Modeling and Simulation 建模与仿真, 2023, 12(2), 1185-1198
Published Online March 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.122112

Hans汉斯

古代玻璃制品的成分分析与鉴别的模型研究

刘泊志[1]，吕珊珊[2]，刘姝邑[2]，李田丰[1]

[1]上海理工大学机械工程学院，上海
[2]上海理工大学光电信息与计算机工程学院，上海

收稿日期：2022年12月5日；录用日期：2023年3月10日；发布日期：2023年3月17日

摘　要

为帮助古代玻璃制品鉴别者通过化学成分的定量分析判断玻璃文物类型，解决如何对玻璃制品准确分类的问题，对五十余件玻璃样品多处采样，以采样得到的六十余份数据进行量化分析，使用K均值聚类模型、随机森林模型对其种类进行分析与验证，筛选出对玻璃样品是否风化以及玻璃样品的种类划分有显著性区别的化学成分，建立数学模型并通过将已知数据分为训练集和验证集交叉验证去评价模型准确度，并对灵敏度进行分析，得到判断玻璃制品是否风化以及分类玻璃制品的最优模型。

关键词

古代玻璃制品分类，配对t检验，K均值聚类，随机森林，交叉验证

Research on the Model of Ancient Glass Products Classification and Identification

Bozhi Liu[1], Shanshan Lv[2], Shuyi Liu[2], Tianfeng Li[1]

[1]School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai
[2]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Dec. 5th, 2022; accepted: Mar. 10th, 2023; published: Mar. 17th, 2023

Abstract

In order to help archeologist to quantitatively classify the type of relics of glass by the difference of chemical composition and solve the problem of Ancient glass products classification accurately, this paper proposes to make a quantitative analysis of the more than sixty pieces of data obtained from sampling more than 50 ancient glass samples, analyzing and verifying the species by using K-means Clustering model and Random forest model, screening the chemical composition which make a significant role in classification of whether weathering or type of glass, building a mathematical model, valuing the accuracy of the model by dividing data into two categories of test and

文章引用：刘泊志，吕珊珊，刘姝邑，李田丰．古代玻璃制品的成分分析与鉴别的模型研究[J]．建模与仿真，2023, 12(2): 1185-1198. DOI: 10.12677/mos.2023.122112

validation and analyzing its sensitivity, then acquiring the optimal model. The purpose of this study is to help archeologist to quantitatively classify the type of relics of glass by the difference of chemical composition.

Keywords

Ancient Glass Products Classification, Matched Samples T-Test, K-Means Clustering, Random Forest, Cross-Validation

1. 引言

玻璃是一种硅酸盐物质，主要成分为石英，熔点较高，由于条件限制，必须加入助溶剂降低熔化温度才能进行炼制。不同的助溶剂有多种不同的矿物组成。古代玻璃会因为各种原因而风化，造成其组成成分发生变化，风化的玻璃表面会出现斑点或各种颜色的沉积垢层，从而影响对其性质类别的判断[1]。

现如今，各种机器学习的算法广泛用于数据的分析中，而传统的统计学算法也经过不断使用和优化，证明其对数据的分析有很不错的效果。对玻璃制品进行分类时，认为能够通过对玻璃样品的化学成分进行定量分析来建立分类模型。分类模型的建立，对于玻璃文物的研究有着重要的意义，意味着人们不仅可以通过专家鉴别，也可以通过文物的化学成分定量分析去进行分类。

本文基于对文物的数据分析，量化分析玻璃文物的种类，合理地建立数学模型，同时将机器学习算法与统计学算法结合对分类模型的准确性以及灵敏度进行验证，得到的结果更加可靠，对玻璃文物的分类与鉴别有一定实际应用价值。

2. 玻璃文物的分类

2.1. 基于指标筛选的 K 均值聚类

1) 求解过程：如图 1

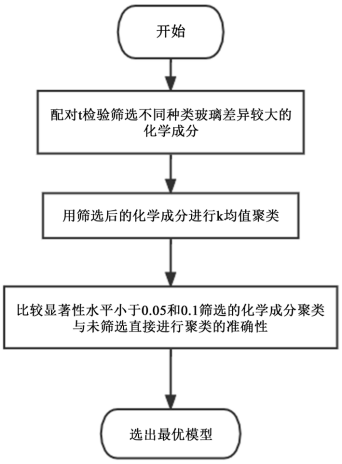

Figure 1. Algorithm flow
图 1. 算法流程

首先针对有无风化的高钾玻璃和铅钡玻璃四种研究对象，就其各个化学成分的平均值加减标准差进行配对 t 检验，通过 t 检验的参数分析数据对应的化学成分是否具有显著性差异，筛选出其中显著性水平较高的成分，认为可能是这些化学成分反映了玻璃的分类规律，通过建立模型与对照组(选用所有十四种化学成分进行 K 均值聚类)的正确率进行比较。

原假设 H0：两组化学成分含量的均值相同。

备择假设 H1：两组化学成分含量的均值不相同。

通过 t 检验，得到有风化高钾玻璃以及铅钡玻璃、无风化的高钾玻璃和铅钡玻璃的各个化学成分含量的均值差有显著性差异的几个成分，通过筛选这些成分，一方面能通过更少的指标进行分类，另外一方面可能有助于得到准确的高钾玻璃和铅钡玻璃的分类规律，并与对照组进行对比。

其次分别对高钾玻璃有无风化、铅钡玻璃有无风化的各个化学成分的平均值加减标准差进行配对 t 检验来分析高钾玻璃和铅钡玻璃有无风化的各化学成分的平均值加减标准差是否有显著性差异，进而推断高钾玻璃和铅钡玻璃有无风化的各化学成分是否有显著性差异，可能正是这些均值有显著性差异的化学成分反映了玻璃是否风化的分类规律，从而分别选出高钾玻璃和铅钡玻璃有无风化的各化学成分变化明显的几个，即分别区分高钾玻璃和铅钡玻璃是否风化的主要化学成分。

