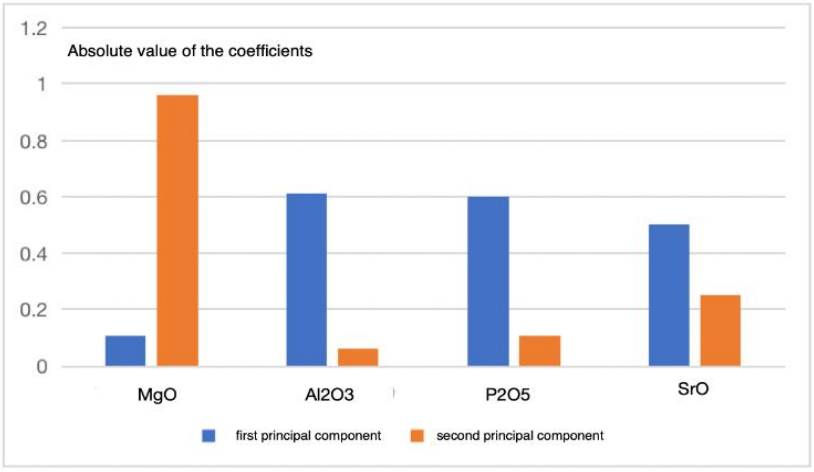

Figure 3. Diagram of lead-barium coefficient

Table 2. Classification results by principal component analysis based on K-means clustering

	Category	Number
High Potassium Glass	Category I (Acidic)	01,031,07,09,10,12,13,16,22,27
	Category II (Neutral)	61,062
	Category III (Alkaline)	032,04,05,14,18,21
Lead barium glass	Category I (Acidic)	02,28*,29*,311,33,42*1,42*2,44*,45,46,48,49*,53*
	Category II (Neutral)	11,19,41,432,49,50,511,512,54, 54!, 58
	Category III (Alkaline)	8, 8!, 20,23*,24,25*,26,26!,302,31,32,34 − 40,431,47,50 *,52,55 − 57

Note: The numerical footnotes are part numbers, ! indicates a severely weathered site, * is an unweathered site.

5. Model Testing

5.1. Rational analysis

To analyze the rationality of the subclass classification results of this question, firstly, the K-means clustering results are evaluated using the contour coefficient, which is based on the following principle: in a set of samples, let c be the average distance between samples of the same class and d be the average distance between samples of different classes, and the formula for calculating the contour coefficient is shown in (14).

$$S = \frac{d - c}{\max(c, d)}, S \in [-1,1] \tag{14}$$

The closer the samples of the same class and the more distant the samples of different classes are, the higher the score of the classification method. The resulting contour coefficients are plotted in Fig. 4 and Fig.5.

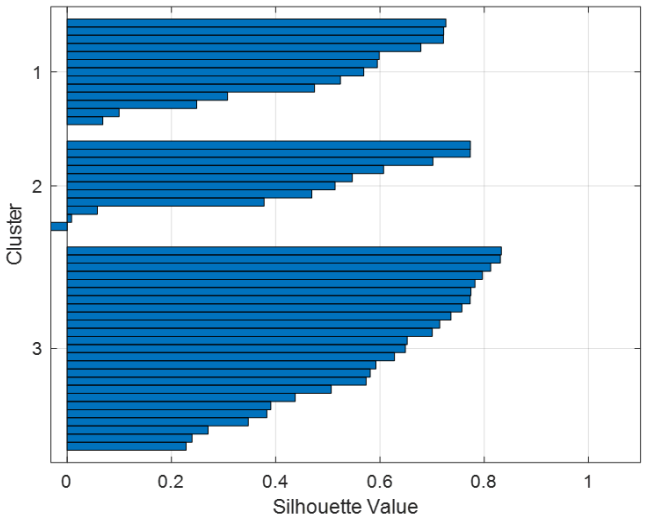

Figure 4. Diagram of lead-barium contour coefficient

The higher the average contour coefficient, the better the quality of clustering. The analysis of the contour coefficients shows that the classification of the subclasses in the second subquestion is reasonable.

