教学与竞赛

古代玻璃成分分析与亚类划分方法研究

楼 阳，窦 雷，卓朝阳，鲁 萍

（西安建筑科技大学 理学院，陕西 西安 710399）

摘 要：基于古代玻璃制品的化学成分数据分析其分类规律以及亚类划分方法. 首先对原始数据进行中心对数比变换后分析风化和无风化数据的转换关系；然后，通过构建有监督特征选择方法探索高钾、铅钡类别规律，进而基于无监督特征选择方法探索亚类划分方法，重点构建了过滤式特征选择结合特征 R 型聚类和封装式特征迭代选择两种亚类划分方法，第二种方法可以给出多种划分方案和结果.

关键词：中心对数比变换；有监督特征选择；无监督特征选择；过滤式；封装式；亚类划分

中图分类号：O29　　**文献标志码**：A　　**文章编号**：2095-3070(2023)04-0073-11

DOI：10.19943/j.2095-3070.jmmia.2023.04.10

0 引言

玻璃是我国丝绸之路早期贸易往来的宝贵物证，其中由石英砂和铅矿石等制成的铅钡玻璃与石英砂和草木灰制成的高钾玻璃[1]最具代表性.

由于玻璃成分复杂，且易被所处环境影响，导致玻璃风化且大量内部元素与外部元素交换，极大地影响了考古工作者对其类型的判断，因此古代玻璃的成分分析、鉴别及亚类划分成为了考古工作者们研究的巨大障碍. 在目前的研究中，将高钾玻璃和铅钡玻璃分为 4 类：高钾风化、高钾无风化、铅钡风化和铅钡无风化[2]. 可以利用 LA-ICP-MS(激光剥蚀电感耦合等离子体质谱仪)技术准确测定古代文物成分[3]，并且王承遇等[4]研究了影响玻璃风化的因素及风化机理，对于成分预测有一定的参考意义；董涵等[5]基于主成分分析方法研究了高钾、铅钡类玻璃的分类规律，利用层次聚类方法选择主导化学成分，对两类玻璃进行亚类划分；宛惠等[6]基于古代玻璃制品化学成分的数据分析，探索了玻璃文物的分类方法. 但是目前对高钾玻璃和铅钡玻璃亚类划分提出的方法较为单一，本文基于 2022 年“高教社杯”全国大学生数学建模竞赛 C 题[7]“古代玻璃制品的成分分析与鉴别”相关数据，针对基于过滤式特征选择结合特征 R 型聚类的亚类划分方法、封装式特征迭代选择的亚类划分方法，对两类玻璃的亚类划分进行具体研究，为技术工作人员提供指导意见.

1 风化预测

1.1 数据预处理

1.1.1 异常值处理

因为玻璃样品中各成分占比之和在 85%～105%之间的数据视为有效数据，因此对附件表单 2 中各样品的成分占比进行分析，删除第 15、第 17 组的成分占比数据.

收稿日期：2023-10-10

通讯作者：鲁萍，E-mail：lping@xauat.edu.cn

引用格式：楼阳，窦雷，卓朝阳，等. 古代玻璃成分分析与亚类划分方法研究[J]. 数学建模及其应用，2023，12(4)：73-83.

LOU Y，DOU L，ZHUO ZH Y，et al. Study on the compositional analysis and subcategory classification method of ancient glass(in Chinese)[J]. Mathematical Modeling and Its Applications，2023，12(4)：73-83.

1.1.2 对原始数据进行中心对数比变换[8]

因为针对玻璃样品各化学成分比例的累加和应为 100%，所以本文通过中心对数比变换(CLR)进行数据转换，以消除定和限制对后续分析的影响. 假设 p 维成分数据为 $\boldsymbol{x}=(x_1, x_2, \cdots, x_p)$，则中心对数比变换公式为：

$$\mathrm{CLR}(\boldsymbol{x})=\left\{\ln\frac{x_1}{g(\boldsymbol{x})}, \ln\frac{x_2}{g(\boldsymbol{x})}, \cdots, \ln\frac{x_p}{g(\boldsymbol{x})}\right\},$$

其中，$g(\boldsymbol{x})=\sqrt[p]{x_1x_2\cdots x_p}$ 表示文物表面是否风化. 通过对初始成分数据与中心对数比变换后数据制作 qq 图，发现中心对数比变换后数据更加满足正态性.

1.2 基于方差分析的玻璃文物风化前后的成分分析

1.2.1 中心对数比变换后的成分数据的统计特征

分别针对中心对数比变换后的有风化、无风化的高钾玻璃和铅钡玻璃计算各化学成分含量的均值与方差. 发现风化对高钾玻璃的 SiO_2、Na_2O、K_2O、CuO、Fe_2O_3 与 BaO 等化学成分与铅钡玻璃的 SiO_2、Na_2O、K_2O、CaO 与 Fe_2O_3 等化学成分有较大影响.

1.2.2 中心对数比变换数据方差分析[9]的统计特征

通过对经过中心对数比变换后的数据进行方差分析，得到高钾类玻璃有无风化情况下各成分方差分析的 p 值，其中 SiO_2 的 p 值远远小于 0.001，可以得到高钾类玻璃风化会引起 SiO_2 成分比例的极其显著变化，且 K_2O 与 CuO 的 p 值也较小，说明风化也会引起 K_2O 与 CuO 成分一定程度的变化.