两样本 t 检验模型的建立，以有风化高钾玻璃以及铅钡玻璃的各个化学成分的平均值加减标准差为例，对无风化高钾玻璃以及铅钡玻璃的各个化学成分的平均值加减标准差进行配对 t 检验来分析无风化的高钾玻璃和铅钡玻璃的各化学成分的平均值加减标准差是否有显著性差异同理求解(分别对高钾玻璃有无风化、铅钡玻璃有无风化的各个化学成分的平均值加减标准差进行配对 t 检验来分析高钾玻璃和铅钡玻璃有无风化的各化学成分的平均值加减标准差是否有显著性差异亦同理求解)：

a) 设总体 $\{(ave\_\sigma)\_YZ_{Ji}\} \sim N(\mu_1, \sigma^2{}_1)$，$(YZ_{J_1}, YZ_{J_2}, \cdots, YZ_{J_i})$ 是来自有风化的高钾玻璃的各个化学成分的平均值加减标准差样本；总体 $\{(ave\_\sigma)\_YZ_{Bi}\} \sim N(\mu_2, \sigma^2{}_2)$，$(YZ_{B_1}, YZ_{B_2}, \cdots, YZ_{B_i})$ 是来自有风化的铅钡玻璃的各个化学成分的平均值加减标准差样本，同时设这两个样本一一对应；

b) 设一个变量 $W = \{(ave\_\sigma)\_YZ_{Ji}\} - \{(ave\_\sigma)\_YZ_{Bi}\}$ 对应得到 $(w_1, w_2, \cdots, w_i)$，其中 $w_i = YZ_{J_i} - YZ_{B_i}$ $(i = 1, 2, \cdots, 14)$，即转化成单样本 t 检验，即检验 W 的均值是否与 0 有显著性差异；

c) 假设 H_0： $\mu_W = 0$；

d) 构造统计量 $t_i = \dfrac{\bar{W}}{S_w / \sqrt[2]{14-1}} \sim t(14-1)$；

e) 计算 t_i 值与之对应的 p_i 值；

h) 判断是否拒绝原假设：

若 $p_i < 0.05$，则拒绝原假设，即认为有风化的高钾玻璃和铅钡玻璃的第 i 号化学成分的平均值加减标准差存在显著性差异。

若 $p_i > 0.05$，则无法拒绝原假设，即认为有风化的高钾玻璃和铅钡玻璃的第 i 号化学成分的平均值加减标准差不存在显著性差异。

2) 求解结果

共有 14 组(14 种化学成分)配对数据如表 1 所示，其中 6 组配对数据呈现出差异性($p < 0.05$)，分别为 SiO_2、K_2O、CaO、MgO、Al_2O_3、Fe_2O_3，分析如下：

经过分析可知用于区分高钾玻璃有无风化的主要化学成分有 SiO_2、K_2O、CaO、MgO、Al_2O_3、Fe_2O_3，并且差别明显，如表 1 所示。

有无风化的铅钡玻璃的各个化学成分的配对 t 检验分析结果可知，共有 14 组(14 种化学成分) 配对数据，其中 6 组配对数据呈现出差异性($p < 0.05$)，分别为 SiO_2、Na_2O、CaO、PbO、P_2O_5、SrO，即用于区

分无风化的高钾玻璃与铅钡玻璃的主要化学成分有 SiO_2、Na_2O、CaO、PbO、P_2O_5、SrO，并且差别很显著，如表 2 所示。

Table 1. Paired t-test results of mean chemical composition of lead-barium glass with or without weathering
表 1. 有无风化的高钾玻璃的化学成分的均值配对 t 检验结果

名称			配对(平均值 ± 标准差)		差值(配对 1-配对 2)	t	*p*
			配对 1	配对 2			
风化 1	配对	未风化 1	93.96 ± 1.73	68.87 ± 9.44	25.10	6.561	0.001**
风化 2	配对	未风化 2	0.00 ± 0.00	0.00 ± 0.00	0.00	null	null
风化 3	配对	未风化 3	0.54 ± 0.45	9.26 ± 2.58	−8.71	−8.701	0.000**
风化 4	配对	未风化 4	0.87 ± 0.49	4.78 ± 3.04	−3.91	−3.224	0.023*
风化 5	配对	未风化 5	0.20 ± 0.31	1.22 ± 0.72	−1.02	−4.479	0.007**
风化 6	配对	未风化 6	1.93 ± 0.96	6.43 ± 2.69	−4.50	−4.721	0.005**
风化 7	配对	未风化 7	0.27 ± 0.07	1.83 ± 0.94	−1.56	−3.995	0.010*
风化 8	配对	未风化 8	1.56 ± 0.93	2.95 ± 1.48	−1.39	−1.895	0.117
风化 9	配对	未风化 9	0.00 ± 0.00	0.31 ± 0.55	−0.31	−1.380	0.226
风化 10	配对	未风化 10	0.00 ± 0.00	0.71 ± 1.19	−0.71	−1.454	0.206
风化 11	配对	未风化 11	0.28 ± 0.21	1.41 ± 1.37	−1.13	−2.078	0.092
风化 12	配对	未风化 12	0.00 ± 0.00	0.05 ± 0.05	−0.05	−2.117	0.088
风化 13	配对	未风化 13	0.00 ± 0.00	0.00 ± 0.00	0.00	null	null
风化 14	配对	未风化 14	0.00 ± 0.00	0.20 ± 0.23	−0.20	−2.207	0.078

*$p < 0.05$，**$p < 0.01$。

Table 2. Paired t-test results of mean chemical composition of lead-barium glass with or without weathering
表 2. 有无风化的铅钡玻璃的化学成分均值配对 t 检验结果

