基于逐层分类模型的古代玻璃制品成分分析与鉴别

摘 要

古代玻璃制品是我国的瑰宝，它既是丝绸之路贸易往来的宝贵物证，也是体现匠人精湛手工艺的特色文物。而玻璃和大气接触时会发生风化现象，风化对玻璃的化学性质起着很大的影响作用。分析风化后玻璃的化学成分含量联系与变化规律、划定玻璃的分类标准在制造学和考古学上均有重要意义。

针对问题一，对表单数据进行补充与删除的预处理后，分别对表面分化和三种无序分类变量进行卡方检验，粗略判断相关性。在卡方检验的基础上计算 Cramer_V 值，依次为：类型-0.5735；颜色-0.3412；纹饰-0.2932。最终得到表面风化与三变量间的关系：类型相较颜色与纹饰，对表面风化起显著性作用。颜色的显著性大于纹饰。

接着，根据高钾、铅钡玻璃有无风化时的频数直方图，得出有无风化时化学成分含量的初步规律。建立 Point-Biserial 点二列相关性分析模型，依次得出各成分与高钾、铅钡玻璃间的二列相关系数并归纳规律。最后，通过 Shapiro-Wilk 检验正态性、箱线图保证无异常值的情况下进行点二列显著性 t 检验，检验结果证明模型显著性良好。

最后，根据风化进程将文物采样点分成四类风化点，并将无风化文物采样点具体划分是否风化。分析各项成分，得到成分随风化进程演变的规律。归纳其变化趋势，在此基础上进行趋势平均移动，得到风化前的预测含量。

针对问题二，先采用 SMOTE 采样法平衡倾斜样本数据。接着，建立数值型特征逐层筛选模型，首先用 Filter 方差过滤法筛选出方差大于阈值 0.1 的特征；其次用嵌入式的随机森林模型二次过滤特征，以 Gini 系数最小作为其最优划分区域。得到的结果特征集为：高钾玻璃：氧化钾、二氧化硅；铅钡玻璃：氧化铅、氧化钡、二氧化硅、氧化锶、五氧化二磷。并根据上述指标得到三个重要的玻璃分类规律。

接着，采用 2 次 FCM 模糊聚类法进行亚分类。最终分类结果为：SiO_2-K_2O 高钾高硅玻璃、K_2O-SiO_2-CaO 超高钾玻璃、$CaO-K_2O-SiO_2$高钾钙钾玻璃、$PbO-BaO-SiO_2$超铅钡玻璃、$PbO-SiO_2-BaO$ 铅钡高铅玻璃、$SiO_2-PbO-BaO$ 铅钡高硅玻璃。最后，选取 CH、DB 指标进行聚类分析的合理性检验，得到模型合理的结论。在系统中引入噪声点以判断其敏感性，得到模型对噪声几乎不敏感的结论。

针对问题三，建立逐层分类模型。运用问题二的分类规律进行初步分类后，建立基于贝叶斯判别分析的亚分类模型，最终两类玻璃分类结果的准确率分别为 92.839%与 95.812%。接着，运用 KS 曲线和扰动分析两种方法对模型进行敏感性检验。KS 曲线的 TRR-FRP 的差距为 0.8931，扰动分析的敏感性值为 1.219，证明分类模型对样本分割能力较强，模型非常敏感。最后，通过 K 折交叉验证避免因数据集划分而产生的偶然性误差。

针对问题四，首先需要将样本数量较少的超高钾、高钾高硅、高钾高钙玻璃进行类别合并，与铅钡高硅、铅钡高铅、超铅钡玻璃进行成分相关分析。使用 Spearman 相关分析遍历成分得到各成分间的相关系数，并做出热力图进行化学成分之间的关联关系分析与差异分析。

此外，本文着重分析了模型的运用。风化影响模型可以帮助设计出更好的抗风化型玻璃；亚分类模型可以作用于文物验证中判断是否为舶来品，且大致判断玻璃的制造年代；相关分析模型可以在化学领域更好的探索玻璃成分间的未知规律。

关键词：点二列相关性、数值型特征筛选、FCM 模糊聚类、逐层分类模型、敏感性分析

一．问题重述

1.1 问题的背景

玻璃文物是早起丝绸之路贸易往来的宝贵物证[1]，早期玻璃传入我国后，我国古代玻璃的制作技术便生根发芽。这些玻璃制品虽然与舶来品外观相似，但化学成分却大不相同。

玻璃在和大气接触时发生的侵蚀被称做风化。玻璃在风化后表面会产生风化堆积物、风化膜、斑点、雾状物，很大程度的影响了玻璃的外观、使用和化学性质。与此同时，玻璃在风化过程中，内部元素与环境元素进行大量交换，导致其比例发生变化[2]。因此，分析玻璃风化后化学成分的变化规律可以为玻璃原有特性的研究及玻璃所处地域特性的研究提供极大帮助，在考古学上有重大实际意义。

1.2 问题的提出

题目给出 58 件中国古代玻璃制品的相关数据，考古工作者根据文物样品的化学成分和其他检测手段已将其分为高钾玻璃和铅钡玻璃两种类型。同时，题目给出了玻璃制品的其他分类信息、表面化学成分所占比例及 8 件未分类的古代玻璃制品信息。基于上述条件及数据，建立模型并分析以下四个问题：

（1）根据表单数据对玻璃文物的表面风化与其玻璃纹饰、类型和颜色的关系进行分析；再根据玻璃类型，分析玻璃样品表面有无风化化学成分含量的规律；最后根据风化点检测数据，预测此点风化前各化学成分的含量。

（2）根据表单数据分析高钾玻璃和铅钡玻璃的分类规律；并对每个类别选择合适的特征指标进行亚类划分，给出具体划分方法及结果；并对结果的合理性和敏感性进行分析。

（3）通过分析表单 3 中未知类别玻璃文物样本的化学成分鉴别其所属玻璃类型，并分析分类结果的敏感性。

（4）针对不同类别的玻璃文物，分析其化学成分之间的关联关系；比较不同类别的玻璃文物之间化学成分的关联关系的差异性。

二．问题分析

2.1 问题一的分析

进行问题一的分析前，需要对数据进行预处理。主要有以下两步：1.根据未检测成分空缺的数据特质，补充或删除文件中的缺失数据。2.根据题中允许成分累加和的条件要求，对各数据进行排序，删除文件中的异常数据。

首先，要探究表面风化与类型、纹饰、颜色的关系，就要研究变量之间的相关性。由于探究的是无序二分类变量与连续类变量间的关系，故使用卡方检验和 Cramer_V 值分析其规律。先进行卡方检验，粗略的分析各变量与表面风化的关系。并在其基础上进行 Cramer_V 值分析，最终得到风化与三类变量间的 Cramer_V 值，得出其相关规律。并对模型进行 Cramer_V 显著性检验。

其次，要探究表面有无风化化学成分含量的统计规律。先初步的根据有无风化时化学成分的频数直方图看出其规律得到初步结论。接着进行 Point-Biserial 点二列相关性分析，依次得出各成分与高钾、铅钡玻璃间的二列相关系数并归纳规律，并进行模型的检验。

最后，需要预测风化前的化学成分含量。先根据风化进程将文物采样点分成“无风化文物普通点”、“风化文物未风化点”、“风化文物普通点”、“风化文物严重风化点”四类。导入各项成分数据，得到其随着风化进程演变的规律。归纳其变化趋势，在此基础上进行趋势移动，得到风化前的预测含量。

2.2 问题二的分析

问题二要分析高钾和铅钡玻璃的分类规律。在研究规律前，发现高钾玻璃的有效样本远小于铅钡玻璃，故采用 SMOTE 采样法平衡倾斜数据。研究基于高钾和铅钡玻璃依据玻璃特征的分类，需要建立数值型特征筛选模型。首先用 Filter 方差过滤法初步选出方差大于阈值

的特征；其次用嵌入式的随机森林以基尼系数最小的属性最为其最优划分区域。

接着，需要设计进一步的亚分类模型。出于玻璃样本化学成本连续且界限不明的特征，采用 FCM 模糊聚类进行多次分类，并选取 CH、DB 指标进行合理性分析。接着在系统中引入噪声点，观察聚类结果是否发生改变，以此反应模型的灵敏性。

2.3 问题三的分析

问题三要求对表单 3 的文物进行分类。由于遍历算法复杂度高，建立逐层分类模型。首先基于问题二的分类规律进行初步分类，用逻辑回归中的梯度下降法和损失函数做出其估计函数，检验其精确度。其次建立基于贝叶斯判别分析的亚分类。运用 KS 曲线和扰动分析两种方法对模型进行敏感性检验。最后，通过 K 折交叉验证避免因测试集、训练集的划分而产生的偶然性误差。

2.4 问题四的分析

问题四需要探究不同类别化学成分间的关系与关联关系的差异性。将样本数量不足的玻璃亚类进行类别合并，与其他类一起进行成分相关分析。由于数据的正态分布性不足，使用 Spearman 相关分析遍历每类玻璃中每两个成分得到相关系数表，并做出热力图进行相关与差异分析。最后以宏观、特征、特殊三个角度分别分析化学成分之间的关联关系与关联关系的差异性分析。

三．模型假设

1.玻璃文物的考察数据真实精准，不考虑数值的系统误差。

2.玻璃的状态比较稳定，没有出现如黏滞、脱水等状态。

3.玻璃在存储过程良好，不考虑因为外力等偶然因素导致的破损侵蚀。

4.玻璃在风化侵蚀过程中的温度与压强条件比较稳定。

四．定义与符号说明

符号	说明
r_{pb}	点二类相关系数
M_t	移动平均数
a_i	模糊组 I 聚类中心
d_{ij}	欧几里得距离
r_s	单向指标间关联性
u_{ij}	模糊隶属度
λ	全局不确定性比例系数
γ	局部不确定性比例系数
m	加权指数
$DB(K)$	数据点之间相似度
$trP(k)$	类间离差矩阵的迹

五．模型的建立与求解

5.1 数据的预处理

5.1.1 缺失值处理

（A）数据填充

表单 2 中给出的已分类玻璃文物化学成分比例数据中存在部分缺失，结合题目说明，即空白意味着未能检测到该成分，需要对空白数据进行填充。在对附件数据进行清洗筛查时，发现表 2 中存在 694 个 NULL 数据，将其全部用 0 替换，表示该化学成分在检测中所占比例为 0%。

（B）数据删除

表单 1 给出的文物分类信息中，部分文物的颜色信息没有给出。由于颜色和文物编号、类型、风化与否均不成一一对应关系，故颜色信息很难通过预测进行填充。数据筛查时，发现表 1 存在 4 个 NULL 数据，将其对应的玻璃文物数据删除。此外，颜色空缺的文物均为风化的铅钡玻璃。在数据集中，铅钡玻璃占总体的样本的 68.9%，高钾玻璃则占 31.1%，铅钡玻璃所占比例远高于高钾玻璃，存在数据倾斜现象；而在铅钡玻璃中，风化玻璃占总体样本的 70%，无风化玻璃则占 30%，仍存在较严重数据倾斜倾斜现象。基于此数据特质，删除四条风化铅钡玻璃样本有利于平衡数据类型。

5.1.2 异常值处理

表单 2 中的数据代表其成分所占比例，各采样点化学成分之和理论上应该达到 100%，但由于检测手段可能会有一定的系统误差，题中允许成分累加和的范围为 85%至 105%，不在此范围内的数据被视作异常值。在 Excel 中累加各成分所占比例并对累加和进行降序，结果如下：

表 1　成分比例累加和排名表

类型	风化判断	文物取样	成分比例累加和	累加和排名
高钾	无风化	03 部位 1	100	1
高钾	风化	22	100	2
铅钡	无风化	37	99.98	3
.....	.....	.....	.....	.....
铅钡	无风化	20	88.41	67
高钾	无风化	15	79.47	68
高钾	无风化	17	71.89	69

由表可见，数据累加和的最高值尚且未超过 105%，故无需舍弃。而排名较低的数据中，第 68、69 名数据的成分累加和小于 85%，故舍弃文物取样 15 和 17 的采样点数据。

5.2 问题一模型的建立与求解

5.2.1 表面风化与其类型、纹饰、颜色的关系分析

（A）Pearson 卡方检验的初步分析

利用 Pearson 卡方检验可以从统计学角度来描述二分类变量之间的相关性，即玻璃文物特征因素中的分类变量与表面分化之间的关系，其计算公式如下：

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \quad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

其中，f_0表示实际频数，f_e表示期望频数，计算所得的 Pearson 卡方越大代表拒绝其原变量之间的相关性越大，即两变量间的相关性较小。将纹饰、类型、颜色分开交叉计算，采用 XTABS 生成频数表，在此基础上进行卡方拟合优度检验，当 p 值小于 0.05 时，拒绝原假设 H0，被测试变量和风化之间有显著关系。所得到的结果如下：

表 2　Pearson 卡方检验表

	类型与表面风化	颜色与表面风化	纹饰与表面风化
皮尔逊卡方	5.400	6.287	5.747
显著性 p 值	0.02	0.507	0.056
费希尔精度检验	/	5.733	5.806

由上表可以粗略的分析出各变量与表面风化之间的关系：类型与表面分化的显著性 p 值为 0.02 远小于 0.05，拒绝原假设，证明该类型在风化中起显著性作用；颜色与表面分化的显著性 p 值为 0.507 远大于 0.05，接受原假设，证明该类型在风化中不起到显著性作用；

纹饰与表面风化的显著性 p 值为 0.056 接近于 0.05，出于谨慎性原则难以判断，只能初步猜想纹饰在风化中起到作用但不显著。

（B）基于卡方检验的 Cramer_V 值分析

Cramer_V 系数是双变量相关分析中卡方检验后拒绝虚无假设后计算的指标，是用于衡量无序分类变量之间相关程度的重要指标之一，且在其计算过程中需要借助于 Pearson 卡方统计值。而本题研究的变量均为无序分类变量，且上文中已完成初步的卡方检验，基于此可进行进一步的相关分析计算。

