基于多种统计方法对玻璃文物成分的分析和鉴别

摘要

古代玻璃由于历史悠久，容易发生风化作用，这给考古工作者鉴别玻璃类型带来一定的困难。本文通过建立数学模型，综合利用统计方法和机器学习方法，分析了古代玻璃制品的成分、对未知类别的文物进行分类预测。研究该问题，可以发现玻璃制品在风化过程中的规律，对于考古工作者具有参考性意义。

针对问题一，本文同时运用 **Fisher 精确检验**和**卡方检验**对玻璃文物表面有无风化和其他三个因素进行显著性检验；随后对不同玻璃类型，分别利用描述性统计定性分析、弹性网正则化 Logistic 回归定量分析风化前后化学成分含量的统计规律，最后利用数据本身的信息，预测出风化点风化前的化学成分含量。最终得出结论：**玻璃类型和纹饰与文物表面风化是相关的；通过描述性统计分析可以得出化学成分在不同类型的玻璃风化前后含量会有所不同，例如风化后高钾玻璃文物二氧化硅含量均高于风化前，但与之相反的是铅钡玻璃文物在风化后二氧化硅含量急剧下降，且有无风化与化学成分之间存在一定的函数关系。**

针对问题二，我们利用表单二的数据建立随机森林模型，得出每个化学成分含量对分类结果的影响大小，最终发现**氧化铅对高钾玻璃和铅钡玻璃的分类起主导作用**；接下来利用**熵权法**对化学成分含量指标进行赋权，选择出合适的化学成分并以此进行**层次聚类**。最终将**未风化的高钾玻璃分为 4 类，风化的高钾玻璃分为 2 类，未风化的铅钡玻璃分为 4 类，风化的高钾玻璃分为 2 类**。最后本文基于聚类结果，对聚类结果的合理性和聚类模型的敏感性进行了解释：**聚类结果对类别数的敏感性较大**。

针对问题三，我们根据问题二中的随机森林模型对表单 3 的数据进行预测，为了验证结果的正确性，我们同时建立了逻辑回归模型，最后得出了相同的预测结果，按顺序依次为**高钾、铅钡、铅钡、铅钡、铅钡、高钾、高钾、铅钡**；随后利用 ROC 曲线探究分类模型的敏感性，最后得出结论：**该预测模型的稳定性较好，分类结果不受阈值的影响**。

针对问题四，本文先使用 Spearman 相关系数分别对不同类别的玻璃样品化学成分含量之间的关联性进行量化，发现大多数化学成分含量之间具有显著的相关关系，并且通过查阅文献解释部分变量呈显著相关关系的原因；接着利用方差分析，探究不同类别化学成分含量关联性是否有显著差异，最终得出结论：**不同类别化学成分关联性是有显著差异的，且高钾玻璃化学成分之间的关联度要显著大于铅钡玻璃。**

关键词：Fisher 精确检验，卡方检验，随机森林，熵权法，层次聚类，方差分析

一、 问题重述

1.1 问题背景

玻璃作为中外交流文化的媒介之一，在玻璃制造技艺传入中国后，于中国就地取材，在不同的助熔剂下，制造出了化学成分不同的钾玻璃和铅钡玻璃。又因为埋藏地点的环境的不同，不同类型的玻璃制品的风化程度不同，与之带来相应化学成分比例不同。根据如今的化学成分分布分析成分之间以及不同类别的玻璃之间的关系成为一个重要的课题。

1.2 问题的提出

试就附件中给出的不同文物的玻璃品种、化学成分分布、纹饰、颜色等建立数学模型，分析以下问题：

（1）根据附件给出的信息，分析文物风化与该文物的玻璃品种、纹饰、颜色之间存在的关系，并在相同的玻璃类型情况下，分析有无风化化学成分含量规律以及对风化前文物化学成分比例进行预测。

（2）根据化学成分含量构建模型对文物玻璃类型进行分类；并尝试对不同玻璃类别按照成分细分，说明理由和方法，并用敏感性分析进行评价。

（3）基于第二问，对附件中表单三数据对未知玻璃类型的文物进行类型判断，并对其进行敏感性度量。

（4）分析相同类别下，其化学成分之间的关联关系，以及分析不同类别化学成分关联关系是否有显著性差异。

二、 问题分析

2.1 问题一的分析

对于风化与其他玻璃属性之间的关系问题，结合附件给出的定性数据，发现传统的皮尔逊相关系数于本题中并不适用；而斯皮尔曼相关系数虽然可以计算定性数据的相关性分析，但是它使用于分级的定序数据中，于此同样也不适用。于本题中，将问题转化为对风化与玻璃属性的 Fisher 精确检验以及卡方检验进行相关性分析。

