Decision tree based classification of glass artefact types

Hai Yu *

School of Chemistry and Chemical Engineering, Tianjin University of Technology, Tianjin, China

* Corresponding Author Email: yuhai710@126.com

Abstract. This paper uses logistic regression, SVM, and decision trees in machine learning to analyse 67 data items from question C of the 2022 GCSE Cup National Student Mathematical Modelling Competition. The data were analysed using systematic clustering, the clustering coefficients were analysed, the number of clusters K was further determined by the "elbow" method, a clear classification pattern between the glasses was obtained, and a grid search method was used to classify the glasses. The results show that the new excavated glass artefacts have been classified using the "elbow" method. The results show that newly excavated glass artefacts are classified by varying PbO content. If the lead oxide (PbO) content is below 5.46%, they are considered high-potassium glasses and vice versa for lead-barium drinks. The content of silica was further used as a boundary to divide the high-potassium glass into two subclasses. Lead-barium glasses are divided into three subclasses using the lead oxide and silica content as the boundary. Using a series of models and algorithms, the classification patterns of different types of glass and their subclasses were clarified, and the results were tested for reasonableness and sensitivity. Such a model can be used to classify newly excavated glass artefacts and can also be modified based on this model for the identification and analysis of other ancient artefacts.

Keywords: systematic clustering, decision trees, heritage type classification, sensitivity analysis.

1. Introduction

The main component of glass is silica, and the composition of glass can change due to differences in the fluxes added during the forging process. The compositional content of glass excavated in different parts of China can vary, and these ancient glasses mined in other regions can weather due to various external factors [1-4]. Their corresponding compositional content can change significantly before and after weathering. This paper uses decision trees and systematic clustering methods in machine learning to analyse the two types of glass available, intending to classify the excavated artefacts utilising a range of data from the existing excavations and, on this basis, to further investigate and explore the components of the excavated artefacts that have variability.

2. Materials and methods

2.1. Data

The data in this paper are from Question C of the 2022 Higher Education Society Cup National Mathematical Contest in Modeling for College Students. Part of its data is shown in Table 1:

Table 1. Data

Cultural relic number	Type	Silicon dioxide	Sodium oxide	…	Strontium oxide	Tin oxide	Sulfur dioxide
01	High potassium	69.33		…			0.39
02	Lead barium	36.28		…	0.19		
…	…	…	…	…	…	…	…
55	Lead barium	49.01	2.71	…			
56	Lead barium	29.15		…			
57	Lead barium	25.42		…			

2.2. Method introduction

Logistic regression [5]: Logistic regression is a generalized linear regression analysis model. It is a supervised learning in machine learning, mainly solving binary and multi-classification problems. The principle is to train the model through a given n set of data as a training set and classify the given one or more sets of data as a test set after the training. Each of these sets of data is made up of p indicators.

Decision tree [6]: A decision tree is a very common classification method. It is supervised learning in machine learning. Decision tree calculation speed is fast, does not make statistical assumptions, and can process the data expressed on different measurement scales. The software used to implement the decision tree is readily available on the Internet. [7] Decision trees have relatively simple, explicit, and intuitive classifications. [8]

SVM [9]: Support vector machine is a computer algorithm that assigns labels to assigned objects by learning from examples. [10] is supervised learning in machine learning? Like most machine learning algorithms, SVM is usually applied using pre-classified, randomly selected training sets [11].

Systematic clustering [12]: The principle of systematic clustering is to start with each sample as a class, then cluster the nearest sample (i.e., the group with the smallest distance) into subclasses first, and then merge the aggregated subclasses according to the distance between the classes, and continue, and finally aggregate all subclasses into a major class. System clustering will result in a series of classifications, and a suitable classification needs to be determined.

3. Results

3.1. Problem analysis

Firstly, this paper needs to explore the basis for the classification of high potassium and lead-barium glass, and on this basis, select the appropriate components for the classification of subclasses, and finally make a corresponding test of the results of the classification. Therefore this problem is still dealt with in three parts in this paper.