首先，需要根据不同的类别变量列出对应的计算列联表：

表 3 表面风化与颜色、类型、纹饰的计算列联表

		颜色								类型		纹饰			
		黑	浅蓝	蓝绿	紫	深绿	绿	浅绿	深蓝	铅钡	高钾	A	B	C	合计
表面风化	无风化	0	8	6	2	3	1	2	2	12	12	11	0	13	24
	风化	2	12	9	2	4	0	1	0	24	6	9	6	15	30
合计		2	20	15	4	7	1	3	2	36	18	20	6	28	54

其次，计算期望次数，计算 Pearson 卡方统计值，用两层 sum 求和实现矩阵元素之和，得到 Cramer_V 系数，计算公式如下：

$$\text{Cramer\_V} = \sqrt{\frac{\chi_0^2}{n \times \left(\min\{h_i,\ h_j\} - 1\right)}} \quad \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

其中，n 表示数据个数；h_i表示交叉资料表的行数；h_j表示交叉资料表的列数，所得结果如下：

表 4 Cramer_V 系数表

	类型与表面风化	颜色与表面风化	纹饰与表面风化
皮尔逊卡方	5.400	6.287	5.747
显著性 p 值	0.02	0.507	0.056
Cramer_V	0.5735	0.3412	0.2932

Cramer_V 系数取值范围为 0 到 1，两个变量越相关取值越大。0 表示两个变量完全无关，1 则表示完全相关。结合上文卡方检验的结论和上表的数据可分析得出：类型与表面风化间 Cramer_V 系数为 0.5735，证明玻璃类型对表面风化的影响最为显著，而颜色和纹饰与表面风化间 Cramer_V 系数分别为 0.3412 和 0.2932，证明颜色和纹饰对表面风化均起到较小作用，但颜色的重要程度相较纹饰更为显著。

（C）模型结论的说明与验证

上文所述数据分析结论在化学和物理学原理上可以得到解释与验证。影响表面风化最为显著的变量是玻璃类型，它反应了不同玻璃类型的化合物组成成分差异，而化合物组成成分是影响其性质的决定性因素，故该类型指标最为重要。

颜色反应的是玻璃表面的显色物质，其背后代表的是玻璃的**烧制环境和生长环境**。主要的显色离子有Fe^{2+};Cu^{2+}, 在不同的酸碱环境中会呈现不同颜色，如：二者于酸性环境下会呈现蓝色，而碱性环境下呈现绿色。并且在生长过程中显色物质也会从内而外流失，进而将有色离子附着在表面。但是，烧制和生长环境对比起化合物的本身的化学性质依旧是不够直观且影响较弱的变量。

纹饰反应的是玻璃的表面物理结构，不同的物理结构会影响玻璃表面的**强度**。在化学层面上，不同纹理表面的**几何熵值**不同、**表面张力**和**表面能**均不相同。在物理层面上，不同纹路的**玻璃强度**、表面的**微裂纹密度**、凹槽的**积水性质**也各不相同。因此，纹路导致了玻璃中存在了许多物理和化学缺陷，使得玻璃的防护效能降低，进而影响风化进程。但是，纹饰的

影响比起类型和颜色比较微观细致，因此是最弱的变量。

5.2.2 类型对表面有无风化化学成分含量的统计规律

据题意建立数学模型思路如下：

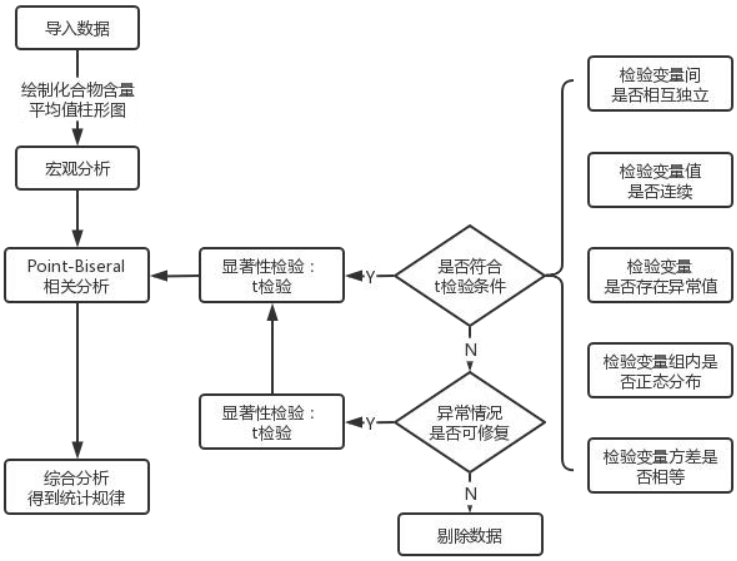


图 1 求解有无风化化学成分含量的统计规律建模思路图

（A） 柱形图的宏观初步分析

根据题干信息，需要研究高钾和铅钡两类玻璃在有风化和无风化时的化学成分含量规律。首先需要大致得出高钾有风化、高钾无风化、铅钡有风化、铅钡无风化两大类四小类之间的分布情况。选定类别不别，在有风化组和无风化组中求出各化学成分的总体均值进行比较，分析结果如下：

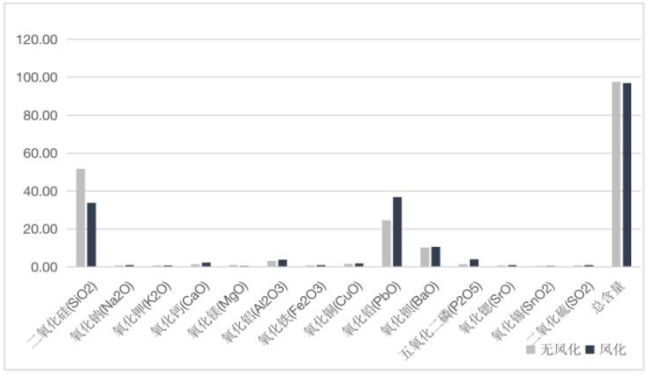


图 2 高钾玻璃风化、无风化直方图

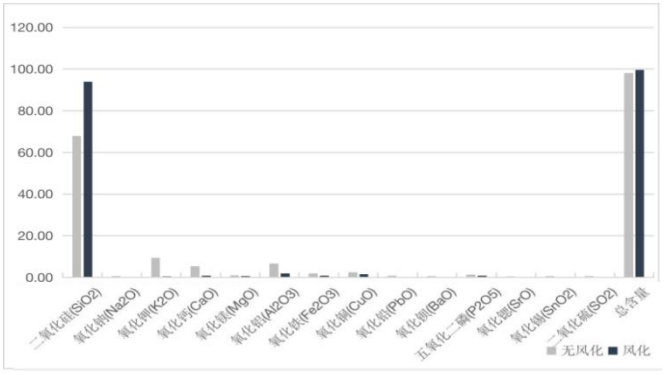


图 3 铅钡玻璃风化、无风化直方图

由上图可以初步分析得到以下结论：

①二氧化硅是最为显著的指标，风化的高钾玻璃中的二氧化硅含量比无风化的高；风化的铅钡玻璃中的二氧化硅含量比无风化的低。

②部分金属氧化物呈现比较明显的差异。风化高钾玻璃的氧化钾、氧化钙、氧化铝含量均低于比无风化玻璃；风化铅钡玻璃的氧化铅、五氧化二磷、氧化钙、氧化铝含量均高于无风化玻璃。

③大部分指标不呈明显差异。如氧化钡、总含量等指标。

（B） Point-Biserial 点二列相关性分析

在得到初步的结论后，需要更系统地定量衡量各化学成分与表面风化之间的相关性。连续数值变量与二分类变量的相关性可以用点二列相关性系数来衡量。算式如下：

$$r_{pb} = \frac{M1 - M0}{S_n}\sqrt{pq} \quad \cdots\cdots\cdots\cdots\cdots\cdots (3)$$

其中，M1 表示二分类变量为 1 的均值，即表面情况判定为风化的玻璃样本；M0 表示二分类变量为 0 的均值，即表面情况判定为无风化的玻璃样本。S_n为连续数值变量的标准偏差，计算各项玻璃样本中化学成分的数值指标与是否风化的相关性。绝对值越大代表其相关性越强，正数代表呈正相关，负数代表呈负相关。将以上数据进行比较并将计算结果进行降序排序如下：

表 5　高钾玻璃点二列相关系数表

高钾玻璃	**二氧化硅**	**氧化锡**	**氧化铜**	**氧化钠**	**二氧化硫**
	0.844402141	-0.16624530	-0.28070711	-0.30017068	-0.30417165
	氧化钡	**氧化铅**	**五氧化二磷**	**氧化锶**	**氧化铁**
	-0.334284874	-0.37649180	-0.41216034	-0.44664239	-0.500428591
	氧化镁	**氧化钙**	**氧化铝**	**氧化钾**	
	-0.582975119	-0.63429395	-0.71678263	-0.77848282	

高钾玻璃中，二氧化硅的r_{pb}正则且非常显著，对风化起到决定性影响。此外，以氧化钾、氧化铝为代表的金属氧化物的r_{pb}均为负，起到负相关影响。可以发现促成高钾玻璃的风化原因比较单一，二氧化硅在其中中占主导作用，其他氧化物均起到抑制作用。

表 6　铅钡玻璃点二列相关系数表

铅钡玻璃	**氧化铅**	**五氧化二磷**	**氧化钙**	**氧化镁**	**氧化锶**
	0.392192819	0.367206067	0.300888241	-0.146121361	0.116389612
	二氧化硫	**氧化铝**	**氧化铜**	**氧化钠**	**氧化钡**
	0.099184803	0.09444974	0.07848814	0.044072899	0.000636358
	氧化锡	**氧化铁**	**氧化钾**	**二氧化硅**	
	0.018202915	0.175604296	-0.185383803	-0.469498147	

铅钡玻璃中，二氧化硅的r_{pb}为负且最为显著；以氧化铅和的五氧化二磷、氧化钙为代表的r_{pb}正则且比较显著；氧化钾和氧化铁的r_{pb}为负则且比较显著；而以氧化钠、氧化钡、氧化锡为代表的氧化物的r_{pb}接近于 0，在风化中不起显著作用。可以发现铅钡玻璃的风化原因比较复杂，依赖许多化合物协同促成风化。

5.2.3 风化前化学成分含量的预测

（A） 按风化进程的数据整理

本题要求根据风化点检测数据对其风化前的化学成分进行预测，需要先将现有数据按风化进程排列。以铅钡玻璃为例，可将其按玻璃风化的时间前后依次排列为“风化文物未风化点”、“无风化文物普通点”、“风化文物普通点”、“风化文物严重风化点”（下文简称“一类点”、“二类点”、“三类点”、“四类点”）。其中，依题目背景可知“二类点”既可能未风化也可能存在较浅风化，所以它所处的时间位置还需要进一步判断。

将表单二铅钡玻璃数据按上述方式排列，并分别将 14 种化合物成分在点类型内部排序以寻找随时间演变规律。观测数据可知，二氧化硅随时间演变规律显著，在可控范围内波动，基本呈现随时间逐渐下降趋势。而“二类点”中的采样点 47、55、45、37、33、24 的二氧化硅含量高于“一类点”二氧化硅含量的平均数，有极大可能为“一类点”采样而得，故将这六个采样点**归于“一类点”**。其余“二类点”被认定为“无风化文物风化点”。据此做出如下铅钡玻璃二氧化硅含量随风化时间变化趋势图：

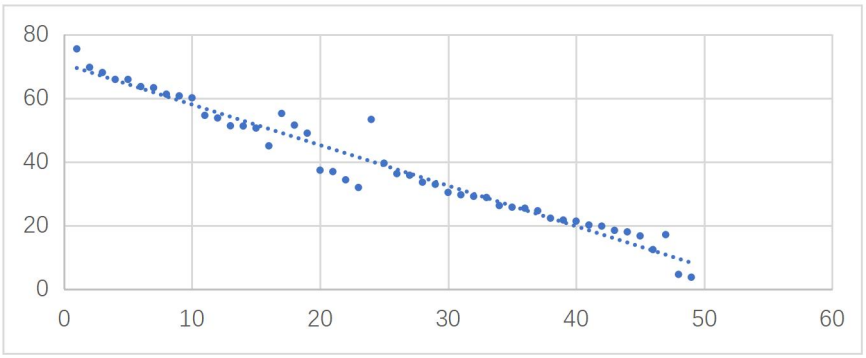


图 4 铅钡玻璃二氧化硅含量随风化时间变化趋势图

由上图可知，铅钡玻璃二氧化硅含量基本随风化呈下降趋势，可以较好地反应风化进程。

按如上排序方式观测铅钡玻璃其他化合物随风化的演变规律，可发现氧化锡与二氧化硫含量在未分化时均为 0，故可直接将该二值在铅钡玻璃未风化时**预测为 0**；发现若干化合物如：氧化钾、氧化铁含量未随风化进程显著性变化，故该二值在铅钡玻璃未风化时**预测为保持不变**。四种化合物含量随风化时间变化趋势图如下：

