

Component Identification and Prediction of Ancient Glass Products Based on Decision Tree Model and SVM Model

Jin Chen ^{*,#}, Ruoyi Jia [#], Yaxin Yan [#]

College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, Sichuan, 610225

* Corresponding Author Email: 19136319921@163.com

[#]These authors contributed equally.

Abstract. The study of the patterns of chemical composition of glass objects is an important research method of classifying ancient glass objects. In this paper, we firstly selected suitable indicators and sought classification boundary lines through a decision tree model to discover the classification pattern of ancient glass and verify its accuracy. Secondly, it was sub-classified and divided by means of hierarchical clustering and k-means mean clustering to realise the work of categorisation of known components. Finally an SVM model was built to obtain a confusion matrix map to achieve the classification prediction of unknown components.

Keywords: Decision Trees, K-Means Clustering, SVM

1. Introduction

The chemical composition of ancient glass in China is somewhat different from that of exotic glass, and the study of the composition of ancient glass can help people to deepen their understanding of glass, which requires us to establish a set of effective classification and prediction schemes [1-2], taking into account many factors, such as weathering effects, composition and distribution patterns. The question focuses on mathematical modelling to develop relevant classification models and predictions for the classification of unknown compositions.

In this paper, based on the known data of high potassium glass and lead-barium glass, the correlation analysis of various compositions is carried out, and then the changes of elements and classification laws after weathering of glass are found through model induction, and suitable chemical compositions are selected for subclass classification and specific classification schemes are given. On this basis, the method of how to make classification predictions for unknown elements is deeply explored, and the analysis of the chemical composition of unknown categories of glass artefacts and the prediction of the types to which they belong are completed.

2. Model building and solving

2.1. Data analysis

Based on the available data we selected different decorations, colors and glass types as indicators for the study, and then used Spearman's correlation coefficient to analyse their correlation. As all the variables in Annex Table 1 are of a definite class, they were first coded for unique heat before the analysis, and the codes are shown in Table 1

Table 1. Annex Form 1 Solo Heat Code Table.

Indicators	Coding	Indicators	Coding
A	001	Blue-green	000000001
B	010	Light blue	000000010
C	100	Black	000000100
Indicators	Coding	Indicators	Coding
		Green	000001000
High Potassium Glass	01	Light green	000010000
Barium Lead Glass	10	Dark blue	000100000
		Dark green	001000000
Weathering	01	Purple	010000000

According to the Spearman correlation coefficient calculation formula;

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (1)$$

The correlation coefficient plot was calculated by MATLAB programming (see Figure 1),