同理对铅钡玻璃有无风化情况下各成分方差分析的 p 值进行分析，发现铅钡玻璃风化也会引起 SiO_2 成分比例的极其显著变化，因为其 p 值为 0.000 554，远远小于 0.001，而 K_2O 与 CaO 的 p 值均小于 0.1，因此风化也会引起其成分一定程度的变化.

1.2.3 结论

1)针对高钾玻璃：风化会引起 SiO_2 成分比例的显著变化，且在一定程度上会引起 K_2O 和 CuO 成分比例的变化；

2)针对铅钡玻璃：风化会引起 SiO_2 成分比例的显著变化，且在一定程度上会引起 CaO 和 K_2O 成分比例的变化.

1.3 基于正态总体特征推测文物风化前成分含量

假设中心对数比变换后的数据服从正态分布，X 表示风化样本集合，Y 表示无风化样本集合，CLR(X)和 CLR(Y)表示分别对 X 和 Y 进行中心对数比变换后的数据集，分别统计其均值与方差：

$$\mathrm{CLR}(X)\sim N(\mu_x, \sigma_x^2), \quad \mathrm{CLR}(Y)\sim N(\mu_y, \sigma_y^2). \tag{1}$$

对风化样本 $\boldsymbol{x}$，其中心对数比变换后为 CLR($\boldsymbol{x}$)，预测该样本无风化的中心对数比变换后的数据 CLR($\boldsymbol{y}$)：

$$\mathrm{CLR}(\boldsymbol{y})=\mu_y+\frac{\sigma_y}{\sigma_x}(\mathrm{CLR}(\boldsymbol{x})-\mu_x). \tag{2}$$

对 CLR($\boldsymbol{y}$)进行中心对数比变换的逆变换可计算得到对该样本预测的无风化状态各成分的含量. 用如上方法对风化高钾玻璃样本进行无风化预测，部分结果如表 1 所示.

表 1 预测风化高钾玻璃文物采样点成分含量部分结果 %

采样点	SiO_2	Na_2O	K_2O	CaO	Al_2O_3
07	66.438 648	0	0	16.291 676	5.968 738
09	70.260 280	0	16.634 269	0.907 423	4.944 362
10	76.363 292	0	14.813 301	0.017 175	1.849 775

同理可以预测得到风化铅钡类玻璃文物采样点各成分的含量.

2 分析高钾玻璃与铅钡玻璃的分类规律

分析高钾、铅钡玻璃的分类规律，需要找到可以区分两类玻璃的特征以及规律. 该问题是有监督

的特征选择，通过方差分析确定不同类别中均值有显著差异的特征，通过随机森林模型确定特征的重要性排序，通过决策树模型确定具体的分类规则.

2.1 变量初筛

本文对原始成分数据统计各成分的均值、极差、极大值、极小值、方差等，发现 SO_2、SrO 和 SnO_2 在高钾玻璃与铅钡玻璃中占比均值均过低，所以筛除掉这些化学成分，便于后续分类.

2.2 方差分析法确定显著分类特征

对无风化、风化的样本分别探究分类规律. 对无风化的高钾、铅钡玻璃的各成分中心对数比变换后的数据进行方差分析，选择在不同分类中有显著差异的特征，由方差分析得到无风化样本中高钾、铅钡玻璃各成分 p 值的结果，其中 PbO、BaO、K_2O 和 Al_2O_3 的 p 值均远远小于 0.01，所以在无风化高钾、铅钡玻璃样本中 PbO、BaO、K_2O 和 Al_2O_3 对分类有显著的统计学差异.

2.3 随机森林模型确定特征重要性[10]

随机森林是由多棵决策树构成的集成学习模型，通过每个特征对袋外误差(OOB error)的影响程度确定特征的重要性. 无风化样本中特征的重要程度如表 2 所示.

表 2 随机森林各成分重要程度(中心对数比变换数据)

成分	特征重要性	排序	成分	特征重要性	排序	成分	特征重要性	排序
K_2O	0.190 174	1	CaO	0.104 466	5	CuO	0.010 149	9
PbO	0.170 732	2	SiO_2	0.030 930	6	MgO	0.009 524	10
BaO	0.159 668	3	P_2O_5	0.023 168	7	Fe_2O_3	0.066 589	11
Al_2O_3	0.110 354	4	Na_2O	0.019 471	8			

由表 2 可知，重要性排序前 4 名的 K_2O、PbO、BaO 和 Al_2O_3 为区别无风化样本高钾、铅钡玻璃分类的最重要特征，与方差分析结果一致.

2.4 决策树模型确定分类规律

建立决策树模型[11]，通过信息熵考查节点不纯度，信息熵计算公式如下：

$$\text{Entropy}(t) = -\sum_{i=1}^{2} p(i|t)\log_2 p(i|t)\,, \tag{3}$$

其中，$p(i|t)$表示节点 t 为分类 i 的概率.

将训练占比设定为 0.7，挑选原样本中的 70%作为训练数据，用剩下的 30%作为测试数据，决策树如图 1 所示.

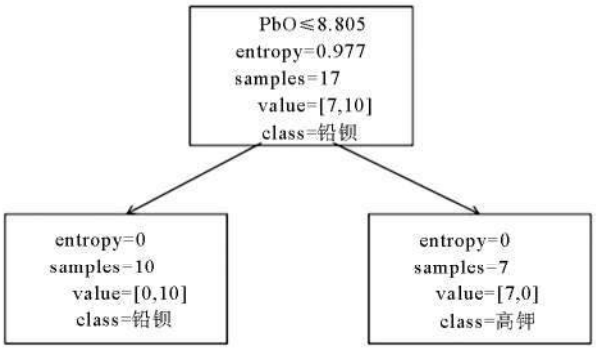


图 1 无风化玻璃分类规律决策树

由图 1 可知：对无风化高钾、铅钡样本，当 PbO 含量小于或等于 8.805 时，判断为高钾玻璃类；当 PbO 含量大于 8.805 时，判断为铅钡玻璃类.