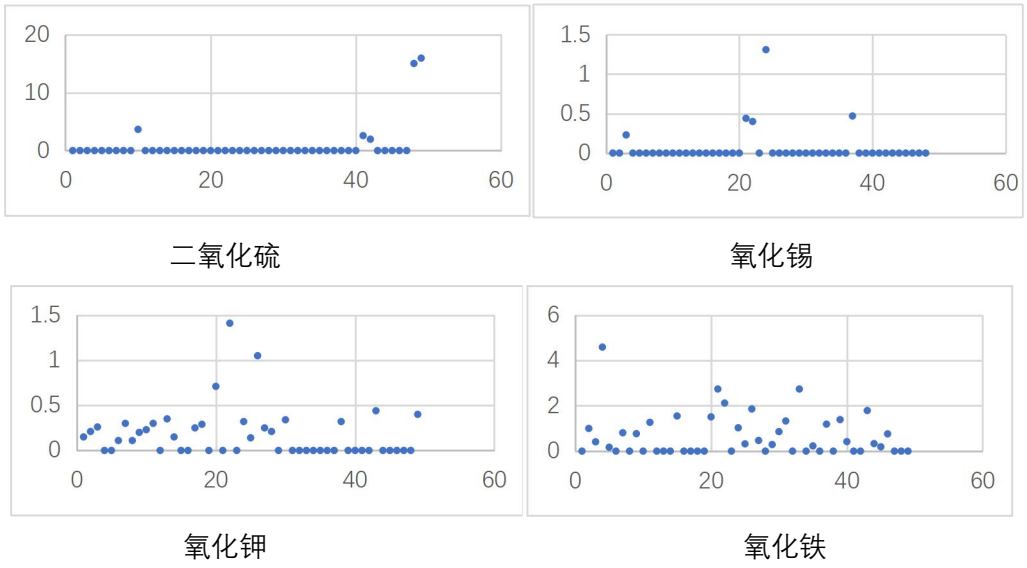


图 5 四种化合物含量随风化时间变化规律

（B） 基于趋势移动平均模型的部分化合物含量预测

随风化进程有显著线性变化的化合物含量可由趋势移动平均法进行预测，本文将以铅钡玻璃的二氧化硅含量为代表进行说明。

趋势移动平均法是将简单移动平均法和加权移动平均法二次移动平均修正得到的模型，其利用移动平均滞后偏差的规律以完成预测。具体步骤如下：

计一次移动的平均数为：

$$M_t^{(1)} = \frac{1}{N}(y_t + y_{t-1} + \cdots + y_{t-N+1}) \cdots\cdots\cdots\cdots\cdots\cdots (4)$$

在此基础上再进行一次移动平均，即二次移动平均：

$$M_t^{(2)} = \frac{1}{N}\left(M_t^{(0)} + \cdots + M_{t-N+1}^{(1)}\right) = M_{t-1}^{(2)} + \frac{1}{N}\left(M_t^{(1)} - M_{t-N}^{(1)}\right) \cdots\cdots\cdots (5)$$

设序列$\{y_t\}$从某时期开始具有直线趋势，且认为未来时期也会按此趋势变化，则可设此直线趋势预测模型为：

$$\hat{y}_{t+T} = a_t + b_t T,\ T = 1,2,\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

其中 t 为当前时期；T为由至预期的时期；a_t为截距；b_t为斜率。二者又称平滑系数。

根据移动平均值确定平滑系数：

$$a_t = y_t\ldots\ldots\ldots\ldots(7)$$

$$y_{t-1} = y_t - b_t\ldots\ldots\ldots\ldots(8)$$

$$y_{t-2} = y_t - 2b_t\ldots\ldots\ldots\ldots(9)$$

$$y_{t-N+1} = y_t - (N-1)b_t\ldots\ldots\ldots(10)$$

所以

$$M_t^{(1)} = \frac{y_t + y_{t-1} + \ldots + y_{t-N+1}}{N} = \frac{Ny_t - [1 + 2 + \ldots + (N-1)]b_t}{N} = y_t - \frac{N-1}{2}b_t\ldots(11)$$

因此

$$y_t - M_t^{(1)} = \frac{N-1}{2}b_t\ldots\ldots\ldots(12)$$

推导可得：

$$y_{t-1} = M_{t-1}^{(1)} = \frac{N-1}{2}b_t\ldots\ldots\ldots(13)$$

所以

$$y_t - y_{t-1} = M_t^{(1)} - M_{t-1}^{(1)} = b_t\ldots\ldots\ldots\ldots(14)$$

推导可得：

$$M_t^{(1)} - M_t^{(2)} = \frac{N-1}{2}b_t\ldots\ldots\ldots(15)$$

基于此，可得平滑系数计算公式：

$$\begin{cases} a_t = 2M_t^{(1)} - M_t^{(2)} \\ b_t = \frac{2}{N-1}\left(M_t^{(1)} - M_t^{(2)}\right) \end{cases}\ldots\ldots\ldots(16)$$

基于上式，用 Python 的移动平均算法程序预测各数据风化前的化学成分，预测部分结果如下，详情见支撑材料。

表 7　风化前各化学成分含量表

类型	文物采样点	二氧化硅(SiO_2)	……	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)
铅钡	38	58.86	……	0.39	21.33	6.25
铅钡	50	50.42	……	0.94	29.87	14.22
高钾	12	73.14	……	2.79	0.00	0.00
铅钡	26 严重风化点	13.9	……	3.99	27.05	47.27

5.3 问题二模型的建立与求解

5.3.1 高钾玻璃、铅钡玻璃的分类规律分析

（A） 基于 SMOTE 采样法的数据倾斜处理

在表单二的所有有效数据样本中，在表单二的所有有效数据样本中，高钾玻璃样本数仅为 18，占比 26.8%，存在数据倾斜现象。若直接用不平衡样本筛选特征，会导致特征无法很好地预测少数类样本，从而降低正确率与模型泛化能力。因此，本文对表面风化为“无风化”的样本数据采用 SMOTE 采样法，以尽可能消除不均衡数据量的影响。

SMOTE 采样法的原理是少数样本之间进行插值来产生额外样本。具体地，计算每个少数类样本的 K 个近邻，再从 K 个紧邻中随机挑选样本进行随即线性插值，从而构造新的少数类

样本，最后将新样本与原数据合成，产生新样本。

（B） 基于 Filter 过滤法的数值型特征初步过滤

本题所给的特征数据均为数值型，而 Filter 方差过滤法是通过计算各数值型特征的方差，在此基础上选出方差大于其阈值的特征。对于方差较小的特征，因为其取值和平均值近似，在模型中的贡献组成度较低，因此在一开始可以过滤掉此类特征。对于方差较大的特征，其分布更加分散，对模型扰动程度高，在分类预测中的效果也更好。方差的计算公式为：

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \ldots\ldots\ldots\ldots (17)$$

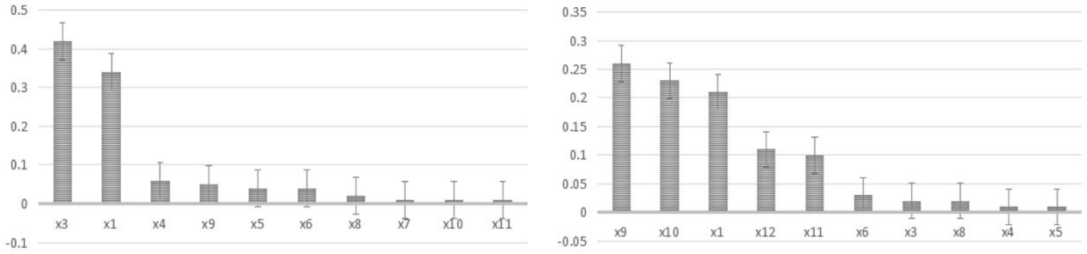
其中，σ^2为方差，n 为其样本数量，x_i为特征 x 的第 i 个取值，$\overline{x}$为特征的期望。在 14 个原始数值型特征的基础之上，通过规定方差阈值 $Q = 0.1$，14 个化合物特征经 Filter 过滤后，高钾玻璃与铅钡玻璃分别得到了 10 个合格特征，得到的特征集如下表所示：

表 8 数值特征初次过滤后的特征集

类型	阈值	数值特征第一次过滤后的特征集									
高钾玻璃	0.1	二氧化硅 (SiO_2)	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)
铅钡玻璃	0.1	二氧化硅 (SiO_2)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al_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)

（C）基于随机森林的数值型特征二次过滤

Filter 过滤法的特征选择与模型训练过程是独立进行的，只能排除部分特征。在特征筛选的第二阶段综合考虑特征选择与模型训练两过程，采用嵌入式的随机森林筛选方法。在前文筛选出的 10 个候选指标中，通过随机森林模型度量出各个特征的重要性程度。随机森林以决策树作为基分类器，构建 CART 二叉树，以 Gini 指数最小的属性作为其最优划分属性。特征的 Gini 指数越小，表示该特征更加重要。通过决策树对各个特征区间的划分得到其对应重要性程度图，按降序作图如下：



高钾玻璃　　铅钡玻璃

图 6 高钾、铅钡玻璃的重要性程度图

本文删除了重要性程度低于 10%的特征，得到了高钾玻璃和铅钡玻璃在二次筛选后存留的特征集，最终得到筛选后的影响指标：

表 9　数值特征二次过滤后的特征集及部分说明

类型	数值特征第二次过滤后的特征集及部分说明
高钾玻璃	**1.氧化钾(K_2O）：重要性程度为 0.42** 高钾玻璃表面的氧化钾成分比例最具特征。从数据观测得出高钾玻璃表面的氧化钾有较大概率数值较高，但也不能因为其值较低就否定其不是高钾玻璃。如文物采样点 07、21 和 27 的氧化钾数值为 0 但依旧判定为高钾玻璃。 **2.二氧化硅(SiO_2)：重要性程度为 0.34** 高钾玻璃表面的二氧化硅成分比例是次要特征，其成分普遍较高，都处于 59%以上。但铅钡玻璃中也有少部分高于 59%，不能因为高二氧化硅比例就否定其不是铅钡玻璃。
铅钡玻璃	**1.氧化铅(PbO)：重要性程度为 0.26** 铅钡玻璃表面的氧化铅成分比例最具特征。从数据观测可得，所有的铅钡样本氧化铅含量均高于 9%且高钾样本氧化铅含量均低于 2%。 **2.氧化钡(BaO)：重要性程度为 0.23** 铅钡玻璃表面的氧化钡成分比例是次要特征。铅钡样本氧化钡的含量普遍较高但也有出现为 0 的情况。 **3.二氧化硅(SiO_2):重要性程度为 0.21** 铅钡玻璃表面的二氧化硅成分比例是第三特征，其成分普遍较低，大多处于 59%以下，但也有部分远大于 59%。 **4.氧化锶(SrO):重要性程度为 0.11** 铅钡玻璃表面的氧化锶是一般特征。对比高钾玻璃，铅钡玻璃的氧化锶成分不为 0 的概率更大，但不显著。 **5.五氧化二磷(P_2O_5）：重要性程度为 0.10** 铅钡玻璃表面的氧化锶是一般特征。对比高钾玻璃，铅钡玻璃的五氧化二磷成分普遍更小，但不显著。

由上表分析可得，区分高钾玻璃和铅钡玻璃的依据可以有以下三类：

①**氧化铅(PbO)的决定性判断**：由于氧化铅是铅钡玻璃最显著的特征，且铅钡玻璃的氧化铅浓度下限与高钾玻璃氧化的上限之间存在 6%的差值，差异显著。因此氧化铅的浓度是区分玻璃类别的决定性因素。规定 5%的铅含量界限，可以非常快速、准确地区分两类玻璃。

②**二氧化硅(SiO_2)的粗略判断**：由于二氧化硅是高钾和铅钡玻璃的**共同稳定特征**，且在分布上呈现完全不同的趋势。因此可以大致判断，二氧化硅浓度较高的为高钾玻璃，较小的为铅钡玻璃。但是，存在钾较少的高钾玻璃。因此，此判断为概率判断，而非必然判断。

③**其他因素（K_2O、BaO、SrO、P_2O_5）的可能性判断**：此类氧化物均是两类玻璃组内较为显著的特征，但均存在特征交叉、特征重合的情况，很难通过范围来确定类标。因此，此类判断不显著，不是良好的初步分类指标。但它反应了两类玻璃更加独特的组内性质，是**良好的亚分类指标。**

5.3.2 亚类划分及其划分结果性质

(A)　基于 FCM 模糊聚类模型的亚类划分

对题中的玻璃样本进行分类时，需要考虑到以下两个方面：

①化学成分的含量是连续性的数值变量，其界限并不分明。

②要对整体的玻璃样本聚类以求得其最优解。基于此背景，我们采用模糊聚类的模糊 C-均值(FCM)模型求解此问题。模糊聚类通过隶属函数来确定数据隶属各个簇的程度，使相近样本尽可能归为一类。以此方法进行聚类，可以避免求解最优解的局部限制性。

FCM 模糊聚类的一般形式为：

$$J(U, a_1, \ldots, a_a) = \sum_{j}^{n} u_{ij}^{m} d_{ij}^{2} \ldots\ldots\ldots\ldots (18)$$

$$d_{ij}=\|a_i - x_j\|.....(19)$$

其中u_{ij}介于 0 到 1 之间，a_i为模糊组 I 的聚类中心，d_{ij}为第 I 个聚类中心与第 j 个数据点间的欧几里得距离；m 是加权指数。构造目标函数，达到最小值的必要条件如下：

$$\bar{J}\left(U, a_1, \ldots, a_a, \lambda_1, \ldots, \lambda_n\right) = \sum_{i=0}^{a} \sum_{j}^{n} u_{ij}^{m} d_{ij}^{2} + \sum_{j}^{n} \lambda_j \left(\sum_{i=1}^{n} u_{ij} - 1\right).....(20)$$

接下来，由于样本数量较小，确定模糊加权指标为 2，隶属度矩阵为 3，迭代计数次数为 50，迭代终止条件为隶属度最小变化量小于 1e-5。设置聚类中心数为 2，即将每一个样本在该化学成分的特征下分为高含量组和低含量组。将各高钾玻璃和铅钡玻璃样本的 14 种化学成分进行聚类。结果如下：

在高钾玻璃中，按照化学成分进行 14 次聚类分析，分别为：高硅组、低硅组；高钠组、低钠组；高钾组、低钾组；.....；高硫组、低硫组。做出模糊聚类结果图，以五氧化二磷为例，聚类结果比较分散，不适合作为聚类指标。对比可得，以钾含量为依据的聚类结果是最优的，因此选用钾作为第一步的亚分类指标，将高钾玻璃分为高钾-高钾型和高钾-低钾型。