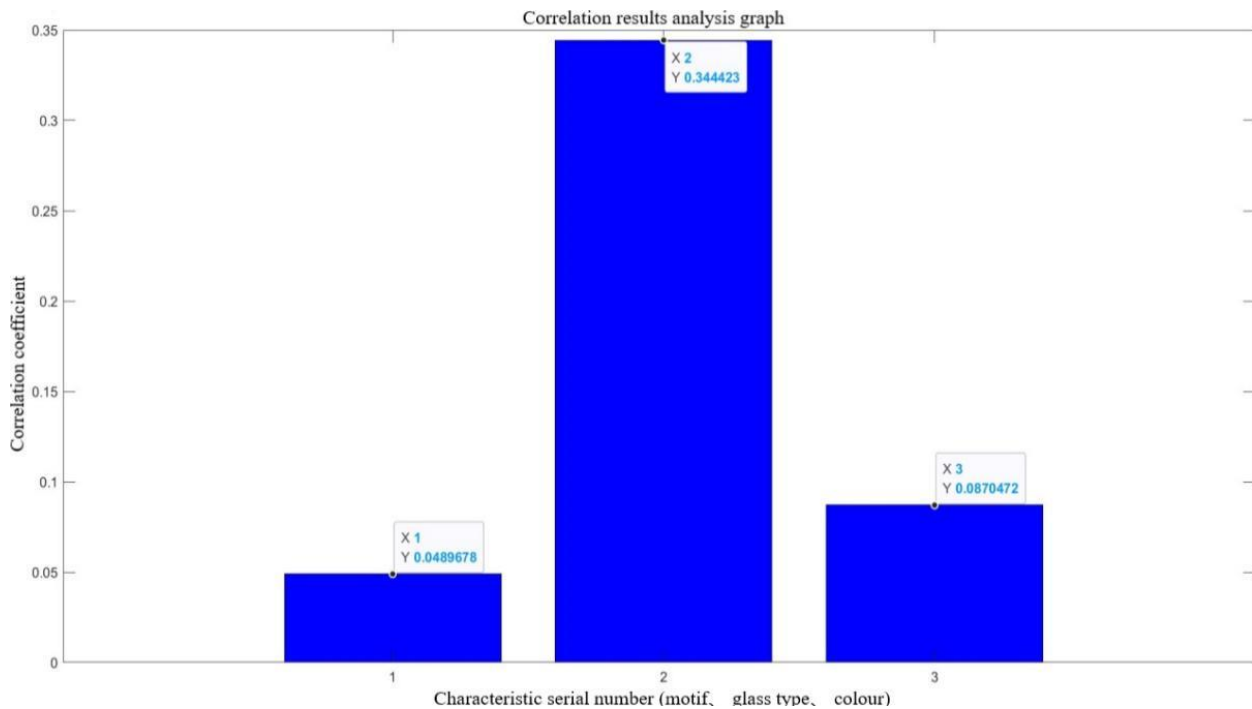


Figure 1. Correlation coefficient graph.

The Kolmogorov-Smirnov test yields a return value result of {Graining: 1, Glass type: 0, Colour: 1},

Ornamentation on surface weathering rejects the original hypothesis at the default 5% significance level.

Glass type on surface weathering accepts the original hypothesis at the default 5% significance level.

Colour rejects the original hypothesis for surface weathering at the default 5% significance level.

2.2. Classification laws based on decision tree models

Firstly, the data onto high-potassium lead-barium glass was organized according to the meaning of the topic, and for the classification law, the decision tree model [3] was used to find the cut-off line of the classification, and the following mapping relationship was established in the preparation work.

{high potassium glass: 0, lead-barium glass: 1}

{silica (SiO₂): x1, sodium oxide (Na₂O): x2, potassium oxide (K₂O): x3, calcium oxide (CaO): x4, magnesium oxide (MgO): x5, aluminum oxide (Al₂O₃): x6, Iron oxide (Fe₂O₃): x7, copper oxide (CuO): x8, lead oxide (PbO): x9, barium oxide (BaO): x10 phosphorus pentoxide (P₂O₅): x11, strontium oxide (SrO): x12, tin oxide (SnO₂): x13, sulphur dioxide (SO₂): x14}.

The classification tree viewer obtained shows the classification pattern as shown in Figure 2:

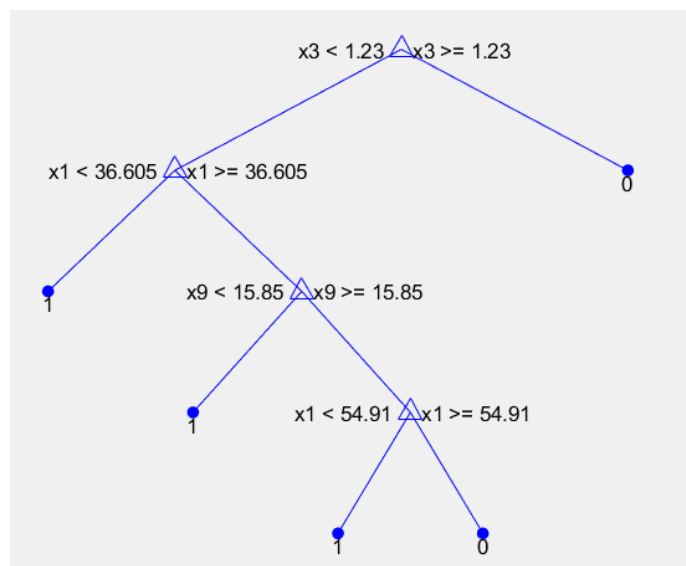


Figure 2. Classification tree viewer.