按同样步骤分析有风化高钾、铅钡玻璃分类规律的结果：当该玻璃 PbO 含量小于或等于 6.161 时，将其归于高钾玻璃类；当该玻璃中 PbO 含量大于 6.161 时，将其归于铅钡玻璃类.

3 两类玻璃的亚类划分

亚类划分首先需要找到可以分类的特征子集，然后基于特征子集进行亚类划分，最后依据亚类的统计学特点明确亚类的分类规律. 该问题是无监督学习问题.

思路 1：使用无监督过滤式特征选择方法确定主要的特征集[12]，用 R 型聚类[13]刻画特征集中特征的相似关系，筛选出不相似的特征构成用于亚类划分的典型特征集，用典型特征集进行亚类划分，并明确划分规则.

思路 2：采用基于封装式(wrapper)特征选择的亚类划分方法，将特征选择过程与模型训练过程结合，根据由特征子集训练得到的模型性能来评价特征子集. 依据给定策略产生候选的特征子集，使用轮廓系数对该特征子集的亚类划分优劣进行评估，不断迭代评估，直到达到停止条件，最后用所选的特征子集进行亚类划分.

思路 3：对无监督样本使用 Q 型聚类对样本进行聚类，初步形成亚类，再使用有监督特征选择方法对该亚类进行分析，找到分类规律. 具体方法详见文献[6].

思路流程图如图 2 所示，思路 3 的详细分析步骤可参考文献[6]，本文仅对思路 1 和思路 2 作详细阐述.

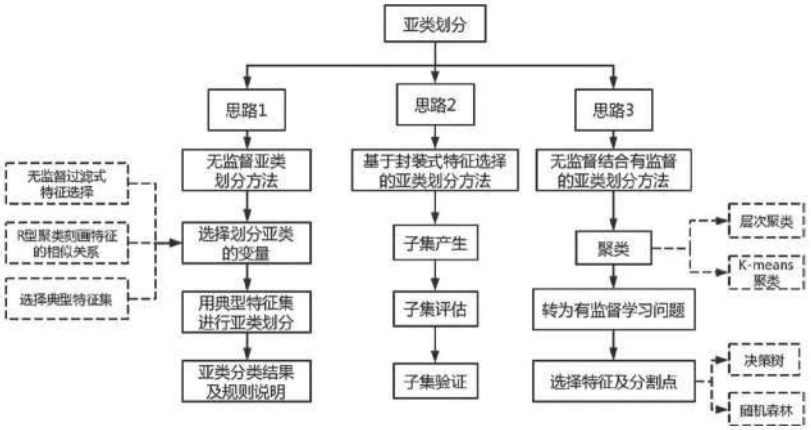


图 2 思路流程图

考虑对风化后的样本进行亚类划分现实意义不强，所以仅针对高钾玻璃与铅钡玻璃无风化的样本进行特征选择以及亚类划分.

3.1 思路 1——无监督亚类划分方法

首先选择划分亚类的变量.

第 1 步：无监督过滤式特征选择；

第 2 步：用 R 型聚类刻画特征集中特征的相似关系；

第 3 步：结合 1、2 步的结果选择典型特征集.

其次用典型特征集对样本集进行亚类划分，最后对亚类分类结果及规则进行说明.

以无风化高钾样本集为例进行分析，用同样的方法可以对铅钡无风化样本进行亚类划分，本文不再详细论述.

3.1.1 无监督过滤式特征选择

过滤式特征选择方法需要确定一个特定统计量来衡量特征的重要性，设定一个重要性阈值，选取那些重要性分量大于该阈值的特征. 此处选择中心对数比变换后特征的方差作为衡量特征重要性的统计量. 在成分数据经过中心对数比变换后，将成分按方差排序，选择方差和大于总方差的 80%的特征. 高钾无风化玻璃各成分的方差如表 3 所示.

表 3 高钾无风化中心对数比变换后数据方差

成分	方差	方差累计占比	排序	成分	方差	方差累计占比	排序
Na_2O	6.617 630	0.153	1	SO_2	2.879 400	0.731	7
CaO	5.457 156	0.279	2	MgO	2.842 339	0.797	8
BaO	5.065 116	0.396	3	Fe_2O_3	2.798 143	0.861	9
K_2O	4.080 777	0.491	4	P_2O_5	2.552 119	0.920	10
PbO	3.859 857	0.580	5	CuO	1.949 150	0.965	11
SnO_2	3.651 824	0.664	6				

由表 3 可知各成分的方差排序，且占总方差 80%的成分是 Na_2O、CaO、BaO、K_2O、PbO、SnO_2、SO_2、MgO 和 Fe_2O_3，这些成分构成重要特征集，其中 Na_2O 在各特征成分中的排序最高.

