Analysis and Identification of the Composition of Ancient Glass Objects based on Statistical Research and Machine Learning Algorithms

Caoyuan Sun [†, *], Zewen Li [†]

Jinan University, University of Birmingham Joint Institute at Jinan University, Jinan University, Guangzhou, China, 511443, China

* Corresponding author email: 10092719@qq.com

[†]These authors contributed equally

Abstract. In order to promote the study of ancient glass artifacts, this paper integrates the weathering principle and the nature of chemical elements in glass, and analyzes the color and chemical composition of glass using statistical and machine learning methods. First, the classification and regression tree featuring Gini index is applied to explore the classification criteria of high potassium glass and lead barium glass, and the accuracy of the model is tested by 10-fold cross-validation. Secondly, subclasses were classified for the collected sample data. In this paper, Two Step Cluster Algorithm is used to divide the high potassium glass into four subclasses and the lead barium glass into five subclasses based on the Euclidean distance between the samples. Finally, the Spearman's rank correlation coefficients between chemical components are computed separately based on the proportion of each chemical component in glass to explore the correlation between them.

Keywords: Classification and Regression Tree; Two Step Cluster Algorithm; Spearman's Rank Correlation Coefficient.

1. Introduction

Machine learning algorithm is recently playing an important role in research in various fields [5], such as disease prediction [6], materials design [7] and glass properties research [8]. In this essay, the goal is to build a classification and identification model to predict various types of chemical composition data before weathering, to analyze the correlation and difference between various types of composition, and to test the accuracy of the model and extend it to analyze and identify newly excavated artifacts by using the existing collected various types of testing data. The mathematical model is a significant study and exploration of the results of the identification and analysis of the chemical composition of glass artifacts, which has important implications for the heritage industry and the application of mathematical methods to other fields. [1-4]

2. The Basic Fundamental of Statistical Research and Machine Learning Algorithms

2.1 The Structure of Classification and Regression Tree

The decision tree algorithm embodies a functional mapping relationship between features and labels. Non-leaf nodes in the tree represent the division of samples on a feature, and the sample set is divided into several subtrees according to the different values taken on the feature until the leaf node or a condition is satisfied then the sample set is no longer split, and each leaf node corresponds to a classification.

(1) Calculate the initial Gini coefficient value. For the sample training set Q, the attributes in this training set are A. Based on each attribute A, calculate the initial Gini coefficients at this point.

(2) Calculate the segmentation Gini coefficient value. For each attribute A in the training set, the training set Q is divided into two subsets Q1 and Q2 with the threshold a as the basis for the

segmentation of attribute A. The segmentation Gini coefficient values are calculated for the two subsets, and the larger the value, the higher the probability of error in the segmented set.

(3) Determine the best attribute and segmentation threshold. For each attribute A, choose the attribute with the minimum segmentation Gini coefficient and its threshold as the best division basis, generate two child nodes, and perform sample division.

$$Gini(Q) = \sum_{i=1}^{n} P_i(1 - P_i) = 1 - \sum_{i=1}^{n} P_i^2 \tag{1}$$

$$Gini(Q, A) = \frac{|Q_1|}{|Q|} Gini(Q_1) + \frac{|Q_2|}{|Q|} Gini(Q_2) \tag{2}$$

2.2 The Structure of Two Step Cluster Algorithm

(1) Pre-clustering, where the sample is roughly divided into several classes using a "through-order" approach. At the beginning, all data will enter a large class, and after entering an observation, it is decided whether the observation should be derived into a new class or merged into one of the existing subclasses based on its "closeness". After this process is performed several times, L classes will be formed. In summary, pre-clustering increases the number of clusters.

