Glass component classification model based on decision tree and cluster analysis

Yunpeng Guo *, Zhangkang Tan, Yujie Zhang

College of Civil Engineering, Xi'an University of Architecture and Technology, Xi'an, China, 710055

* Corresponding Author Email: gyp2002020110@163.com

Abstract. In order to investigate the classification laws of the two types, three machine learning models (decision tree, SVM) were constructed in this paper, and their classification accuracy was 96%, which met the practical requirements. Subsequently, a K-means algorithm was constructed to classify the subclasses, and the high potassium and lead-barium glasses were divided into three subclasses. By descriptive statistics of the differences between the subclasses, the results showed that there existed a better differentiation of the divided subclasses in terms of multiple chemical compositions as well as ornamentation and color, which verified its reasonableness. By setting a perturbation factor (a normally distributed sequence with a mean of 0 and a standard deviation of 3) to test the sensitivity of the classification results, the model classification results did not change after several repetitions of the experiment and showed good robustness.

Keywords: Glass, Chemical composition, Classification, Machine learning models, K-means.

1. Introduction

Since the opening of the Silk Road, a large number of glass products have been introduced into China. The arrival of these different styles of glass products had a great impact on the development of China's glass industry [1]. The ancients gradually mastered the high level of glass making techniques and made modifications and improvements to it, developing it into a unique glass making process with Chinese characteristics.

Ancient glass in China appears to be the same as foreign glass, but in fact the chemical composition is very different, mostly consisting of quartz sand, fluxes and stabilizers [2, 3]. The study and identification of the chemical composition of ancient glass products is a vivid portrayal of the integration and development of modern technology and archaeology, which is of great significance to the development of archaeology and history in China and plays an important role in better understanding the profound Chinese civilization [4].

In this study, the classification rules of high potassium glass and lead-barium glass are analyzed and the two categories are divided into subcategories, in which the specific classification methods and results and the research analysis of the rationality and sensitivity of the classification results are required.

2. Analysis ideas and model selection

For the first sub-question, it is required to analyze the classification law of high potassium glass and lead-barium glass. Two solution methods are used in this question, Decision Tree and Support Vector Machine (SVM). For the Decision Tree model and SVM, we obtain the final learning models after training by means of dividing the data set such as cross-validation. The accuracy of the two learning models is obtained by computing to 96%.

For the second subquestion, it is required to select the appropriate compounds for subclassification of high potassium glass and lead-barium glass, and this problem is solved by cluster analysis and visualization of the results [5-7].

To investigate the sensitivity of the prediction results, we set the measurement error (sampling from a normal distribution with mean 0 and variance 3) and sampled each chemical component of each sample independently, and repeated 100 times to obtain the final results.

3. Data pre-processing

3.1. Missing value handling

In the process of sample acquisition, it is inevitable that the final sample set will have some missing attribute data due to some reasons. Since the missing data are very few, the missing data are discarded directly in this problem.

3.2. Normalization of raw data

Before performing K-means clustering, the original data should be standardized to prevent the influence of different levels on the clustering. In this problem, SPSS software was used to standardize the data.

4. Model building and solving

4.1. Decision tree model

Based on the fact that the size of the information gain and is not affected by the number of attribute values, this problem can be solved by constructing a decision tree for discrete data. Decision Tree is a tree-like structure similar to a flowchart, which divides the data based on the tree structure and classifies them by constant division conditions. The greater the uncertainty of the variables, the greater the corresponding entropy. One of the most critical things is to find out the conditions that have the greatest impact on the results, i.e., low information entropy, and put them in front to save the number of lookups.

MATLAB was used to model the obtained data, and a total of 58 samples of 14 chemical components were involved in the analysis in this paper. We can get the inference rules directly from the decision tree, and get the sample size and confidence level contained in each node.

4.2. SVM model

Support Vector Machine (SVM) is a new technique of data mining, which solves machine learning problems with the help of optimization methods to overcome "dimensionality difficulties". The basic idea is to find a hyperplane that correctly separates as many data points as possible, while keeping the two separated data points as far away from the classification plane as possible [8-10].

4.3. Clustering Analysis

Cluster analysis is a multivariate statistical analysis method for quantitative classification of multiple samples, in which data information is grouped according to information found in the data that describes the objects and their relationships. Objects within a group are related to each other, while objects in different groups are unrelated. The greater the similarity within groups and the greater the gap between groups, the better the clustering effect.

The goal of clustering is to obtain a high intra-cluster similarity and a low inter-cluster similarity, so that the distance between clusters is as large as possible and the distance between samples within clusters and cluster centers is as small as possible. The relevant definitions in clustering analysis are as follows.

