玻璃文物的多组分分析与鉴别研究

摘要

本文对完成预处理后的样本数据运用 Spearman 相关系数，卡方检验等方法对玻璃制品的数据进行挖掘，运用 k-means 聚类，Logistic 回归等方法对玻璃制品进行鉴别和分类，通过 Pearson 相关系数，关联规则挖掘化学成分的关联。

问题 1 中，针对第一个问题，由于表单 1 中的数据为分类变量，所以本文首先使用**卡方检验**分析玻璃文物属性和表面风化情况间的联系，发现玻璃文物的类型与其表面风化情况有较强相关关系，而纹饰和颜色与风化情况关系较弱。针对第二个问题，由于化学成分含量为连续性定量数据，而风化情况为二分类数据，因此我们使用 **Spearman 相关系数**分别挖掘高钾和铅钡两种类型的玻璃与风化相关化学成分的统计规律。针对第三个问题，我们利用风化前后各化学元素的样本均值，构建了风化衰减比例指标，建立了基于**风化衰减比例**的预测模型，较好地预测了风化点在风化前的化学成分含量，结果见附录。

问题 2 中，针对第一个问题，使用 Spearman 相关系数分析得到玻璃类型分类的高依赖性化学成分，在此基础上总结了两类玻璃的分类规律。针对第二个问题，本文采用 **k-means 聚类**，将每个大类中的样本划分成 4 个亚类，随后我们使用十折划分，获得 9 份的数据集与 1 份测试集，通过对比测试集加入前后轮廓系数的改变对聚类模型进行了效果评估与样本敏感性分析。

针对问题 3，我们首先建立 **Logistic 二分类模型**，将未风化样本初步分为高钾与铅钡玻璃两个类型，随后使用问题 1 中基于风化衰减比例的预测模型，将待分类的样本中的部分风化样本化学成分含量还原为未风化时的含量，最后通过计算待分类样本与问题 2 中的聚类中心的距离进一步划分入亚类，分类结果见正文。待分类样本分类完成后，我们对样本数据增加-5%～5%的随机扰动，得到测试样本，使用该模型再次进行分类，与原样本的分类结果完全一致，完成了该分类模型的敏感性检验，该模型具有较好的鲁棒性。

针对问题 4，我们构造了 Pearson 相关系数矩阵和关联规则，得到了两种玻璃中相关性较高的化学成分（详见正文），结合文献分析并解释了同一类别内化学成分的相关性和不同类别间化学成分关联关系的差异性。

关键词：Spearman 相关系数 风化衰减比例 k-means 聚类 Logistic 二分类模型

一、 问题重述

古代玻璃经历埋藏时的风化过程后，其内部元素与环境元素发生大量交换，引起成分比例改变，进而为其分类带来困难。

现有一批已经分为高钾玻璃和铅钡玻璃两类的古代玻璃制品，表单 1 为分类数据，表单 2 为成分比例数据，由于检测方法的误差，存在成分比例累加和不为100%的情况，故将有效数据的累加和范围设定为 85%～105%。

问题一：分析玻璃文物类型、纹饰和颜色与其表面风化间的关系。

问题二：求高钾玻璃与铅钡玻璃分类规律；对它们进行亚类划分，分析结果的合理性与敏感性，并具体阐述分类方法。

问题三：给表单 3 中未知类别的文物进行分类，并进行敏感性分析。

问题四：分析同一类别玻璃文物化学成分间的关联性，比较不同类别间化学成分的差异性。

二、问题假设

1、忽略表面无风化文物可能存在的局部轻微风化

2、忽略风化程度的差异，仅对讨论是否风化

3、假设未分类玻璃只属于铅钡玻璃与高钾玻璃中的一种

4、假设化学成分在玻璃文物分类中其最主要作用

三、符号说明

符号	含义
χ^2	Pearson 卡方值
d_i	Spearman 相关性检验中的秩次差
ρ_s	Spearman 相关系数
σ	风化衰减比例
a_i	风化强相关成分风化点检测值
a_i^*	风化强相关成分风化前含量预测值

b_i	风化强相关成分风化点检测值
b_i^*	风化强相关成分风化前含量预测值
$e(i)$	样本内聚度
$g(i)$	样本分离度
$s(i)$	样本轮廓系数
S	类簇轮廓系数
ρ_p	Pearson 相关系数

四、模型的建立和求解

数据预处理

（1）异常值处理：原始数据中 15 和 17 号的成分比例累加和超出有效数据范围，我们将这两个样本去除；

（2）缺失值处理：原始数据中部分化学成分含量缺失，我们将缺失值用 0 赋值。

4.1 问题一

4.1.1 表面风化与样品属性的关联性分析

4.1.1.1 Pearson 卡方指标

卡方检验是一种适用于分类变量的差异性检验方法，能够对二分类无序变量进行检验

Step1 将随机抽样得到的二分类无序变量 X、Y 的频数用列联表表示：

表 1 2×2列联表

	属性 1	属性 2	总计
分组 1	a	b	$a+b$
分组 2	c	d	$c+d$
总计	$a+c$	$b+d$	$a+b+c+d$

Step2 提出原假设和备择假设：

$$H_0 = X与Y无关$$
$$H_1 = X与Y有关$$

Step3 计算χ^2

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \tag{1}$$

将χ^2与设定的显著性水平α下的临界值χ_α^2进行比较，若$\chi_{pearson}^2 < \chi_\alpha^2$，则接受原假设；否则拒绝原假设，接受备择假设。

4.1.1.2 卡方独立性检验

我们设定显著性水平阈值为0.05，对两种玻璃类型的风化情况进行卡方检验。首先用2×2列联表表示不同类型的玻璃的风化情况：

表2 不同类型玻璃风化情况列联表

	风化	无风化	
高钾	6	12	18
铅钡	28	12	40
总计	34	24	58

提出原假设和备择假设：

$$H_0 = 风化情况与玻璃类型无关$$
$$H_1 = 风化情况与玻璃类型有关$$

使用SPSS软件进行卡方检验，得到结果见下表：

表3 不同类型风化情况卡方检验结果

	值	自由度	渐进显著性（双侧）	精确显著性（双侧）	精确显著性（单侧）
皮尔逊卡方	6.880[a]	1	0.009		
连续性修正[b]	5.452	1	0.020		
似然比	6.889	1	0.009		
费希尔精确检验				0.011	0.010
有效个案数	58				

a. 0 个单元格 (0.0%) 的期望计数小于 5。最小期望计数为 7.45。

b.仅针对 2x2 表进行计算

由表中可知，χ^2为表中的皮尔逊卡方值 6.880，由于最小期望计数为 7.45，P值采用渐进显著性0.009,，因此在99%的置信度下，拒绝原假设，接受备择假设，认为不同的玻璃类型与风化情况有关。

同理，对不同纹饰、不同颜色的风化情况进行卡方检验，得到结果如表

表 4 不同纹饰风化情况卡方检验

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	4.957[a]	2	0.084
似然比	7.120	2	0.028
有效个案数	58		

a. 2 个单元格 (33.3%) 的期望计数小于 5。最小期望计数为 2.48。

表 5 不同颜色风化情况卡方检验

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	9.432[a]	8	0.307
似然比	12.636	8	0.125
有效个案数	58		

a. 14个单元格(77.8%)的期望计数小于5。最小期望计为 .41。

分析表中数据可以得到样品的纹饰和颜色与风化情况无关的显著性水平分别为 P=0.084 和 P=0.307，均大于 0.05，不能拒绝原假设，因此认为纹饰和颜色与风化无关。

综上，我们认为玻璃样品的类型与其风化情况有较强的相关性，纹饰与颜色与其风化情况不存在明显相关性。

4.1.1.3 样本数据直观分析

本文分别对风化样品与未风化样品的数据进行分析，发现是否风化与玻璃类型和纹饰的相关性较显著，而颜色关系相关性较弱。

由表中数据可以发现，铅钡玻璃样品中风化样品比例占 70%，而高钾玻璃样品中风化样品仅占 24%。铅钡玻璃的风化样品比例远高于高钾玻璃，一定程度上说明铅钡玻璃较高钾玻璃更易被风化。