名称			配对(平均值 ± 标准差)		差值(配对 1-配对 2)	t	*p*
			配对 1	配对 2			
风化 1	配对	未风化 1	24.47 ± 11.20	54.66 ± 11.83	−30.19	−8.927	0.000**
风化 2	配对	未风化 2	0.24 ± 0.59	1.68 ± 2.37	−1.44	−2.78	0.011*
风化 3	配对	未风化 3	0.14 ± 0.25	0.22 ± 0.31	−0.08	−1.077	0.293
风化 4	配对	未风化 4	2.79 ± 1.70	1.32 ± 1.28	1.47	3.689	0.001**
风化 5	配对	未风化 5	0.70 ± 0.72	0.64 ± 0.55	0.06	0.333	0.742
风化 6	配对	未风化 6	3.03 ± 2.79	4.46 ± 3.26	−1.43	−1.697	0.104
风化 7	配对	未风化 7	0.62 ± 0.76	0.74 ± 1.15	−0.11	−0.404	0.69
风化 8	配对	未风化 8	2.35 ± 2.97	1.43 ± 1.97	0.92	1.495	0.149
风化 9	配对	未风化 9	43.50 ± 13.00	22.08 ± 8.22	21.41	6.133	0.000**
风化 10	配对	未风化 10	11.59 ± 10.50	9.00 ± 5.83	2.59	1.071	0.296
风化 11	配对	未风化 11	5.46 ± 4.21	1.05 ± 1.85	4.42	4.301	0.000**
风化 12	配对	未风化 12	0.46 ± 0.25	0.27 ± 0.24	0.19	2.869	0.009**
风化 13	配对	未风化 13	0.08 ± 0.29	0.05 ± 0.13	0.03	0.479	0.637
风化 14	配对	未风化 14	1.54 ± 4.45	0.16 ± 0.76	1.39	1.453	0.16

*$p < 0.05$，**$p < 0.01$。

其中风化(未风化) 1~14 分别为玻璃样品的十四种化学成分，依次为二氧化硅(SiO_2)、氧化钠(Na_2O)、氧化钾(K_2O)、氧化钙(CaO)、氧化镁(MgO)、氧化铝(Al_2O_3)、氧化铁(Fe_2O_3)、氧化铜(CuO)、氧化铅(PbO)、氧化钡(BaO)、五氧化二磷(P_2O_5)、氧化锶(SrO)、氧化锡(SnO_2)、二氧化硫(SO_2)。

以下为高钾玻璃风化(未风化)、铅钡玻璃风化(未风化)样本的化学成分所占百分比的箱体图，分别如图 2~5 所示。

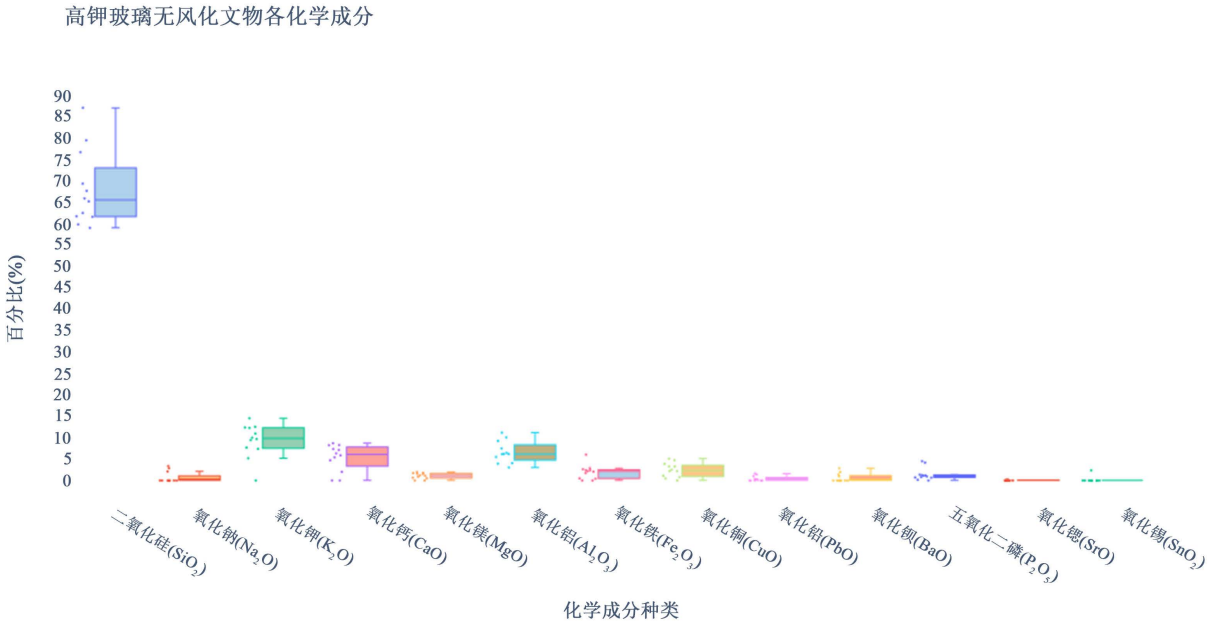

Figure 2. Box-plot of non-weathering chemical composition of high-potassium glass relics
图 2. 高钾玻璃文物无风化化学成分箱型图