The classification results based on this decision tree model [4] were also validated and found to be nearly 90% accurate (see Figures 3 and 4), so that the classification law elaborated in terms of the decision number model is reasonable and reliable.

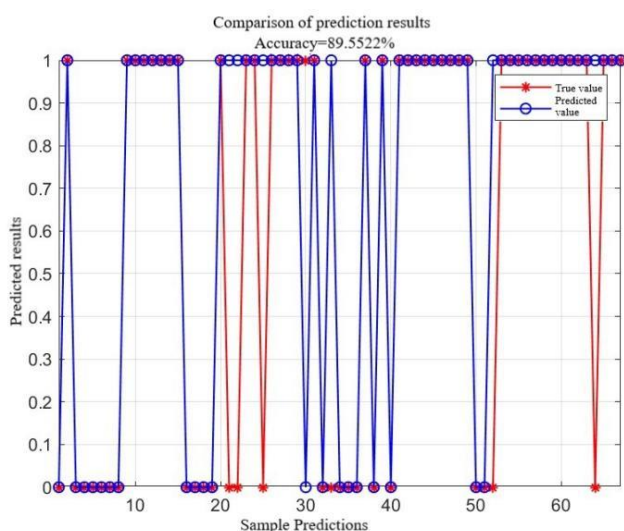


Figure 3. Comparison of forecast results.

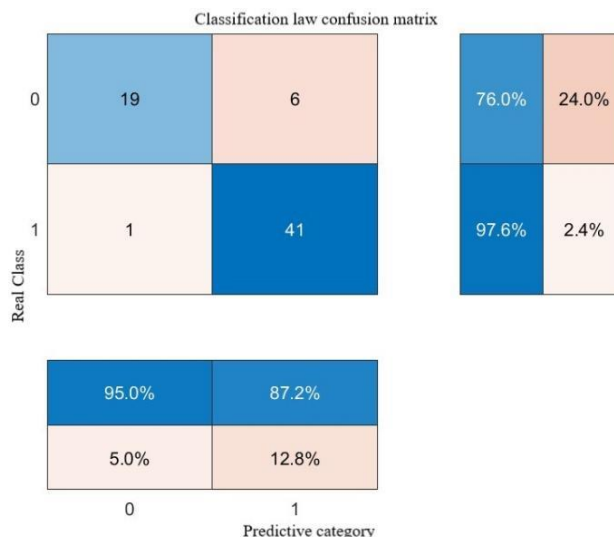


Figure 4. Confusion matrix of classification laws.

2.3. Subclass classification based on clustering methods

BP neural network is back propagating, mainly composed of three parts: input layer, middle layer and output layer. The number of nodes in the input and output layers is relatively easy to determine, but the determination of the number of nodes in the hidden layer is a very important and complex problem.

2.3.1. Hierarchical clustering

For the subclassification of the two types of glass we use hierarchical clustering [5-6], which is a common clustering algorithm. The basic steps are as follows:

Step1 The elements to be clustered are each viewed as a class.

Step2 select the two classes with the smallest distance to merge into a new class, and calculate the distance between the new class and the other classes;

Step3 merge the two closest classes until all the samples have been merged into one class.

This successive merging process can be represented in a tree diagram to give the final clustering situation.

In this way, the clustering results can be obtained for high potassium glass and lead-barium glass respectively.