The first part uses Matlab to classify different categories of data using three supervised learning methods: logistic regression, decision trees and SVM, to select the best method for classification and to use this one method to further find data with representative component changes as the basis for both glass classifications.

The second part builds on the first part by analysing the systematic clustering of different glass types and using the 'elbow' method of clustering coefficients to determine a reasonable number of sub-classifications.

In the third part, the best method from the first part is used in the second part of the solution to find the classification criteria for subclassification, and finally, a sensitivity analysis of the hyperparametric tuning of the model using a grid search method.

3.2. Model assumption

In response to the question, the following assumptions are made in this paper.

(1) Only those components detected in the available data are present in the glass artefacts, i.e. components not detected in the blanks are treated as if they were not present.

(2) Subclassification of glass ignores the influence of colour, decoration and weathering of the surface on the classification results and only considers the type and chemical content.

3.3. Establishment and solution of the model

3.3.1 Data preprocessing

(1) Based on the data form, the presence of invalid data was determined based on the valid data provisions given in the question. After testing, it was found that the proportion of components in

number 15 and number 17 was lower than the valid data provision, so number 15 and number 17 were removed from the subsequent processing, and then 67 valid data were left in the form to be used.

(2) Based on the data form, the data in the blanks are uniformly filled with the value "0" according to the provision that the blanks given in the title indicate that no component was detected.

3.3.2 Establishment of model

For the problem, this paper sub-classifies different glass types. There are very many common algorithms in mathematics for classification, commonly decision trees, k-means, logistic regression, support vector machine SVM, etc. This paper uses Matlab's machine learning toolbox to select three models for initial classification, decision trees, logistic regression, and support vector machine SVM, and select the model with the highest accuracy for sub-classification.

3.3.3 Solving the model

In the first part of this paper, three models are selected to classify glass types using the Matlab machine learning toolbox.

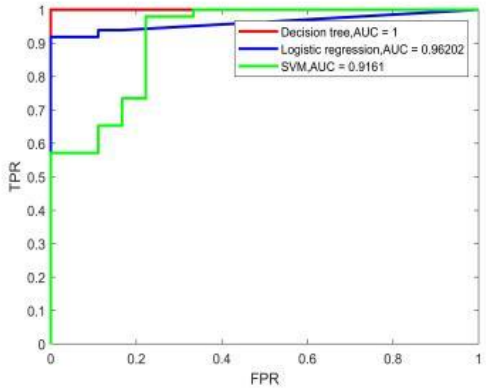


Figure 1. ROC curves of three models

It is generally considered that if the AUC (the area enclosed by the curve and the coordinates below) is less than 0.5, the model cannot be used, and the classification effect is better if the AUC is greater than 0.5. According to the information in Figure 1, all three models have an AUC greater than 0.85, so it is concluded that all three models have a better classification effect, where the decision tree AUC=1 and the ROC curve is closer to the upper left and the area enclosed by the coordinates below is the largest. The decision tree model has a better classification effect than the other two, so the decision tree is chosen for further classification in this paper. The decision tree flow chart is shown in Figure 2 below.

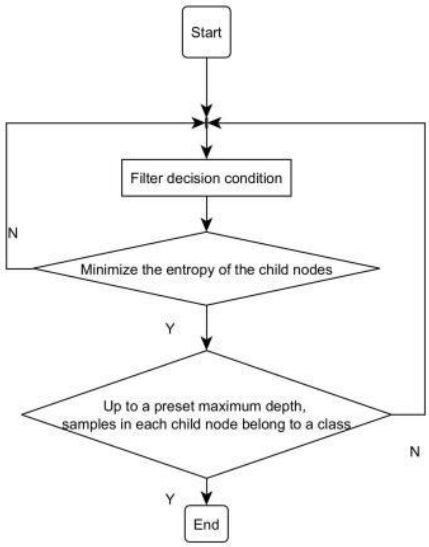


Figure 2. Decision tree flow chart

This paper uses Matlab to write code to implement the type as a definite class variable Y, and the content of each chemical component under each type as a quantitative variable X, and then find the representative component data as the basis for the classification of the two glasses.