3.1.2　用 R 型聚类刻画重要特征集中特征的相似关系

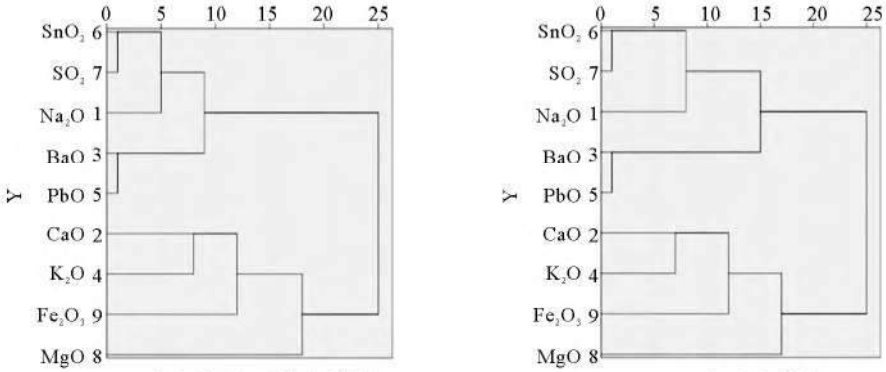
对重要特征集中的特征用 R 型聚类方法刻画相似关系，选择余弦相似度作为特征近似程度的度量指标，分别用组间联接和质心距离法作为簇间距离测算方法，用层次聚类法对特征进行聚类.

余弦相似度是 n 维空间中两个向量 $\boldsymbol{A}=(A_1, A_2, \cdots, A_n)$ 和 $\boldsymbol{B}=(B_1, B_2, \cdots, B_n)$ 之间角度的余弦，计算公式如下：

$$\text{Similarity}(\boldsymbol{A}, \boldsymbol{B})=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\cdot\|\boldsymbol{B}\|}=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}}, \tag{4}$$

该值的范围为[−1, 1]，−1 为完全不相似，1 为完全相似.

根据公式(4)求得高钾玻璃各成分间余弦相似度，并依据余弦相似度对成分进行聚类，如图 3 所示，其中，横轴表示聚类的特征，纵轴表示聚类的水平.



(a) 各成分组间联接聚类结果　　(b) 质心聚类结果

图 3　对高钾无风化样本重要特征集进行 R 型聚类

由图 3 可知，使用组间联接聚类和质心聚类结果的谱系图结构一样，均将特征分为两类，得到的第 1 类为 SnO_2、SO_2、Na_2O、BaO 和 PbO，第 2 类为 CaO、K_2O、Fe_2O_3 和 MgO. 每个类中的特征相似度高，两个类之间相似度较低.

3.1.3　确定典型特征集并进行亚类划分

因为 2 个特征类类内相似度高，类间相似度低，因此分别从 2 个类中选择特征构建划分亚类的典型特征集，结合表 3 中各特征的方差排序，在两类中均选择方差最大的成分，即第 1 类中选择 Na_2O，第 2 类中选择 CaO. 将 Na_2O 和 CaO 作为高钾无风化样本亚类划分的典型特征集合，基于该特征集合对样本进行 Q 型聚类.

用欧式距离测算样本间距离，用组间联接测试簇间距离，用层次聚类方法对高钾无风化样本进行聚类，结果如图 4 所示.

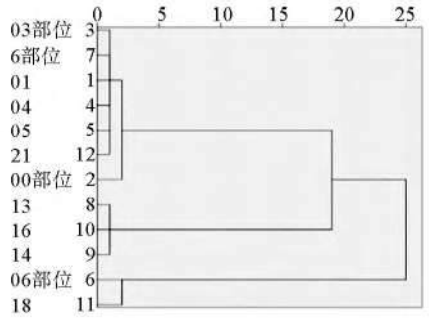


图 4　高钾无风化样本的 Q 型聚类图

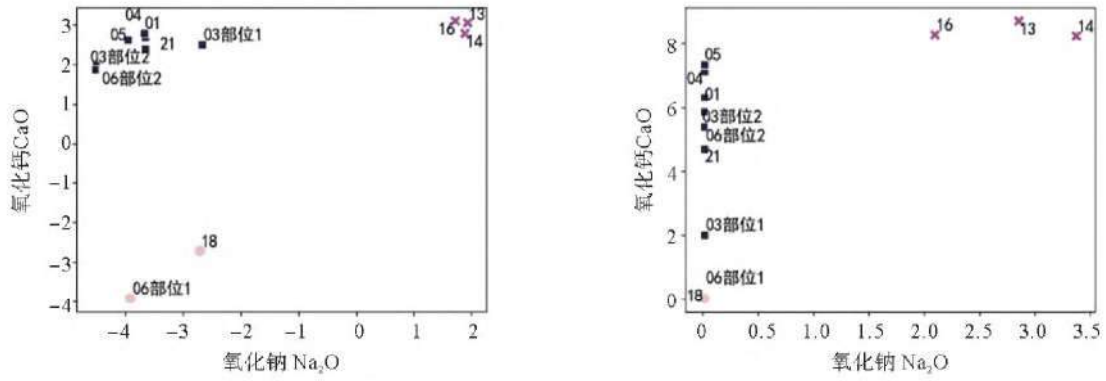
由图 4 可知，样本很清晰地被分为 3 个亚类，簇内距离很小，簇间距离很大，分类边界清晰. 具体亚类划分结果如表 4 所示.

表 4　思路 1 高钾无风化样本亚类划分结果

亚类	文物采样点
亚类 1	01、03 部位 1、03 部位 2、04、05、06 部位 2、21
亚类 2	06 部位 1、18
亚类 3	13、14、16

对高钾玻璃该亚类划分结果对应的样本集

中心对数比变换后数据与原始数据分别作出分类散点图，如图 5 所示，横坐标、纵坐标分别表示该成分的 CLR 变换后数据.