2.3.2. k-means clustering

Through hierarchical clustering, we find that high potassium glass can be clustered into two categories, and lead barium glass can be clustered into three categories. Therefore, k-means clustering method [7-8] is selected. The main steps of k-means clustering method are:

Step 1 First randomly select k initial cluster centers C_i ($1 \leq i \leq k$) from the dataset);

Step 2 Calculate the Euclidean distance between the remaining data objects and the cluster center C_i ;

Step 3 Find the nearest cluster center C_i to the target data object and assign the data object to the cluster corresponding to the cluster center C_i ;

Step 4 Calculate the average of the data objects in each cluster as a new cluster center for the next iteration until the cluster center does not change or the maximum number of iterations stops.

By solving the algorithm, we can obtain the clustering results of high-potassium glass and lead-barium glass, including the average value of each category under the influence of standard deviation, as well as F value and P value, and summarize them into field difference tables. Based on this, we get the pie chart of clustering analysis results.

The clustering result statistical pie charts are shown in Figure 5 and Figure 6.

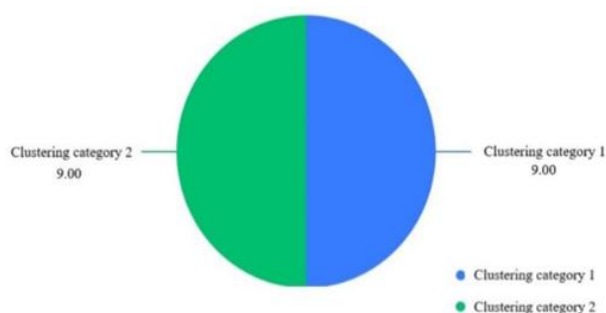


Figure 5. Statistical pie chart of clustering results for high potassium glass.

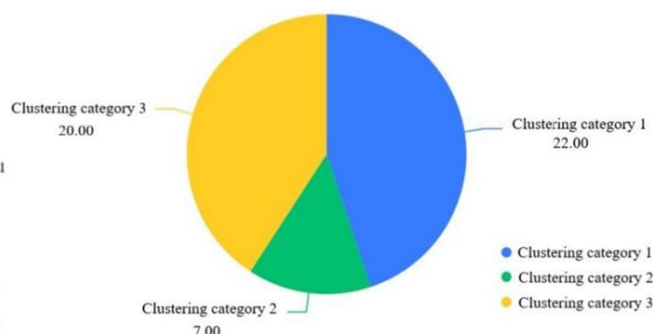


Figure 6. Statistical pie chart of clustering results for lead barium glass.

2.4. Classification prediction based on SVM model

2.4.1. Preparation of the SVM model

Before modelling, some processing of the data was required. The valid data was filtered by combining the attached Form 1 and Form 2 and using the type as the primary key as the training test set, and Form 3 was created with a new primary key type and used as the prediction set. For better classification, it was also noted that the surface weathering column was also given in Form 3, so surface weathering was also added as an element to the classification prediction, where the surface weathering column was given the mapping {no weathering: 0, weathering 1}.

Before selecting the model, a simple analysis of the data was carried out and it was found that the data sample size was small and the data elements were large, so a support vector machine (SVM for

short) classifier model [9-11] was used to classify and predict the data for the purpose of classification and regression of the small sample data.

Support vector machines (SVMs) are a class of generalized linear classifiers that binary classify data in a supervised learning fashion, with a decision boundary that is a maximum margin hyperplane solved for the learned samples.

The mathematical theory of its model is as follows:

What the SVM wants is to find the furthest distance from each type of sample point to the hyperplane, i. e. to find the maximum interval hyperplane.

The arbitrary hyperplane can be described by the following linear equation.

$$\omega^T x + b = 0 \quad (2)$$

Establish the optimization problem as follows:

$$\min_{\omega, b} \left(\frac{1}{2} \omega^T \times \omega + C \sum_{i=1}^N \xi_i \right) \quad (3)$$

$$s.t. \begin{cases} y_i (\omega^T \cdot x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, 2, \dots, N \end{cases} \quad (4)$$

The formula: is the normal vector; b is the constant term; C is the penalty factor; ξ_i is the relaxation variable.