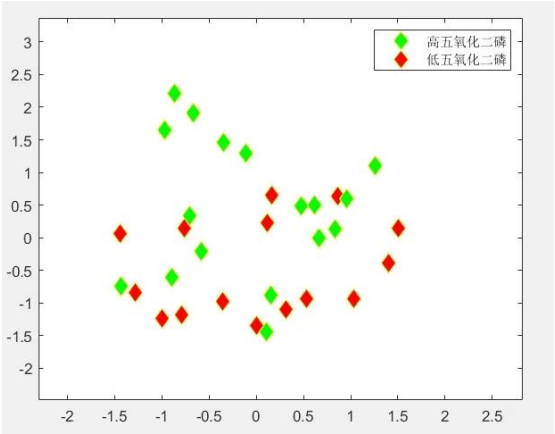


图 7 高低五氧化二磷聚类图

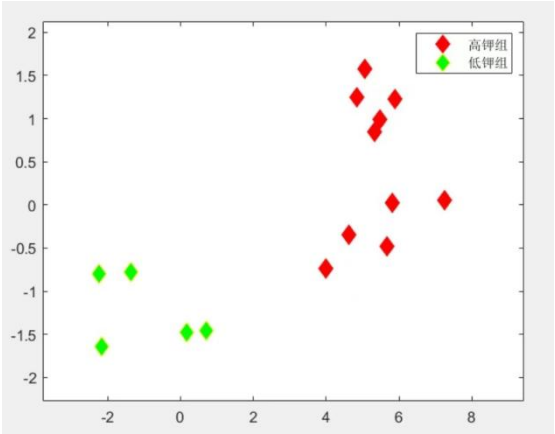


图 8 高低钾聚类图

低钾型的玻璃样本中二氧化硅均作为最主要的成分，化学成分结构比较稳定。因此，低钾型的分类最终结果命名为：高钾-高硅主导型玻璃$Si0_2 - K_2O$。（**简称为高钾高硅玻璃**）为而在高钾型中，玻璃的化学成分结构不统一。因此，再次按照上述思路进行聚类。选取最为显著的指标--钙进行聚类。

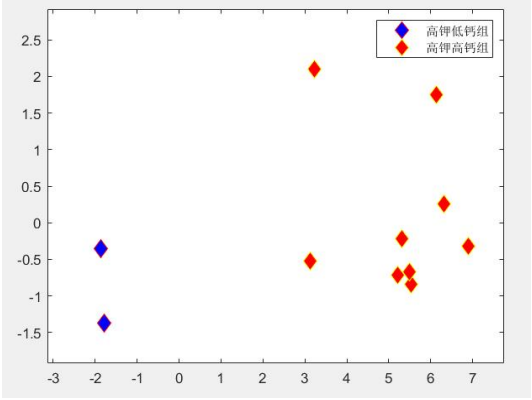


图 9 高钾高低钙聚类图

基于此，最终将高钾-高钾型分为：①含钙量较低的高钾-高钾型：高钾-高钾主导型玻璃 $K_2O - Si0_2 - CaO$（**简称为超高钾玻璃**）②含钙量较低的高钾-高钾型：高钾-钙钾主导型玻璃 $CaO - K_2O - Si0_2$（**简称为高钾钙钾玻璃**）

接着，在铅钡玻璃中，重复上述思路进行聚类分析。对比发现，以铅和钡含量为依据的聚类结果是最为显著的。结合表中数据与铅钡玻璃的性质，将铅和钡共同作为第一步的亚分类指标。

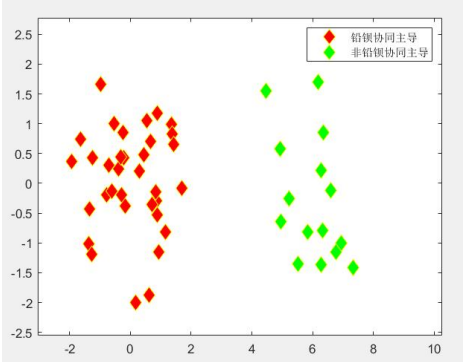


图 10 铅钡协同与否聚类图

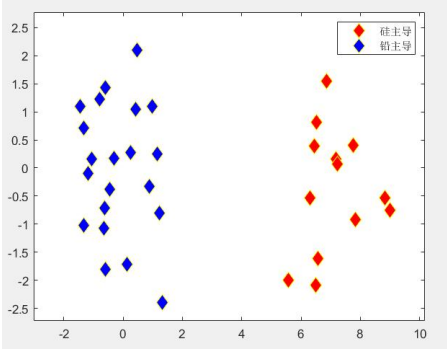


图 11 铅钡不协同-铅硅聚类图

依据铅钡含量聚类，最终将铅钡玻璃分为：①铅和钡含量均较大的铅钡玻璃：铅钡-铅钡协同主导型玻璃 $PbO-BaO-Si0_2$（**简称为超铅钡玻璃**）②铅和钡含量最少有一个含量较小的铅钡玻璃：铅钡-非协同主导型。此类玻璃通常钡的含量非常小，部分玻璃含有大量铅，部分玻璃含有大量二氧化硅少量铅。因此，最终以铅的含量进行聚类。

基于此，将铅钡-高钾组最终分为：①含铅量较高的铅钡-非协同主导型：铅钡-高铅主导型玻璃 $PbO-Si0_2-BaO$（**简称为铅钡高铅玻璃**）②含铅量较低的铅钡-非协同主导型：铅钡-高硅主导型玻璃 $Si0_2-PbO-BaO$（**简称为铅钡高硅玻璃**），分类思维导图如下：

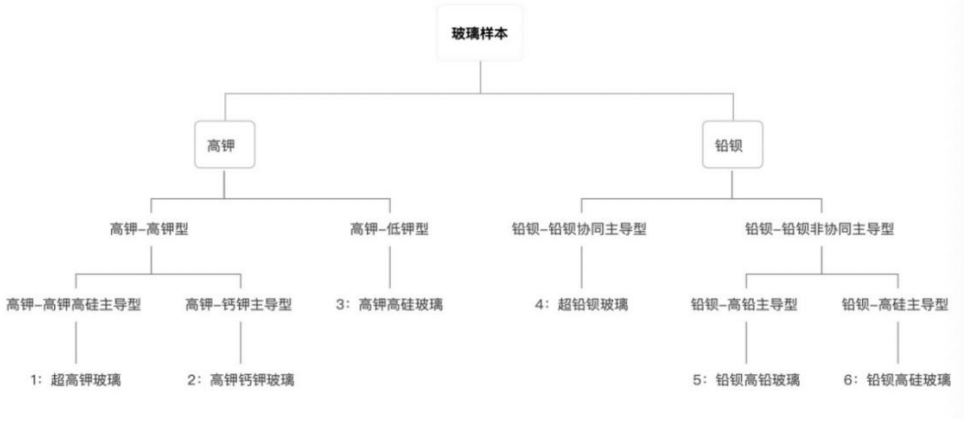


图 12 亚分类模型思维导图

根据上述亚分类模型进行分类，只选取部分展示，详见附录，结果如下：

表 10 亚分类结果表

文物编号	亚分类	文物编号	亚分类	文物编号	亚分类
01	高钾高钙	20	超铅钡	53 未风化点	铅钡高硅
02	铅钡高铅	21	高钾高硅	54	铅钡高铅
03 部位 1	超高钾	22	高钾高硅	54 严重风化点	铅钡高铅

(B) FCM 模糊聚类模型合理性分析

为了验证聚类模型的合理性，保证得到有效的分类结果，对以上四次聚类情况的 FCM 算法进行有效性分析。本文采用内部标准指标 Calinski-Harabasz；Davies-Bouldin 评判同一聚类算法在同一聚类数 2 下聚类结果的优良程度，算式如下：

$$CH(K)=\frac{trP(k)/(k-1)}{trQ(k)/(2-k)}\cdots\cdots\cdots\cdots(21)$$

其中，k 表示当前类， trP(k)表示类间离差矩阵的迹， trQ(k)表示类间离差矩阵的迹。CH 反应的是数据类之间的紧密，数值越大代表类之间越紧密，效果较好。反之代表之间比较分散。效果较差。

$$DB(K)=\frac{\sum_{i=0}^{k}\max_{i=1\ldots k,j=i}\left(\frac{D_i+D_j}{C_{ij}}\right)}{K}\ldots\ldots\ldots\ldots\ldots(22)$$

其中，D_i表示类C_i中所有样本到聚类中心的平均距离，D_j表示类C_j中所有样本到聚类中心的平均距离，C_{ij}表示类C_i和C_j中心之间的距离。DB 值反应数据点间的相似度，是平方类变量，其值越小类与类之间的相似度就更低，是更好的聚类结果。通过上述有效性检验，得到 CH=0.894，DB=112.2 **聚类有效性分析结果较好**，中心数为 2 的**聚类分析有效**。

(C) FCM 模糊聚类模型敏感性分析

为了验证模型的敏感性，需要引入噪声点、离群点，分析其对聚类系统的影响，以此作为聚类模型敏感性的判断依据。噪声[3]是一个测量变量中的随机错误或偏差。如果数据点周围存在较多样本点，聚类系统将其判断为噪声点，评判误差较大。反之，如果样本点远离大部分样本点，且其欧氏距离较大时。视为噪声点更为合理，评判误差较小。

由于模型中模糊隶属度之和为 1 的约束条件。噪声点的存在会带来聚类中心的**部分偏移**。其表现为：算法如果将噪声归为聚类中心，会往噪声处偏移。当噪声点离数据簇比较远时会将此点判别成离群点，造成数据簇的类内聚程度降低，削弱了数据簇之间的可分性。

在此基础上，单独地在聚类系统中引入十次噪声点，点的个数在 1 至 6 个不等。记录观察结果。只选取第 3 次和第 10 次的结果展示，如下：

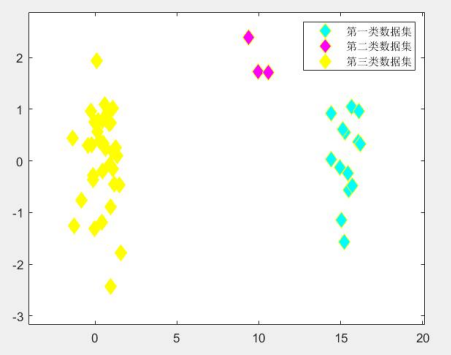


图 13 第三次引入噪点数据集

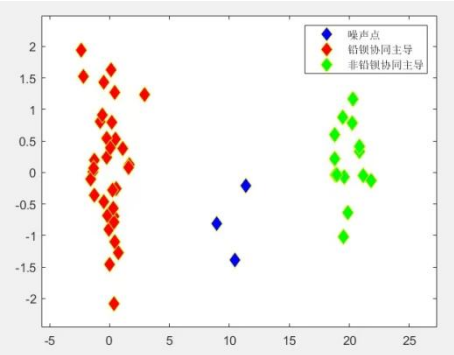


图 14 第三次引入噪点聚类图

由上图可知，第 3 次观测时，在数据集中引入 3 个噪声点。但没有影响 FCM 聚类模型的聚类结果，聚类系统在聚类时可以把噪声点全部判别出来。**在 10 次观测中，有 9 次噪声点的引入没有影响到了聚类结果**。可以得出聚类模型对于噪声的敏感性较差，鲁棒性较强。

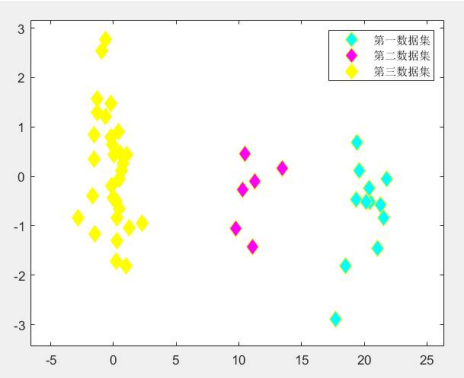


图 15 第十次引入噪点数据集

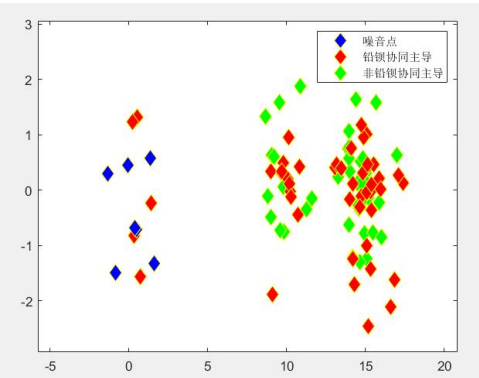


图 16 第十次引入噪点聚类图

上图是第 10 次失败结果的观测过程。在数据集分布引入了 6 个噪声点，但改变了模型的聚类结果。由此可以得出，虽然模型的鲁棒性较强，但是依旧不能抵御多噪声低的干扰，需要在后续对模型进行**优化设计**。

(D) 模型结论的现实应用

接下来对分类结果进行现实意义说明。

①高钾玻璃的亚分类结果：

1. 高钾-高硅主导型玻璃$Si0_2-K_2O$。（**简称为高钾高硅玻璃**）
2. 高钾-高钾主导型玻璃 K_2O-Si0_2-CaO（**简称为超高钾玻璃**）
3. 高钾-钙钾主导型玻璃 $CaO-K_2O-Si0_2$。（**简称为高钾钙钾玻璃**）

此结果符合高钾玻璃的历史发展。我国与西亚、埃及等丝绸之路邻国虽在古代交往密切，但玻璃的化学组成成分各大相径庭。古代中国会使用草木灰作为传统的助熔剂，所以草木灰中的主要成分K_2O 便会体现在玻璃中，形成初始的K_2O-Si0_2-CaO 体系的超高钾玻璃。战国之后，CaO 为助熔剂的冶炼技术不断发展，逐渐转向了 $CaO-K_2O-Si0_2$体系的高钙钾钙玻璃。

基于此论证，可以通过K_2O **判断丝绸之路时期的玻璃文物是否制作于中国或是舶来品**。也可以大致判断出各高钾玻璃样本制造的历史年代。

②铅钡玻璃的亚分类结果：

1. 铅钡-铅钡协同主导型玻璃 $PbO-BaO-Si0_2$（**简称为超铅钡玻璃**）
2. 铅钡-高铅主导型玻璃 $PbO-Si0_2-BaO$（**简称为铅钡高铅玻璃**）
3. 铅钡-高硅主导型玻璃$Si0_2-PbO-BaO$。（**简称为铅钡高硅玻璃**）