(2) The clustering is based on the pre-clustering, and then the "closeness" determines which subclasses can be merged to form the L class. It can be seen that through this step, the number of clusters will be reduced and the differences between different classes will be gradually expanded, in this paper, we will take the Euclidean distance to describe the distance between cases, the formula (with two data (x_1, y_1, z_1),(x_2, y_2, z_2)as an example).

$$Euclid(1,2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \tag{3}$$

2.3 Spearman's Rank Correlation Coefficient

In correlation analysis, Spearman's rank correlation coefficient is a measure of dependence between two variables, with the value between -1 and 1. For two n-dimensional random variables $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_n)$, the Spearman's rank correlation coefficient r_s between them can be computed from formula (4) below.

$$r_s = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\Sigma_{i=1}^{n}(p_i - \bar{p})^2 \Sigma_{i=1}^{n}(q_i - \bar{q})^2}} \tag{4}$$

where p_i and q_i are rankings of x_i and y_i. If some variables have the same value, the rank corresponding to this value is the average of rankings of these values [4].

If $0 < r_s \leq 1$, we have there is a positive correlation between variables; if $-1 \leq r_s < 0$, we have there is a negative correlation between them; if $r_s = 0$, we say that there is no correlation between variables.

3. Results

3.1 The Basis for Distinguishing between High Potassium Glass and Lead Barium Glass

By regarding proportion of chemical composition content and type of glass as independent variable and dependent variable respectively, we obtain that whether the glass is high potassium or lead barium can be determined by content of lead(II) oxide in it. As shown in Figure 1, if the content proportion is less than or equal to 5.460%, the sample will be classified among high potassium glass, otherwise it will be judged as lead barium glass.

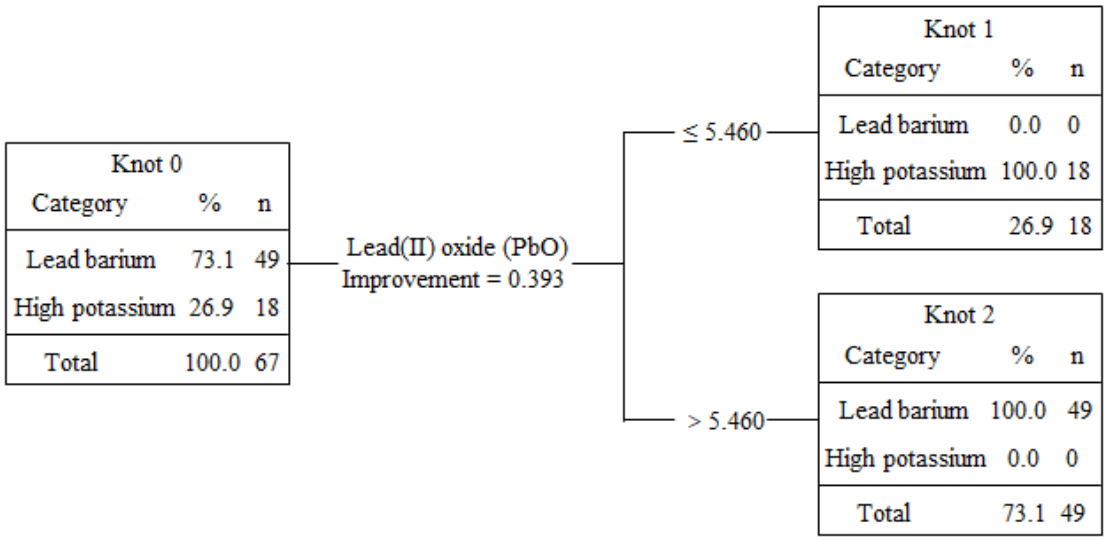

Figure 1. Decision tree for glass classification

To verify the accuracy of the above rule, we adopt 10-fold cross-validation and can see that CART model allows high classification accuracy. The result is shown in Table 1.

Table 1. Analysis of accuracy of classification

Category			
	Projection		
Real test	Lead barium	High potassium	Percentage correct
Lead barium	49	0	100.00%
High potassium	0	18	100.00%
Overall percentage	73.10%	26.90%	100.00%
Growth method: CRT			
Dependent variable: Type			

In addition, we derive the order of importance of chemical composition that can be used for classification. We can notice from Figure 2 that the content of lead(II) oxide is of vital importance when distinguishing between high potassium and lead barium glass, which accounts for nearly 100%. And potassium oxide in sample artifacts also plays an important role in determining the glass type.