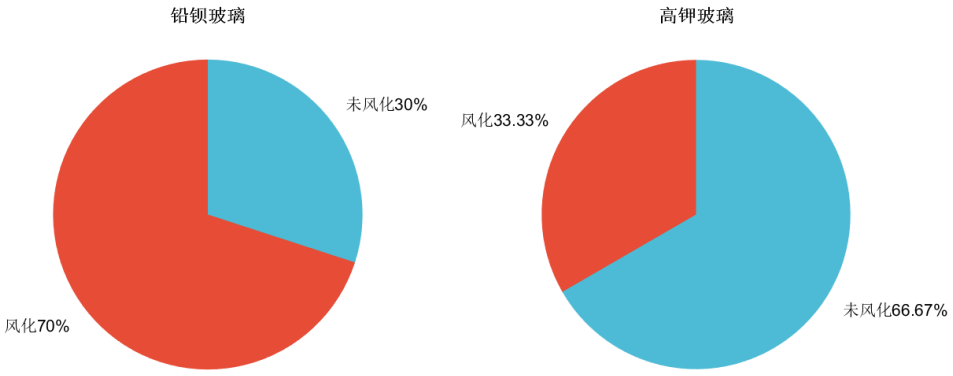


图 1 风化样品占比对比

4.1.2 化学成分含量统计规律

4.1.2.1 Spearman 相关系数

Spearman 相关系数是利用数据的排序来衡量变量相关性的方法，能够体现定序变量与连续性定量变量之间的相关关系，具体计算公式为

Spearman 相关系数是利用数据的排序来衡量变量相关性的方法，能够体现定序变量与连续性定量变量之间的相关关系。首先求出待分析的连续型变量数据的秩次，随后求出两个变量对应数据的秩次差：

$$d_i = d_i^x - d_i^y \tag{2}$$

代入公式计算相关系数，其中 n 为样本数：

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3}$$

本文将化学成分含量转化为秩次，计算与风化情况的 Spearman 相关系数。

4.1.2.2 风化相关成分

由于各化学成分在不同类型玻璃中含量差异较大且与样品风化情况的相关性不一，本文使用 Spearman 相关系数分别对高钾玻璃和铅钡玻璃的化学成分进行筛选，选择相关系数 $\rho_s > 0.5$ 的化学成分作为与玻璃风化情况相关程度高的化学成分，如下表所示：

表 6 高钾玻璃风化相关成分

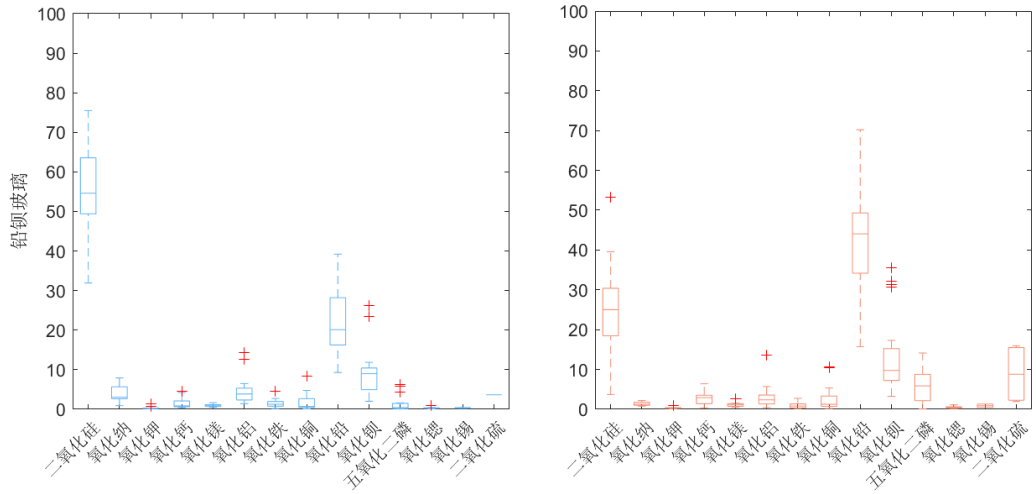
化学成分	Spearman 相关系数
二氧化硅	-0.8349
氧化纳	-0.6667
氧化铅	0.6938
氧化锡	0.8660
二氧化硫	0.6324

表 7 铅钡玻璃风化相关成分

化学成分	Spearman 相关系数
二氧化硅	0.8177
氧化钾	-0.7676
氧化钙	-0.8401
氧化铝	0.7950
氧化铁	-0.8401

4.1.2.3 正态性检验

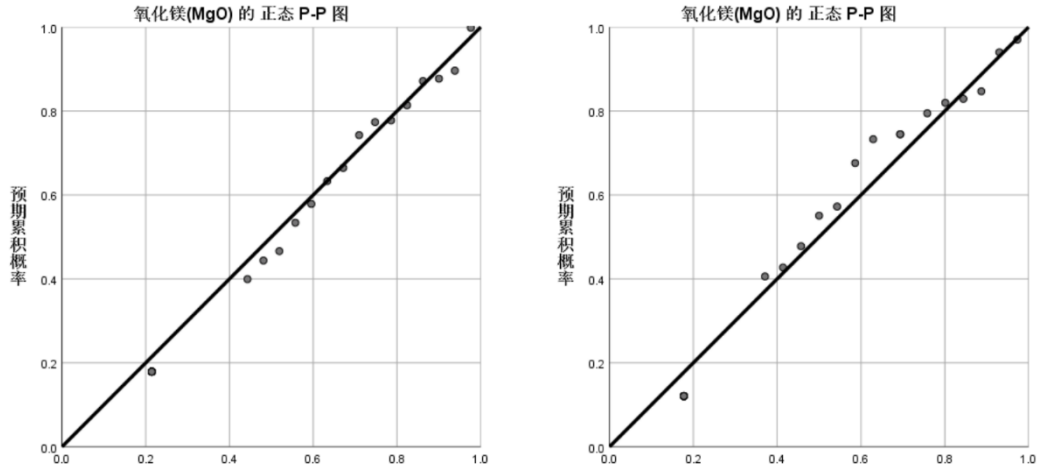
为探究玻璃文物表面风化情况对化学成分含量的分布情况的影响，我们首先作出风化前后各化学成分的箱形图，以铅钡玻璃为例：



a)未风化的样本各化合物数据箱形图　　b)风化的样本各化合物数据箱形图

图 2 铅钡玻璃风化前后各化学成分的箱形图

对风化前后的化学成分含量进行正态性检验，以铅钡玻璃中 MgO 为例：



(a)风化前　　(b)风化后

图 3 不同风化情况下 MgO 分布情况对比图

上述检验结果表明，风化前后的化学成分含量的正态性检验结果相同，说明表面风化情况不会影响的化学成分含量的分布情况。

4.1.3 化学含量预测模型

4.1.3.1 风化衰减比例模型

为预测风化样品在风化前的化学成分，我们构造了风化衰减比例σ。

对于风化相关成分，我们计算每种成分在风化后与风化前平均值的比值，得到风化衰减比例σ：

$$\sigma = \frac{average(value_{风化前})}{average(value_{风化后})} \tag{4}$$

随后根据风化衰减比例与风化点检测值 a_i 作比，得到该点风化前的化学成分含量 a_i^*：

$$a_i^* = \sigma_a \cdot a_i \tag{5}$$

对于风化弱相关成分，首先计算的成分的权重：

$$w_i = \frac{b_i}{\sum_{i=1}^{n} b_i} \tag{6}$$

$$b_i^* = w_i \left(1 - \sum_{j=1}^{n} a_j^*\right) \tag{7}$$

分配完风化强相关成分后，以成分比例和为 100%为准，根据各风化弱相关成分的权重分配比例。

根据上述方法我们得到了 32 个风化检测点的化学成分预测值，下表展示部分风化检测点的二氧化硅预测值（详见附录）：

表 8 风化点二氧化硅原含量预测值

风化检测点	预测值	风化监测点	预测值
02	79.5999	36	86.8183
08	44.1880	38	72.2498
08 严重风化点	10.1145	39	57.5936
11	73.6979	40	36.6624
19	65.0314	41	40.5020
26	43.4201	43 部位 1	27.2280
26 严重风化点	8.1618	43 部位 2	47.6107
34	78.5029	48	117.0084