此结果符合铅钡玻璃的历史发展。铅钡玻璃是我国最早烧制、成型、流行的玻璃。它始于工匠有意识的利用方铅矿中混有众多含铅化合物的重晶石研制初级玻璃，因此玻璃中会含有 BaO 成分，因此在西周时期最初形成了 $PbO-BaO-Si0_2$体系的玻璃。至东汉时期，为了使得玻璃的烧制更加纯净，开始使用铅丹和黄丹来烧制玻璃，此类烧制方法在冶炼铅的过程中不可能留有过多的 BaO 成分。因此，唐宋以后，玻璃体系逐渐朝 $PbO-Si0_2-BaO$ 的高铅体系靠拢。

$Si0_2$的特征是熔点高、溶液粘度大。出于其特性，$Si0_2$在玻璃熔制中的添加与应用难度巨大。后来，随着玻璃制造业的发展，出现了在熔制时添加硫磺、硝石、硼砂等配方以提高效率，改善了其缺陷，也出现了以$Si0_2$为主导的$Si0_2-PbO-BaO$ 体系。

基于此论证，可大概**判断出各玻璃样本所制造的年代**。$PbO-BaO-Si0_2$体系的超铅钡玻璃大致制造于西周及春秋战国时期；$PbO-Si0_2-BaO$ 体系的铅钡高铅玻璃大致制造于东汉唐宋时期；$Si0_2-PbO-BaO$ 体系的铅钡高硅玻璃大致制造于为明清时期。

5.4 问题三模型的建立与求解

5.4.1 玻璃文物逐层分类模型的建立

问题三要求根据表单三所给的 8 件古代玻璃文物信息进行分类，如直接遍历上述亚分类，算法时间复杂度较高，故建立**逐层分类模型**，先进行大类初筛选，后完成亚分类。

（A） 基于问题二分类规律的初分类

前文已经论述，玻璃的分类大致有三种规则。而氧化铅(PbO)在其中起到决定性判断作用，因此采用氧化铅成分作为初分类指标。规定 5%的铅含量界限，在界限之上的判定为铅钡玻璃，反之则判定为高钾玻璃。得出分类结论，结果如下：

表 11 初分类结果表

初分类	文物编号
高钾玻璃	A1、A6、A7
铅钡玻璃	A2、A3、A4、A5、A8

（B） 基于贝叶斯判别分析模型的亚分类

①贝叶斯判别分析模型的建立

经过二分类逻辑回归模型的大类划分，已得到 8 件文物的“高钾”、“铅钡”玻璃大类归属，现求解其亚类型归属情况。贝叶斯判别分析为多类别识别的判别分析模型，具有灵活的框架和较好的预测能力。基本步骤如下：

设待判别样本$\mathrm{B}=(\mathrm{B}_1,\ \mathrm{B}_2,\ \ldots\ldots\mathrm{B}_8)$，三个已知总体$\Pi_1$，$\Pi_2$，$\Pi_3$，并且这三部分具有如下特性：$\Pi_1\cup\Pi_2\cup\Pi_3=\Pi^n$，$\Pi_1\cap\Pi_2\cap\Pi_3=\emptyset$。三个总体的密度函数为$\mathrm{P1}(\cdot)$，$\mathrm{P2}(\cdot)$，$\mathrm{P3}(\cdot)$，样本B属于总体$\Pi_1$的概率为$q_1$，[4]属于总体$\Pi_2$的概率为$q_2$，属于总体$\Pi_3$的概率为$q_3$，设立原属于 i 被误判为 j 所带来的损失函数$\mathrm{C(j|i)}$以衡量样本被误判时所导致的损失程度，产生错误的概率记为$\mathrm{P}(\mathrm{j}|\mathrm{i},\ \mathrm{R})$。以总体$\Pi_1$，$\Pi_2$为例，每次误判发生的概率可用下述公式表达：

$$\mathrm{P}(2|1,\ \mathrm{R})=\int_{\mathrm{R}_2}\mathrm{P}_1(\mathrm{x})\mathrm{dx}\ldots\ldots\ldots\ldots\ldots(23)$$

$$\mathrm{P}(1|2,\ \mathrm{R})=\int_{\mathrm{R}_1}\mathrm{P}_2(\mathrm{x})\mathrm{dx}\ldots\ldots\ldots\ldots\ldots(24)$$

那么每做一次判断的平均错误程度就可以用以下的公式衡量：

$$\begin{aligned}\mathrm{G}(\mathrm{R}_1,\ \mathrm{R}_2,\ \mathrm{R}_3)=&\ q_1\cdot C(2|1)\cdot\mathrm{P}(2|1,\ \mathrm{R})+q_1\cdot C(3|1)\cdot\mathrm{P}(3|1,\ \mathrm{R})+q_2\cdot C(1|2)\cdot\mathrm{P}(1|2,\ \mathrm{R})\\&+q_2\cdot C(3|2)\cdot\mathrm{P}(3|2,\ \mathrm{R})+q_3\cdot C(1|3)\cdot\mathrm{P}(1|3,\ \mathrm{R})\\&+q_3\cdot C(2|3)\cdot\mathrm{P}(2|3,\ \mathrm{R})\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(25)\end{aligned}$$

以上述平均错误程度最小为目标可得具体分类标准。

②贝叶斯判别分析模型的求解

分别随机抽取 35%上文已分类的高钾玻璃类、铅钡玻璃类样本数据作为测试集，剩余 65%数据作为特征集导入 SPSS 软件，得到预测结果以及准确率如下表所示：

表 12 高钾玻璃贝叶斯预测结果表

实际值 \ 预测值		高钾玻璃亚分类			总计
		高钾高硅	超高钾	高钾钙钾	
高钾玻璃亚分类	高钾高硅	7	0	0	7
	超高钾	1	2	0	3
	高钾钙钾	0	1	7	8
%	高钾高硅	100	0	0	100
	超高钾	33.3	66.7	0	100
	高钾钙钾	0	12.5	87.5	100
总体百分比					92.839

表 13 铅钡玻璃贝叶斯预测结果表

实际值 \ 预测值		铅钡玻璃亚分类			总计
		铅钡高铅	超铅钡	铅钡高硅	
铅钡玻璃亚分类	铅钡高铅	13	0	2	15
	超高铅	1	19	0	20
	铅钡高硅	1	1	12	14
%	铅钡高铅	86.7	0	0.13	100
	超高铅	5	95	0	100
	铅钡高硅	7.1	7.1	85.8	100
总体百分比					95.812

由表可得，该模型分类准确度较高，高钾玻璃亚分类准确度为 92.839%，铅钡玻璃亚分类准确度为 95.812%。

将 8 件待亚分类文物信息导入 SPSS 软件，操作步骤同上，得到如下亚分类结果：

表 14 待分类文物预测结果表

文物编号	初分类	亚分类
A1	高钾	高钾高硅
A2	铅钡	铅钡高铅
A3	铅钡	铅钡高铅
A4	铅钡	铅钡高硅
A5	铅钡	铅钡高硅
A6	高钾	高钾高硅
A7	高钾	高钾高硅
A8	铅钡	铅钡高硅

5.4.2 分类模型的敏感性分析

（A） 基于洛伦兹曲线（KS 曲线）的模型分割样本能力检验

KS 曲线是用来衡量分类型模型分割样本能力的工具。问题三中先对未知类别的玻璃文物进行二分初分类，再进行三分亚分类，下面以二分初分类为例探究其分割样本能力，三分亚分类则同理。

KS 曲线横轴为阈值，纵轴为“在所有真实值为 Negative 的数据中，被错误判断的比例”与“在所有真实值为 Positive 的数据中，被正确判断的比例”两条曲线。KS 值即两曲线相距最远的距离，最远距离越大即模型分割样本能力越强，对样本数据越敏感。KS 曲线如下图：

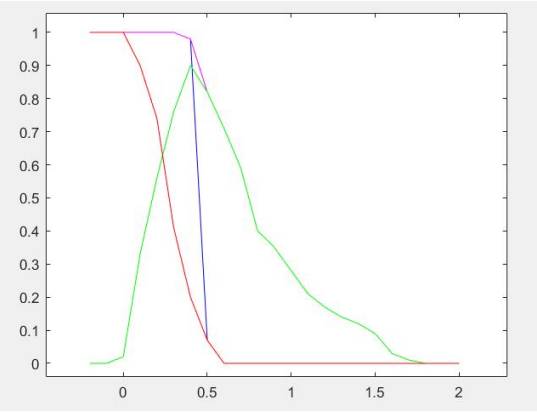


图 17 KS 曲线图

在阈值为 0.5122 的时候，TRR-FRP 的差距最大，为 0.8931。因此认为是很好的输出结果，即**模型分割样本能力强，模型敏感性高。**

（B）基于扰动分析的敏感性检验

根据分类规律选择氧化钾、氧化铅、氧化钡、二氧化硅四个对模型输出结果有重要影响的关键成分作为参数。分别在不同类别的玻璃中进行局部敏感性分析。采用扰动分析的相对敏感性检验方法。即在其余参数不改变的情况下，仅使 1 个参数在一定范围内扰动。扰动幅度为-20%、-15%、-10%、-5%、0、5%、10%、15%、20%。接着进行逐次模拟并计算出参数的敏感度 S 和其扰动结果变化率γ，算式如下：

$$S=\left|\frac{\sum_{i=0}^{n}\frac{(M_{i+1}-M_i)/M_a}{(P_{i+1}-P_i)/P_a}}{n-1}\right|,\ \gamma=\frac{M_p-M_b}{M_b}\qquad(26)$$

其中，P_i和P_{i+1}表示第次参数的第 i 和 i+1 次的输入值；P_a为两次输入均值。M_i和M_{i+1}表示第 i 和第 i+1 次模型产量的模拟值；M_a为两次模拟值均值。M_p为扰动后的模拟值，M_b为基准参数下的参量模拟值。根据 S 可以将[5]敏感度划分**五个范围**。

表 15 敏感度划分表

级别	范围	敏感程度
1	$S<0.10$	不敏感
2	$0.10\le S<0.25$	弱敏感
3	$0.25\le S<0.50$	一般敏感
4	$0.50\le S<1.00$	比较敏感
5	$S\ge 1.00$	非常敏感

经过模拟计算，综合得出敏感度值 S 为 1.219；**模型非常敏感**。

综以上两种判别方法，**模型的敏感性强**。

5.5 问题四模型的建立与求解

5.5.1 类别的合并与选用说明

在对不同类别的玻璃进行组内成分的相关关系分析时，需要保证各个指标的样本量较为充足。本文针对**玻璃类别和玻璃颜色**两种分类方式进行关系研究。

在此特别说明的玻璃类别的选用方式。模型在亚分类后各类型样本的个数从小到大依次为：超高钾（3 个）、高钾高硅（7 个）、高钾高钙（8 个）、铅钡高硅（14 个）、铅钡高铅（15 个）、超铅钡（20 个）。可见，在样本较少的高钾玻璃中由于亚分类的细化，各类别样本数极少，且难以用补值插值的方式弥补缺陷，不具有研究意义。所以，将高钾玻璃作为整体进行研究。而铅钡玻璃样本较多，且在亚分类后每类的样本数均较为充足，可以单独进行相关分析。基于此，选择高钾玻璃、超铅钡玻璃、铅钡高硅玻璃、铅钡高铅玻璃进行相关分析。

5.5.2 Spearman 相关分析模型的建立

由于题中所给数据中正态分布性和线性关系度明显不足，如若使用皮尔逊相关分析则模型的精度不高。故采用效率稍低，但精度更高的 Spearman 相关性分析进行插层并计算各单向指标之间的关联性r_s。首先，定义a_i为各单向指标插层后的数值，b_i为每组的插层率，计算 Spearman 相关系数。

$$r_s=\frac{\sum i\,(a_i-\overline{a})-\left(b_i-\overline{b}\right)}{\sqrt{\sum i\,(a_i-\overline{a})^2\left(b_i-\overline{b}\right)^2}}\qquad(27)$$

其次，根据原始数据在总数中降序排列。由于每个单向指标的含插层率样本数都接近大

样本情况，定义统计量g_{ij}和误差函数p，算式如下：

$$g_{ij}=\begin{cases} r_s\sqrt{n-1}\sim N(0,1) \\ H_0: r_s=0\ ; H_1: r_s\neq 0 \end{cases}，\ p=r_s\sqrt{n-1} \quad \cdots\cdots\cdots\cdots (28)$$

5.5.3 Spearman 相关分析模型的求解

根据上述模型，依次在高钾玻璃、超铅钡玻璃、铅钡高硅玻璃、铅钡高铅玻璃中进行相关分析。将组间成分的相关系数进行可视化处理，做出相关性系数热力图，只呈现一张其余在支撑材料中显示：

图 18　斯皮尔曼相关性热力图

当$p<0.05$时，认为p值显著，可判断两变量之间存在相关性。

5.5.4 关联关系与差异性分析

根据热力图标记相关系数绝对值大于 0.5 的成分组合，在此基础上初步分析成分关系。在各类玻璃类型中，符合要求的成分组按照相关系数从大到小的排序的表格如下，并且每个表格中都随机选取几个相关性较小、不符合要求的成分组，并用*标记。