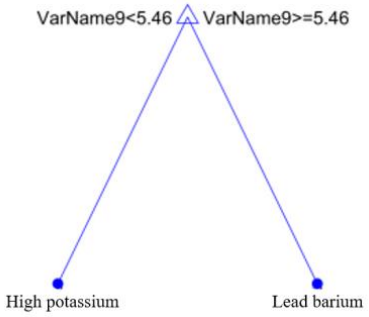


Figure 3. Decision tree for glass classification

As can be seen in Figure 3, the results are classified in terms of lead oxide (PbO) as indicated by the decision tree [13]. If the lead oxide (PbO) content is below 5.46%, the glass is considered to be high potassium glass, and vice versa for lead-barium glass.

The second part of this paper separates high potassium and lead barium glass for further subclassification. Firstly, the optimal number of groups for subclassification needs to be determined. This paper selected a systematic clustering method to cluster high potassium and lead barium glass, respectively, the flow chart of which is shown in Figure 4 below.

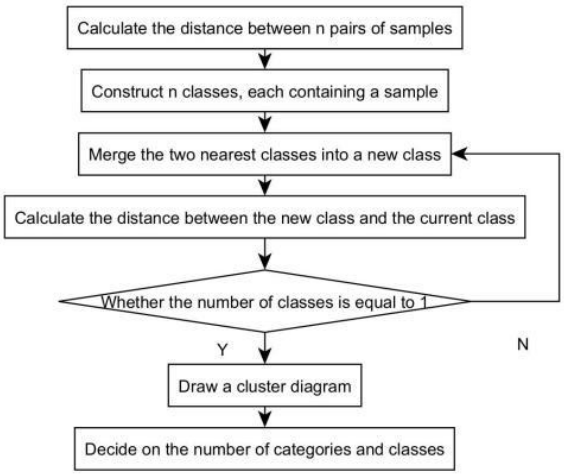


Figure 4. System clustering flowchart

In this paper, the aggregation coefficients are analysed using SPSS for the systematic clustering of different glass types separately.

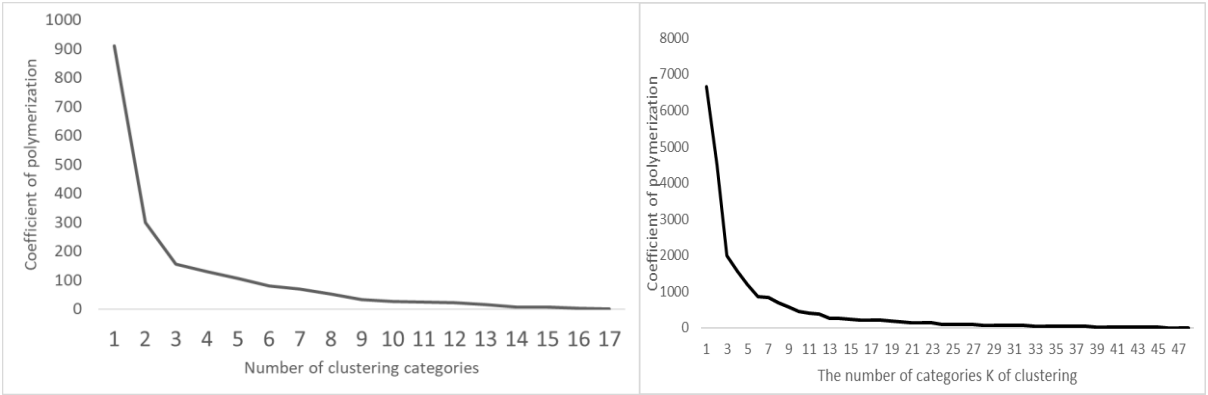


Figure 5. Line diagram of polymerization coefficient of high potassium, lead barium glass

According to the coefficient of aggregation line graph in Figure 5, the coefficient of aggregation starts to decrease slowly when the category number of high potassium glass is 2, and the coefficient of aggregation starts to decrease slowly when the category number of lead-barium glass is 3. In this paper, SPSS was used to generate the category numbers for the different types of glass. Again, a decision tree was chosen for the compositional analysis.