The optimal classification surface can be obtained by finding the solution of the optimal normal vector ω and the constant term b . In order to transform the above equation into a quadratic programming problem, the corresponding Lagrange function is introduced and the classification problem becomes:

$$L(\omega, b, \lambda) = \frac{1}{2} \omega^T \cdot \omega + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i (\omega^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^N \beta_i \xi_i \quad (5)$$

The formula: α_i, β_i are Lagrange multipliers, $\alpha_i \geq 0, \beta_i \geq 0$

According to the principle of duality, the above equation can become:

$$\max_{\alpha} L(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^N \alpha_i \alpha_j y_i y_j (\phi_{xi} \times \phi_{xj}) \quad (6)$$

$$s.t. \sum_{i=1}^N y_i \alpha_i = 0, 0 \leq \alpha_i \leq C \quad (7)$$

$$\phi_{xi} \times \phi_{xj} = K(x_i, y_j) \text{ is a kernel function} \quad (8)$$

For linearly indistinguishable problems, support vector machines need to introduce kernel functions to project data from low-dimensional space to higher-dimensional space to achieve linearly divisible purposes. Commonly used kernel functions are linear kernel functions, polynomial kernel functions, radial basis kernel functions and perceptron kernel functions, and different kernel functions yield different forms of support vector machines.

2.4.2. Solution of the SVM model

After preparing the SVM classifier model, we start to model the SVM model to solve it.

For the SVM model solving steps are as follows.:

Step1 prepare the training test data, shuffle the data and divide the training set and test set according to the ratio of 7:3;

Step2 select the linear kernel, the penalty coefficient is set to 1, the kernel function is scale, and the radial basis function parameter is set to 0.01;

Step3 conduct training and testing, and analyses the training and testing results;

Step4 bring the data to be predicted into the model for classification prediction, and get the classification prediction results.

From the results of the training test, the confusion matrix plot (see Figure 7 and Figure 8) through the training set test set shows that:

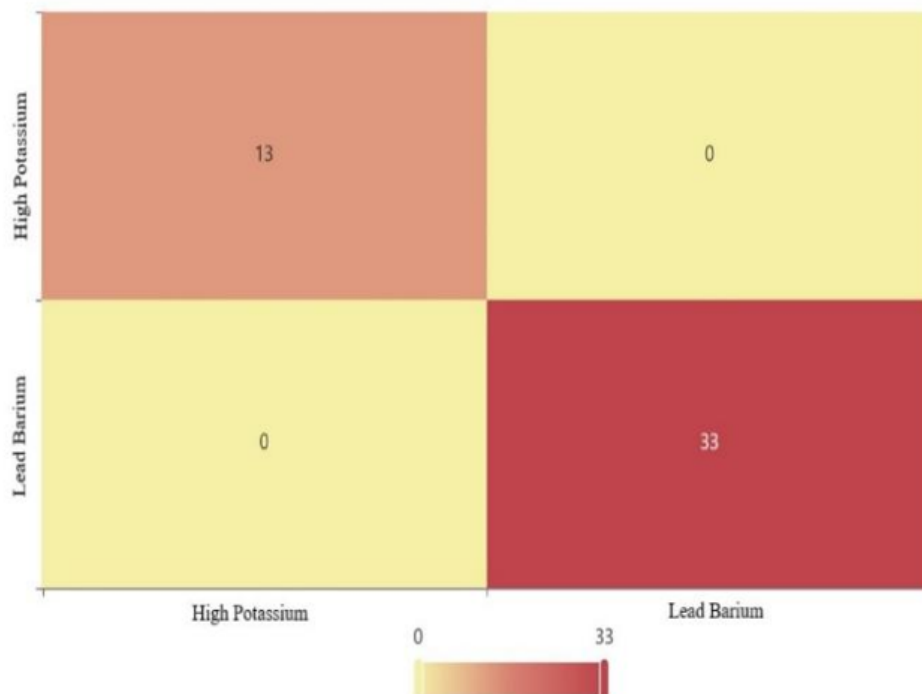


Figure 7. Confusion matrix of the raining set.