表 16 部分成分列举表

高钾玻璃	超铅钡玻璃	铅钡高铅玻璃	铅钡高硅玻璃
二氧化硅 氧化铝：-0.831	氧化钡 二氧化硅：-0.7	氧化钙 氧化镁：0.81	氧化钙 氧化铝：0.77
二氧化硅 氧化铁：-0.783	氧化铜 氧化钡：0.69	氧化铅 五氧化二磷：-0.76	氧化镁 氧化铝：0.71
二氧化硅 氧化钾：-0.781	二氧化硅 氧化铝：0.66	氧化铁 氧化铅：-0.67	氧化钾 氧化铝：0.7
二氧化硅 氧化钙：-0.752	氧化锡 氧化铁：0.64	氧化铝 氧化铁：0.62	氧化铝 氧化铅：-0.68
氧化钾 氧化钙：0.69	氧化纳 五氧化二磷：-0.61	氧化铁 氧化锶：-0.62	氧化纳 五氧化二磷：-0.67
氧化铅 氧化钡：0.682	二氧化硅 氧化铅：-0.59	氧化铅 氧化锶：0.6	氧化纳 氧化钡：0.62
氧化镁 氧化锶：0.666	----------	五氧化二磷 氧化锶：-0.6	二氧化硅 氧化钡：-0.6
氧化铜 氧化铁：0.643	*氧化钙 氧化镁 -0.08	----------	氧化铝 氧化锶：0.58
氧化镁 五氧化二磷 0.643	*氧化铜 氧化锶： 0.25	*氧化铜 氧化锶 -0.14	----------
氧化钙 氧化纳：0.64	*氧化钡 氧化锶： 0.29	*氧化钡 氧化锶： 0.27	*氧化钙 氧化镁：0.49
氧化钾 氧化纳：0.606			*氧化铜 氧化锶 0.09
氧化铝 氧化铅：0.602			*氧化钡 氧化锶： 0.12
*氧化钙 氧化镁：0.087			
*氧化铜 氧化锶： 0.141			
*氧化钡 氧化锶： 0.565			

由上表可以从宏观、特征、特殊三个角度分析出成分间的相关关系：

①从**宏观**来看所有成分组合，可以得到各类玻璃**每两种成分的相关关系**。以高钾玻璃的部分成分组为例作说明：二氧化硅和氧化铝之间存在高度负相关性；氧化钡和氧化锶之间存

在中度的正相关性；氧化铜和氧化锶之间存在轻度的正相关性；氧化钙和氧化镁之间不存在相关性。其他组间内各个元素组的相关性不再赘述。（此类方法可以**从总体角度解释关联**）

②从**单一成分**发散来看，根据某成分与不同化合物的相关系数正负性可以判断**该成分与另一类成分的共性关系**。以二氧化硅为例，在高钾玻璃中二氧化硅与氧化铝、氧化铁、氧化钾、氧化钙的相关系数均为很大的负数。可以判断出二氧化硅与金属氧化物之间存在相互抑制关系。玻璃在风化进程中金属氧化物可以有效防止风化。同时，在风化后二氧化硅增大逐渐取代了玻璃本身的金属氧化物。此类关联现象也**与第一问中点二列相关性分析的结论吻合。**（此类方法一般选取比较高或低的相关系数的成分组合进行分析，**从特征的角度解释关联**）

③对若干含有**特殊**金属氧化物的成分组合，在不同玻璃类型中观测其分布差异，可以反应该成分组合在不同玻璃类型中配比情况，进而得到该**玻璃在烧制时的配料比**。以氧化钡和氧化锶为例，在高钾玻璃中其相关系数为 0.565 呈中高度正向相关，但在其他三类玻璃中只呈现微弱相关。此类分析具有很好的化学意义。在玻璃的烧制中，引入氧化钡的目的与氧化锶类似。均为提高玻璃的折射率、密度、光泽。因此在玻璃成分中往往会同时引入氧化钡和氧化锶。此技术在唐宋后才逐渐发展起来，而氧化钡和氧化锶的相关性可以适时的**反应其烧制配比的方式，推断玻璃烧制的时代。**（此类方法一般选取比较特别的成分组进行分析，从特殊的角度解释关联。）

反之，也可以从宏观、特征、特殊三个角度分析出不同类别成分间相关关系的差异：

①从**宏观**来看，每种玻璃成分的组间相关系数值差异巨大。以氧化钙和氧化镁为例，在铅钡高铅玻璃中的相关系数为 0.81，而在铅钡高硅玻璃中为 0.49，在其他两类中分别为 0.087 和-0.08，可以看出同一成分组在不同类型的玻璃中的相关程度差异大。

②从每一类玻璃的成分组合相关系数的**最值**来看，可以得出不同玻璃体系背后**主导成分**的差异性。在高钾玻璃中，最值均从二氧化硅与其他氧化物的组合中取得，证明在高钾玻璃中是以二氧化硅作为**自发性影响因素**，主导其他氧化物的改变。而在铅钡高硅玻璃中是以氧化铝作为自发性影响因素，主导其他氧化物的改变。

③从同一成分组合在不同类别玻璃的**不同表现**来看，可以看出此类玻璃体系的**颜色等物理属性差异**。当常规高相关性成分组合在某类玻璃中异常地出现低相关性或者常规低相关性成分组合在某类玻璃中异常地出现高相关性时，可分析造成差异的原因。以氧化锶为例，在各类玻璃中氧化锶与其他成分间的相关系数均比较小，但在超铅钡玻璃中却和 5 种化合物的相关系数大于 0.5。调取数据发现，在超铅钡玻璃中氧化锶含量较高的均为**紫色玻璃**。可以得到氧化锶在玻璃熔制成色环节中，玻璃类型的不同和化学成分的不同会起到重要影响，关联关系差异性大。此类分析实际意义较强。因为含锶的超铅钡玻璃在熔制中会与其他氧化物结合变为紫色，进而抵挡 X 辐射。

六、模型的检验

6.1 问题一 Point-Biserial 点二列显著性检验

为了保证模型的精度，需要对点二列相关性系数 r 的显著性[6]进行检验。将上文公式中 M1、M0 进行 t 检验，若差异显著，表明 r 显著。如差异不显著，则说明 r 不显著。但对独立样本进行 t 检验需要满足以下假设：观测变量为连续变量、观测值之间相互独立、符合正态分布、观测变量间不存在显著的异常值。正态性可以使用 Shapiro-Wilk 检验来判断，而箱线图选择异常值比较客观，在识别异常取值上具有优越性。因此，用 SPSS 输出 Shapiro-Wilk 检验结果并在导入表 2 数据，得到各化学成本对应的箱线图。本文只选取有代表性的数据进行论述，其余详见支撑材料。

所得正态分布的部分结果如下：

表 17 部分正态分布检验表

类型	化学成分	检验-风化	检验-无风化
铅钡	二氧化硅	0.088	0.362
铅钡	氧化铅	0.385	0.546
高钾	二氧化硅	0.304	0.072
高钾	氧化钠	0	0
高钾	氧化钾	0.175	0.255

数据接近正态分布，那么 Shapiro-Wilk 检验的 P 值就大于 0.05；反之如果数据并不接近正态分布 P 值就小于 0.05。从上表得，仅以氧化钠为代表的成分不呈现正态分布，将此类指标排除后进行 t 检验。所得箱线图的部分结果如下：

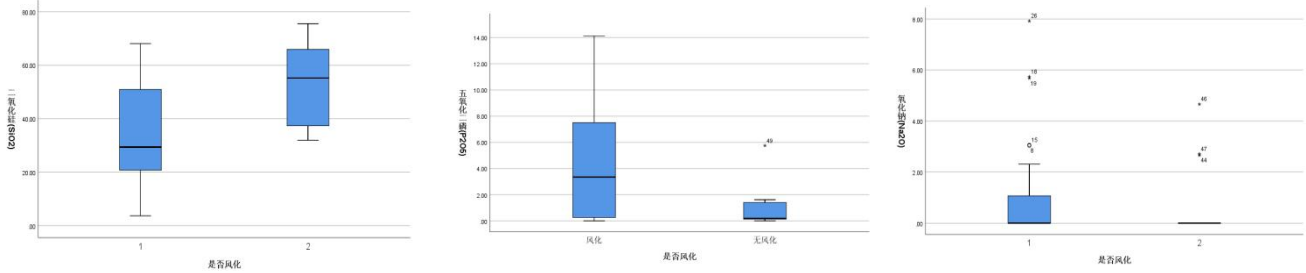


图 19　二氧化硅、五氧化二磷、氧化纳箱线图

如上图所示，用*标出的代表异常取值。在高钾玻璃的化学成分中，以二氧化硅为代表的箱线图非常集中，没有出现异常值。以五氧化二磷为代表的箱线图比较集中，但出现了零星的异常值，此时需要将其异常值剔除并利用均值加以修正。在铅钡玻璃的化学成分中，以氧化钠为代表的箱线图非常分散，异常值很多，将此类指标排除后进行 t 检验。

利用 Excel 和 Spss 软件进行 t 检验，所得结果如下，其中√表示显著，×表示不显著。

表 18 两步检验结果表

氧化铅		五氧化二磷		氧化钙		氧化镁		氧化锶	
×	√×	√ √	×	√ √	√ √	√ √	√×	×	√×
二氧化硫		**氧化铝**		**氧化铜**		**氧化钠**		**氧化钡**	
×	×	√ √	√×	√×	×	×	×	×	×
氧化锡		**氧化铁**		**氧化钾**		**二氧化硅**			
×	×	√ √	×	√ √	×	√ √	√ √		

在上表中，每一个化学成分的左侧格为高钾玻璃的 t 检验结果，右侧格为铅钡玻璃的 t 检验结果。检验结果分为两步，第一步为检测该成分数据在该类型中是否符合 t 检验条件，即是否出现异常值或不符合正态分布。第二步为检测该成分数据在该类型中的 t 检验结果是小于规定阈值。未通过第一步的成分会被标记一个×，不再进入第二步判定。未通过第二步判定的会被标记为√×，通过的被标记为√ √，只有被√ √记号成分的点二列相关性系数影响较为显著，其余均判定为不显著。

综上，在高钾玻璃中五氧化二磷、氧化钙、氧化镁、氧化铝、氧化铁、二氧化硅及铅钡玻璃中氧化钙、二氧化硅的点二列相关性系数影响比较显著。**和模型结论中所得结果较为一致，模型结论的显著性较强**。并且证明后续模型中选取的二氧化硅、氧化钙两个指标具有良好的统计学意义。

6.2 问题三二分类规律准确度检验—二分类逻辑回归模型

在问题三中对 8 份玻璃文物的初步二分类是基于问题二总结的分类规律得出的，其准确性有待验证，现用二分类逻辑回归模型对其进行检验。具体地，将特征向量输入分类模型，

先将特征向量的各元素加权求和，后输入 Sigmoid 激活函数，最后得到一个 0 到 1 之间的值，该值用于表示样本属于其中一类的概率。逻辑回归训练的过程是通过创造损失函数，并利用**梯度下降法**更新模型中的权重，使得损失函数达到最小。其中，损失函数最小值可使用**负向后的极大似然估计函数**表示。得到如下分类概率对比图：

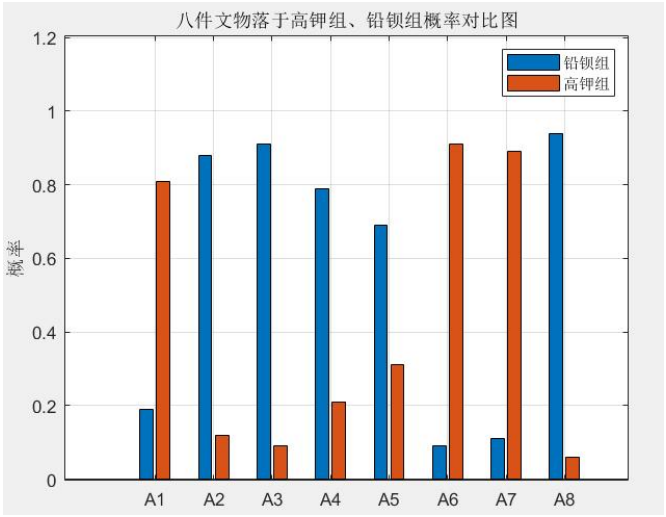


图 20　分类概率对比图

通过上图可显然得出分类结论，结果如下表，**与上文所得结论完全一致**。

表 19 分类结果表

初分类	文物编号
高钾玻璃	A1、A6、A7
铅钡玻璃	A2、A3、A4、A5、A8

6.3 问题二三亚类划分准确度检验

混淆矩阵是评判模型结果的指标，多用于判断分类器的优劣。问题二、问题三中玻璃文物的亚类划分准确度可用混淆矩阵见下图：

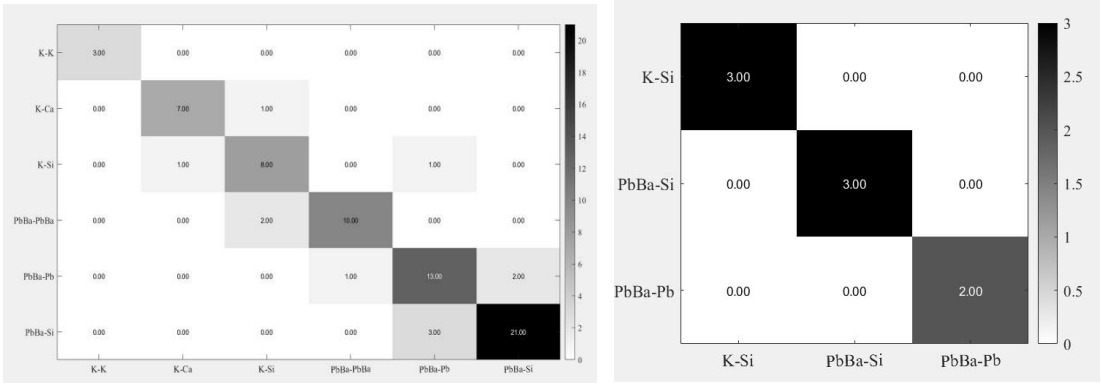


图 21　亚类划分混淆矩阵图

其中，混淆矩阵数据集中分布在对角线上，说明真实值与预测值具有较高的相同数量，**亚分类模型准确度良好**。

6.4 问题三数据集分割灵敏度检验

在解决问题三划分玻璃文物亚类时，是随机取 65%的已分类数据做训练集、35%的数据做测试集输入贝叶斯判别分析模型。因本题数据体量较小，为排除因数据集切割方式不同产生的偶然性因素，现进行数据集分割灵敏度检验。