Based on the sub-classification results in the second subquestion, a random forest-based test model is established. Random Forest is a common Bagging model, which is based on the principle of "random data" and "random features", and randomly selects sample points from the original data set to obtain n sample data sets (different from each other) [10]. Then we build the corresponding decision tree model based on the formed dataset, and obtain the final result according to the mean voting situation. The series of lead-barium glass chemical composition data were selected as the test samples, of which the first 6 data sets basically contained all the weathering degree and classification cases as the sensitivity samples for the final detection; the latter 43 data sets classification cases were used as the training samples for the random Sen collar model. When the number of decision trees reached 100, the error converged to about 0.15 and the model training was finished (as shown in Fig.6).

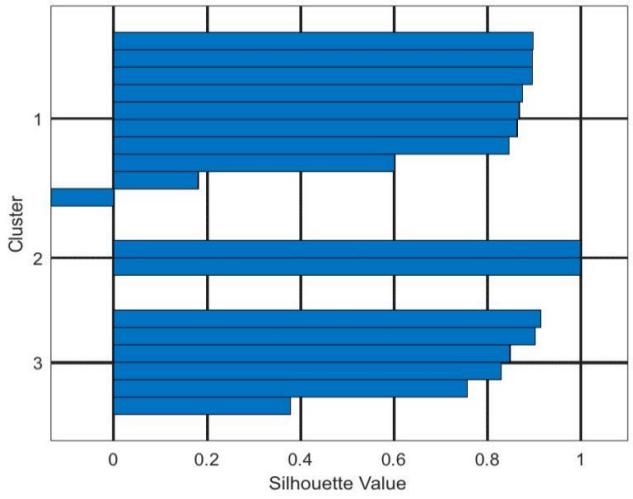

Figure 5. Diagram of high potassium contour coefficient

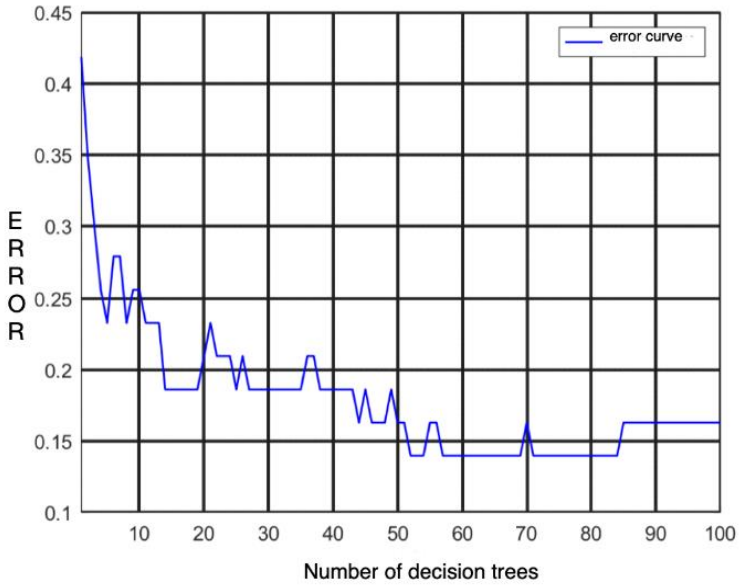

Figure 6. Random forest training error values

The first six sets of data were tested with the subclassification results of the second subquestion as the true classification results, as shown in Fig. 7. The obtained 100% of the test classification results are consistent with the real classification, which proves that the subclassification pattern is stable and reasonable, and is satisfactory.