4.2 问题二：亚类划分

4.2.1 分类规律分析

4.2.1.1 未风化样本规律分析

本文使用 Spearman 相关系数筛选出与高钾、铅钡玻璃分类相关度高的化学成分，得到结果见下表

表 9 类别相关化学成分

化学成分	Spearman 相关系数
氧化钾	0.7154
氧化钙	0.5311
氧化铅	-0.8236
氧化钡	-0.8213
氧化锶	-0.5080

4.2.1.2 风化样本规律分析

从数据中我们发现高钾玻璃和铅钡玻璃风化样本中二氧化硅的变化趋势相差较大，对两类型玻璃的风化样本的二氧化硅含量进行独立样本 T 检验：

表 10 风化样本独立性 T 检验

		莱文方差等同性检验		平均值等同性 t 检验						
		F	显著性	t	自由度	Sig.（双尾）	平均值差值	标准误差差值	差值 95% 置信区间 下限	差值 95% 置信区间 上限
二氧化硅 (SiO2)	假定等方差	3.482	.071	8.305	33	.000	64.61000	7.77961	48.78227	80.43773
	不假定等方差			19.956	31.977	.000	64.61000	3.23760	58.01504	71.20496

平均值等同性 t 检验中 $p<0.05$,拒绝原假设，两类型玻璃风化样本的均值存在显著性差异。

进一步计算两类型玻璃的风化样本均值得到右图,由图中可以发现铅钡玻璃的风化样本的二氧化硅含量均值为 24.9127 远低于高钾玻璃的二氧化硅含量均值 93.9633，因此可以使用该指标作为类别划分的依据。

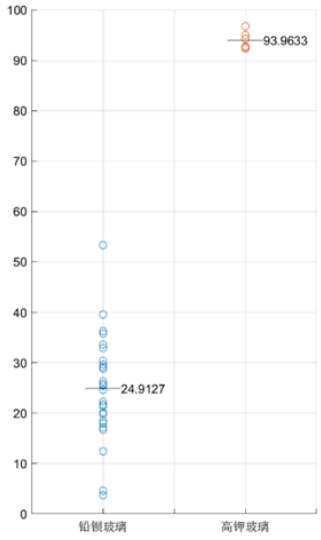


图 4 风化样本 SiO_2 硅均值对比

4.2.2 亚类划分

4.2.2.1 聚类指标筛选

由于化学成分较多，且不是所有化学成分都对亚类划分起作用，因此我们使用 Spearman 相关分析筛选出相关性较强的 5 个化学成分分别为：二氧化硅，氧化钾，氧化铅，氧化钡，氧化锶。下文我们根据这位五项化学成分建立 k-means 聚类模型。

4.2.2.2 k-means 聚类

k-means 聚类是一种常用的动态聚类算法，该算法通过计算元素间的距离，不断更新聚类中心以提升聚类效果。本文中具体操作步骤如下：

Step1 随机生成 K 个聚类中心

Step2 计算样本点与聚类中心的距离，将样本点划分到距离最近的类簇

Step3 根据当前类簇，更新聚类中心

Step4 判断新的聚类中心是否发生改变，若改变则回到 Step2,否则输出聚类结果

4.2.2.3 轮廓系数

轮廓系数是通过计算聚类的类簇中样本点的内聚度和类簇间样本点的分离度来表示聚类效果的指标。其中，内聚度 $e(i)$ 表示一个样本点 i 与类簇中其他样本点 j 之间的距离：

$$e(i)=\frac{1}{n-1}\sum_{j\neq i}^{n}d_{ij} \tag{8}$$

分离度则为一个样本点 i 与邻居类簇样本点 u 间的距离，其中存在一非己类簇使得该样本点到这一类簇内所有样本点平均距离最短，该类簇即为邻居类簇。分离度计算公式为：

$$g(i)=\frac{1}{n}\sum_{u=1}^{n}d_{iu} \tag{9}$$

公式（7）（8）中的 n 为类簇中样本点的个数。

单一样本 i 的轮廓系数为：

$$s(i)=\frac{g(i)-e(i)}{\max\{e(i),g(i)\}} \tag{10}$$

一个类簇总体的轮廓系数为：

$$S=\frac{1}{N}\sum_{i=1}^{N}s(i) \tag{11}$$

其中 N 为类簇内样本总数，观察公式不难发现，轮廓系数的取值范围为[-1，1]，数值越接近 1 表明聚类的效果越好。

4.2.2.4 k-means 聚类模型求解

将风化样本的化学成分含量还原为风化前含量后，根据筛选出的五个化学成分，计算样本点间的欧式距离，在铅钡玻璃和高钾玻璃两个原有分类中划分亚类，并通过计算类簇的轮廓系数评判聚类的效果。

使用 MATLAB 编程求解，发现当聚类中心数量为 4 时聚类的效果最好，并得到此时的轮廓系数：

表 11 类簇轮廓系数

	类簇 1	类簇 2	类簇 3	类簇 4
高钾玻璃	0.7130	0.6370	1	1
铅钡玻璃	0.5346	0.6792	0.6673	0.6114

得到亚类如下表所示：

表 12 高钾玻璃分类结果

高钾玻璃			
类簇 1	类簇 2	类簇 3	类簇 4
06 部位 1	01	03 部位 1	21
18	03 部位 2		
	04		
	05		
	06 部位 2		
	13		
	14		
	16		

表 13 铅钡玻璃分类结果

铅钡玻璃			
类簇 1	类簇 2	类簇 3	类簇 4
25 未风化点	20	23 未风化点	24
30 部位 1		28 未风化点	
30 部位 2		29 未风化点	
50 未风化点		31	
		32	
		33	
		35	
		37	
		42 未风化 1	
		42 未风化 2	
		44 未风化点	
		45	
		46	
		47	
		49 未风化点	

4.2.2.5 基于十折划分的样本敏感性分析

本文将参与聚类的数据集均等划分为 10 份，其中 9 份作为训练集，1 份作为验证集。使用训练集得到聚类中心后，加入验证集，计算聚类中心的变化距离。

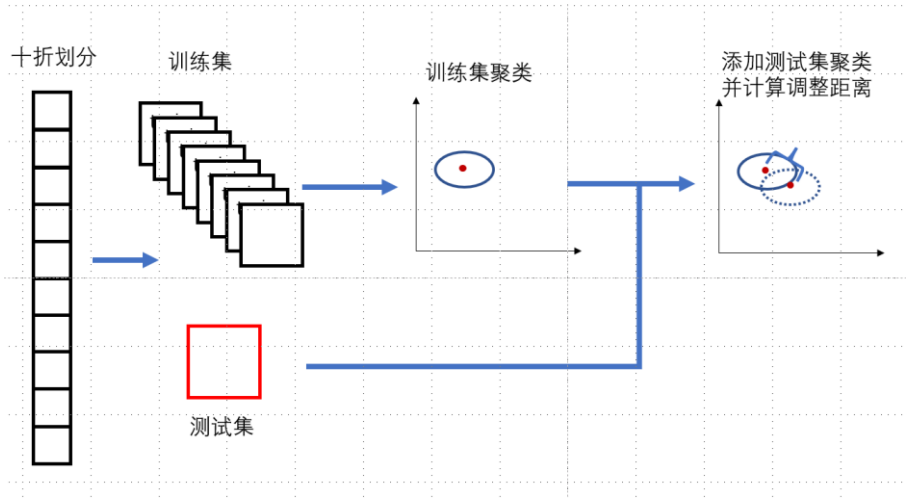


图 5 敏感性分析流程图

使用 MATLAB 软件计算得到训练集类簇的的轮廓系数，高钾玻璃为 0.9429，铅钡玻璃为 0.9156。将测试集加入后，重新计算得到高钾玻璃轮廓系数为 0.9529，

铅钡玻璃的轮廓系数为 0.9051。可以发现轮廓系数的变化较小，因此我们认为该聚类模型的对于新数据的鲁棒性较好。

4.3 问题三：玻璃文物类型鉴别

4.3.1 二分类模型

4.3.1.1 Logistic 二分类模型

Logistic 函数在当自变量 $x>0$ 时，因变量 y 快速趋近于 1，自变量 $x<0$ 时，因变量 y 快速趋近于 0，因此可以利用这种性质解决连续型变量的二分类问题。