K 折交叉验证是使数据集 D 划分为 K 份互斥的数据集D_K，满足其 $D = D_1 \cup \ldots\ldots \cup D_K$，平均分配使每份数据量接近并且使数据分布尽可能一致。每次用一份数据测试，其余数据训练，最后将 K 份数据结果汇总，从而评估指标。K 值太小将导致实验数据稳定性偏低，太高将导致实验成本大幅增加，故 K 值通常取 5。K 折交叉验证能够尽可能避免因数据集划分的偶然性产生误差。5 折交叉检验后，计算结果如下：

表 20 K 折交叉验证结果表

类型 / 序号	高钾玻璃亚类划分准确度	铅钡玻璃亚类划分准确度
1	92.839	95.812
2	90.329	91.234
3	94.129	96.126
4	90.384	94.235
5	97.238	98.564

由上表可知，两类玻璃亚类划分准确度均大于 90%，故该模型有**很好的鲁棒性**，对**数据集的分割方式不灵敏**。

七、模型的评价与推广

6.1 模型的优点

6.1.1 模型的现实推广意义强

本文对模型的推广及运用做出了较大篇幅的论证：

针对问题一，不同指标对风化的影响。本文论证了玻璃的类型在风化进程相较颜色对应的酸碱性烧制环境和纹饰对应的玻璃表面强度、裂纹密度与积水能力结构对玻璃的风化具有更显著的影响。**此结论有助于设计出更好地抗风化型玻璃**。

针对问题二三，亚分类模型可以应用于考古学领域的**文物验证**。由于我国早期助熔材料的独特性，可以通过氧化钾判断丝绸之路时期的玻璃文物**是否为舶来品**。而铅钡玻璃的发展周期短，技术更替快，不同朝代隶属于不同的铅钡体系，因此可以通过亚分类大概判断**玻璃的制造年代。**

针对问题四，相关分析模型的结论可以用于**探索玻璃成分间的未知规律**。通过不同的成分关系，可以有效推测玻璃在烧制时的配料比与烘焙环境。也可以很直观在某些特殊情况下的异常关系，进而推断玻璃更为边缘与极端的性质，例如玻璃的无色与黑色呈色原因。

6.1.2 模型的分析方法多样、可进行互相验证

本文在分析中结合了多种数学方法。在数值处理、分类与筛选中运用了 SMOTE 采样、Filter 过滤、移动平均法、随机森林、FCM 模糊聚类、贝叶斯判别等方法。在相关性分析中采用了 Pearson 卡方检验、Cramer_V 值分析、Point-Biserial 点二列相关性分析法、Spearman 相关分析并且借助了混淆矩阵、K 折交叉验证、t 检验、Shapiro-Wilk 检验、逻辑回归、KS 曲线、误差函数、Calinski-Harabasz 指标、Davies-Bouldin 指标、等常见数学工具进行检验，融合了噪声点进行参数敏感性分析。数学模型之间可进行相互检验，模型比较完整、准确度较高。

6.1.3 模型的层次丰富、逻辑紧密

模型在设计时遵循从易到难、层次清晰的原则。问题一在 Pearson 卡方检验的基础上进一步进行 Cramer_V 值分析；在频数分布图的基础上进一步 Point-Biserial 点二列相关性分析。问题二在 Filter 过滤的基础上进一步进行随机森林特征值筛选。问题三在问题二模型得出的规律基础上通过贝叶斯判别进行第二步亚分类。主要模型均进行准确度、敏感性检验，逻辑完整。

6.2 模型的缺点与改进

6.2.1 FCM 模糊聚类模型

（A）问题：FCM 模糊聚类模型鲁棒性不够强

在问题二敏感度引入噪声点的检验中，模型没有办法抵御 6 个噪声点的扰乱。在此需要对模型的鲁棒性进行增强，以抵御更多噪声点。

（B）改进：基于可靠性的鲁棒模糊聚类算法（RRFCM）

RRFCM 算法首先根据初始聚类中心判断全局可靠性，然后在阈值范围内，根据数据点的标签信息，判断数据点的局部可靠性，将两个信息整合到 FCM 损失函数中，既可解决上文中噪声点过多带来的扰动，也可解决数据簇不均衡问题。

原 FCM 算法损失度函数为：

$$J=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|x_j-v_i\right\|_2^2 \cdots\cdots\cdots\cdots (29)$$

改进后的 RRFCM 算法损失度函数为：

$$J=\sum_{i=1}^{c}\sum_{j=1}^{n}\left(u_{ij}^{m}+\lambda\ \alpha_j\right)\left[\left\|x_j-v_i\right\|_2^2+\gamma\left\|x_j-v_j^{local}\right\|_2^2\right]\cdots\cdots\cdots\cdots (30)$$

其中，u_{ij}表示模糊隶属度，v_i表示聚类中心，α_j表示数据点的模糊不确定性，v_j^{local}表示数据点近邻约束中心，λ 表示全局不确定性比例系数，γ 表示局部不确定性比例系数。

改进后再次引入 6 个噪声点，聚类的结果如下：

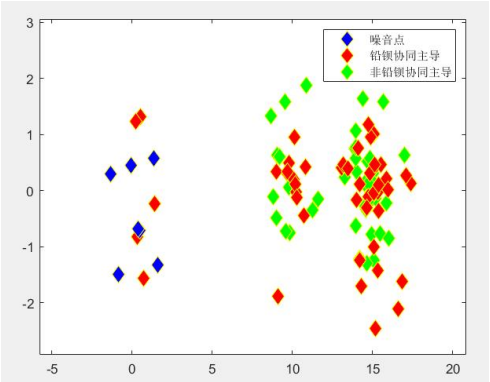


图 22　改进前聚类情况图

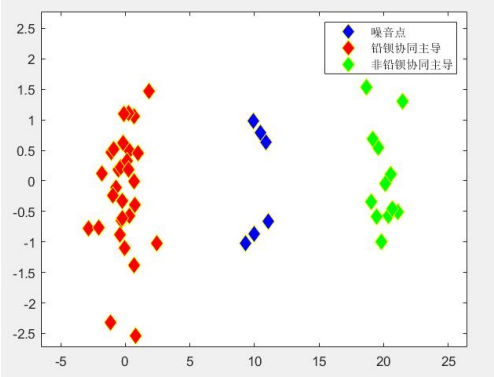


图 23　改进后聚类情况图

如上所示，改进后在数据集中引入噪声点后，没有影响 FCM 聚类模型的聚类结果，聚类系统在聚类时可以把噪声点全部判别出来。可以得出聚类模型对噪声的敏感程度降低，鲁棒性得到了提升。

6.2.2 模型对部分特殊情况欠缺考虑

（A）问题四中成分分析中潜在的双成分协同影响规律

在双成分系统的相关分析中，只考虑到了两个成分之间的相关关系。然而在化学物质反应中，因素的相关性远远不局限于双成分。例如在高钾玻璃和铅钡玻璃中但看碳酸锶、碳酸钾以及碳酸锶、碳酸钙成分组合，均不构成显著的相关关系。但如果联合碳酸锶和碳酸钾、碳酸钙两者，便可以发现其规律。碳酸锶与碳酸钾和碳酸钙之和具有负相关性。这也是因为碳酸锶属于碱性金属氧化物[7]，在配方中引入的主要目的是为了弥补钾钙含量和较小而造成的玻璃粘滞。所以，碳酸锶虽然与单一的碳酸钾、碳酸钙不成相关关系；但却**与碳酸钾-碳酸钙系统成相关关系**。

（B）玻璃颜色不同的原因

题中附件按颜色把玻璃划分为：浅蓝、浅绿、蓝绿、深绿、深蓝、紫色、黑色。其中除了浅蓝与蓝绿以外，其他颜色样本数量均比较小，分析不具有普世性。但此现象源于玻璃呈色的复杂性。故本文将从以下原因解释其样本分布并试图还原影响玻璃呈色的原因：

①每个金属元素都有其**光谱特征**，不同的金属氧化物会呈现不同的颜色。在熔制阶段，古人会在玻璃中会添加画碗石[8]这种含有氧化锰（使其着紫色）、氧化钴（蓝色）、氧化铬（绿色）等染色物质的矿石使其呈色。虽然着色效果良好，**但其中金属氧化物添加非常微弱**，很难在数据中作为共性的变量显示并加以研究。因此，只能通过数据中部分显色金属来还原其颜色规律。以氧化锡为例，氧化锡含量较高的普遍可以呈现更深的蓝色，如下表：

表 20 氧化锡、氧化铜降序表

颜色	类型	是否风化	氧化锡(Sn02)
深蓝	高钾	无风化	2.36
无	铅钡	风化	1.31
浅蓝	铅钡	风化	0.47
深蓝	铅钡	无风化	0.44
深蓝	铅钡	无风化	0.4

颜色	类型	是否风化	氧化铜(CuO)
紫	铅钡	风化	10.57
紫	铅钡	风化	10.41
紫	铅钡	无风化	8.46
浅蓝	铅钡	风化	5.35
蓝绿	高钾	无风化	5.09
浅蓝	铅钡	风化	4.93
浅蓝	铅钡	无风化	4.78
浅蓝	高钾	无风化	4.73
蓝绿	高钾	无风化	3.87
紫	铅钡	风化	3.6

②金属元素在高温下的反应会带来颜色的变化。有时，烧熔一次还不能使玻璃显示出色来，再要在**第二次加热的环境中**才能使玻璃显色。而高温熔制的环境在各个朝代甚至同时代的不同区域里大相径庭。因此，易受高温影响的显色金属成为玻璃呈色的极不稳定因素。

③同一金属化合物也**有多重染色机制**。例如玻璃中的铜，以高浓度铜析出时呈紫色。以高价氧化铜存在时，玻璃呈蓝绿色；而以低价的氧化亚铜存在时，玻璃不呈蓝绿色。如上表，氧化铜成分的降序颜色结果与此结论吻合。

④部分玻璃中的氧化物具有褪色作用。首先，部分像砒霜这样作为**脱色剂**、澄清剂的材料在熔制时就会加速玻璃的褪色性能。并且，普通的颜料会受阳光中**紫外线的照射**而褪色。

以上均为所给数据之外的**外生性变量**，很难通过本题提供的表单数据得到变化规律。

八、参考文献

[1]刁云超. 钠钙平板玻璃表面化学处理与特性研究[D].大连工业大学,2012.
[2]王洪成,刘启明.高压电场对钠钙硅酸盐玻璃风化性能的影响[J].武汉理工大学学报,2009,31(22):63-65+70.
[3]齐济,古丽斯坦,王承遇,宁桂玲.硅酸盐玻璃表面析碱的研究[J].玻璃与搪瓷,2006(03):9-13.
[4]王承遇,陶瑛.硅酸盐玻璃的风化[J].硅酸盐学报,2003(01):78-85.
[5]陶瑛,薄学微,王继红,王承遇.热处理对P_2O_5-B_2O_3-Al_2O_3-MgO-K_2O玻璃风化的影响[J].玻璃,1999(03):1-4
[6]刘元新,蒋荃.平板玻璃风化程度评价方法的选用[J].中国建材科技,1992(06):29-33.
[7]王承遇,黄明.二氧化硫气氛对浮法玻璃风化性能的影响[J].硅酸盐通报,1987(01):7-11.DOI:10.16552/j.cnki.issn1001-1625.1987.01.002.
[8]周良知.影响硅酸盐玻璃风化的主要因素[J].大连轻工业学院学报,1984(01):34-44.

九、附录

9.1 表单 2 文物亚类划分结果

文物编号	亚分类	文物编号	亚分类	文物编号	亚分类	文物编号	亚分类
01	高钾高钙	20	超铅钡	38	铅钡高铅	53 未风化点	铅钡高硅
02	铅钡高铅	21	高钾高硅	39	铅钡高铅	54	铅钡高铅
03 部位 1	超高钾	22	高钾高硅	40	铅钡高铅	54 严重风化点	铅钡高铅
03 部位 2	高钾高钙	23 未风化点	超铅钡	41	铅钡高铅	55	铅钡高硅
04	高钾高钙	24	超铅钡	42 未风化点 1	超铅钡	56	超铅钡
05	高钾高钙	25 未风化点	铅钡高硅	42 未风化点 2	超铅钡	57	超铅钡
06 部位 1	超高钾	26	超铅钡	43 部位 1	铅钡高铅	58	铅钡高铅
06 部位 2	高钾高钙	26 严重风化点	超铅钡	43 部位 2	铅钡高铅		
07	高钾高硅	27	高钾高硅	44 未风化点	铅钡高硅		
08	超铅钡	28 未风化点	铅钡高硅	45	超铅钡		
08 严重风化点	超铅钡	29 未风化点	铅钡高硅	46	超铅钡		
09	高钾高硅	30 部位 1	超铅钡	47	铅钡高硅		
10	高钾高硅	30 部位 2	超铅钡	48	铅钡高硅		
11	超铅钡	31	铅钡高硅	49	铅钡高铅		
12	高钾高硅	32	铅钡高硅	49 未风化点	铅钡高硅		
13	高钾高钙	33	铅钡高硅	50	超铅钡		
14	高钾高钙	34	超铅钡	50 未风化点	铅钡高硅		
16	高钾高钙	35	铅钡高硅	51 部位 1	铅钡高铅		
18	超高钾	36	超铅钡	51 部位 2	铅钡高铅		
19	铅钡高铅	37	超铅钡	52	铅钡高铅		

9.2 模型代码

问题 1 python 代码(移动平均预测）

```
def weighting_shifts(x,n,w,m):
    num=0
    sum=0
    for i in range(n):
        num=w[i]+num
        sum=w[i]*x[m-i-2]+sum
    y=sum/num
    return y
for i in range(6,16):
    list_y.append(weighting_shifts(y,5,w,i))
y=y[5:15]
def mean_shift(list_y,y):
```

```
    sum1=0
    sum2=0
    y=list(y)
    for i in range(len(list_y)):
        sum1=sum1+list_y[i]
        sum2=sum2+y[i]
    error_mean=(1-sum1/sum2)
    return error_mean
mean_shift(list_y,y)
#输入时间窗口
def get_error(x,n,w):
    y_error=[]
    for i in range(x.size-n):
        y=weighting_shift(x,n,w,n+i+1)
        y_error.append((x[n+i]-y)/x[n-1+i])
    return y_error
```