(1) Cluster center: the mean of all sample points in a cluster (center of mass).

(2) Cluster size: the number of samples contained in a cluster.

(3) Cluster density: the closeness of the sample points in a cluster.

(4) Cluster description: the operational characteristics of the samples in the cluster.

The K-means algorithm, a division-based clustering, is used in this problem. The algorithm clusters objects by dividing them into mutually exclusive clusters, each object belongs to one cluster and only one cluster, calculates the distance between the sample points and the center of mass of the class cluster, and the sample points close to the center of mass of the class cluster are divided into the

same class of clusters, and the similarity between them is measured by the distance between the samples. The closer the two samples are, the higher the similarity is.

The steps of K-means clustering are shown as Figure 1.

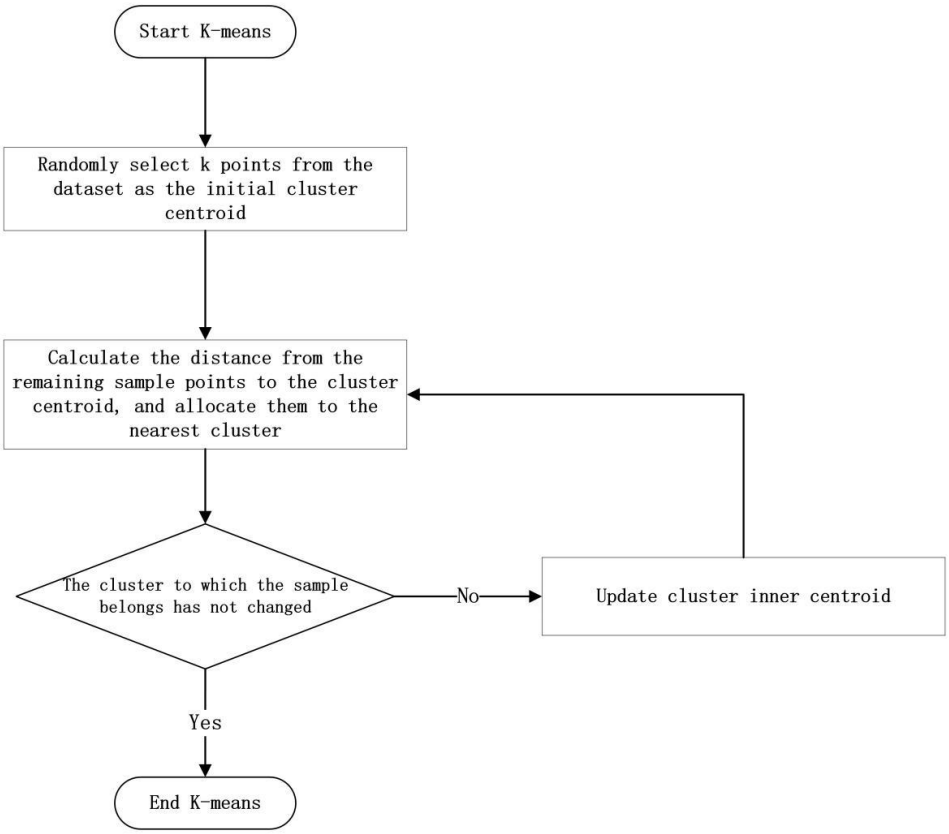


Figure 1. Cluster analysis flow chart

4.4. Analysis and testing of results

The confusion matrix, ROC curve of this model is obtained by building a decision tree classification model as shown in Figure 2 and Figure 3.