对于附件中存在的数据缺失值现象，因为颜色无法预测，于本题中，我们选择对缺失值删除的数据预处理操作。

化学成分含量的统计规律将分为描述性统计分析和定量分析，通过散点图和均值方差等分析相同类别下风化前后化学成分含量的差异，此外，运用弹性网正则化 Logistic 模型拟合不同类别下有无风化与化学成分之间的函数关系，得出最终的统计规律结论。

在预测风化后的数据在风化前的化学成分含量时，通过观察数据特点，结合风化前后不同化学含量之间的差别，得出最终的预测结论。

2.2 问题二的分析

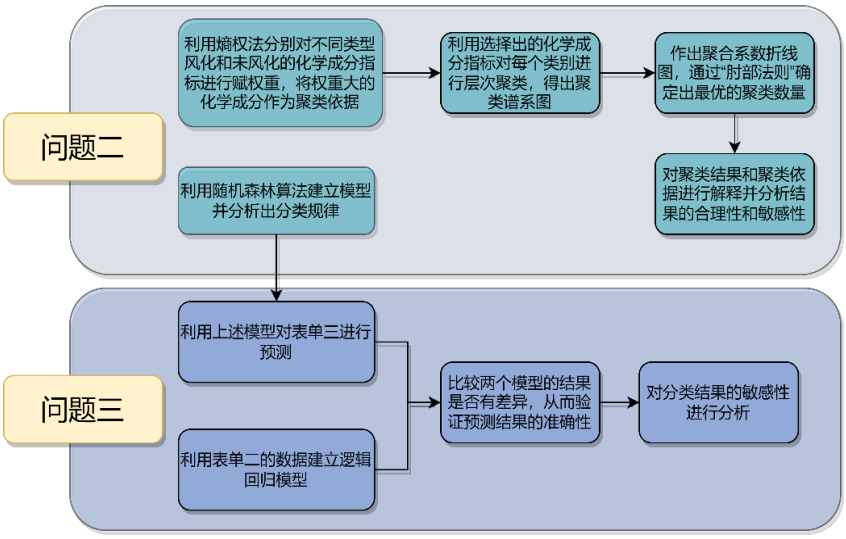


图 1 问题二、问题三流程图

问题二可以分成三个小问，首先需要利用数据建立合适的模型进行分类，发现其中的分类规律，之后需要选择合适的化学成分指标进行聚类并分析聚类的结果，最后需进行敏感性分析。

本题所要求的分类规律，我们可以根据不同化学成分的含量并考虑风化对文物化学成分含量的影响，构建化学成分对文物类型的分类预测模型，那么我们的难点就是：如何选择分类模型。

模型算法的选择，我们可以结合预测模型特点和数据特征特点以及问题本身的复杂性综合考量。

从数据特点出发，我们发现化学成分含量还受到风化作用的影响，若采用如 KNN（最邻近分类）可能容易受到风化作用带来的数据噪声干扰；另外在观察数据特点时，我们可以明显看出在不同玻璃类型中化学成分含量和各种化学成分之间含量具有显著的差异，也就是指标对分类的作用具有层次性，这一点类似于树的层次结构，于是我们首先对决策树模型进行几番试验，发现得到的模型欠拟合（相关支撑见附录 4），因此使用更具说服力的随机森林集成学习方法；再者，考虑到本题所给出的数据量太小，并且特征变量也较少，如果使用深度学习容易因模型训练不充分造成过拟合，直接影响问题 3 的分类结果，因此放弃使用神经网络。

之后是确定合适的化学成分指标和聚类，我们使用较为客观的熵权法，确定重要指标并分别对高钾风化、高钾未风化、铅钡风化、铅钡未风化的数据进行层次聚类，最后对聚类结果进行解读，最后考虑到不同因素对聚类结果的影响大小，本文对聚类结果进行了敏感性分析。