问题 2 python 代码（基于 SMOTE 采样法的数据倾斜处理）

```
class SMOTE():
    def _fit_(self,
                    K_neighbors = 5,
                    N_need = 3,
                    random_state = 42):
        self.K_neighbors = K_neighbors
        self.N_need = N_need
        self.random_state = 42
        def div_data(self, x_data, y_label):
        print('y_label',y_label)
        tp = set(y_label)
        print('tp',tp)
        #找出结果值出现次数较少的一类
        tp_less = [a for a in tp if sum(y_label == a) < sum(y_label != a)][0]
        print('tp_less',tp_less)
        data_less = x_data.iloc[y_label == tp_less,:]
        print('tp',tp)
        print('data_less',data_less)
        data_more = x_data.iloc[y_label != tp_less, :]
        tp.remove(tp_less)
        return data_less, data_more, tp_less, list(tp)[0]
        def get_SMOTE_sample(self, x_data, y_label):
        data_less, data_more, tp_less, tp_more = self.div_data(x_data, y_label)
        n_integ = self.N_need
        data_add = copy.deepcopy(data_less)
        print('data_less', data_less.shape)
```

```
if n_integ == 0 :
    print('WARNING: PLEASE RE-ENTER N_need')
else:
    for i in range(1,n_integ):   #扩充少数类的倍数
        data_less = data_less.append(data_add)
data_less.reset_index(inplace = True, drop = True)
print('data_out', data_less.shape)
return data_less, tp_less
def over_sample(self, x_data, y_label):
sample, tp_less = self.get_SMOTE_sample(x_data, y_label)
knn = NearestNeighbors(n_neighbors = self.K_neighbors ,n_jobs = -1).fit(sample)
n_atters = x_data.shape[1]
label_out = copy.deepcopy(y_label)
new = pd.DataFrame(columns = x_data.columns)
print('new', new)

for i in range(len(sample)): #选择一个正样本
    #选择少数类中最近的 K 个样本
    k_sample_index = knn.kneighbors(np.array(sample.iloc[i, :]).reshape(1, -1),
                                    n_neighbors = self.K_neighbors + 1,
                                    return_distance = False)
   print('k_sample_index',type(k_sample_index),k_sample_index)
    print('np.array(sample.iloc[i, :])
    .reshape(1,-1)', np.array(sample.iloc[i, :])
    .reshape(1, -1))
    # 计算插值样本
    np.random.seed(self.random_state)
    choice_all = k_sample_index.flatten()
     print('choice_all',choice_all)
     print('choice_all[choice_all != 0]',choice_all != 0)
    choosed = np.random.choice(choice_all[choice_all != 0])
    print('choosed',choosed)
    #在正样本和随机样本之间选出一个点
    diff = sample.iloc[choosed,] - sample.iloc[i,]
    print('diff',type(diff), diff)
    gap = np.random.rand(1, n_atters)
    print('gap', gap)
    print('sample.iloc[i,]', sample.iloc[i,])
    new.loc[i] = [x for x in sample.iloc[i,] + gap.flatten() * diff]
    print('new',new)
    label_out = np.r_[label_out, tp_less]   #新增数据添加 label 标签
print('new', new)
new_sample = pd.concat([x_data, new])
new_sample.reset_index(inplace = True, drop = True)
```

```
            return new_sample, label_out
from sklearn.neighbors import NearestNeighbors
import numpy as np
import pandas as pd
import copy
from sklearn.datasets import load_iris
#数据处理
iris= load_iris()
data = pd.DataFrame(data = iris.data, columns = iris.feature_names)
data['label'] = iris.target
data0 = data.loc[data['label'] == 0,]
data1 = data.loc[data['label'] == 1,][:10]
data = pd.concat([data0, data1],axis = 0)
data.reset_index(inplace=True,drop=True)
label = np.r_[label[label == 0], label[label == 1][:10]] #np.r_函数作用是将两个数组首尾相连
#使用类
smt = SMOTE() #实例化
new_sample, new_label = smt.over_sample(data, label)
```

问题 2 python 代码（方差过滤法筛选特征值）

```
import pandas as pd
data=pd.read_csv("D:/num1.csv",encoding='gbk')
data.head()
data=data.drop(['label'], axis=1)
data.shape
data.describe()
des1 = data.describe()
std1 = des1.loc['std'].sort_values(ascending=False)
X_16_1 = data[std1[std1>0.14].index]
X_16_1.shape
X_16_1.to_csv("D:/var.csv",index=True,encoding='gbk')
```

问题 2 python 代码（随机森林法筛选特征值）

```
from sklearn.ensemble
import RandomForestClassifier
import pandas as pd
data=pd.read_csv("D:h1.csv",encoding='gbk')
import matplotlib.pyplot as plt
from sklearn.feature_selection
import SelectFromModel data=pd.
data.head()
data.shape
X=data.drop(['label'], axis=1)
y=data['label']
```

```
from imblearn.over_sampling import ADASYN
X_resampled, y_resampled = ADASYN().fit_sample(X, y)
print('采样结果为:')
print(y_resampled.value_counts())
clf = RandomForestClassifier()#基于随机森林度量各个变量的重要性
clf = clf.fit(X_resampled, y_resampled)
importance1 = np.mean([tree.feature_importances_ for tree in clf.estimators_], axis=0) std1 =
np.std([tree.feature_importances_ for tree in clf.estimators_], axis=0)
indices = np.argsort(-importance1) #返回由大到小的数据之前的索引
choose_num = 10
suoyin = indices[0:choose_num][::-1]
range_ = range(choose_num)
columns=X.columns
feature_name = pd.Series(columns)
plt.figure(figsize=(12, 9))
plt.barh(range_, importance1[suoyin], color='r', xerr=std1[suoyin], alpha=1, align='center')
plt.xticks(fontsize=20)
plt.yticks(range(choose_num), feature_name[suoyin],fontsize=20)
plt.ylim([-1, choose_num])
```

问题 2 matlab 代码（模糊聚类）

```
X=[
16.71    70.21    6.69     1.77     0.68
26.25    61.03    7.22     1.16     0.61
12.41    59.85    7.29     0        0.64
17.11    58.46    0        14.13    1.12
22.28    55.46    7.04     4.24     0.88
21.35    51.34    0        8.75     0
32.93    49.31    9.79     0.48     0.41
36.28    47.43    0        3.57     0.19
25.74    47.42    8.64     5.71     0.44
35.78    46.55    10       0.34     0.22
25.42    45.1     17.3     0        0
21.7     44.75    3.26     12.83    0.47
18.46    44.12    9.76     7.46     0.47
17.98    44       14.2     6.34     0.66
29.64    42.82    5.35     8.83     0.19
39.57    41.61    10.83    0.07     0.22
29.15    41.25    15.45    2.54     0
24.61    40.24    8.94     8.1      0.39
30.39    39.35    7.66     8.99     0.24
34.34    39.22    10.29    0        0.35
36.93    37.74    10.35    1.41     0.48
28.79    34.18    6.1      11.1     0.46
```

```
49.01    32.92    7.95     0.35     0
4.61     32.45    30.62    7.56     0.53
50.61    31.9     6.65     0.19     0.2
45.02    30.61    6.22     6.34     0.23
3.72     29.92    35.45    6.04     0.62
19.79    29.53    32.25    3.13     0.45
31.94    29.14    26.23    0.14     0.91
20.14    28.68    31.23    3.59     0.37
51.54    25.4     9.23     0.1       0.85
33.59    25.39    14.61    9.38     0.37
55.21    25.25    10.06    0.2      0.43
54.61    23.02    4.19     4.32     0.3
65.91    22.05    5.68     0.42     0
51.26    21.88    10.47    0.08     0.35
51.33    20.12    10.88    0         0
69.71    19.76    4.88     0.17     0
60.12    17.24    10.34    1.46     0.31
68.08    17.14    4.04     1.04     0.12
53.79    16.98    11.86    0         0.33
65.91    16.55    3.42     1.62     0.3
75.51    16.16    3.55     0.13     0
61.28    15.99    10.96    0         0.23
53.33    15.71    7.31     1.1      0.25
63.66    13.66    8.99     0         0.27
60.74    13.61    5.22     0         0.26
63.3     12.31    2.03     0.41     0.25
37.36    9.3      23.55    5.75     0 ]';
Y=pdist(X);
SF=squareform(Y);
Z=linkage(Y,'average');
dendrogram(Z);
T=cluster(Z,'maxclust',3)
x1=randn(34,1);
x2=randn(14,1)+5;
x3=randn(1,1)+5;
x=[x1;x2;x3];
y=randn(49,1);
T=clusterdata([x,y],3)
temp1=find(T==1)
plot(x(temp1),y(temp1),'yd','markersize',10,'markerfacecolor','b')
hold on
temp1=find(T==2)
plot(x(temp1),y(temp1),'yd','markersize',10,'markerfacecolor','g')
hold on
```

```
temp1=find(T==3)
plot(x(temp1),y(temp1),'yd','markersize',10,'markerfacecolor','r')
legend('高钡低铅','高铅低钡','钡铅均衡')
```

问题 3 matlab 代码（初分类）

```
num = 8; %共八组数据待分类
while(num)
    dt=zeros(1,Features);
    J=0;
    for i=1:m1
        xx=X1(i,1:Features);
        yy=Y1(i,1);
        h=1/(1+exp(-(theta1 * xx')));
        dt=dt+(h-yy) * xx;
        end
J=J+ yy*log(h)+(1-yy)*log(1-h);%损失函数
J=-J/m1;%代价函数
L=[L,J];
theta2=theta1 - delta*dt/m1;%更新 theta theta1=theta2;
num = num - 1;
if J<0.01
break;
end
end
```

问题 2，3 matlab 代码（混淆矩阵合理性检验）

```
function draw_cm(mat,tick)
mat=[ 3 0 0 0 0 0
      0 7 1 0 0 0
      0 1 8 0 1 0
      0 0 2 10 0 0
      0 0 0 1 13 2
      0 0 0 0 3 21]; %基于亚分类的六类划分偏差参数矩阵
imagesc(mat);
colormap(flipud(gray));
num_class=size(mat,1);
 set(gca,'xtick',1:6)
set(gca,'xticklabel',{'K-K','K-Ca','K-Si','PbBa-PbBa','PbBa-Pb','PbBa-Si'},'XTickLabelRotation',1
)
set(gca,'ytick',1:6)
set(gca,'yticklabel',{'K-K','K-Ca','K-Si','PbBa-PbBa','PbBa-Pb','PbBa-Si'})
set(gca,'FontSize',14,'Fontname', 'Times New Roman');
colorbar
textStrings = num2str(mat(:), '%0.2f' );
```

```
textStrings = strtrim(cellstr(textStrings));
[x,y] = meshgrid(1:num_class);
hStrings = text(x(:),y(:),textStrings(:), 'HorizontalAlignment' , 'center' );
midValue = mean(get(gca, 'CLim' ));
textColors = repmat(mat(:) > midValue,1,3);
set(hStrings,{ 'Color' },num2cell(textColors,2));
set(gca, 'xticklabel' ,tick, 'XAxisLocation' );
rotateXLabels(gca, 1 );
set(gca, 'yticklabel' ,tick);
```

问题 4 matlab 代码（斯皮尔曼相关系数-热力图）

```
X=[45.02 0 0 3.12 0.54 4.16 0 0.7 30.61 6.22 6.34 0.23
0 0
54.61 0 0.3 2.08 1.2 6.5 1.27 0.45 23.02 4.19 4.32 0.3 0 0
65.91 0 0 1.6 0.89 3.11 4.59 0.44 16.55 3.42 1.62 0.3 0
0
53.33 0.8 0.32 2.82 1.54 13.65 1.03 0 15.71 7.31 1.1 0.25
1.31 0
68.08 0 0.26 1.34 1 4.7 0.41 0.33 17.14 4.04 1.04 0.12
0.23 0
65.91 0 0 0.38 0 1.44 0.17 0.16 22.05 5.68 0.42 0 0
0
63.3 0.92 0.3 2.98 1.49 14.34 0.81 0.74 12.31 2.03 0.41
0.25 0 0
49.01 2.71 0 1.13 0 1.45 0 0.86 32.92 7.95 0.35 0 0
0
50.61 2.31 0 0.63 0 1.9 1.55 1.12 31.9 6.65 0.19 0.2 0
0
69.71 0 0.21 0.46 0 2.36 1 0.11 19.76 4.88 0.17 0 0
0
75.51 0 0.15 0.64 1 2.35 0 0.47 16.16 3.55 0.13 0 0
0
51.54 4.66 0.29 0.87 0.61 3.06 0 0.65 25.4 9.23 0.1
0.85 0 0
63.66 3.04 0.11 0.78 1.14 6.06 0 0.54 13.66 8.99 0
0.27 0 0
60.74 3.06 0.2 2.14 0 12.69 0.77 0.43 13.61 5.22 0 0.26
0 0]
R=corr(X, 'type', 'Spearman') %求斯皮尔曼相关系数矩阵
xvalues = {'二氧化硅(SiO2)','氧化钠(Na2O)','氧化钾(K2O)','氧化钙(CaO)','氧化镁(MgO)','氧化
铝(Al2O3)','氧化铁(Fe2O3)','氧化铜(CuO)','氧化铅(PbO) ','氧化钡(BaO)','五氧化二磷(P2O5)','
氧化锶(SrO)','氧化锡(SnO2)','二氧化硫(SO2)'};
 yvalues = {'二氧化硫(SO2)','氧化锡(SnO2)','氧化锶(SrO)','五氧化二磷(P2O5)','氧化钡(BaO)','
氧化铅(PbO)','氧化铜(CuO)','氧化铁(Fe2O3)','氧化铝(Al2O3)','氧化镁(MgO)','氧化钙(CaO)','
```

```
氧化钾(K2O)','氧化钠(Na2O)','二氧化硅(SiO2)'};
cdata = R
h = heatmap(xvalues, yvalues, cdata);
```