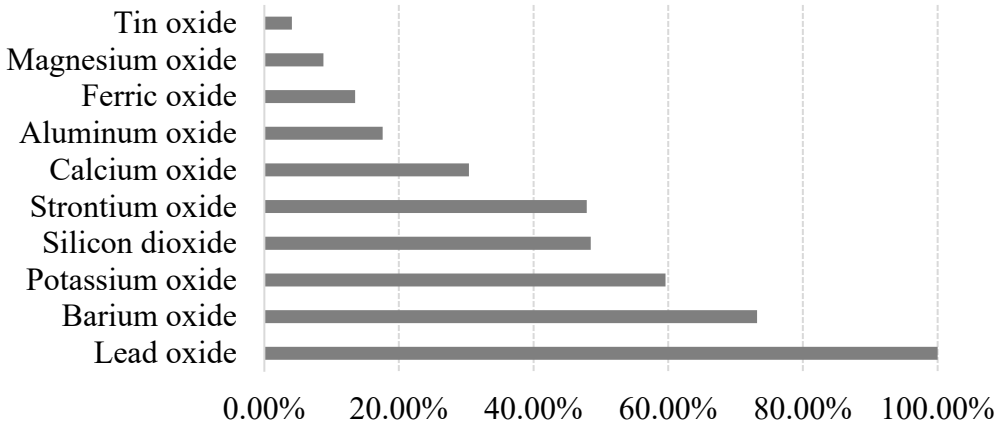

Figure 2. The order of importance of chemical composition when considering type of glass

3.2 The Basis for Identifying Subcategories of High Potassium Glass and Lead Barium Glass

After splitting the collected data into two groups according to glass type, we apply Two Step Cluster Algorithm to each group. By changing the number of clusters and comparing the clustering quality, we finally set the number of clusters to be 4 for high potassium and 5 for lead barium glass.

3.2.1 Subcategories of High Potassium Glass

Figure 3 below shows criterion of classifying high potassium glass into four smaller subcategories. We first consider the proportion of the content of silicon dioxide, that is, if it is greater than the value 83.255%, the sample will be classified into Subcategory 3 of high potassium glass. By further comparing the proportion of phosphorus pentoxide in this sample with 1.315%, we can draw a conclusion that whether it is more probable to be classified into Subcategory 1 or Subcategory 2. In other words, when classifying high potassium glass into subcategories, the most two important components that should be considered is silicon dioxide and phosphorus pentoxide, which is shown in Figure 4.