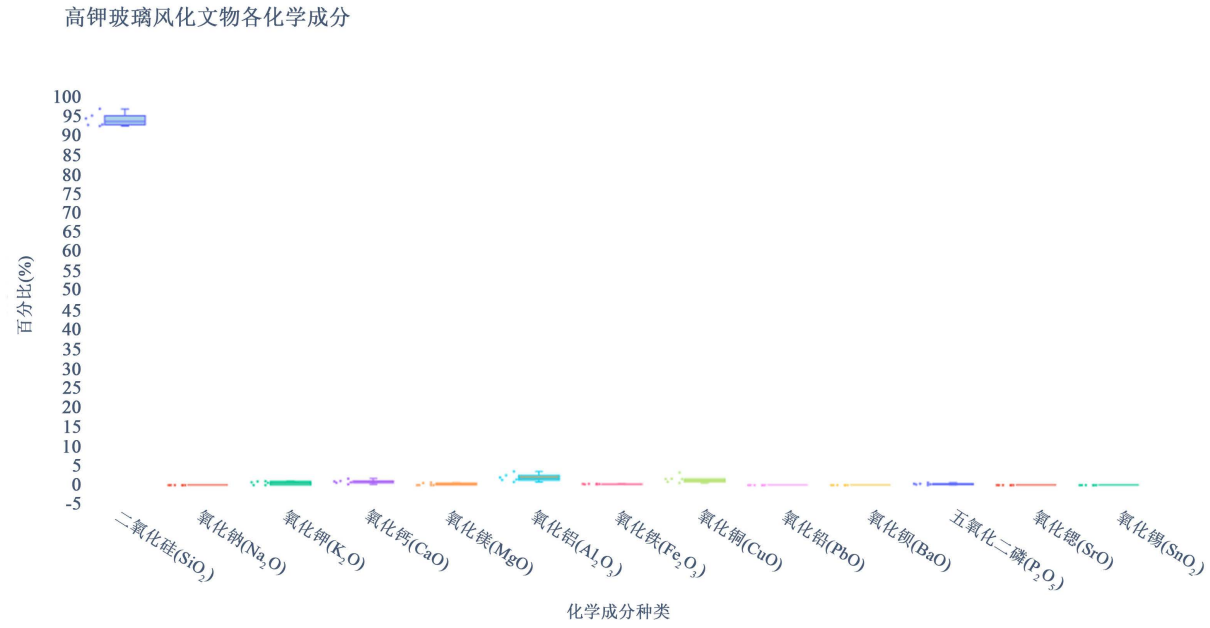

Figure 3. Box-plot of Weathering chemical composition of high potassium glass relics
图 3. 高钾玻璃文物风化化学成分箱型图

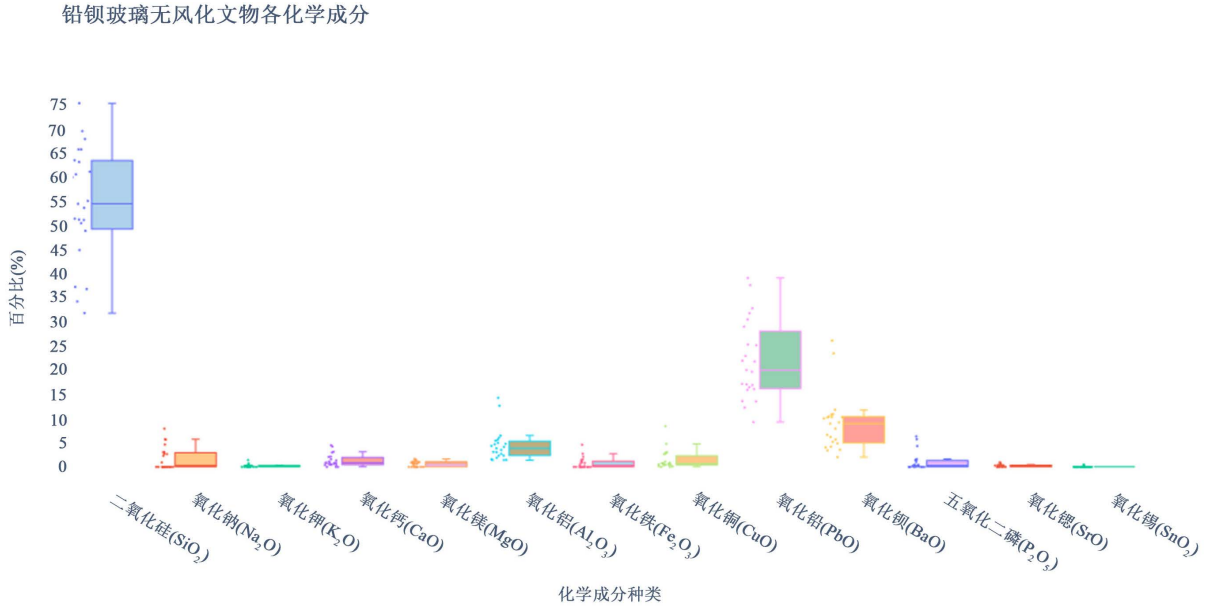

Figure 4. Box-plot of Non-weathering chemical composition of lead-barium glass relics
图 4. 铅钡玻璃文物无风化化学成分箱型图

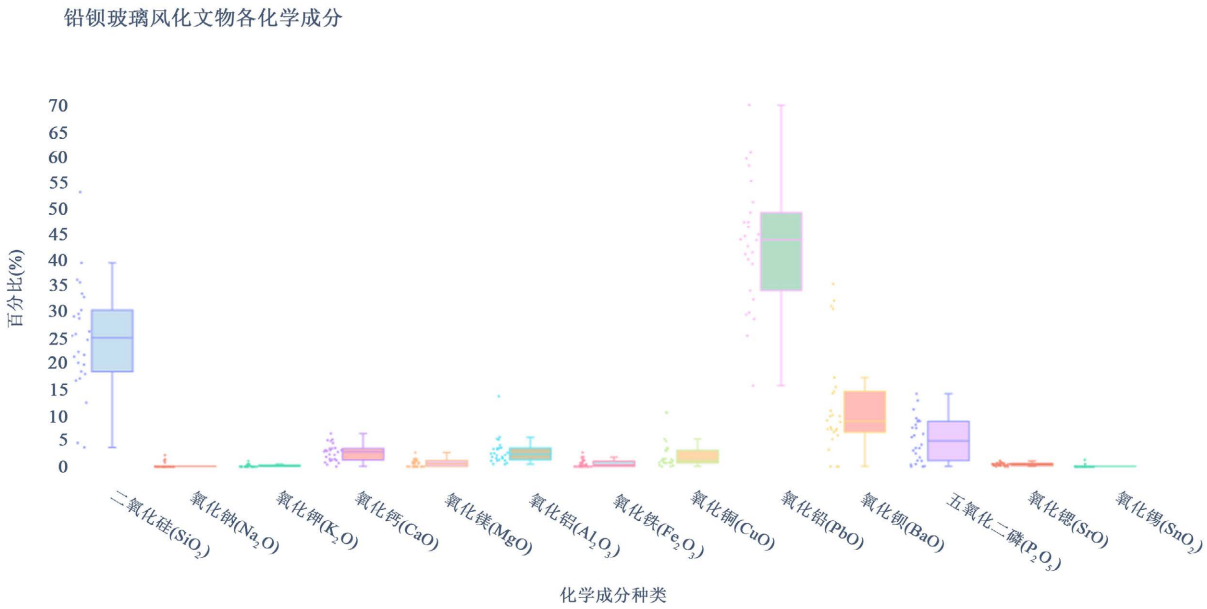

Figure 5. Box-plot of Weathering chemical composition of lead-barium glass relics
图 5. 铅钡玻璃文物风化化学成分箱型图