利用 Logistic 函数解决二分类首先需要保证因变量 y 符合二项分布，在此基础上可以写出 Logistic 初始模型：

$$\log_e(\frac{\pi}{1-\pi}) = \omega_0 + \sum_{i=1}^{n} \omega_i X_i \tag{12}$$

其中 X_i 为样本自变量矩阵，n 为矩阵的维度即自变量的个数。

类别的划分显然属于二项分布，我们使用两种类型的未风化样本拟合出 Logistic 模型对待分类样本中的未风化样本进行类别划分，得到类别分类为：

表 14 未风化样本类别划分结果

编号	A1	A3	A4	A8
类别	高钾	铅钡	铅钡	铅钡

4.3.1.2 风化样本二分类

我们使用问题 2 中得到的风化样本规律，定义高钾玻璃和铅钡玻璃风化样本的 SiO_2 标准点为它们的均值，分别记为 $O_k = 93.9633$，$O_{Pb} = 24.9172$。

计算待分类风化样本的 SiO_2 到两 SiO_2 标准点的差值：

$$\begin{aligned} D_i^k &= |I_i - O_k| \\ D_i^{Pb} &= |I_i - O_{Pb}| \end{aligned} \tag{13}$$

表 15 风化样本类别划分结果

编号	A2	A5	A6	A7
类别	铅钡	高钾	高钾	高钾

4.3.2 亚类划分模型

由于未分类数据中存在部分风化数据，而风化数据的数据特征不明显，不利于进行亚类划分，因此我们使用问题二中的风化衰减比例模型，还原出风化前各化学成分的含量，下表展示还原得到的二氧化硅的含量，其余成分详见附录：

表 16 风化前二氧化硅含量预测

编号	A1	A2	A3	A4	A5	A6	A7	A8
二氧化硅	78.45	27.31	31.95	35.47	46.51	67.41	65.71	51.12

在二分类的基础上，使用问题二中的聚类模型，通过计算待分类样本与聚类中心间的距离，再次对待分类模型进行亚类划分，得到的亚类划分见下表：

表 17 亚类划分结果

编号	A1	A2	A3	A4	A5	A6	A7	A8
亚类	高钾 1	铅钡 1	铅钡 1	铅钡 1	高钾 4	高钾 4	高钾 4	铅钡 1

4.3.3 分类模型敏感性分析

为检验分类模型的敏感性，我们对 8 个样本的各化学成分施加-5%～5%的随机扰动，生成得到新的 8 个测试样本标记为 B1～B8，下表仅展示 SiO_2（其余数据详见附录）：

表 18 SiO_2 随机扰动后

编号	A1	A2	A3	A4	A5	A6	A7	A8
亚类	78.61	84.60	32.09	35.85	46.61	67.80	65.88	51.56

再次代入分类模型进行分类，得到分类结果与原分类结果完全相同，证明本模型的对小幅度随机波动并不敏感，鲁棒性较好。

表 19 敏感性分析分类结果

编号	A1	A2	A3	A4	A5	A6	A7	A8
亚类	高钾 1	铅钡 1	铅钡 1	铅钡 1	高钾 4	高钾 4	高钾 4	铅钡 1

4.4 问题四 化学成分关联关系分析

4.4.1 Pearson 相关系数矩阵

Pearson 相关系数是衡量连续性定距变量之间线性相关性的常用指标，也适用于本题中化学成分相关关系的计算，具体计算公式如下：

$$\rho_p = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y - \overline{Y})^2}} \tag{14}$$

计算得到 14 种化学成分含量之间的 Pearson 相关系数矩阵：

图 6 高钾玻璃 Pearson 相关系数矩阵

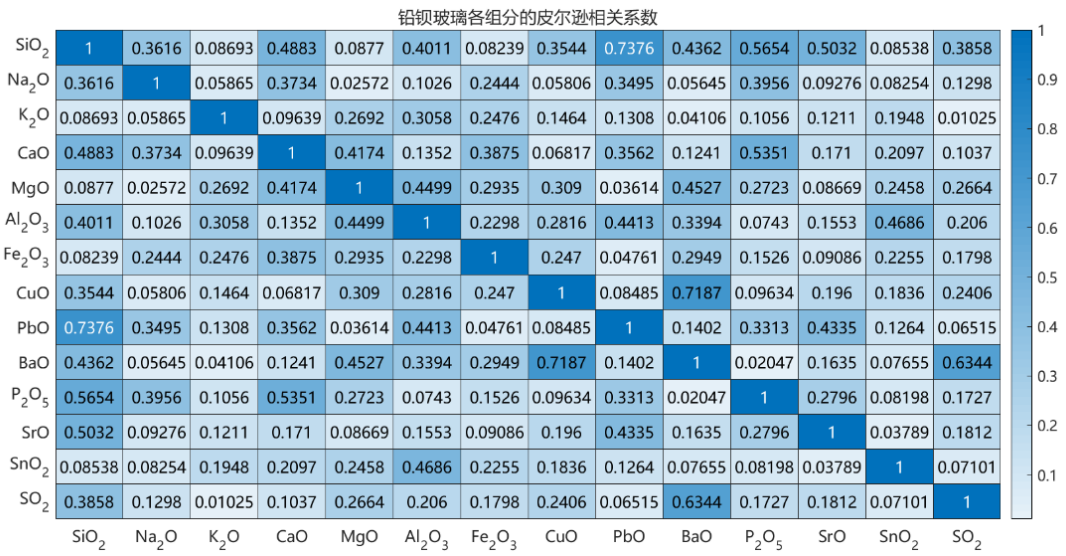


图 7 铅钡玻璃 Pearson 相关系数矩阵

分析图中数据可以发现在铅钡玻璃中，SiO_2 与 PbO，BaO 与 SO_2 和 CuO，三组化学成分的相关性较高；在高钾玻璃中，SiO_2 与 K_2O、CaO、Al_2O_3，P_2O_5 与 SrO、Fe_2O_3，MgO 和 SrO 的相关性比较高。

比较两类玻璃化学成分关联性发现，与二氧化硅关联程度高的化学成分有较大差异，二氧化硅在铅钡玻璃中主要与氧化铅存在相关关系，而在高钾玻璃中主要与氧化钾、氧化钙存在相关关系。该差异性符合两种玻璃由于制作工艺不同导致的差异。

4.4.2 关联规则

关联规则分析属于无监督算法的一种，通过查找存在于事物集合之间的频繁度、相关性或因果结构。反映一个事物与其他事物之间的相互依存性和关联性。

关联规则算法主要有频繁项集、支持度、置信度、提升度等指标，各个指标定义如下：

(1) 频繁项集

对于一个数据表而言，每个字段的不同取值称为“项”，而该数据表内各项的任意组合称为项集，支持度大于最小支持度的项集称为频繁项集。

(2) 支持度 sup()

支持度 $\sup(A \to B)$ 表示 A 与 B 同时出现的概率，公式为：

$$\sup(A \to B) = P(A \cup B) = \frac{count(A \cup B)}{N} \tag{15}$$

count 表示计数，N 表示所有事务个数。

如果 A 与 B 同时出现的概率小，说明 A 与 B 的关系不大；如果 A 与 B 同时出现的非常频繁，则说明 A 与 B 总是相关的。

(3) 置信度 *confidence*()

置信度 $confidence(A \to B)$ 表示在 A 发生的条件下，B 发生的概率，公式为：

$$confidence(A \to B) = P(B \mid A) = \frac{\sup(A \cup B)}{\sup(A)} \tag{16}$$

置信度揭示了 A 出现时，B 是否也会出现或有多大概率出现。如果置信度为 100%，则说明 A 和 B 必然存在联系。如果置信度太低，则说明 A 的出现与 B 是

否出现关系不大。

(4) 提升度 *lift*()