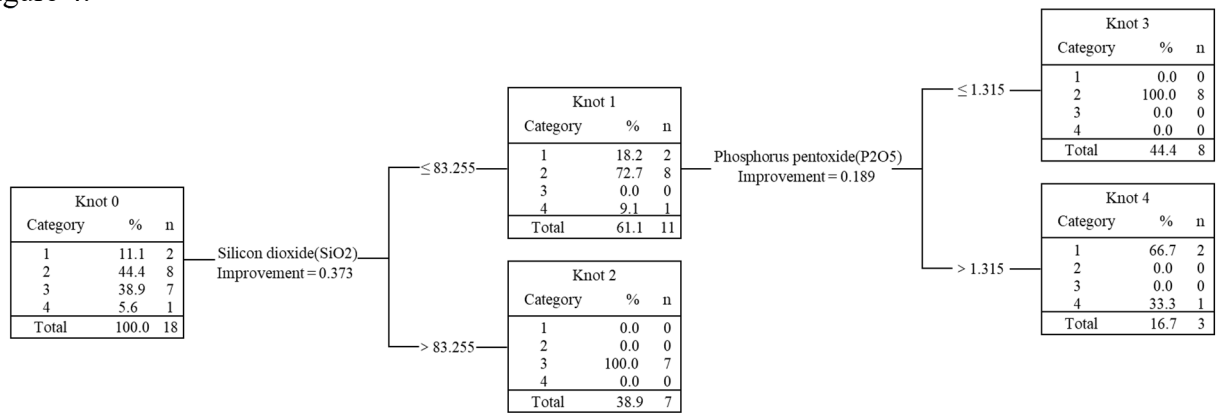

Figure 3. Decision tree for subcategories of high potassium glass

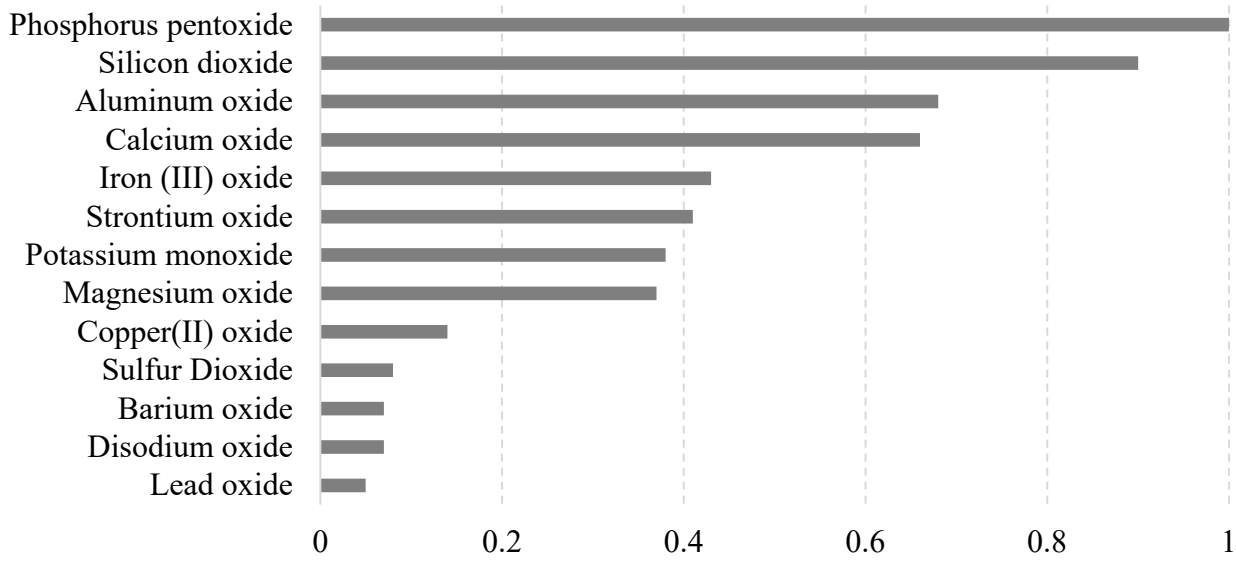

Figure 4. The order of importance of chemical composition when considering subcategories of high potassium glass

3.2.2 Subcategories of Lead Barium Glass

Figure 5 below shows basis of classification when considering subcategories of lead barium glass. If the proportion of barium oxide in the sample is less than or equal to 24.890%, we will regard it as Subcategory 2 of lead barium glass. And then if the proportion of silicon dioxide is greater than the value 12.20%, there is a high probability that this sample is of Subcategory 4. Besides, the experiment

results that sulfur dioxide in glass is the most important component used to determine subcategories of lead barium glass can be seen in Figure 6.