根据上述得到的高钾玻璃和铅钡玻璃的主要化学成分利用 K 均值聚类(欧氏距离)分别对高钾玻璃和铅钡玻璃聚类以及有无风化的玻璃进行聚类。

以对高钾玻璃(铅钡玻璃)进行分类的步骤进行阐述，在分类时，对有风化玻璃(无风化玻璃)的分类步骤相同，只是样本间距离的计算维度与之前不同种分类方法筛选出的指标数有关。

a) 定义以高钾玻璃(铅钡玻璃)主要化学成分的个数——多维度的空间里的欧氏距离[2]；

b) 计算高钾玻璃(铅钡玻璃)样本两两之间的距离：

$$d\left(J_p, J_q\right)=\sqrt[2]{\sum_{i\in e1}\left(J_{pi}-J_{qi}\right)^2} \tag{1}$$

其中，$e1$ 为高钾玻璃主要化学成分编号集合， p、$q\in r1$ ($r1$ 为高钾玻璃标号的集合且 $p\neq q$)

$$d\left(B_p, B_q\right)=\sqrt[2]{\sum_{i\in e1}\left(J_{pi}-J_{qi}\right)^2} \tag{2}$$

其中，$e2$ 为铅钡玻璃主要化学成分编号集合， p、$q\in r2$ ($r2$ 为铅钡玻璃标号的集合且 $p\neq q$)

K 均值聚类(K-means++)基本步骤如图 6 所示：

1) 随机选取一个样本作为第一个聚类中心；

2) 计算每个样本到已选择的聚类中心的距离，$D(X)$表示，$D(X)$越大，表示被选取作为聚类中心的概率较大；

3) 用轮盘法的方式选出下一个聚类中心($D(X)$越大，被选取聚类中心的概率越大)；

4) 重复步骤 2，直到选出 k 个聚类中心；

5) 选出 k 个聚类中心后，即选出初始点后就可以使用标准的 K-means 算法进行聚类[3]。

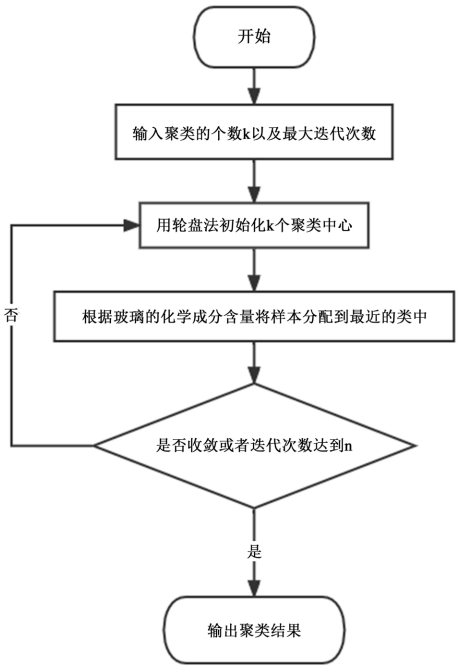

Figure 6. K-means Clustering algorithm flow
图 6. K 均值聚类算法流程图

基于以上筛选后的具有显著差异的化学成分对高钾玻璃(铅钡玻璃)的样本进行 K 均值聚类，建立的高钾玻璃和铅钡玻璃有无风化的分类模型分类中心分别如表 3、表 4 所示：

Table 3. Classification center of high potassium glass with or without weathering
表 3. 有无风化的高钾玻璃的分类中心

	风化	未风化
二氧化硅	89.66	63.62
氧化钾	1.99	10.82

Continued

氧化钙	1.33	6.36
氧化镁	0.44	1.13
氧化铝	2.76	7.35
氧化铁	0.44	2.31

Table 4. Classification center of lead-barium glass with or without weathering
表 4. 有无风化的铅钡玻璃的分类中心

	风化	未风化
二氧化硅	24.91	57.49
氧化钠	0.17	1.88
氧化钙	2.73	1.14
氧化铅	43.45	19.88
五氧化二磷	4.92	1.13
氧化锶	0.44	0.22

利用上述公式(1) $d\left(J_p,J_q\right)=\sqrt[2]{\sum_{i\in e1}\left(J_{pi}-J_{qi}\right)^2}$ 和公式(2) $d\left(B_p,B_q\right)=\sqrt[2]{\sum_{i\in e2}\left(B_{pi}-B_{qi}\right)^2}$ 即可得出分类结果，其中 J_{qi} 和 B_{qi} 为聚类中心的值，J_{pi} 和 B_{pi} 为要预测玻璃样品采样的值，根据六维空间样本距离分类中心的距离进行分类，结果如图 7 和图 8 所示：

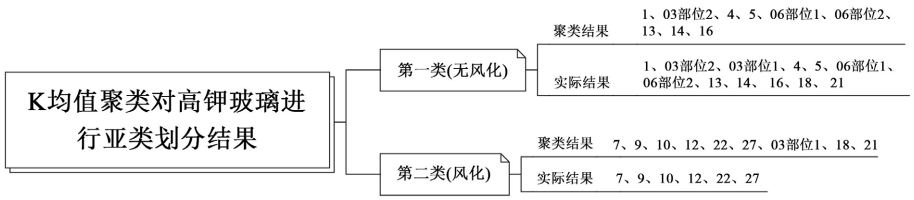

Figure7. K-means clustering results of sub-classification of high potassium glass
图 7. K 均值聚类对高钾玻璃亚分类结果