The results of the clustering of high potassium glass and lead-barium glass were analysed by Matlab using a decision tree in a similar way to the first part, with the ultimate aim of still identifying the most representative component data as a criterion for the classification of the two types of glass subclasses, which led to the following conclusions.

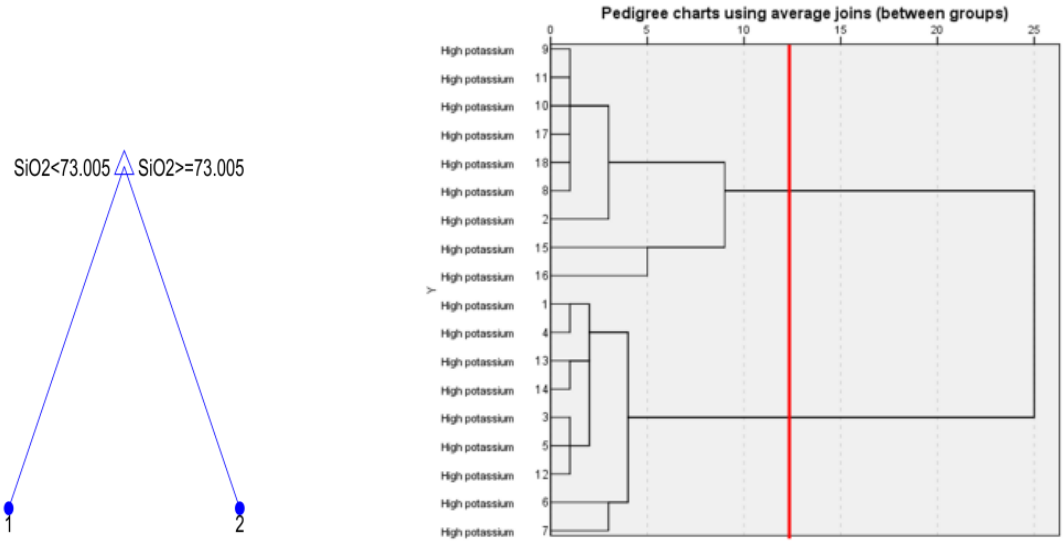


Figure 6. Subclassification of high potassium glass

As can be seen from Figure 6, the subclasses of high potassium glasses are divided by the content of silica, with compositions above 73.005% as one of the subclasses, and if below 73.005% as the other, a total of two subclasses are differentiated.