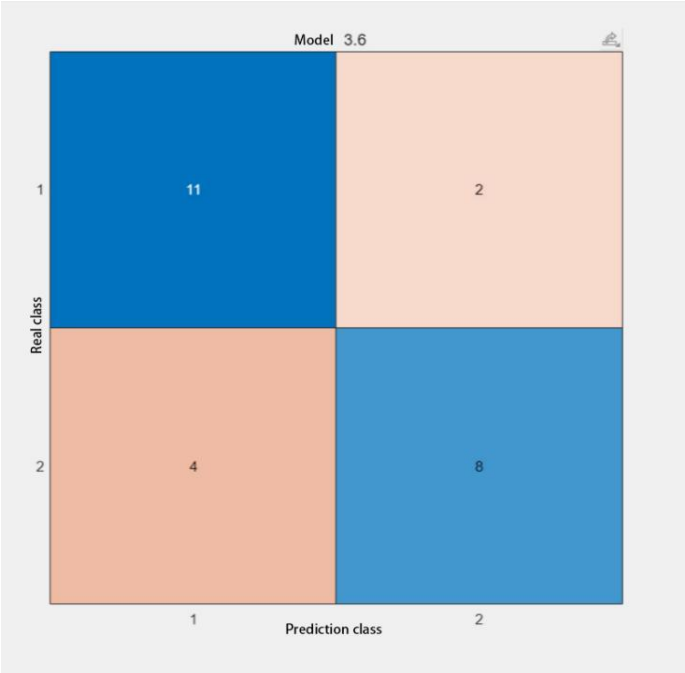


Figure 2. Confusion matrix of decision tree model

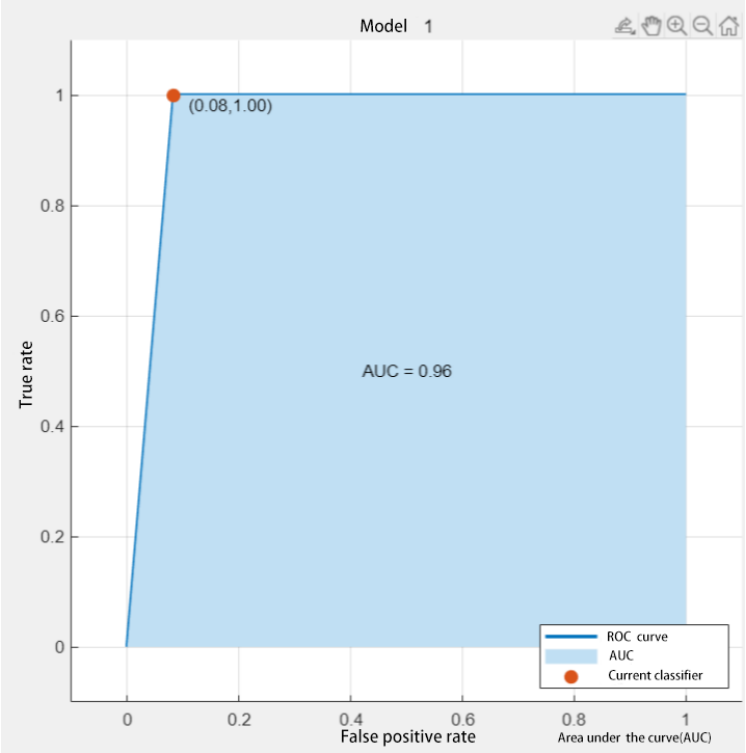


Figure 3. ROC curve of decision tree model

The decision tree model applied in this problem was cross-validated to achieve 96% accuracy for the solution.

The confusion matrix, ROC curve obtained by SVM are shown in Figure 4 and Figure 5.