提升度用于度量规则是否可用的指标，描述的是相对于不用规则，使用规则可以提高多少，有用的规则的提升度大于 1，提升度的公式为：

$$lift(A \to B) = \frac{confidence(A \to B)}{\sup(B)} \tag{17}$$

当要寻找多个指标间的关联关系时，指标间的组合会达到指数级别，因此需要一种算法来减少搜索空间的大小以及扫描数据集的次数。

Apriori 算法是经典的挖掘频繁项集的算法，其核心思想为：

① 某个项集是频繁的，那么它的所有子集也是频繁的

② 如果一个项集是非频繁项集，那么它的所有超集也是非频繁项集

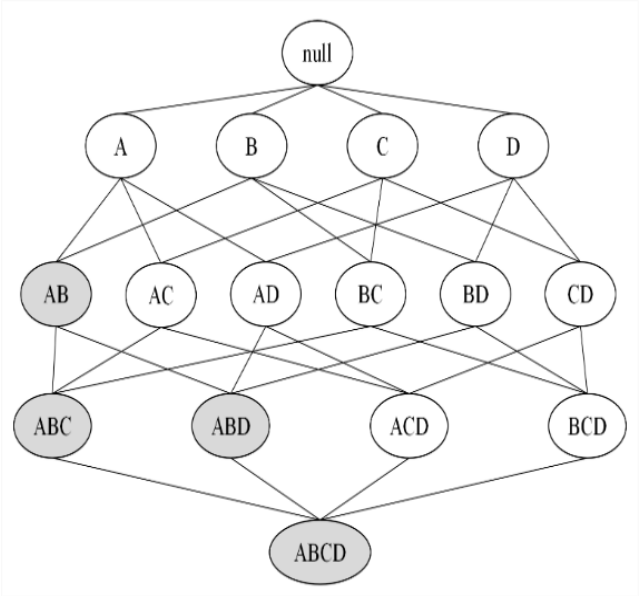


图 8 Apriori 算法示意图

当 AB 的支持度小于最小支持度时，即为非频繁项集，那么它的超集 ABC，ABD，ABCD 均为非频繁项集，这些就可以忽略不计，大大简化计算。

关联规则在挖掘单个样本对多个样本、多个样本对多个样本之间的关系有良好的性能，因此本题采用关联关系进一步挖掘化学成分间的关联性。

4.4.3 关联规则求解结果

表 20 关联规则结果

高钾玻璃		铅钡玻璃	
规则	置信度(%)	规则	置信度(%)
K_2O, Al_2O_3 =>SiO_2	100	SiO_2, Al_2O_3 =>PbO, BaO	100
Al_2O_3, P_2O_5 =>SiO_2	100	SiO_2, Al_2O_3, CuO=>PbO, BaO	100
SiO_2, CaO=>Al_2O_3, CuO	100	SiO_2, CaO, Al_2O_3=>PbO, BaO	100
SiO_2, Al_2O_3=>Fe_2O_3, CuO	100	Al_2O_3, PbO=> SiO_2, BaO	100

通过查阅文献我们可以对以上关联规则做出解释：对于高钾玻璃，古代炼制高钾玻璃的基本原料为石灰石，草木灰，石英，分别含有 Ca，K，Si，而石灰石中含有杂质 SiO_2, Al_2O_3，二者化合成硅酸铝（粘土、长石、云母），因此每块玻璃的成分均含有 CaO, K_2O, SiO_2, Al_2O_3，所以 K_2O，Al_2O_3，SiO_2 均呈强相关性。因为草木灰中主要成分为 K，次要成分中占比最大的为 P，因此每块玻璃均含有 SiO_2，K_2O，Al_2O_3，P_2O_5，四者均呈强相关性。

对于铅钡玻璃，古代炼制铅钡玻璃的基本原料为石灰石，石英，铅矿石，分别含有 Ca，Al，Si，Pb，铅矿石通常与重晶石（主要成分为 $BaSO_4$）形成共生矿，因而原料中也含有 Ba，同时，铅矿石的杂质中含有 Cu，所以 SiO_2，CaO，Al_2O_3，PbO，BaO，CuO 均呈强相关性。

五、模型的评价和推广

5.1 模型优点

合理使用 Pearson 相关系数和 Spearman 相关系数，较好地处理了定距变量之间，定距变量与定类变量之间的相关性问题

使用 Logistic 回归，以较高的分类精度解决了问题 3 中连续变量二分类的问题，模型简单，直接使用拟合得到的权重对各指标定权，避免了人工筛选分类指标的困难。

5.2 模型不足

在分析玻璃类型的分类规律时，由于数据较少，仅能得到类型与化学成分相关性规律，难以确定两种玻璃类型的准确数值划分标准。

5.3 模型改进与推广

对于问题 2、3 中的二分类问题，可以使用 SVM 等机器学习算法，进一步挖掘化学含量间的关系，提高分类精度，改进后模型将对多维二分类问题有广泛的适应度。该模型能对其他类型的文物进行分类。

六、参考文献

[1]董西明.对两个定类变量间相关系数计算方法之分析[J].江苏统计,1997(05):15-16.

[2] 赵匡华.试探中国传统玻璃的源流及炼丹术在其间的贡献[J].自然科学史研究,1991(02):145-156.

[3] 高卫东.基于 Logistic 回归模型的膨胀土判别与分类[J].长江科学院院报,2020,37(06):153-155.

[4]卓金武.MATLAB 在数学建模中的应用[M].北京：北京航空航天大学出版社，2011

[5]韩中庚.数学建模方法及其应用[M].北京：高等教育出版社，2009

[6]谢金星，薛毅。优化建模与 lingo 软件[M].北京：清华大学出版社，2005

[7]司守奎.数学建模算法与程序[M].北京：国防工业出版社，2011

七、附录

问题 1 程序

```
% T1
clear;clc
load data.mat
%  斯皮尔曼相关系数
%  第一问
name1=[{"纹饰"} {"种类"} {"颜色"}];
for i=1:3
    t=[cell2mat(data1(:,i)) cell2mat(data1(:,4))];
    R=corr(t(:,1) , t(:,2) , 'type' , 'Spearman');
    fprintf("%s\t 与风化程度之间的斯皮尔曼相关系数为%f\n",name1{i},R)
end
disp(' ')

%%  第二问
name1=[{"二氧化硅"} {"氧化纳"} {"氧化钾"} {"氧化钙"} {"氧化镁"} {"氧化铝"}
{"氧化铁"} ...
    {"氧化铜"} {"氧化铅"} {"氧化钡"} {"五氧化二磷"} {"氧化锶"} {"氧化锡"}
{"二氧化硫"} ];
name2=[{"铅钡玻璃"} {"高钾玻璃"}];
for i=1:2
    t1=cell2mat(data2(:,[1:2,4:end]));
    t1=t1(t1(:,1)==i,:);
    fprintf("\t---%s---\n",name2{i});
    for j=3:16
        t2=[t1(:,2) t1(:,j)];
        t2(t2(:,2)==0,:)=[];
```

```
        R=corr(t2(:,1) , t2(:,2) , 'type' , 'Spearman');
        if abs(R)>0.5
            fprintf("%s\t 与风化程度之间的斯皮尔曼相关系数为%f\n",name1{j-2},R)
        end
    end
    disp(' ')
end

t1=cell2mat(data2(:,[1:2,4:end]));
xg=[1 2 9 11 14];wg=[3:8 10 12 13];
roc=zeros(2,14);yc=[];
for i=1:2
    yct=[];
    t2=t1(t1(:,1)==i,:);
    for j=xg
        a=mean(t2(t2(:,2)==0,j+2));
        b=mean(t2(t2(:,2)~=0,j+2));
        roc(i,j)=a/b;
        if roc(i,j)==inf;roc(i,j)=0;end
    end
    yct(:,xg)=t1(t1(:,1)==i & t1(:,2)~=0,xg+2).*roc(xg);
    w=t1(t1(:,1)==i & t1(:,2)~=0,wg+2)./sum(t1(t1(:,1)==i & t1(:,2)~=0,wg+2),2);
    yct(:,wg)=(100-sum(yct,2)).*w;
    yct=[data2(t1(:,1)==i & t1(:,2)~=0,3) num2cell(yct,size(yct))];
    t=[yc;yct]; yc=t;
end
yc=[[{"检测点"},name1];yc];
save data_predict.mat roc xg
```

writecell(yc,"T1 预测结果.xlsx")