5.2. Sensitivity Analysis

For the subclass classification, we took the chemical components related to the subclass classification (MgO, P_2O_5, Al_2O_3 and SrO) in the unknown samples, and according to the absolute value of their principal component coefficients, we obtained the chemical component with the strongest explanatory effect on the classification of high potassium glass as K_2O, and the chemical component with the strongest explanatory effect on the classification of lead-barium glass as PbO, and changed the data of these two groups, i.e., increased the deflection, and changed the values are 0.9, 0.95, 1.05, 1.1 times the original values. The classification results were obtained again using the principal component analysis model based on K-means clustering. The analysis revealed that the correctness of the obtained results is shown in Fig. 8 and Fig. 9, which proves that the sensitivity of the classification law of the subclasses obtained in the second subquestion is weak and the classification model is more stable than the sex.

6. Model Evaluation, Improvement and Extension

6.1. Model Evaluation

The combination of principal component analysis and K-means clustering method improves the efficiency of analyzing the actual problem on the basis of ensuring the majority of information of the original data, and the clustering using principal component scores reduces the variability of data and makes the classification results more reliable. The random forest algorithm is further used to mine the data to ensure the accuracy of the classification results and the applicability of the model.

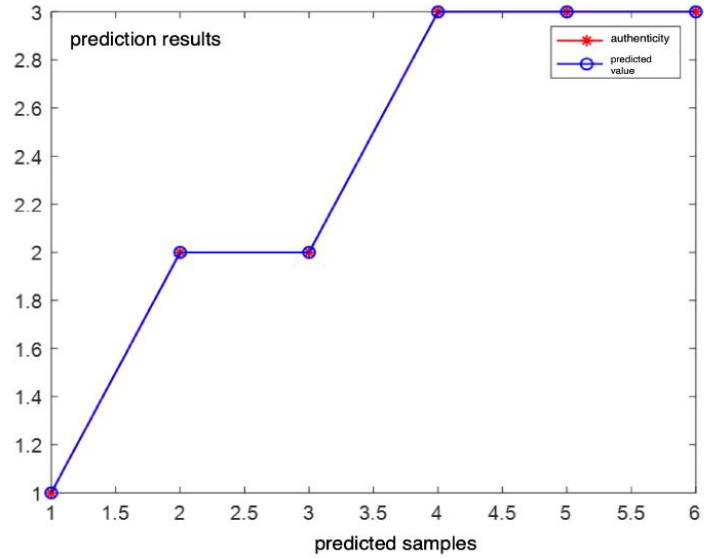

Figure 7. Random forest test results graph

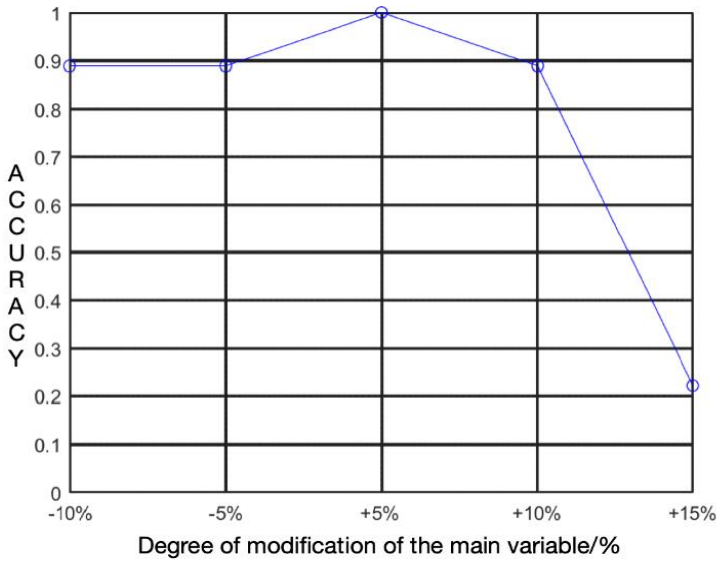

Figure 8. Diagram of lead-barium contour coefficient

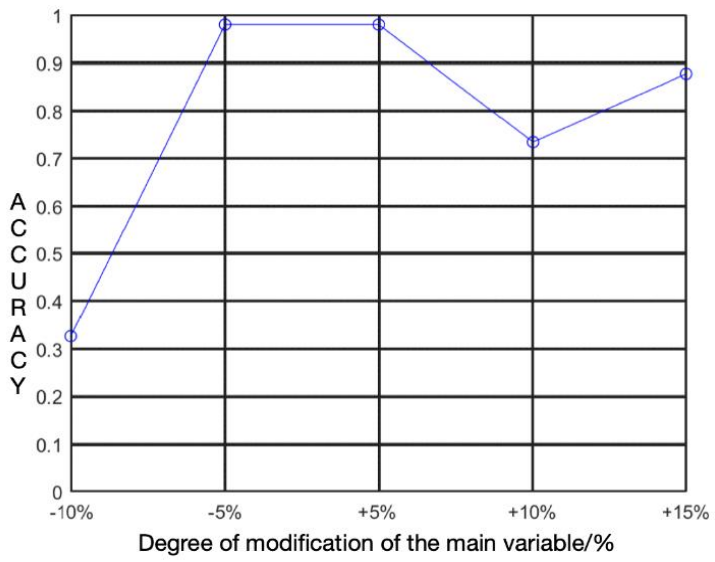

Figure 9. Diagram of high potassium contour coefficient

6.2. Model improvement and extension

The use of mathematical methods to analyze and classify ancient artifacts has become a research trend with the continuous development of science and technology. It is of great practical importance to use statistical methods to change the study and classification of ancient glass-making techniques from relying on superficial characteristics such as color and texture to using modern technology for classification and identification. Further research is needed on the analysis of trace components and the influence of chemical reactions on the test results of the chemical components contained in glass products.

7. Conclusion