(a) 中心对数比变换数据分类散点图　　(b) 原始数据分类散点图

图 5　高钾玻璃分类结果散点图

由图 5 可以明显发现该分类方法将样本分为 3 个亚类，分类效果较好.

3.1.4　亚类分类规律及判别方法

亚类分类规律应简单、清晰、可用性强，其目的是可以对未知文物样本依据规则进行亚类判别. 采用 $\mu+2\sigma$ 估计特征取值区间，但由于亚类 1 中 CaO 标准差大，变异系数大于 15%，所以该特征取值区间调整为 $\mu+\sigma$. 对高钾无风化样本可得到以下分类规律：

亚类 1：Na_2O 含量约为 0，同时 CaO 含量约 1.87～9.47；

亚类 2：Na_2O 含量、CaO 含量均约为 0；

亚类 3：Na_2O 含量约 0.885～4.755，同时 CaO 含量约 7.635～9.405，如表 5 所示.

表 5　高钾无风化样本亚类成分取值区间表

亚类	Na_2O			CaO		
	平均值	标准值	取值范围	平均值	标准值	取值范围
亚类 1	0	0	0	5.67	1.900	1.87～9.47
亚类 2	0	0	0	0	0	0
亚类 3	2.82	0.645	0.885～4.755	8.52	0.295	7.635～9.405

基于亚类成分均值、标准差及取值区间建立对未知高钾无风化样本的亚类两种判别方法：

1)区间判别法. 判断未知样本中 Na_2O 和 CaO 含量取值是否属于 3 个亚类的取值区间，如果是则判别结果为对应亚类即可；

2)最近区间判别法. 未知样本中 Na_2O 和 CaO 含量取值不在 3 个亚类的区间内，而是介于两个亚类之间，则分别计算样本值距离两个亚类的最近取值的距离，选择距离最近的亚类作为判别结果.

用上述判别方法，对本题中 12 个高钾无风化样本进行判别，其中采样点 03 部位 1 介于亚类 1 和亚类 2 之间，用最近区间判别法，其余 11 个样本均使用方法 1 区间判别法，所有样本与亚类分类结果一致.

3.2　思路 2：基于封装式特征选择[14]的亚类划分方法

通过一定的策略选取特征子集，对特征子集进行评估，迭代计算不断选择最优的特征子集，迭代结束后对所选的特征子集进行验证[15]，依据这个流程设计亚类划分方法.

3.2.1　构建封装式特征选择的亚类划分算法

1)子集产生

依据排名法产生候选的特征子集. 理论上可以穷举所有的特征子集，但是计算量大，采用贪心算法在每一步迭代中都采纳当前条件下最好的特征子集，可以同时获得较好的运算效率和较高的特征子

集质量.

单特征采用基于排名的无监督特征选择方法，首先通过统计指标筛除占比过低不重要的成分，再依据信息论评估特征重要性，如熵和方差，离散度越高认为对样本的区分越明显，重要性越高，得到特征重要性序列. 单特征子集根据该顺序选择即可.

多特征使用特征之间的相关性辅助选择当前最优的特征子集. 应选择重要性高且与已选特征相关性小的特征. 本文通过熵和方差定义一个特征备选排名公式，如下：

$$\mathrm{ord}_i=\mathrm{ent}_i+\mathrm{var}_i, \tag{5}$$

其中：ord_i 为经过熵的秩次 ent_i 和方差的秩次 var_i 计算的特征成分的新秩次；i 为具体特征. 再通过对 ord_i 进行比较，越小的特征秩次越靠前，据此确定备选特征.

多特征子集产生算法：

①对单特征按照重要性由高到低排序，按序选择；

②计算其余每个特征关于该已选特征的备选排名；

③按备选排名添加最高排名的特征构成特征子集.

2)子集评估

使用轮廓系数对特征子集的优劣进行评估，用特征子集对样本进行 K-means 聚类，计算聚类结果的轮廓系数，使用轮廓系数作为该特征子集的评价函数，轮廓系数越大评估分数越高. 轮廓系数 $S(i)$ 的计算方法如下：

$$S(i)=\frac{b(i)-a(i)}{\max\{b(i),a(i)\}}, \tag{6}$$

其中：$a(i)$为样本 i 与其所属簇内其他文物样本的平均距离，若簇内仅 i 一个样本，则 $S(i)=0$；$b(i)$为样本 i 与其他簇的样本的平均距离最小值.

3)停止条件

本文选择迭代次数作为特征选择算法停止的判断条件，对保留的子集进行组合不断迭代.

4)子集验证

对所选的特征子集验证其有效性，将所选特征子集和全部特征子集进行亚类划分，评估聚类的合理性.

综合以上步骤，构建基于封装式特征选择的亚类划分算法框架，如图 6 所示.