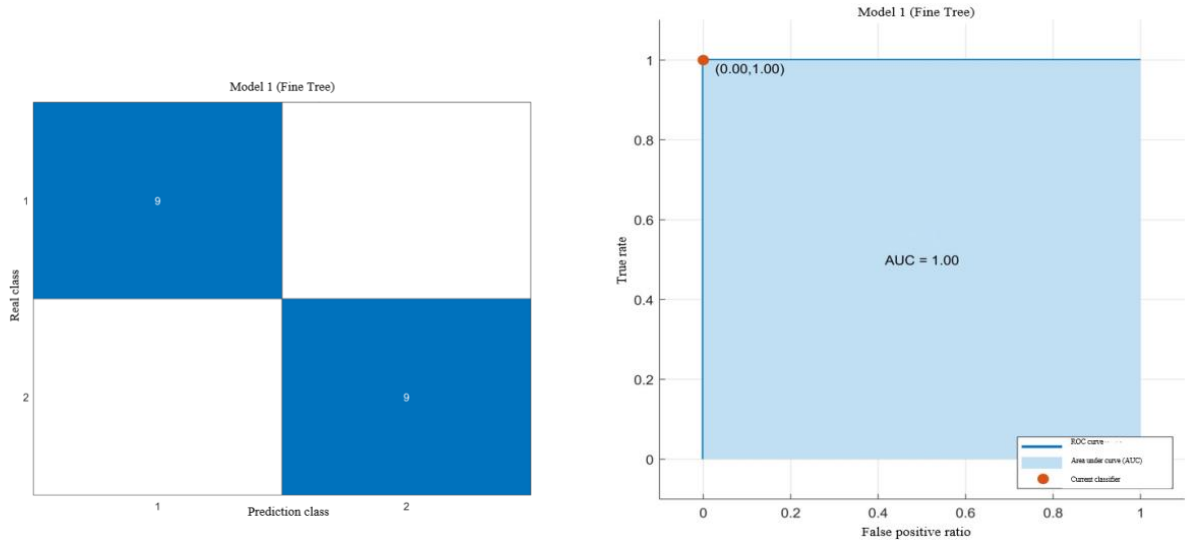


Figure 7. The decision tree generates confusion matrix and ROC curve of high potassium glass

In this paper, the decision tree model is tested for classifying high potassium glass. As can be seen from the confusion matrix in Figure 7, this model does not predict classification sample errors, and the area of the curve enclosed by the ROC curve and the lower axis is even 1. Therefore, this decision tree model classifies well.

Repeat the above steps for sub-classifying lead barium glass.

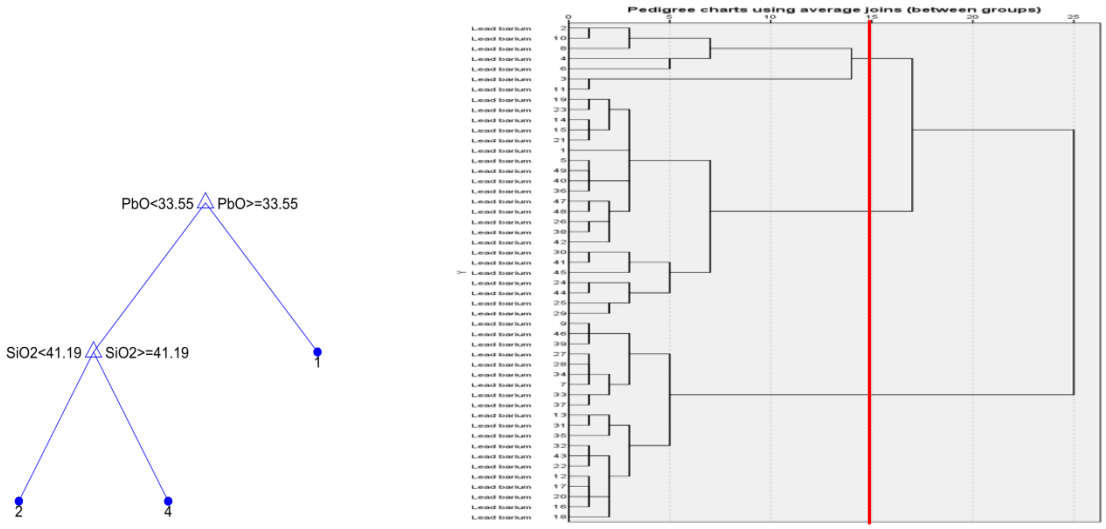


Figure 8. Subclassification of lead barium glass

From Figure 8 it is known that if the lead oxide content is higher than 33.55%, the glass is classified as one subclass; if the lead oxide content is lower than 33.55%, the glass is classified as silica content, and if the lead oxide content is higher than 41.19%, the glass is classified as one of the subclasses, and if not, the glass is classified as another subclass.