In recent years, with the widespread application of glass materials in various industries, classifying them has become an important research direction. However, traditional classification methods require a lot of manual experience and time, resulting in low efficiency and inconsistent results, making it difficult to meet practical needs. Therefore, machine learning has become an effective classification method for glass materials. Conducting machine learning research on glass material classification can improve classification accuracy, improve classification efficiency, promote progress in glass material research, and enhance the safety of glass products. In summary, the importance of machine learning in glass material classification research is self-evident. With the development of technology and artificial intelligence, machine learning will become the mainstream technology for glass material classification, bringing more refined and safe glass products to various industries.

References

[1] L. Dussubieux et al, “The trading of ancient glass beads: new analytical data from south asian and east african soda-alumina glass beans,” Archaeometry, vol. 50, (5), pp. 797-821, 2008.

[2] Y. Guo, T. Wu and IOP, "Analysis on the development of ancient glass technology in the Central Plains of China in the pre Qin Period," 2020 6th International Conference on Energy, Environment and Materials Science, vol. 585, (1), pp. 12206, 2020. Cancel

[3] N. Welter, U. Schüssler and W. Kiefer, “Characterisation of inorganic pigments in ancient glass beads by means of Raman microspectroscopy, microprobe analysis and X-ray diffractometry,” Journal of Raman Spectroscopy, vol. 38, (1), pp. 113-121, 2007.

[4] S. Saminpanya et al, “Shedding New Light on Ancient Glass Beads by Synchrotron, SEM-EDS, and Raman Spectroscopy Techniques,” Scientific Reports, vol. 9, (1), pp. 16069-12, 2019.

[5] A. Gallo et al, "Use of principal component analysis to classify forages and predict their calculated energy content," Animal (Cambridge, England), vol. 7, (6), pp. 930-939, 2013.

[6] N. O’Rourke, L. Hatcher and E. J. Stepanski, “A Step-by-step approach to using SASR for univariate and multivariate statistics,” 2nd edition. SAS Institute Inc., SAS Campus Drive, Cary, North Carolina, USA. 2005.

[7] J. P. Stevens, “Applied Multivariate Statistics for the Social Sciences”, 5th edition. Routledge, Taylor & Francis Group, New York, USA. 2009.

[8] S. Lloyd, “Least squares quantization in PCM,” IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[9] J. A. Hartigan, Clustering Algorithms. New York, NY, USA: Wiley, 1975.

[10] C. Strobl et al, "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics, vol. 8, (1), pp. 25-25, 2007.