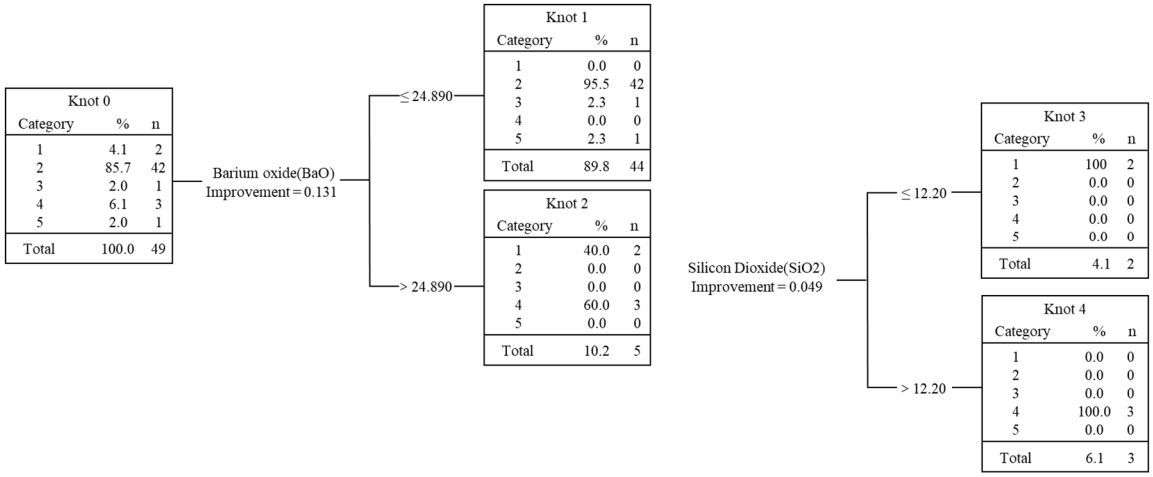

Figure 5. Decision tree for subcategories of lead barium glass

3.2.3 Sensitivity

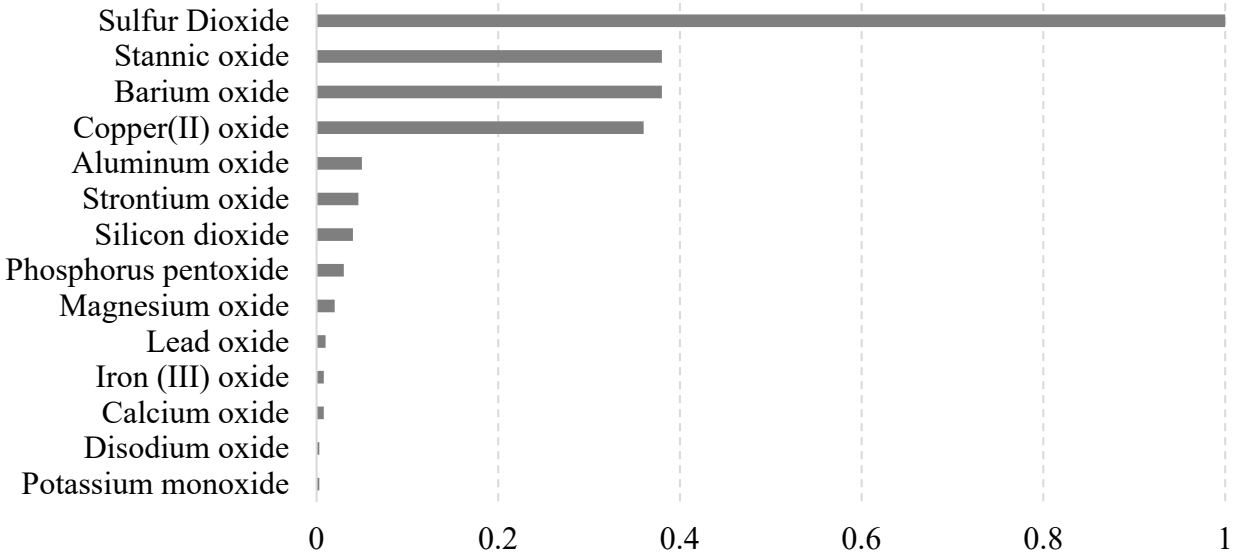

Figure 6. The order of importance of chemical composition when considering subcategories of lead barium glass

To test sensitivity of our model in classifying samples into subcategories, we add figures for phosphorus pentoxide in high potassium glass and sulfur dioxide in lead barium glass by 5% respectively and observe the impact of this change on results. It is shown that there is no difference between the new classification and previous ones.

3.3 Correlation between Chemical Components in Glass

According to the method mentioned in this paper, we can obtain the Spearman's rank correlation coefficient between each chemical component. For high potassium glass, the proportion of silicon dioxide shows a negative correlation with all other components, among which the coefficient between it and aluminum oxide is $r_s = -0.812$. That is, there is a relatively stronger correlation between the two chemical components. On the contrary, there is no significant correlation between some components in high potassium glass. For example, the correlation coefficient between barium oxide

and potassium oxide is only $r_s = -0.026$, which indicates a weak correlation. Details can be seen from Table 2.