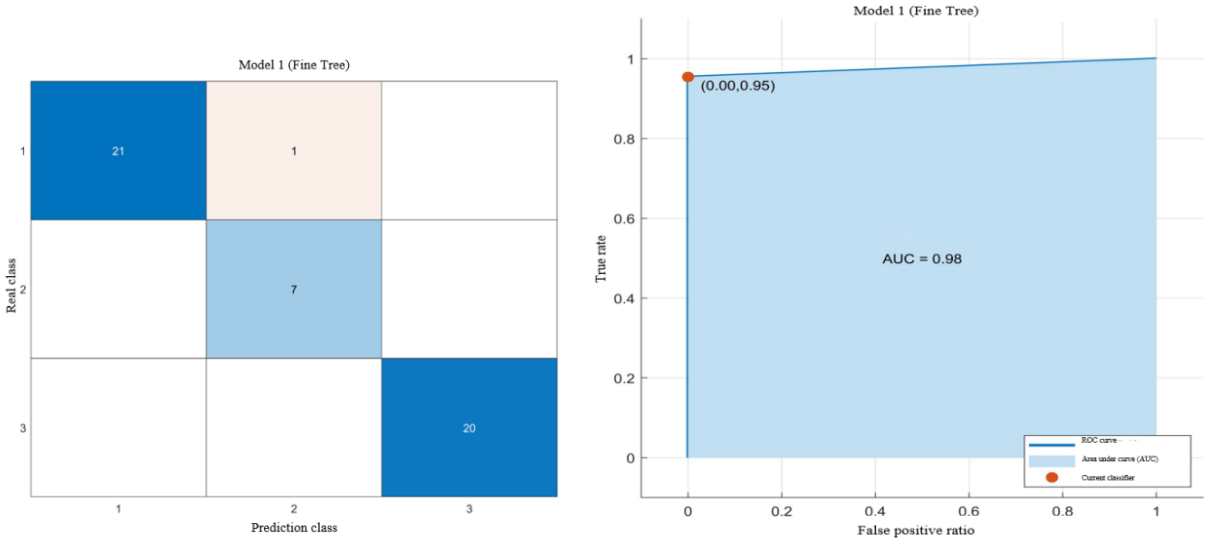


Figure 9. The decision tree generates confusion matrix and ROC curve for lead-barium glass

The model was also tested in this paper, and from Figure 9, it can be concluded that the test found one classification error this time, but the decision tree model classification result still had 98% accuracy, so the model was considered to be good.

Table 2. Members of each subclass

Type	Subclass	Subclass member
High potassium **(PbO<5.46%)**	1	1,3,4,5,6,6,13,14,16
	2	3,7,9,10,12,18,21,22,27
Lead barium **(PbO≥5.46%)**	1	2,19,30,30,34,36,38,39,40,41,43,43,49,50,51,51,52,54,54,56,57,58
	2	8,8,11,20,24,26,26
	3	23,25,28,29,31,32,33,35,37,42,42,44,45,46,47,48,49,50,53,55

The resulting subclass classification members for each glass are shown in Table 2.

The purpose of the third part is to verify the reasonableness of the subcategories and their sensitivity. In the second part of the paper, the K value is determined by the "elbow" method. For example, according to the aggregation coefficient fold diagram for lead barium, when the number of categories is 3, the decline of the fold tends to be slow, so the elbow is K=3, thus verifying the reasonableness (high potassium is verified in the same way as lead barium). The grid search method was then used to conduct a sensitivity analysis of the maximum number of splits in the decision tree and the splitting criterion tuning parameters.