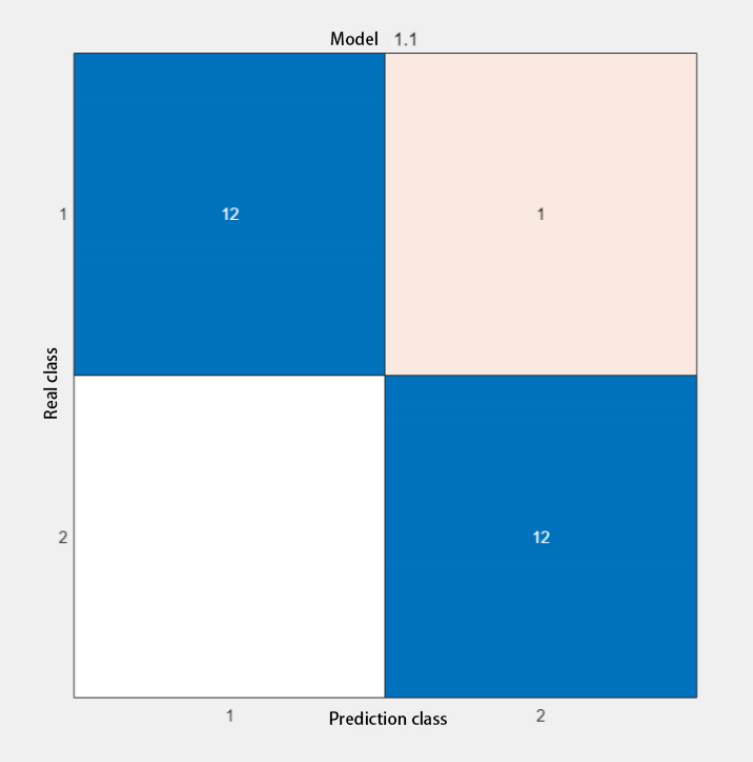


Figure 4. Confusion matrix of SVM algorithm

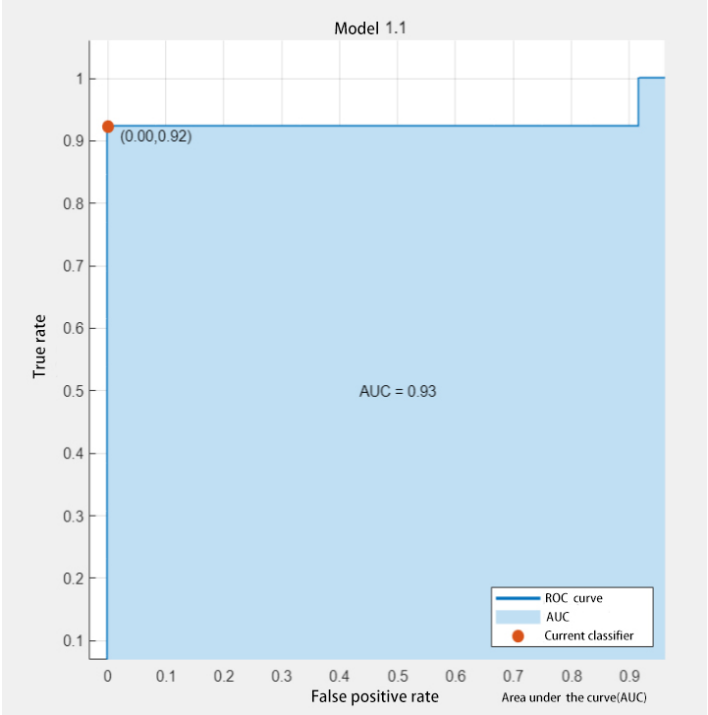


Figure 5. ROC curve of SVM algorithm

After cross-validation, the SVM algorithm is evaluated as follows.
The results of clustering analysis are as follows.
(1) The contour plot of clustering results is shown in Figure 6.

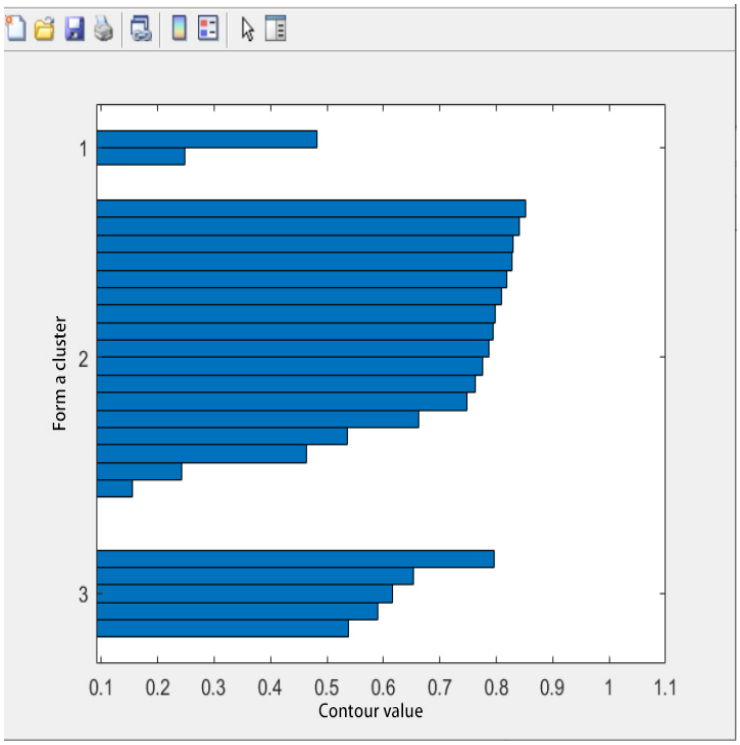


Figure 6. Clustering analysis contour diagram

(2) The scatter plot of clustering results is shown in Figure 7.