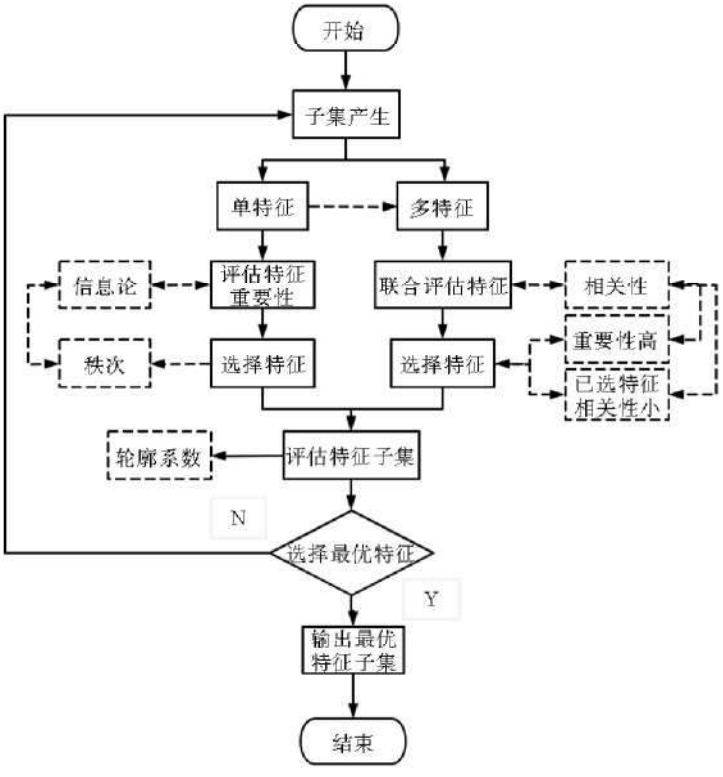


图 6　特征选择算法流程图

3.2.2　高钾无风化样本的亚类划分

1)子集产生

依据统计指标筛除占比极低的不重要成分 SO_2、SrO 和 SnO_2，对其余单特征用公式(5)计算重要性，并按照值越小的特征秩次越靠前进行排序，结果如表 6 所示.

2)子集迭代评估

通过对产生的子集迭代评估，计算每个特征子集亚类划分的轮廓系数，得到特征子集 CaO、Na_2O、(Na_2O，BaO)和(Na_2O，Fe_2O_3)的评估排名，具体如表 7 所示.

由表 7 的排名可知，当用单特征分类时可选择 CaO 和 Na_2O，当用多特征分类时可选择特征子集(Na_2O，BaO)和(Na_2O，Fe_2O_3)．实际问题中可依据排名给出多种亚类划分方案．

表 6　高钾玻璃单特征选择秩次结果

成分	方差秩次	熵秩次	综合秩次	成分	方差秩次	熵秩次	综合秩次
Na_2O	1	2	1	Fe_2O_3	7	4	6
BaO	3	1	2	MgO	6	6	7
PbO	5	3	3	P_2O_5	8	8	8
CaO	2	7	4	CuO	9	9	9
K_2O	4	5	5				

表 7　单特征、多特征亚类划分时高钾玻璃轮廓系数结果

特征子集	轮廓系数	排序	特征子集	轮廓系数	排序
CaO	0.8929	1	BaO	0.8531	5
Na_2O	0.8810	2	(SiO_2，Na_2O)	0.8500	6
(Na_2O，BaO)	0.8610	3	(K_2O，CaO)	0.8449	7
(Na_2O，Fe_2O_3)	0.8548	4	(Na_2O，K_2O)	0.8371	8

3)特征子集 Na_2O 和(Na_2O，BaO)的亚类划分结果

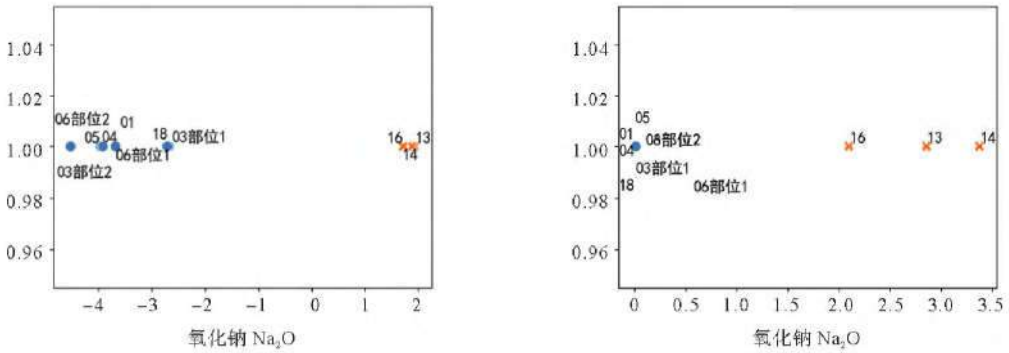
当以 Na_2O 作为亚类划分依据时，高钾无风化样本划分亚类成分范围如表 8 所示．

表 8　单特征亚类划分-高钾玻璃亚类划分规则表(原数据)

划分依据	Na_2O		
	平均值	标准差	取值范围
亚类 1	2.78	0.53	(2.25～3.31)
亚类 2	0	0	0

据此分类规律绘制样本集在中心对数比变换与原数据集下的分类散点图[16]，如图 7 所示，横坐标、纵坐标分别表示该成分的 CLR 变换后数据．可以看出，样本被明显分为 2 个亚类，且分类效果较好．

当以(Na_2O，BaO)组合作为亚类划分依据时，高钾无风化样本划分亚类成分范围如表 9 所示．



(a) 中心对数比变换数据分类散点图　　(b) 原始数据分类散点图

图 7　高钾无风化玻璃亚类划分结果图(特征子集：Na_2O)

表 9　多特征高钾玻璃亚类划分规则表(原数据)