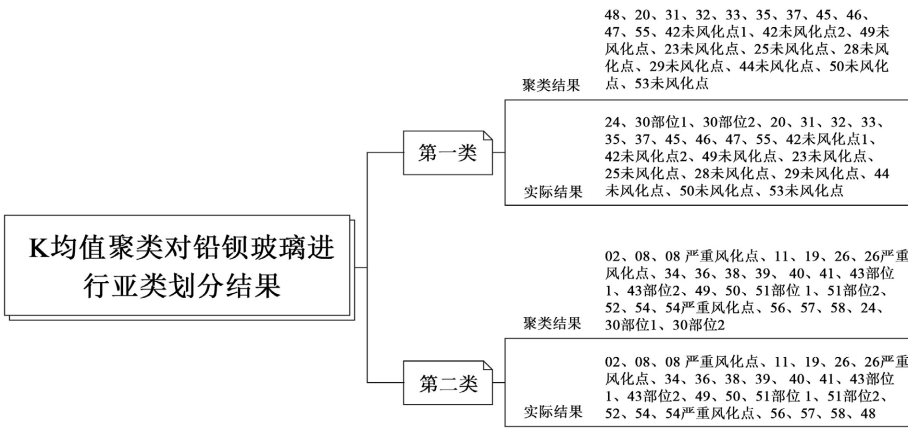

Figure 8. K-means clustering results for subclassification of lead barium glasses
图 8. K 均值聚类对铅钡玻璃亚分类结果

对比实际分类可以知道，按此方法对高钾玻璃和铅钡玻璃进行亚类划分，聚类结果显然是将高钾玻璃和铅钡玻璃分为有无风化的部分，现在已知每个高钾玻璃和铅钡玻璃样本实际的风化情况，所以可以依次对聚类结果的准确性进行验证，其中铅钡玻璃分类模型的准确性稍差，以显著性水平调整为小于 0.01 为指标重新筛选化学成分进行聚类后，准确性有了提高，如上图知，对两种玻璃亚类划分的正确数分别为 15、45，准确率分别达到了 83.3%、91.8%，可以推出此种方法建立的聚类模型对高钾玻璃和铅钡玻璃是否风化的判断具有较高的准确率。

$$\text{准确率} = \frac{\text{无扰动分类结果与扰动后分类结果相同数}}{\text{样本总数}} * 100\%$$

通过上述的 K 均值聚类，已经建立了可靠的模型，得到了高钾、铅钡玻璃的化学成分含量的聚类中心，以欧式空间距离作为分类依据，将要预测的样本的化学成分含量与聚类中心之间距离的远近来对样本进行预测。

$$d_{i,j} = x\left(x_{i,j} - \overline{x}_i\right) \tag{3}$$

$d_{i,j}$ 为第 j 个样本到聚类中心的距离，$x_{i,j}$ 为第 j 个样本的第 i 个化学成分含量，$\overline{x}_i$ 为第 i 个化学成分含量的聚类中心值。

同样，同时为了避免过拟合现象的出现，找到八组已知风化情况的样本采样点化学成分含量，通过对这部分学习的样本外的数据按照高钾、铅钡玻璃两类进行预测，既对模型进行了验证，也对过拟合是否发生进行了判断，对八项样本预测的结果如表 5 所示：

Table 5. Results of eight sample forecasts
表 5. 八项样本预测的结果表

分类	预测
高钾玻璃	A1、A5、A6、A7
铅钡玻璃	A2、A3、A4、A8

其中，根据前面对高钾和铅钡玻璃的化学成分分析，A5 与铅钡玻璃的成分含量更相似，于是为了结果的可靠性和正确性，选择使用随机森林学习玻璃分类规律，使用随机森林分类的方法对样本进行再次预测，提高结果的稳健性，也进一步得到更合适的模型。

2.2. 随机森林分类

随机森林算法是一种有监督算法，基于多个决策树的一种弱分类算法。每次在样本总体中有放回的随机选择数量相同的一部分样本进行训练，每次训练都建立一个决策树，模型中包含多个决策树，且随机森林算法在处理高维数据时具有很快的速度，玻璃文物中含有多种化学成分，因此这种算法十分适合对玻璃文物进行分类[4]。

对于随机森林的方法，不需要在使用前像聚类一样对成分进行初步筛选，在模型建立时，对每棵决策树进行训练时，不仅对样本进行随机选取，而且对样本的化学成分也进行随机选取，能够在对样本的不同种化学成分的选取大量训练中分析出每个成分对于分类结果的重要性的差别，即每次训练的样本选取的均为原始样本的一个子集，建立多个训练集矩阵[5]。

$$\begin{bmatrix} A_{i_{11}} & \cdots & A_{i_{1n}} \\ \vdots & \ddots & \vdots \\ A_{i_{m1}} & \cdots & A_{i_{mn}} \end{bmatrix}$$

对于以分类为目的的随机森林模型，按照投票产生结果，即投票最多的选项为分类结果。使用此模型的目的是对样本之外的玻璃文物类型进行预测，所以按照平均法计算所有随机数的结果的均值作为结论。通过建立的随机森林计算特征重要性，特征的重要性越强，对预测结果的影响越大，特征重要性分别如图 9 和图 10 所示。