问题一风化点预测值

检测点	二氧化硅	氧化纳	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
02	79.59994856	0	0.435753138	0.275671655	0.037948472	0.171422406	0.075896943
08	44.18806406	0	0.695342732	0	0	5.337489988	1.478827501
11	73.69796781	0	5.458174506	2.113859491	0	4.269785837	0
19	65.0314905	0	4.543805492	2.156195492	0.407730323	2.020285384	0.793421169
26	43.42014834	0	13.39105253	9.8598029	2.160293894	8.918136331	2.852695783
34	78.50292611	0	0.87773091	0.589161844	0.141879791	0.601185555	0.210014154
36	86.81835624	17.28116974	-4.818977733	0	-1.294650734	-7.29058368	-1.562734977
38	72.24989818	10.74234875	-0.622329623	-0.438385841	-0.14018623	-0.814376655	-0.489436318
39	57.59367832	0	0	1.708437337	0	3.161407408	0.271433969
40	36.66249009	0	0	0.773132141	1.09784764	2.195695279	0
41	40.50206864	0	0	0.460079127	0	1.556437899	0
43 部位 1	27.22809707	0	5.248602392	5.51548048	0	11.74263586	2.846699603
43 部位 2	47.61077408	0	8.94947549	2.042815058	0	7.879429508	2.529199595
48	117.0084139	6.227448553	0	-1.396717897	0	-4.212323815	-3.436369428
49	63.16655234	0	3.712096879	2.64624728	0	5.36600143	1.065849599
50	39.4489271	0	13.16787803	9.142900152	0	6.473593671	3.026615223
51 部位 1	53.99544471	0	9.802065036	6.569299287	0.526821085	7.367513052	0.399106883
51 部位 2	46.84285837	0	12.62893559	7.192926812	0.45227593	5.375125478	0.36529979
52	56.47471543	9.496859044	6.013418415	0	0.976701717	1.947019763	0
54	48.88332012	0	0.448496841	2.31148372	1.724987851	8.107442897	0.707245019
56	63.9564085	0	0.192862341	1.915765923	0.957882962	9.218819912	0.520728321
57	55.77262107	0	0	4.798566884	1.242940891	6.306396817	2.414565502
58	66.67702416	0	1.282596214	2.87717529	1.109272401	6.066333443	0.606633344
08 严重风化点	10.11454694	0	4.207618931	13.39872979	2.924444363	12.98095202	6.326349031

26 严重风化点	8.161846985	0	0	13.43081269	1.615498696	12.22710778	8.679345936
54 严重风化点	37.54010804	0	0	3.271375513	1.819702629	6.358736154	9.384758504
07	67.01947595	0	0	6.558362069	2.539775396	6.791502591	1.23918239
09	68.74868406	0	10.1322546	3.800172414	0	4.527668394	2.332578616
10	70.01484063	0	15.79944785	1.287155172	0	2.778341969	1.895220126
12	68.22051589	0	17.34504601	4.413103448	0.285582182	5.007875648	2.113899371
22	66.8168909	0	12.70825153	10.17465517	0.237274146	12.00518135	2.551257862
27	67.08459257	0	0	5.761551724	4.645655654	8.609430052	1.457861635

续上表

检测点	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
02	0.16880527	24.18351941	0	0.051034151	0	0	0
08	4.681321494	14.62330459	23.06383289	5.631296777	0	0	0.300519976
11	0.820303683	12.94580556	0	0.694103116	0	0	0
19	1.869682292	21.83298126	1.0505484	0.257127231	0.036732462	0	0
26	3.018872236	15.05670099	0	1.093994985	0	0	0.228301998
34	0.262116902	23.73482666	0	0.07534859	0.004809484	0	0
36	-1.641198658	21.21602873	-0.90233233	-2.73315155	-0.071925041	0	0
38	-0.176650857	25.14209028	-0.078601528	-0.364646263	-0.0097239	0	0
39	5.173212122	31.11786189	0	0.973968949	0	0	0
40	4.6233302	35.79854307	18.33869437	0	0.510267213	0	0
41	8.281424292	22.49582282	25.67633087	0.137044846	0.890791502	0	0
43 部位 1	13.7887012	30.51620571	0	3.11357769	0	0	0
43 部位 2	8.17126023	22.81704604	0	0	0	0	0
48	-2.483054038	8.010185324	-14.74313335	-0.421232382	-0.443402507	0.89018	0
49	6.064316684	17.42763427	0	0.551301517	0	0	0
50	4.97079514	22.43463744	0	1.334653241	0	0	0
51 部位 1	0.37516047	20.5174957	0	0.127714202	0	0.31937	0

51 部位 2	0.930644702	26.17714287	0	0	0.034790456	0	0
52	0	24.17842063	0	0.868179304	0.044685699	0	0
54	0.569245991	28.27784074	6.968950916	1.793987365	0.206998542	0	0
56	0.475727108	21.0324726	1.305035176	0.263578533	0.160718618	0	0
57	3.34167715	22.99550338	2.007043898	1.12068441	0	0	0
58	0.95328097	20.06370417	0	0.363980007	0	0	0
08 严重风化点	0	16.54554512	30.70666582	0	1.044444415	0	1.750703584
26 严重风化点	0	15.25555346	32.7851206	4.466378748	1.520469361	0	1.857865746
54 严重风化点	0.899628266	29.80747511	6.99256516	3.312267707	0.613382909	0	0
07	11.29762159	0	0	3.415560015	0	0	1.138520005
09	1.152427578	2.242170811	5.677755119	1.386288405	0	0	0
10	3.796151193	1.216715126	0	3.212127932	0	0	0
12	1.309561536	0.362766555	0.73582436	0.180096871	0.025728124	0	0
22	0.331575409	0	0	0.120158061	0	0	0.054755572
27	8.582651972	0	0	2.467184359	0.157479853	0	1.233592179