Table 2. Spearman's rank correlation coefficient between components of high potassium glass

	SiO_2	MgO	Al_2O_3	CuO	CaO	BaO	P_2O_5	SrO	SnO_2	SO_2	PbO	Fe_2O_3	K_2O	Na_2O
SiO_2	1.000	-0.478	-0.812	-0.438	-0.761	-0.242	-0.405	-0.426	0.084	-0.345	-0.436	-0.774	-0.802	-0.539
MgO	-0.478	1.000	0.731	0.131	0.016	0.451	0.611	0.595	0.256	0.477	0.338	0.457	0.198	-0.250
Al_2O_3	-0.812	0.731	1.000	0.232	0.518	0.367	0.424	0.404	-0.140	0.332	0.59	0.756	0.504	0.380
CuO	-0.438	0.131	0.232	1.000	0.340	0.492	0.487	0.093	-0.420	0.341	0.100	0.653	0.174	-0.134
CaO	-0.761	0.016	0.518	0.340	1.000	-0.136	-0.063	-0.150	-0.392	0.367	0.224	0.564	0.654	0.669
BaO	-0.242	0.451	0.367	0.492	-0.136	1.000	0.367	0.477	-0.123	-0.228	0.703	0.457	-0.026	-0.228
P_2O_5	-0.405	0.611	0.424	0.487	-0.063	0.367	1.000	0.383	0.364	0.298	-0.007	0.450	0.080	-0.196
SrO	-0.426	0.595	0.404	0.093	-0.150	0.477	0.383	1.000	0.375	0.016	0.378	0.247	0.423	-0.094
SnO_2	-0.084	0.256	-0.140	-0.420	-0.392	-0.123	0.364	0.375	1.000	-0.123	-0.170	-0.420	0.028	-0.123
SO_2	-0.345	0.477	0.332	0.341	0.367	-0.228	0.298	0.016	-0.123	1.000	-0.315	0.380	0.268	-0.228
PbO	-0.436	0.338	0.59	0.100	0.224	0.703	-0.007	0.378	-0.170	-0.315	1.000	0.386	0.301	0.378
Fe_2O_3	-0.774	0.457	0.756	0.653	0.564	0.457	0.450	0.247	-0.420	0.380	0.386	1.000	0.507	0.263
K_2O	-0.802	0.198	0.504	0.174	0.654	-0.026	0.080	0.423	0.028	0.268	0.301	0.507	1.000	0.627
Na_2O	-0.539	-0.250	0.380	-0.134	0.669	-0.228	-0.196	-0.094	-0.123	-0.228	0.378	0.263	0.627	1.000

Table 3. Spearman's rank correlation coefficient between components of lead barium glass