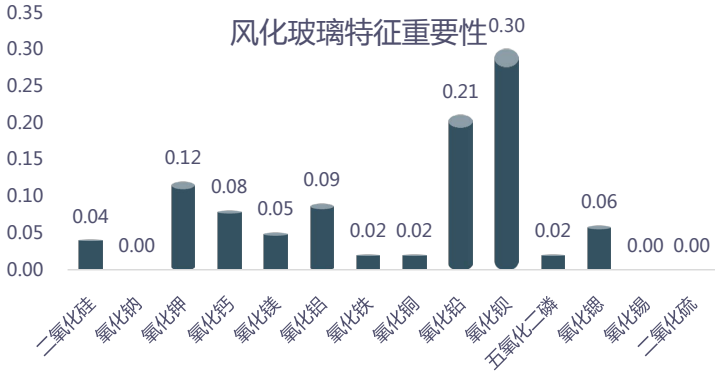

Figure 9. Importance of weathered glass type characteristics
图 9. 风化的玻璃类型特征重要性

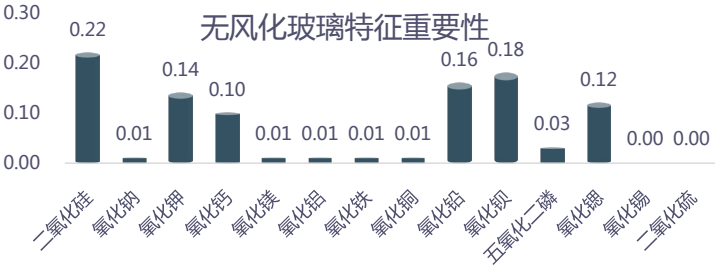

Figure 10. Importance of non-weathering glass type characteristics
图 10. 无风化的玻璃类型特征重要性

同样，同时为了避免过拟合现象的出现，找到八组已知风化情况的样本采样点化学成分含量，通过对这部分学习的样本外的数据按照高钾、铅钡玻璃两类进行预测，既对模型进行了验证，也对过拟合是否发生进行了判断，实验结果证明随机森林分类模型具有较好的准确性。随机森林模型对已知风化情况的高钾、铅钡玻璃样品的分类结果准确率在训练集和测试集中均在 95%以上，在交叉验证是也分别能达到 81%、83.3%。

随机森林模型分类预测结果如表 6 所示：

Table 6. Classification and prediction of stochastic forest model
表 6. 随机森林模型分类预测

分类	预测
高钾玻璃	A1、A6、A7
铅钡玻璃	A2、A3、A4、A8、A5

发现与第一个分类模型预测结果不同的是 A5 样本，结合不同类别玻璃的描述统计以及 t 均值差检验中的参数，认为随机森林分类模型对 A5 的预测结果更加可靠。

以对已知风化情况的玻璃分为高钾玻璃、铅钡玻璃两类的步骤进行阐述，在分类时，对已知是高钾玻璃或铅钡玻璃的风化情况的分类步骤以及方法相同，只是在训练前对样本进行了重新的划分，即可得到适用于不同分类条件的模型。但随机森林模型不能得到确定的方程，所以作为验证模型准确性的一种方法。

2.3. 模型敏感性的测试

经验证了上述两种方法构建的模型具有较高的准确率，但在实际应用时仍需要对模型的可靠性进行进一步验证。一件玻璃文物不同地方采样获得的各项化学成分含量可能不同，为了验证以上两种方法建立的模型在出样本中包含的采样点外仍具有一定的适用性，需对其进行敏感性测试。同一件文物，大概率具有相似的化学成分含量，对采样点处的数据进行扰动处理后产生的数据认为是该件文物别处的化学成分含量，这种方法一方面避免了需再次采样数据的问题，另外一方面这种方法具有一定的可行性。

首先通过 matlab 生成多组扰动数据，14 个化学成分含量的均值、波动范围以及最大值和最小值是不同的，所以对于随机生成相同范围的扰动值显然是不合理的，尤其是有多组成分含量接近于 0，如果以相同的扰动范围去加入干扰值，一方面可能会出现负数的情况，而成分含量值的定义是这个成分在总成分含量中所占的百分比，显然不符合实际情况，另一方面，如果扰动值以相同的范围加入，对于含量接近 0 的那部分成分，干扰会是几倍甚至数十倍，产生很大程度的失真，以此数据去测试敏感度大小的可靠性也不具有足够的可信度，于是考虑以各项指标的三倍标准差或最大值与最小值的差值作为扰动范围，加以合适的裕度去限定扰动后的数据范围，经过多次尝试后，认为以加入裕度最大值与最小值的差值作为扰动范围最为合适，效果最佳。首先生成一组−1~1 之间的随机数向量，并定义一个扰动系数，在多次实验后，选择以 10%、50%、100%扰动后的数据对分类情况的敏感性进行判断，对于高钾风化、高钾无风化、铅钡风化、铅钡无风化四组数据，他们的各项化学成分的含量的均值以及范围不相同，所以生成不同的干扰数据时，选择的最大值、最小值、平均值为对应的数据，生成的扰动值为−1~1 之间的随机数向量乘以扰动系数再乘以对应数据波动的范围，将此扰动值加上原数据，产生加入扰动后的数据。

在多次实验后，发现当加入较大扰动系数(50%，100%或更大) 时，一些数据虽然通过 K 均值聚类，仍然得到了不错的效果，但是经过统计分析，发现很多组数据的含量小于 85%或大于 105%，即实际上已经成为无效数据，一部分接近于 0 的数据加入负的干扰后，成为了负数，但也在 0 附近，将其重新赋值为 0，一方面没有对数值有过大的改动，另一方面，也保留了数据的特征，所以认为这种方法具有可行性，考虑到对其他均值显著大于 0 的化学成分含量的扰动具有足够的合理性，便不对加入扰动的方法进行更改，而对扰动后的数据进行优化，通过求和得到加入扰动后每组样本化学成分含量的总和，用每项化学成分含量除以所在样本化学成分含量的总和，进行了归一化处理，处理后的数据基本都为有效数据，也让结果具有更高的可信度。

分别对 K 均值聚类建立的分类模型和随机森林分类的预测模型进行测试，结果如表 7~9 所示。

Table 7. K-means clustering sensitivity test results of high potassium glass
表 7. 高钾玻璃的 K 均值聚类敏感性测试结果

文物采样点	实际类型	10%扰动分类结果	50%扰动分类结果	100%扰动分类结果
7	风化	风化	风化	风化
9	风化	风化	风化	风化
10	风化	风化	风化	风化
12	风化	风化	风化	风化