问题二程序

```
% T2
clear;clc
load data.mat
x0=cell2mat(data2(:,[1:2,4:end]));
%  用斯皮尔曼相关系数挑选合适的指标
zb=[];zb_xishu=[];
x00=x0(x0(:,2)==0,:);
for m=1:14
    R=corr(x00(:,1) , x00(:,m+2) , 'type' , 'Spearman');
    if abs(R)>0.5
        t=[zb,m];zb=t;
        t=[zb_xishu,R];zb_xishu=t;
```

```
    end
end

% 分类规律 1 的可视化

% 散点图
x=cell(1,2);y=cell(1,2);
for i=1:2
    y{i}=[x0(x0(:,1)==i & x0(:,2)~=0,3)];
    x{i}=linspace(i,i,length(y{i}));
end
axe1=axes; hold on; grid on
fig=gcf;
for i=1:2
    plot(x{i},y{i},'o')
    x{i}=linspace(i-1/8,i+1/8,100);
    y{i}=linspace(mean(y{i}),mean(y{i}),100);
    line(x{i},y{i},'Color','k')
    str=y{i}(end);
    text(x{i}(end),y{i}(end),num2str(str))
end
xlim([0.5,2.5]);
axe1.XTickLabel=[{' '} {'铅钡玻璃'} {' '} {'高钾玻璃'} {' '}];
fig.Position=[1209 257 402 676];
saveas(axe1,"pictrue\两类风化玻璃中二氧化硅含量的差异.png")

%% 多次动态聚类取最好的情况
tic
t_R=zeros(2,1);t_R_test=zeros(2,1);
```

```
% 用 Kmeans 法聚类
name1=[{"检测点"} {"亚类"} {"种类"} {"风化程度"} {"二氧化硅"} {"氧化纳"}...
    {"氧化钾"} {"氧化钙"} {"氧化镁"} {"氧化铝"} {"氧化铁"} ...
    {"氧化铜"} {"氧化铅"} {"氧化钡"} {"五氧化二磷"} {"氧化锶"} {"氧化锡"}
{"二氧化硫"} ];
num=4; % 几类

for kk=1:1000

    center=zeros(2*num,length(zb));
    center_test=zeros(2*num,length(zb));
    julei=name1;julei_test=name1;
    for m=1:2
        a=x0(x0(:,1)==m & x0(:,2)==0,:);
        b=data2(x0(:,1)==m & x0(:,2)==0,3);

        [index_km,center(num*(m-1)+1:num*m,:)]=kmeans(a(:,zb+2),num);
        [B,I]=sort(index_km);
        t=[julei;[b(I) num2cell([B a(I,:)])]];
        julei=t;

        % 敏感性分析
        [la,~]=size(a);
        t=randperm(la);a=a(t,:);b=b(t,:);  % 打乱顺序
        a(end-5:end,:)=[];  % 去掉最后五个样本
        b(end-5:end,:)=[];
        [index_km_test,center_test(num*(m-
1)+1:num*m,:)]=kmeans(a(:,zb+2),num);
        [B,I]=sort(index_km_test);
```

```
        t=[julei_test;[b(I) num2cell([B a(I,:)])]];
        julei_test=t;
    end
    % 用轮廓系数检验全集
    t0=cell2mat(julei(2:end,2:end));
    S=zeros(2,num);
    a=zeros(1,num);
    b=linspace(inf,inf,num);
    for m=1:2   % 铅钡和高钾大类
        % 计算亚类内距离
        t1=t0(t0(:,2)==m,:);   %当前是哪一大类的玻璃
        for n=1:num   % 铅钡和高钾其中的亚类
            t2=t1(t1(:,1)==n,:);   % 当前亚类中的所有样本
            [t2l,~]=size(t2);

            for k=1:t2l   % 遍历每一个样本
                p=1:num;p(n)=[];   %去掉自身这个亚类
                for i=p   % 遍历其它亚类
                    t3=t1(t1(:,1)==i,:);   % 找出其他亚类中的所有样本
                    [t3l,~]=size(t3);
                    for j=1:t3l
                        d=dist(t2(k,zb+3),t3(j,zb+3)');
                        if b(i)>d;b(i)=d;end
                    end
                end
            end

            if t2l==1;a(i)=0;continue;end
            for i=1:t2l
```

```
            t3=t2;t3(i,:)=[];
            a0=0;
            for j=1:t2l-1
                d=dist(t2(i,zb+3),t3(j,zb+3)');
                a0=a0+d;
            end
            a(i)=sqrt(a0)/(t2l-1);
        end

    end
    for n=1:num
        S(m,n)=(b(n)-a(n))/max(a(n),b(n));
    end
end

% 测试集
t0=cell2mat(julei_test(2:end,2:end));
S_test=zeros(2,num);
a_test=zeros(1,num);
b_test=linspace(inf,inf,num);
for m=1:2   % 铅钡和高钾大类
    % 计算亚类内距离
    t1=t0(t0(:,2)==m,:);   %当前是哪一大类的玻璃
    for n=1:num   % 铅钡和高钾其中的亚类
        t2=t1(t1(:,1)==n,:);   % 当前亚类中的所有样本
        [t2l,~]=size(t2);

        for k=1:t2l   % 遍历每一个样本
            p=1:num;p(n)=[];   %去掉自身这个亚类
```

```
            for i=p    % 遍历其它亚类
                t3=t1(t1(:,1)==i,:);    % 找出其他亚类中的所有样本
                [t3l,~]=size(t3);
                for j=1:t3l
                    d=dist(t2(k,zb+3),t3(j,zb+3)');
                    if b_test(i)>d;b_test(i)=d;end
                end
            end
        end

        if t2l==1;a_test(i)=0;continue;end
        for i=1:t2l
            t3=t2;t3(i,:)=[];
            a0=0;
            for j=1:t2l-1
                d=dist(t2(i,zb+3),t3(j,zb+3)');
                a0=a0+d;
            end
            a_test(i)=sqrt(a0)/(t2l-1);
        end

    end
    for n=1:num
        S_test(m,n)=(b_test(n)-a_test(n))/max(a_test(n),b_test(n));
    end
end

R=mean(S,2);    % 轮廓系数得分
R_test=mean(S_test,2);
```

```
    if sum(R)>sum(t_R)
        t_center=center;
        t_R=R;
        t_julei=julei;
    end

    if sum(R_test)>sum(t_R_test)
        t_center_test=center_test;
        t_R_test=R_test;
        t_julei_test=julei_test;
    end

end

center=t_center;julei=t_julei;R=t_R;
center_test=t_center_test;julei_test=t_julei_test;R_test=t_R_test;

save data_center0.mat center zb num
writecell(julei,"result\T2 聚类结果.xlsx")
disp("输出完成~")
toc
```

问题三程序

```
% T3
clear;clc;
load data.mat
load data_center0.mat
load data_predict.mat
```

```
[data3l,~]=size(data3);
data3=[data3(:,1:2),zeros(data3l,2),data3(:,3:16)];
data3_test=data3;

n=20220918;
[t1,t2]=size(data3_test);
for i=1:t1
    % 产生一组和为扰动值的随机数
    s = RandStream("mcg16807","Seed",n+i);
    sum0 = rand(s)*15-10;   % 指定的和
    N = sum(data3_test(i,5:18)~=0);        % 随机数个数
    r = zeros(1, N);     % 生成的随机数
    sum0temp = sum0/N;     % 每生成一个随机数后，剩余的和
    for j=1:(N-1)
        r(j) = sum0temp.*rand(s);
        sum0temp = (sum0 - r(j))/(N-j);
    end
    r(N) = sum0 - sum(r(1:N-1));


    % 增加扰动
    p=0;
    for j=5:t2
        if data3_test(i,j)~=0
            p=p+1;
            data3_test(i,j)=data3_test(i,j)+r(p);
            if data3_test(i,j)<0;data3_test(i,j)=0.01;end
        end
    end
```

```
end

for k=1:2
    % 玻璃初分类（法 2）未风化的用 logit 回归，风化的用二氧化硅欧式距离
    t0=cell2mat(data2(:,[1:2,4:end]));
    xishu =glmfit(t0(t0(:,2)==0,zb+2),t0(t0(:,2)==0,1)-1, 'binomial');
    fit = glmval(xishu,data3(:,zb+2), 'logit');
    fit_test = glmval(xishu,data3_test(:,zb+2), 'logit');

    t(1)=mean(t0((t0(:,1)==1 & t0(:,2)~=0),3));
    t(2)=mean(t0((t0(:,1)==2 & t0(:,2)~=0),3));
    for i=1:data3l
        if data3(i,2)==0
            data3(i,3)=(fit(i)>0.5)+1;
        else
            [~,I]=min(abs(t-data3(i,5)));
            data3(i,3)=I;
        end

        if data3_test(i,2)==0
            data3_test(i,3)=(fit_test(i)>0.5)+1;
        else
            [~,I]=min(abs(t-data3_test(i,5)));
            data3_test(i,3)=I;
        end
    end

    % 将风化变回未风化
    for i=1:data3l
```

```
    if data3(i,2)~=0
        m=data3(i,3);   % 找出这是哪一类的玻璃
        t1=data3(i,:);
        data3(i,5:18)=0;
        data3(i,5:18)=t1(5:18).*roc(t1(3),:);
        w=t1(roc(m,:)==0)./sum(t1(roc(m,:)==0));
        t2=logical(linspace(0,0,4));
        data3(i,[t2,roc(m,:)==0])=(100-sum(data3(5:18))).*w;
    end

    if data3_test(i,2)~=0
        m=data3_test(i,3);   % 找出这是哪一类的玻璃
        t1=data3_test(i,:);
        data3_test(i,5:18)=0;
        data3_test(i,5:18)=t1(5:18).*roc(t1(3),:);
        w=t1(roc(m,:)==0)./sum(t1(roc(m,:)==0));
        t2=logical(linspace(0,0,4));
        data3_test(i,[t2,roc(m,:)==0])=(100-sum(data3_test(5:18))).*w;
    end
end

% 判断样本属于哪一亚类
d=zeros(1,4);
for i=1:data3l
    p=(data3(i,3)~=1)*4;
    for j=1:4
        d(j)=dist(data3(i,5:18),center(p+j));
    end
    [~,I]=min(d);
```

```
        data3(i,4)=I;

        p=(data3_test(i,3)~=1)*4;
        for j=1:4
            d(j)=dist(data3_test(i,5:18),center(p+j));
        end
        [~,I]=min(d);
        data3_test(i,4)=I;
    end
end

name1=[{"样本编号"} {"风化程度"} {"种类"} {"亚类"} {"二氧化硅"} {"氧化纳
"}...
    {"氧化钾"} {"氧化钙"} {"氧化镁"} {"氧化铝"} {"氧化铁"} ...
    {"氧化铜"} {"氧化铅"} {"氧化钡"} {"五氧化二磷"} {"氧化锶"} {"氧化锡"}
{"二氧化硫"} ];

out=[[{'无随机扰动'},cell(1,17)];name1;num2cell(data3);...
    [{'增加随机扰动后'},cell(1,17)];name1;num2cell(data3_test)];

writecell(out,"result\T3 分类结果.xlsx")
disp("输出完成~")
```