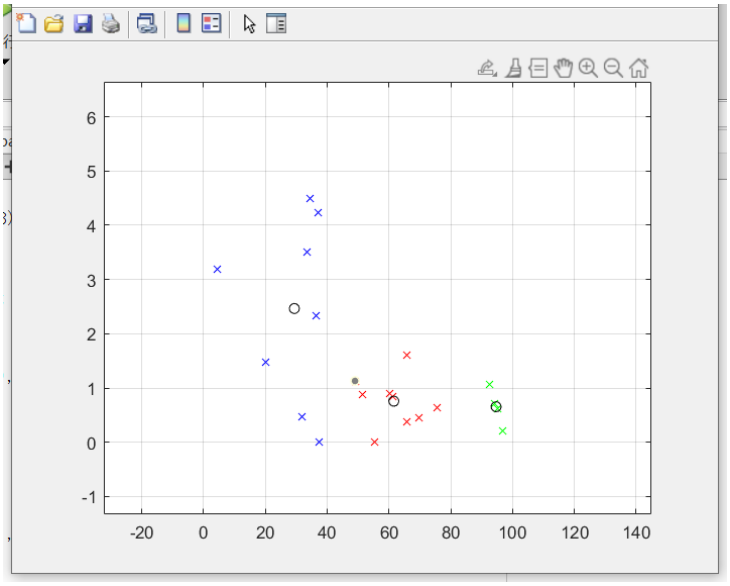


Figure 7. Scatter plot of K-means algorithm results

According to Figure 6 and Figure 7, it can be seen that in the case of k=3, i.e., clustering into three categories, the clusters of sample points have a good value of the result in the category assignment, i.e., based on the 58 sample data, it is possible to classify these 58 samples into three types of glass, i.e., the two glass artifacts, high potassium glass and lead-barium glass, are classified into three categories by this algorithm.

5. Conclusions

In order to investigate the classification laws of the two types, three machine learning models (decision tree, SVM) were constructed in this paper, and their classification accuracy was 96%, which met the practical requirements. Subsequently, a K-means algorithm was constructed to classify the subclasses, and the high potassium and lead-barium glasses were divided into three subclasses. By descriptive statistics of the differences between the subclasses, the results showed that there existed a better differentiation of the divided subclasses in terms of multiple chemical compositions as well as ornamentation and color, which verified its reasonableness. By setting a perturbation factor (a normally distributed sequence with a mean of 0 and a standard deviation of 3) to test the sensitivity of the classification results, the model classification results did not change after several repetitions of the experiment and showed good robustness.

References

[1] He Hao,Li Panpan,Liu Xiaoqing,Xu Shiqing,Zhang Junjie. Crystallization characteristics of stress-induced CsPbCl3-xBrx (X ≤ 3) quantum dots phosphate glasses[J]. Optical Materials,2023,136.

[2] Xu Guilong,Wang Lingxiao,Xie Zijie,Xu Buqin,Tang Min,Liang Yun,Yang Jin,Hu Jian. Functional polymer coating on glass-fiber fabric for de-emulsification and coalescence of water-in-oil emulsion[J]. Reactive and Functional Polymers,2023,182.

[3] Krzyżanowska Marta,Syta Olga,Kępa Luiza,Wagner Barbara. Small and overlooked: Roman glass counters from archaeological sites in Poland[J]. Journal of Archaeological Science: Reports,2023,47.

[4] Polski Ashley,Kennedy Brandon,Bair Christopher,Simpson Rachel. Acute iris vascular tuft hemorrhage treated successfully with intravitreal bevacizumab and pressure patching in a patient with branch retinal vein occlusion[J]. American Journal of Ophthalmology Case Reports,2023,29.

[5] Matsusako Shunsuke,Masuda Sayaka,Matsuo Saeko,Imamura Ryota,Sakuragi Takahisa,Inada Honami,Hayakawa Hikaru,Nishida Tetsuaki,Oka Nobuto. CrIII-doped vanadate glass: characterization and application as a cathode active material for lithium-ion batteries[J]. Hyperfine Interactions,2022,244(1).

[6] Hardy Penelope K. The art of aquarium scienceOceans Under Glass Samantha Muka University of Chicago Press, 2022. 240 pp.[J]. Science (New York, N.Y.),2022,378(6626).

[7] Capitão Jeffrey,Martins Jorge,Emami Seyedali,Ivanou Dzmitry,Mendes Adélio. Fully glass frit encapsulated dye-sensitized solar cells: Challenges for hermetical sealing of electrolyte injection holes[J]. Solar Energy,2023,249.

[8] Wang Huajun,Shao Yuanhai. Fast truncated Huber loss SVM for large scale classification[J]. Knowledge-Based Systems,2023,260.

[9] Upadhya Ankita, Thakur M.S.,Sihag Parveen,Kumar Raj,Kumar Sushil,Afeeza Aysha,Afzal Asif,Ahamed Saleel C. Modelling and prediction of binder content using latest intelligent machine learning algorithms in carbon fiber reinforced asphalt concrete[J]. Alexandria Engineering Journal,2023,65.

[10] Belghit Amar, Lazri Mourad,Ouallouche Fethi,Labadi Karim,Ameur Soltane. Optimization of One versus All-SVM using AdaBoost algorithm for rainfall classification and estimation from multispectral MSG data[J]. Advances in Space Research,2023,71(1).