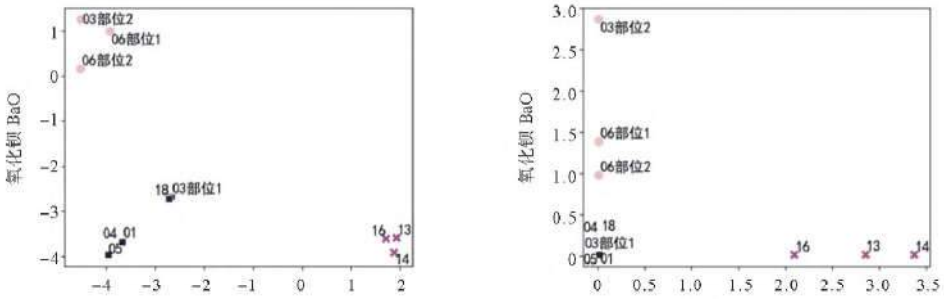
亚类	Na_2O			BaO		
	平均值	标准值	取值范围	平均值	标准值	取值范围
亚类 1	0	0	0	1.74	0.81	(0.93～2.55)
亚类 2	0	0	0	0	0	0
亚类 3	2.78	0.53	(2.25～3.31)	0	0	0

据此分类规律绘制高钾玻璃样本集在中心对数比变换与原数据集下的分类散点图，如图 8 所示，可以看出样本被明显分为 3 类，分类结果也较清晰．

3.2.3　对铅钡玻璃进行封装式特征选择

同 3.2.1 与 3.2.2，计算铅钡玻璃单特征选择下的秩次顺序，并对产生的子集迭代评估，计算每个特征子集亚类划分的轮廓系数，得到特征子集的评估排名．由排名知用单特征分类时，可以选择 Na_2O

或 CaO，用多特征分类时可以选择特征子集(Na_2O，CaO)或(Na_2O，Fe_2O_3)，本文在此不再展示具体结果. 在实际问题中可依据排名给出多种可选的亚类划分方案.



(a) 中心对数比变换数据分类散点图　　(b) 原始数据分类散点图

图 8　高钾无风化玻璃亚类划分结果图(特征子集：(Na_2O，BaO))

3.3　特征子集对比分析

表 10 列出了思路 1 和思路 2 下的比较适合划分亚类的特征子集.

表 10　思路 1、2 特征子集对比分析

类型	思路 1	思路 2
高钾玻璃	(Na_2O，CaO)	CaO、Na_2O、(Na_2O，BaO)、(Na_2O，Fe_2O_3)
铅钡玻璃	(Na_2O，Fe_2O_3)	Na_2O、CaO、(Na_2O，CaO)、(Na_2O，Fe_2O_3)

由表 10 可以看出，思路 2 可供选择的特征子集更多，且分为单特征与多特征两类，这是由于思路 2 特征子集产生方式是通过遍历导致的，这样产生了更多可能的特征子集. 比较思路 1、2 中的特征子集进一步发现，两种思路产生的特征子集的特征具有高度的相关性，表明两种思路的结果可以相互印证，具有合理性.

3.4　拓展：重构特征，探索亚类划分方法

探索亚类划分方法，考虑添加成分常有满足某两种成分和、成分比的要求，因此尝试用两种成分和、成分比重构特征进行亚类划分探索.

以高钾玻璃为例，将重构成分比作为新特征，因为 SiO_2 作为玻璃的主要制成材料，不适宜配比或是添加来进行亚类划分，所以本文不考虑 SiO_2 构成的重构特征. 用思路 2 方法求解处理，得到 CaO/Al_2O_3 轮廓系数最大，所以选择 CaO/Al_2O_3 对高钾无风化玻璃进行亚类划分，得到表 11 和表 12.

表 11　特征重构时高钾玻璃无风化样本亚类划分结果表

亚类	文物采样点
亚类 1	01、03 部位 1、03 部位 2、04、05、06 部位 2、13、14、16
亚类 2	06 部位 1、18

表 12　高钾玻璃无风化样本单特征 CaO/Al_2O_3 亚类划分规则表

亚类	CaO/Al_2O_3		
	平均值	标准差	取值范围
亚类 1	1.05	0.35	0.35～1.75
亚类 2	0	0	0

4　未知样本判别

4.1　大类判别

基于 2.4 节得到的无风化高钾、铅钡玻璃分类规律，以及有风化的高钾、铅钡玻璃分类规律，可以对未知样本进行大类判别.

4.2　亚类判别

1)无风化的未知样本的亚类判别

针对无风化的样本直接利用表 5 及铅钡无风化玻璃亚类划分的判别规则进行判别.

2)有风化的未知样本的亚类判别

针对有风化的样本，先使用 1.3 节的模型，将对风化样本预测其无风化时各成分的含量，如表 13 所示，再使用无风化亚类划分规则进行判别.

表 13 有风化的未知样本预测其无风化的数据（仅列出部分重要特征）

文物编号	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3
A6	50.23	0	37.10	0.46	2.62	5.80
A7	46.01	0	26.29	5.87	0	19.51
A2	62.18	0	0	12.21	0	1.88
A5	74.65	0.52	0.51	0.28	0.93	5.94

根据思路 1 和思路 2 下的特征子集(Na_2O)及高钾玻璃特征子集(Na_2O 与 BaO)和铅钡玻璃特征子集(Na_2O 与 CaO)分别进行亚类划分得到表 14.

由表 14 可以看出，在亚类划分规则不同的前提下，亚类划分的结果也略有不同，可以为该问题提供不同的亚类划分思路.