问题 3 风化前化学含量预测值

编号	A1	A2	A3	A4	A5	A6	A7	A8
二氧化硅 (SiO_2)	78.45	27.31281	31.95	35.47	46.51497	67.41018	65.71714	51.12
氧化钠 (Na_2O)	0	0.776379	0	0	0.653983	0.649478	0.646045	0
氧化钾 (K_2O)	0	1.552758	1.36	0.79	1.307967	1.298957	1.292091	0.23

氧化钙(CaO)	6.08	0	7.19	2.89	0	0	0	0.89
氧化镁(MgO)	1.86	29.30832	0.81	1.05	42.04459	60.51191	58.68031	0
氧化铝(Al_2O_3)	7.23	0	2.93	7.07	0.78478	0	0	2.12
氧化铁(Fe_2O_3)	2.15	0	7.06	6.45	0.241974	0.876796	0.633125	0
氧化铜(CuO)	2.11	5.923774	0.21	0.96	1.072533	0.415666	0.723571	9.01
氧化铅(PbO)	0	0	39.58	24.28	1.530321	0.13639	0	21.24
氧化钡(BaO)	0	1.808964	4.69	8.31	8.338288	0.987207	3.26899	11.34
五氧化二磷(P_2O_5)	1.06	71.47741	2.68	8.45	0.951696	1.051875	0.651161	1.46
氧化锶(SrO)	0.03	0	0.52	0.28	0.614744	1.123598	0.755873	0.31
氧化锡(SnO_2)	0	26.62981	0	0	7.998216	0	0	0
二氧化硫(SO_2)	0.51	0	0	0	1.412604	0	0	2.26

样本编号	1	2	3	4	5	6	7	8
风化程度	0	1	0	0	1	1	1	0
种类	2	2	1	1	1	2	2	1
亚类	1	2	2	2	2	2	2	2
二氧化硅	78.32	60	31	35.60	102.028	48.87	47.2715	51.263
氧化纳	0	0.3	0	0	5.43388	0.43	0.42680	0
氧化钾	0	23	1	0.96	0.63114	446.44	165.669	0.2709
氧化钙	5.633	0	6.8	3.08	1.26228	39.96	41.9010	0.9321
氧化镁	1.543	32.	0.8	1.181	29.3496	29.23	27.8859	0
氧化铝	7.1055	0	2.8	7.351	0.44057	19.75	51.9015	2.2044
氧化铁	1.4151	0	6.6	6.680	0.21357	55.81	0.53134	0
氧化铜	1.8011	0	0	1.210	4.85242	2.82	2.91774	9.36051
氧化铅	0	20	39	24.59	0.74356	26.80	26.6491	21.2819

氧化钡	0	0	3.0	8.315	26.724	2.49	6.45823	12.0401
五氧化二磷	0.5345	0	1.5	9.535	1.97361	3.31	0.03111	1.4726
氧化锶	0.01	4.0	0	0.047	0.55263	0.30	0.32914	0.54511
氧化锡	0	6.9	0	0	5.78563	0.10	0	0
二氧化硫	0.01	1.4	0	0	0.17180	0.48	1.30184	3.94824

问题四

```
% T4
clear;clc;
load data.mat
load data_center0.mat

% 对大类的关联分析
t0=cell2mat(data2(:,[1:2,4:end]));
R1=ones(2*14,14);R2=zeros(2*14,14);
for m=1:2
    k=(m-1)*14;
    R2(k+1:k+14,:)=corrcoef(t0(t0(:,1)==m,3:end));
end
name1=[{"二氧化硅"} {"氧化纳"} {"氧化钾"} {"氧化钙"} {"氧化镁"} {"氧化铝"} {"氧化铁"} ...
    {"氧化铜"} {"氧化铅"} {"氧化钡"} {"五氧化二磷"} {"氧化锶"} {"氧化锡"} {"二氧化硫"} ];
out=cell(31,15);
out(1,2:15)=name1; out(2:15,1)=name1'; out(2:15,2:15)=num2cell(R2(1:14,:));
out(17,2:15)=name1;                                    out(18:31,1)=name1';
out(18:31,2:15)=num2cell(R2(15:28,:));
writecell(out,"result\T4 相关系数矩阵.xlsx")
disp("输出成功~")
```

```
% 提取上三角
R2l=width(R2);
n=(1+R2l)*R2l/2-14;
x=1:n;
t=zeros(2,n);
p=0;
for m=1:2
    k=(m-1)*14;
    for i=1:R2l
        for j=i+1:14
            p=p+1;
            t(m,p)=R2(k+i,j);
        end
    end
end

% 分别找出两类玻璃中关联程度前三的指标，并画图
y=zeros(2,n);
y(1,:)=abs(t(1,1:n));
y(2,:)=abs(t(2,n+1:end))*(-1);
t1=abs(y);
t2=1:n;
[~,I1]=sort(t1(1,:),'descend');
[~,I2]=sort(t1(2,:),'descend');
p(1,1:n)=t2(I1);
p(2,1:n)=t2(I2);
out=[[{"铅钡"};{"高钾"}] num2cell(p(:,1:3))];
disp(out)
```

```
h1 = axes;
fig = gcf;
bar(x,y(1,:),'EdgeColor','w','FaceColor',[0.3843 0.7098 0.9608])
hold on
bar(x,y(2,:),'EdgeColor','w','FaceColor',[0.9569 0.5647 0.4627])
ylim([-1.25,1])
legend('铅钡玻璃','高钾玻璃','Location','north');
h1.XTick=[];h1.YTick=[];
fig.Position= [117 370 1652 607];
text(p(1,1)-0.2,y(1,p(1,1))+0.05,"\downarrow SiO_2 and PbO",'Color','r')
text(p(1,2)-0.2,y(1,p(1,2))+0.05,"\downarrow BaO and CuO",'Color','r')
text(p(1,3)-0.2,y(1,p(1,3))+0.05,"\downarrow BaO and SO_2",'Color','r')

text(p(2,1)-0.3,y(2,p(2,1))-0.165,"\uparrow SiO_2 and K_2O",'Color','b')
text(p(2,2)-0.3,y(2,p(2,2))-0.125,"\uparrow SiO_2 and CaO",'Color','b')
text(p(2,3)-0.3,y(2,p(2,3))-0.045,"\uparrow SiO_2 and Al_2O_3",'Color','b')

saveas(fig,"pictrue\T4 不同类别之间的化学成分关联关系的差异性分析.png")
```

问题四 热图

```
xiangguan1=abs(xlsread('热力图.xlsx',1,'B2:O15'));
xiangguan2=abs(xlsread('热力图.xlsx',1,'B18:O31'));
x_name={'SiO_2','Na_2O','K_2O','CaO','MgO','Al_2O_3','Fe_2O_3','CuO','PbO'
,'BaO','P_2O_5','SrO','SnO_2','SO_2'};
y_name=x_name;
figure
H = heatmap(x_name,y_name,xiangguan1,'FontSize',12, 'FontName','微软雅黑');
H.Title = '铅钡玻璃各组分的皮尔逊相关系数';
figure
```

```
H = heatmap(x_name,y_name,xiangguan2,'FontSize',12, 'FontName','微软雅黑');
H.Title = '高钾玻璃各组分的皮尔逊相关系数';
```

问题四 关联规则

```
clc;clear  % 高钾无风化共 12 组，铅钡无风化共 23 组
data1=xlsread('关联.xlsx',2,'E8:R19');
for j=1:14
    for i=1:size(data1,1)
        if data1(i,j)>0
            data1(i,j)=1;
        end
    end
end
% guanlian(data1)
data2=xlsread('关联.xlsx',2,'E46:R68');
for j=1:14
    for i=1:size(data2,1)
        if data2(i,j)>0
            data2(i,j)=1;
        end
    end
end
guanlian(data2)
```