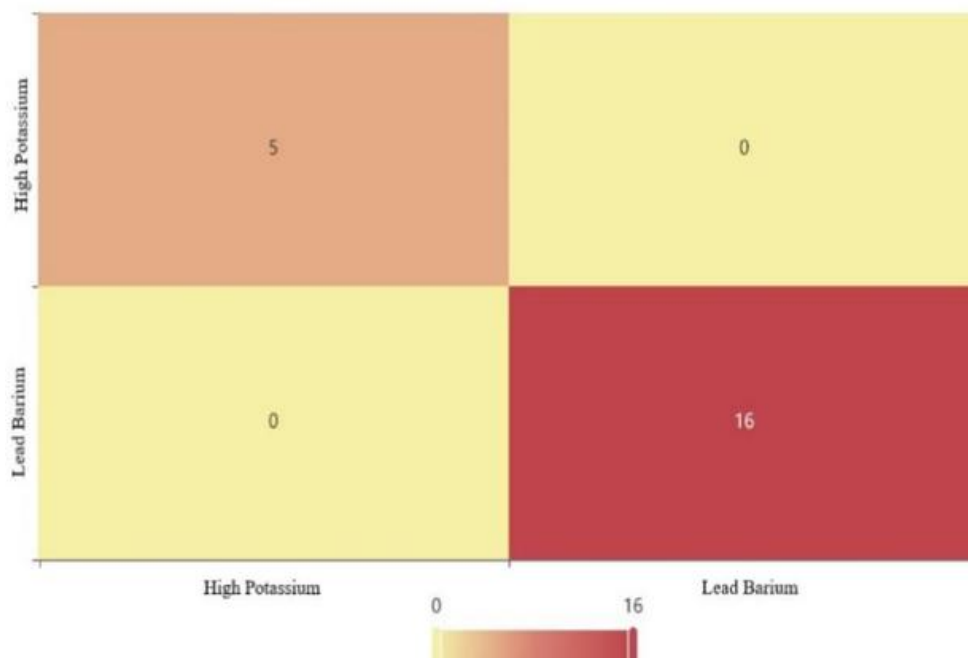


Figure 8. Test set confusion matrix.

As can be seen from the results, the training results are more than satisfactory for the expected balls and the results table for the model evaluation is obtained as shown in Table 2.

Table 2. Table of results of the model evaluation.

	Accuracy	Recall rate	F1
Training set	1	1	1
Test set	1	1	1

For this purpose, the predicted data to be predicted was brought into the model and the predicted results were obtained as shown in Table 3,

Table 3. Table of predicted results.

Predicted results -Y	Predicted Outcome Probability - Lead Barium	Predicted outcome probability - high potassium
High Potassium	0.11842828575478606	0.8815717142452141
Lead Barium	0.9943101251710296	0.005689874828970424
Lead Barium	0.9983225522802769	0.001677447719723093
Lead Barium	0.9794814421880564	0.020518557811943446
Lead Barium	0.7002219568704884	0.2997780431295115
High Potassium	0.0983016514209515	0.9016983485790483
High Potassium	0.10369728817632334	0.8963027118236767
Lead Barium	0.9552800486277604	0.04471995137223955

3. Results

By comparing the results of the hierarchical clustering of surface weathering data and the hierarchical clustering of weathering point data indicated that the classification accuracy of both high potassium glass and lead barium glass was above 95%, for this reason this classification division was considered reasonable, while the high accuracy could still be reached after washing the data, indicating good sensitivity.

As for the sensitivity of the classification results, as mentioned above, the data was shuffled before the training test and the training test has a certain randomness, for this reason, we retrained the model several times to get better results, so the classification prediction model is not affected by the sensitivity element.