Continued

22	风化	风化	风化	风化
27	风化	风化	风化	风化
1	无风化	无风化	无风化	无风化
03 部位 1	无风化	风化	风化	风化
03 部位 2	无风化	无风化	无风化	无风化
4	无风化	无风化	无风化	无风化
5	无风化	无风化	无风化	无风化
06 部位 1	无风化	无风化	无风化	无风化
06 部位 2	无风化	无风化	无风化	无风化
13	无风化	无风化	无风化	无风化
14	无风化	无风化	无风化	无风化
16	无风化	无风化	无风化	无风化
18	无风化	风化	风化	风化
21	无风化	风化	风化	风化

Table 8. K-means clustering sensitivity test results of lead-barium glass
表 8. 铅钡玻璃的 K 均值聚类敏感性测试结果

文物采样点	实际类型	10%扰动结果	50%扰动结果	100%扰动结果
02	风化	风化	风化	风化
08	风化	风化	风化	风化
08 严重风化点	风化	风化	风化	风化
11	风化	风化	风化	风化
19	风化	风化	风化	风化
26	风化	风化	风化	风化
26 严重风化点	风化	风化	风化	风化
34	风化	风化	风化	风化
36	风化	风化	风化	无风化
38	风化	风化	风化	风化
39	风化	风化	风化	风化
40	风化	风化	风化	风化
41	风化	风化	风化	风化
43 部位 1	风化	风化	风化	风化
43 部位 2	风化	风化	风化	风化
49	风化	无风化	无风化	无风化
50	风化	风化	风化	风化
51 部位 1	风化	风化	风化	风化

Continued

51 部位 2	风化	风化	风化	风化
52	风化	风化	风化	风化
54	风化	风化	风化	风化
54 严重风化点	风化	风化	风化	风化
56	风化	风化	风化	风化
57	风化	风化	风化	风化
58	风化	风化	风化	风化
48	风化	风化	风化	风化
24	无风化	无风化	无风化	无风化
30 部位 1	无风化	风化	风化	风化
30 部位 2	无风化	风化	风化	风化
20	无风化	风化	风化	风化
31	无风化	无风化	无风化	无风化
32	无风化	无风化	无风化	无风化
33	无风化	无风化	无风化	无风化
35	无风化	无风化	无风化	无风化
37	无风化	无风化	无风化	无风化
45	无风化	无风化	无风化	无风化
46	无风化	无风化	无风化	无风化
47	无风化	无风化	无风化	无风化
55	无风化	无风化	无风化	无风化
42 未风化点 1	无风化	无风化	无风化	无风化
42 未风化点 2	无风化	无风化	无风化	无风化
49 未风化点	无风化	无风化	无风化	无风化
23 未风化点	无风化	无风化	无风化	无风化
25 未风化点	无风化	无风化	无风化	无风化
28 未风化点	无风化	无风化	无风化	无风化
29 未风化点	无风化	无风化	无风化	无风化
44 未风化点	无风化	无风化	无风化	无风化
50 未风化点	无风化	无风化	无风化	无风化
53 未风化点	无风化	无风化	无风化	无风化

Table 9. Sensitivity test results of random forest classification of high potassium and lead barium glasses
表 9. 高钾、铅钡玻璃的随机森林分类敏感性测试结果

文物采样点	10%扰动预测结果	10%扰动预测结果	100%扰动预测结果
A2	铅钡	铅钡	铅钡
A5	铅钡	铅钡	铅钡

Continued

A6	高钾	高钾	高钾
A7	高钾	高钾	高钾
A1	高钾	高钾	高钾
A3	铅钡	铅钡	铅钡
A4	铅钡	铅钡	铅钡
A8	铅钡	铅钡	铅钡

对 K 均值聚类模型和随机森林分类模型分类结果的敏感性分别验证，以无扰动的分类结果作为对照组计算准确度，得到了 K 均值对高钾玻璃亚类分类的准确率分别为 100% (10%扰动)，100% (50%扰动)，100% (100%扰动)，对铅钡玻璃亚类分类的准确率分别为 91.8367% (10%扰动)，91.8367% (50%扰动)，89.7959% (100%扰动)；随机森林分类模型对已知风化情况的玻璃分类在 10%、50%、100%扰动下的准确度均为 100%。扰动数据的生成具有随机性，因此多次实验的结果可能不完全相同，但都有较好的稳定性。

3. 结论

通过对五十余件玻璃文物采样的化学成分的研究，运用 K 均值聚类和随机森林分类模型对玻璃进行亚分类，建立模型后对样本进行敏感性分析，使得分类结果十分可靠。对于偏离样本程度不同的扰动后的数据进行了优化处理，让更多扰动生成的数据为有效数据，对于随机产生数据的偶然性进行了优化，同时对于预测结果有争议的样本，通过两种方法的比较进行再次验证，如将两种模型预测的概率取均值参考，得到了正确率和稳健性更好的模型，具有一定实际应用价值。

参考文献

[1] 朱瑛培. 新疆鄯善县洋海墓地出土玻璃珠的成分体系和制作工艺研究[D]: [硕士学位论文]. 西安: 西北大学, 2018.

[2] 聂益芳, Mbugua Allan Wainaina, 李余, 姚行艳, 蔡雪松. 无线信道建模中二分 K 均值聚类径分簇算法[J/OL]. 电波科学学报: 1-8. https://kns.cnki.net/kcms/detail/41.1185.tn.20220613.1209.002.html, 2022-09-18.

[3] 杨文君. 入侵检测技术中 k-means 聚类算法综述[J]. 科学技术创新, 2018(36): 65-66.

[4] 李嘉康, 陶智麟, 徐波, 徐大勇, 堵劲松, 李华杰. 基于随机森林的烟叶纹理定量分析[J]. 湖北农业科学, 2022, 61(14): 155-159.

[5] 吕广旭, 卢加奇, 魏先燕, 王小英. 基于随机森林-聚类混合方法的多分类入侵检测研究[J]. 现代信息科技, 2022, 6(16): 165-167.