	SiO_2	MgO	Al_2O_3	CuO	CaO	BaO	P_2O_5	SrO	SnO_2	SO_2	PbO	Fe_2O_3	K_2O	Na_2O
SiO_2	1.000	0.319	0.105	-0.377	-0.759	-0.557	-0.51	0.057	-0.318	-0.182	-0.046	0.260	0.356	-0.494
MgO	0.319	1.000	0.663	-0.257	-0.412	0.049	-0.112	0.359	-0.325	-0.277	0.395	0.553	0.080	0.32
Al_2O_3	0.105	0.663	1.000	-0.279	-0.159	0.174	0.235	0.362	-0.33	-0.295	0.262	0.445	-0.002	0.464
CuO	-0.377	-0.257	-0.279	1.000	0.073	0.280	0.144	-0.301	0.52	0.412	-0.326	-0.267	-0.104	-0.062
CaO	-0.759	-0.412	-0.159	0.073	1.000	0.364	0.312	-0.110	-0.002	-0.106	0.085	-0.240	-0.310	0.311
BaO	-0.557	0.049	0.174	0.280	0.364	1.000	0.226	-0.012	-0.185	0.142	0.315	-0.080	-0.569	0.606
P_2O_5	-0.51	-0.112	0.235	0.144	0.312	0.226	1.000	-0.023	0.162	0.152	-0.109	-0.087	-0.006	0.329
SrO	0.057	0.359	0.362	-0.301	-0.110	-0.012	-0.023	1.000	-0.067	-0.095	0.312	0.252	-0.095	0.300
SnO_2	-0.318	-0.325	-0.33	0.52	-0.002	-0.185	0.162	-0.067	1.000	0.377	-0.486	-0.072	0.080	-0.299
SO_2	-0.182	-0.277	-0.295	0.412	-0.106	0.142	0.152	-0.095	0.377	1.000	-0.284	-0.110	-0.192	-0.007
PbO	-0.046	0.395	0.262	-0.326	0.085	0.315	-0.109	0.312	-0.486	-0.284	1.000	0.241	-0.191	0.368
Fe_2O_3	0.260	0.553	0.445	-0.267	-0.240	-0.080	-0.087	0.252	-0.072	-0.110	0.241	1.000	-0.042	0.097
K_2O	0.356	0.080	-0.002	-0.104	-0.310	-0.569	-0.006	-0.095	0.080	-0.192	-0.191	-0.042	1.000	-0.287
Na_2O	-0.494	0.32	0.464	-0.062	0.311	0.606	0.329	0.300	-0.299	-0.007	0.368	0.097	-0.287	1.000

As shown in Table 3, for lead barium glass, silicon dioxide and calcium oxide have the strongest correlation relationship, with the coefficient $r_s = -0.759$. And we notice that there is a relatively strong positive correlation between aluminum oxide and magnesium oxide, with $r_s = 0.663$. However, there is no significant relationship between silicon dioxide and lead(II)oxide.

4. Conclusion

In this essay, we find the basis for distinguishing between high potassium glass and lead barium glass by applying CART algorithm. By producing a decision tree for classifying glass based on Gini Index, we draw a conclusion that whether the sample is high potassium glass or not can be determined through the proportion of lead(II)oxide in it.

Furthermore, we apply two-step clustering algorithm to identify subcategories of high potassium and lead barium glass and find that the quality of clustering is highest when setting the number of clusters to be 4 and 5 respectively. For the classification criterion, silicon dioxide and phosphorus pentoxide are the two most important components when studying with high potassium glass; and the proportion of barium oxide and silicon dioxide allow us to classify lead barium into subcategories.

Also, by calculating Spearman's rank correlation coefficients, we obtain correlation relationships between chemical components in the sample. There is a great difference in correlation among components between glass of two different types. For high potassium glass, silicon dioxide has a strong negative correlation with aluminum oxide while with calcium oxide in lead barium glass.

References

[1] Zhang Liang, Ning Qian. Two improvements and applications of CART decision tree [J]. Computer Engineering and Design, 2015, 26(5): 1209-1213.

[2] Yadav S, Shukla S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification [C].2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016: 78-83.

[3] Jiang Hong, Fu Junze, Yang Jun. Infrared spectrum identification of sports shoe sole materials based on Two-step Clustering and RBFNN [J]. Leather Science and Engineering, 2022, 32(5): 51-56.

[4] Fan Rong, Meng Dazhi, Xu Dashun. Advances in statistical correlation analysis methods [J]. Mathematical Modeling and Its Applications, 2014, 3(01).

[5] Wei Jing, Chu Xuan, Sun Xiangyu, et al. Machine learning in materials science [J]. InfoMat. 2019, 1(3): 338-358.

[6] Liu Yuan, Yang Xiaowen, Li Lezhi. Research advances in the application of machine learning for disease prediction [J]. Journal of Nursing (China), 2021, 28(7): 30-34.

[7] Liu Yue, Zhao Tianlu, Ju Wangwei, et al. Materials discovery and design using machine learning [J]. J Materiomics, 2017, 3(3): 159-177.

[8] Zhang Liyan, Li Hong, Chen Shubin, et al. Simulation methods of glass composition and properties: a short review [J]. Journal of The Chinese Ceramic Society, 2022, 50(8): 2338-2350.