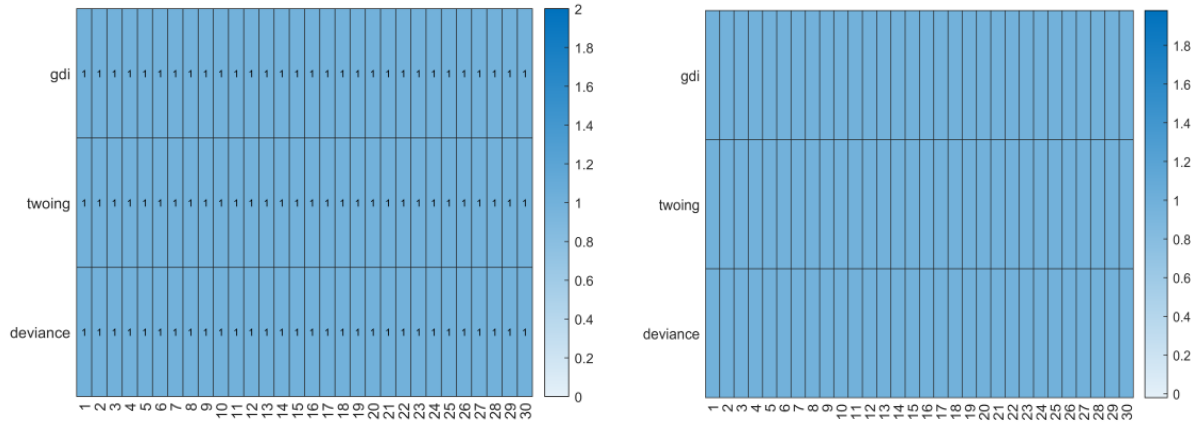


Figure 10. Grid search generates heat maps

From the heat map generated by Figure 10 on the grid search results, it can be seen that the maximum number of splits and splitting criterion of this decision tree model vary arbitrarily without significant changes in the classification results; the decision tree model does not require high sensitivity to parameters and the model has good classification effect. The model established in this paper has good accuracy and stability.

4. Conclusions

This paper uses a decision tree to more intuitively and easily understand the analysis of the chemical composition content that has a significant impact on the glass classification results and also verifies the chemical composition content that changes significantly before and after glass weathering. After adjusting the maximum number of splits and splitting criterion of the decision tree, the classification accuracy of the model was obtained to be over 95% for both, indicating that the model is universal and the prediction accuracy is relatively high. However, due to the ancient process problem, the distribution of relevant chemical components in the artefacts may not be too uniform. In this regard, the paper assigns a value of '0' to undetected data in the blanks during data pre-processing, reducing the difficulty of dealing with the problem. It should be recognised that there are differences in the content of components at different sampling points of the same artefact and that components not detected at a particular sampling point should also be taken into account, for which the model can be improved by filling in the selection of indicators with vertical, median, mean and interpolation methods, and can be corrected at the blank data of non-detected components so that the effects and relationships between the various component data can be further considered.

References

[1] ssubieux L, Golitko M, Gratuze B. Recent advances inlaser ablation ICP-MS for archaeology[M]Heidelberg: Springer,2016: 137-211.

[2] Humphry D.Some experiments and observations on the colours used in painting by the ancients [J].Philosophical Transactions of the Royal Society,1815,105(1) : 97-124.

[3] Seligman C G, Beck H C. Far eastern glass: Some westernorigins [J]. The Bulletin of the Museum of Far Eastern Antiquities,1938(10): 1-64.

[4] Gratuze B,Giovagnoli A, Barrandon J N, et al.Apport dela méthode ICP - MS couplée à l'ablation laser pour lacaractérisation des archéomatériaux [J]. Revue d 'Archéométrie,1993 (17) : 89-104.

[5] Cucchiara, Andrew. "Applied Logistic Regression." Technometrics 34.3(2012):358-359.

[6] Utgoff, P. E., N. C. Berkman, and J. A. Clouse. "Decision Tree Induction Based on Efficient Tree Restructuring." Machine Learning 29.1(1997):5-44.

[7] Pal, M., and P. M. Mather. "An assessment of the effectiveness of decision tree methods for land cover classification." Remote Sensing of Environment 86.4(2003):554-565.

[8] Shuang, L. I., S. Ding, and L. Qian. "The Decision Tree Classification and ItsApplication Research in Land Cover." Remote Sensing Technology & Application 17.1(2002).

[9] Hastie T, Rosset S, Tibshirani R, et al. The Entire Regularization Path for the Support Vector Machine[J]. The Journal of Machine Learning Research, 2004.

[10] Noble W S. What is a support vector machine? [J]. Nature Biotechnology.

[11] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. The Journal of Machine Learning Research, 2002, 2(1):999-1006.

[12] D'Andrade R G. Hierarchical Clustering[J]. Psychometrika, 2011, 43(1):59-67.

[13] HU Zhi-zhong,LI Pei,JIANG Lu-man,et al. Application of LA-ICP-MS in the Analysis of Archaeological Glass and Source Discrimination[J] Rock and Mineral Analysis,2020,39(4) : 505 - 514.