4. Model evaluation and extension

In this paper, a decision tree model was first used to find the dividing line of the classification, and a mapping relationship was established by the data of high potassium and lead barium glass, and a classification tree viewer was obtained to show the classification laws. The advantage is that the results are very intuitive and concise, the data preparation is simple, and it is suitable for analysing the classification laws of high potassium and lead barium glass. The disadvantage is that it is only suitable for problems with a small number of variables and is not capable of handling missing values.

Also, in performing subclassification, we used hierarchical clustering and k-means mean clustering, from the results it can be seen that the clustering algorithm has low time complexity and space complexity, excellent rationality and sensitivity, but is sensitive to outliers and cannot solve non-convex data.

For classification prediction, this paper uses the SVM support vector machine model for classification prediction, which has better generalization performance and is less prone to overfitting, and can achieve good performance with less data, although it is difficult to implement in large-scale training samples.

The paper selects suitable indicators to establish several models such as decision trees, k-means mean clustering and SVM, and integrates factors such as whether or not they have undergone weathering to solve the classification problem of ancient glass objects in depth, and quantitatively studies the statistical patterns of chemical composition in glass objects, which provides implications for studies related to the identification of the type of unknown cultural objects to which they belong.

5. Conclusions

The trend of mass data in power system provides a basis for load characteristic analysis and prediction model establishment, but the classical load forecasting method cannot afford such a huge time and computing resource consumption. The problem of over fitting in large sample set will affect the prediction accuracy. In this paper, a power load forecasting model is built by using the BP neural network model, making full use of the powerful data processing function of Clementine and preventing the over fitting function. The experimental results show that the BP neural network model has good predictability and robustness, and has a certain practical application value.

References

- [1] Zhizhong Hu et al. Application of LA-ICP-MS in the Analysis of Archaeological Glass and Source Discrimination [J]. Rock and Mineral Analysis, 2020, 39(04): 505-514.
- [2] Song Liu, Qinghui Li, Fuxi Gan. Influence of Ancient Glass Samples Surface Conditions on Chemical Composition Analysis Using Portable XRF [J]. Spectroscopy and Spectral Analysis, 2011, 31(07): 1954-1959.
- [3] Deger Zeynep Tuna and Taskin Kaya Gulsen. Glass-box model representation of seismic failure mode prediction for conventional reinforced concrete shear walls [J]. Neural Computing and Applications, 2022, 34(15): 13029-13041.
- [4] Solomon O. Akinola and Oluwatobi I. Raji. Relationship between structural complexity and performance of data mining classification algorithms [J]. Journal of Applied Science, Engineering and Technology, 2017, 17(1): 90-97.
- [5] Carlsson Gunnar et al. Hierarchical clustering of asymmetric networks [J]. Advances in Data Analysis and Classification, 2018, 12(1): 65-105.
- [6] Saumya Singh and Smriti Srivastava. Review of Clustering Techniques in Control System [J]. Procedia Computer Science, 2020, 173(C): 272-280.
- [7] Gong, Jianping et al. GNSS Multipath Mitigation Method Based on K-means Classification in Urban Environment [J]. Journal of the Indian Society of Remote Sensing, 2022: 1-9.
- [8] Mohamed Ouhda et al. A Content Based Image Retrieval Method Based on K-Means Clustering Technique [J]. Journal of Electronic Commerce in Organizations (JECO), 2018, 16(1): 82-96.
- [9] Mingliang Li et al. Study on the Samples of Weathered Red-Bed Mudstone Based on THz-TDS [J]. Spectroscopy and Spectral Analysis, 2016, 36(04): 929-933.
- [10] He Jin xin et al. Automatic Classification of Rock Spectral Features Based on Fusion Learning Model [J]. SPECTROSCOPY AND SPECTRAL ANALYSIS, 2021, 41(1): 141-144.
- [11] Andrade Renata et al. Micronutrients prediction via pXRF spectrometry in Brazil: Influence of weathering degree [J]. Geoderma Regional, 2021, e00431-