表 14 有风化的未知样本大类划分与亚类划分结果表

文物编号	表面风化	大类划分	思路 1 亚类划分	思路 2 亚类划分（Na_2O）	思路 2 亚类划分（Na_2O 与 BaO）、（Na_2O 与 CaO）
A1	无风化	高钾	高钾亚类 1	高钾亚类 2	高钾亚类 2
A2	风化	铅钡	铅钡亚类 1	铅钡亚类 1	铅钡亚类 2
A3	无风化	铅钡	铅钡亚类 2	铅钡亚类 1	铅钡亚类 2
A4	无风化	铅钡	铅钡亚类 2	铅钡亚类 1	铅钡亚类 2
A5	风化	铅钡	铅钡亚类 2	铅钡亚类 2	铅钡亚类 3
A6	风化	高钾	高钾亚类 2	高钾亚类 2	高钾亚类 2
A7	风化	高钾	高钾亚类 1	高钾亚类 2	高钾亚类 2
A8	无风化	铅钡	铅钡亚类 1	铅钡亚类 1	铅钡亚类 2

5 结语

5.1 优点

1)用多个思路探索亚类划分，对比分析优缺点，确保划分的合理性；

2)亚类判别规则简单，划分结果可解释性好，使得划分后亚类易于用 1～2 个特征进行清晰的解释，同时实现快速亚类判别，在实际应用中可操作性强，有实际应用价值；

3)得到的亚类划分方法具有普适性. 形成一套完整的亚类划分方法，可以提供不同角度的多种亚类划分结果，供该领域专家进行研究. 该方法可适用于同类基于成分数据划分亚类的应用，如材料的化学成分分析，也可稍加修改后用于一般的物种亚类划分问题.

5.2 展望

本文基于轮廓系数等指标作为亚类划分合理性的评价指标的评价模型还有一定的进步空间，可以在未来继续不断修订亚类划分合理性的评价模型.

参考文献

[1]干福熹.中国古代玻璃的起源和发展[J]. 自然杂志，2006，4：187-193+184.

[2]Liu H，He D，Li S. Chemical subclasses of ancient glass based on K-means[J]. Advances in Engineering Technology Research，2023，6(1)：63.

[3] Dussubieux L，Robertshaw P，Glascock M D. LA-ICP-MS analysis of African glass beads：laboratory inter-comparison with an emphasis on the impact of corrosion on data interpretation[J]. International Journal of Mass Spectrometry，2009，284(1/3)：152-161.

[4]王承遇，陶瑛，陈敏，等. 钠钙铝镁硅酸盐玻璃和碱铅硅酸盐玻璃的风化[J]. 硅酸盐通报，1989，6：1-9.
[5]董涵，邹明华，李露，等. 古玻璃成分的分类预测研究[J]. 咸阳师范学院学报，2023，38(4)：31-37.
[6]宛惠，邓明华. 古代玻璃制品成分分析与鉴别的统计建模[J]. 数学建模及其应用，2023，12(2)：27-40.
[7]全国大学生数学建模组委会. 2022“高教社杯”全国大学生数学建模竞赛赛题[Z/OL]. [2022-09-15]. http://www.mcm.edu.cn/html_cn/node/388239ded4b057d37b7b8e51e33fe903.html.
[8]李春轩，罗毅，包安明，等. 基于对数比转换的成分数据空间插值研究[J]. 中国农业科学，2012，45(4)：648-655.
[9]司守奎，孙玺菁. 数学建模算法与应用[M]. 北京：国防工业出版社，2021.
[10]路佳佳. 基于集成特征选择和随机森林的古代玻璃分类模型[J]. 硅酸盐学报，2023，51(4)：1060-1065.
[11]李航. 统计学习方法[M]. 2 版. 北京：清华大学出版社，2019.
[12]Lim H，Kim D W. Pairwise dependence-based unsupervised feature selection[J]. Pattern Recognition，2021，111：107663.
[13]陈志豪，季晶敏. 基于层次聚类模型的古代玻璃制品成分分析与鉴定[J]. 现代信息科技，2023，7(8)：122-125.
[14]欧高炎，朱占星，董彬，等. 数据科学导引[M]. 北京：高等教育出版社，2017.
[15]Dash M，Liu H. Feature selection for classification[J]. Intelligent Data Analysis，1997，1(3)：131-156.
[16]Hancer E，Xue B，Zhang M. A survey on feature selection approaches for clustering[J]. Artificial Intelligence Review，2020，53：4519-4545.

Study on the Compositional Analysis and Subcategory Classification Method of Ancient Glass

LOU Yang，DOU Lei，ZHUO Zhaoyang，LU Ping
(College of Science，Xi′an University of Architecture and Technology，Xi′an，Shaanxi 710399，China)

Abstract：Analyze the classification rules and subclass classification methods of ancient glass products based on their chemical composition data. Firstly，perform a central logarithmic ratio transformation on the original data，and then analyze the conversion relationship between weathered and unweathered data；. Secondly，by constructing a supervised feature selection method to explore the rules of high potassium and lead-barium categories. Thirdly，based on unsupervised feature selection methods，we explore subcategory partitioning methods. We have focused on constructing a subcategory partitioning method using filtered feature selection combined with feature R-type clustering，as well as a subcategory partitioning method based on wrapper feature iterative selection. The second method can provide multiple subcategory partitioning schemes and results.

Key words：central log-ratio transformation；supervised feature selection；unsupervised feature selection；filtered；wrapper；subcategory partitioning

作者简介

楼　阳(2002－)，男，西安建筑科技大学数据科学与大数据技术专业本科生.
窦　雷(2002－)，男，西安建筑科技大学数据科学与大数据技术专业本科生.
卓朝阳(2002－)，男，西安建筑科技大学数据科学与大数据技术专业本科生.
鲁　萍(1979－)，女，副教授，主要研究方向为数据挖掘与分析.