2.3 问题三的分析

本题要求对未知文物化学成分分析并鉴别类型，又由于在问题 2 中我们已经建立了强硬的分类预测模型，结果具有极其好的表现，因此直接使用该模型进行预测结果是可信的；除此之外，为了提高结果可信度，我们又建立了逻辑回归模型对结果进行加强论证，同时为了展示我们随机森林模型的鲁棒性，这里继续使用随机森林模型。

2.4 问题四的分析

问题四分为两个小问，先针对不同类型的玻璃文物样品，分别探究样品中化学成分

之间的关系，得出两两指标间的相关系数，又因为皮尔逊相关系数只能描述线性关系，故而本题中使用斯皮尔曼相关系数；再对两个类型的相关系数进行方差分析，分析其差异性，得出结论。

三、 模型假设

1、假设本题中所提供的 67 个样本数据以及需要预测的 8 个样本数据的化学元素所提供的含量是准确的，不存在录入错误的数据；

2、假设本题中所提供的文物纹饰、类型、颜色、表面风化的信息是正确的；

3、除了题目给定的化学元素成分含量外，不存在影响最终的预测结果或者分类的其他化学元素，即使存在，含量微乎其微，可以不纳入考量；

4、问题中所使用的算法模型、预测模型结果是正确的。

四、 符号说明

表 1 符号说明

指标	符号	指标	符号
二氧化硅(SiO_2)	x_{i1}	氧化铜(CuO)	x_{i8}
氧化钠(Na_2O)	x_{i2}	氧化铅(PbO)	x_{i9}
氧化钾(K_2O)	x_{i3}	氧化钡(BaO)	x_{i10}
氧化钙(CaO)	x_{i4}	五氧化二磷(P_2O_5)	x_{i11}
氧化镁(MgO)	x_{i5}	氧化锶(SrO)	x_{i12}
氧化铝(Al_2O_3)	x_{i6}	氧化锡(SnO_2)	x_{i13}
氧化铁(Fe_2O_3)	x_{i7}	二氧化硫(SO_2)	x_{i14}
有无风化	z	铅钡玻璃/高钾玻璃	y

五、 模型的建立与求解

5.1 数据预处理

对表单二进行了整合，剔除不满足文物成分比例 85%~105%的数据，最终生成一列对应的玻璃类型指标，此外，对于表单一中缺失的颜色信息，则将其剔除。

5.2 问题一模型的建立与求解

对于附件中存在的数据缺失值现象，因为颜色无法预测，于本题中，我们选择对缺失值删除的数据预处理操作。

5.2.1 Fisher 精确检验和卡方检验

对于非定序数据来说，本题中的定性数据之间的相关性分析可以使用 Fisher 精确检验和卡方检验来评估风化和玻璃属性之间的关系。该检验设定显著性水平$\alpha=0.10$。

5.2.1.1 Fisher 精确检验步骤及结果

由于篇幅所限，将 Fisher 精确检验具体步骤放在附录 2 中。

根据题目要求，结合 Fisher 精确检验，设定如下假设：

H_0:玻璃类型、纹饰、颜色与有无风化的是独立的

H_1:玻璃类型、纹饰、颜色与有无风化是相关的

并通过 R 语言程序，得到玻璃类型与纹饰、颜色和玻璃文物有无风化的频数数据如下：

表 2　玻璃类型和有无风化的频数分布表

类型 / 有无风化	高钾	铅钡	行总数
有风化	6	24	30
无风化	12	12	24
列总数	18	36	54

表 3　玻璃颜色和有无风化的频数分布表

颜色 / 有无风化	黑	蓝绿	绿	浅蓝	浅绿	深蓝	深绿	紫	行总数
有风化	2	9	0	12	1	0	4	2	30
无风化	0	6	1	8	2	2	3	2	24
列总数	2	15	1	20	3	2	7	4	54

表 4　玻璃纹饰和有无风化的频数分布表

类型 / 有无风化	A	B	C	行总数
有风化	9	6	15	30
无风化	11	0	13	24
列总数	20	6	28	54

通过 R 语言程序，计算得出相应的 P-value，结果如下:

表 5 Fisher 精确检验

P-value：其他属性 / 表面风化	表面风化
玻璃类型	**0.04047**
纹饰	**0.0588**
颜色	**0.616**

根据所得到的 P 值，在给定的显著性水平下，文物表面风化与玻璃类型和纹饰相关。

5.2.1.2 卡方检验步骤及结果

为了保证结论的可靠性，在研究定性因素之间的关系时，我们同样也采取了卡方检验，与 Fisher 精确检验不同的是，卡方检验计算出相应交叉分类数据出现的频率，计算出相应的检验统计量，得出相应结论，但由于篇幅所限，将 Fisher 精确检验具体步骤放在附录 2 中。

根据题目要求，结合卡方检验，设定如下假设：

H_0:玻璃类型、纹饰、颜色与有无风化的是独立的

H_1:玻璃类型、纹饰、颜色与有无风化是相关的

并通过 R 语言程序，得到玻璃类型与纹饰、颜色和玻璃文物有无风化的频率数据：

表 6 玻璃类型和有无风化的频率分布表

类型 / 有无风化	高钾	铅钡	行总数
有风化	0.11	0.44	0.56
无风化	0.22	0.22	0.44
列总数	0.33	0.67	1

表 7 玻璃颜色和有无风化的频率分布表

颜色 / 有无风化	黑	蓝绿	绿	浅蓝	浅绿	深蓝	深绿	紫	行总数
有风化	0.04	0.17	0	0.22	0.02	0	0.07	0.04	30
无风化	0	0.11	0.02	0.15	0.04	0.04	0.06	0.04	24
列总数	0.04	0.28	0.02	0.37	0.06	0.04	0.13	0.08	54

表 8 玻璃纹饰和有无风化的频率分布表

类型 / 有无风化	A	B	C	行总数
有风化	0.17	0.11	0.28	0.56
无风化	0.20	0	0.24	0.44
列总数	0.37	0.11	0.52	54

最后，通过 R 语言程序，计算得出相应的 P-value，结果如下:

表 9 Chisq 检验

表面风化 / P-value / 其他属性	表面风化
玻璃类型	**0.04202**
纹饰	**0.0565**
颜色	**0.5066**

根据所得的到的 P 值，在给定的显著性水平下，玻璃类型和纹饰与文物表面风化是相关的。

5.2.1.3 结论

基于结果，结合相关文献，周良知[1]于 1984 年就已提出，在相同风化条件下，不同玻璃其风化程度与玻璃的化学成分有着密切的关系，同时他也提出，风化玻璃表面形貌也与风化程度有着密切的联系；王承遇 [2]（2003）得出，玻璃的耐风化不仅与玻璃成分有关，还与其形貌有关。

因而我们可以分析，不同类型的玻璃文物中存在着不同的化学因素，必然会使得其化学因素与周围环境发生化学风化，而不同化学因素的性质又决定了化学风化的速度，可知玻璃类型与文物表面风化相关；不同玻璃文物的纹饰由于其表面形态的差异，与周围环境接触面积也不同。因此，不同的纹饰也导致不同玻璃文物的表面风化不同。

5.2.2 基于玻璃类型下的统计规律分析

5.2.2.1 对数据的描述性统计分析

表 10 高钾玻璃的各化学因素风化前后均值和标准差数值

指标	风化前均值	风化后均值	风化前标准差	风化后标准差	指标	风化前均值	风化后均值	风化前标准差	风化后标准差
SiO_2	67.98	93.96	3.9203	0.4452	CuO	2.4525	1.562	0.9821	0
Na_2O	0.695	0	3.0925	0.4878	PbO	0.4117	0	1.4340	0.2100
K_2O	9.331	0.5433	0.6761	0.3063	BaO	0.5983	0	0.0484	0
CaO	5.332	0.870	2.4915	0.9645	P_2O_5	1.403	0.2800	0.6813	0
MgO	1.079	0.1967	1.6667	0.0695	SrO	0.04167	0	0.1855	0
Al_2O_3	6.620	1.930	1.6600	0.9348	SnO_2	0.1967	0	0	0
Fe_2O_3	1.932	0.2650	0.5890	0	SO_2	0.1017	0	0	3

表 11 铅钡玻璃各化学因素风化前后均值和标准差数值

指标	风化前均值	风化后均值	风化前标准差	风化后标准差	指标	风化前均值	风化后均值	风化前标准差	风化后标准差
SiO_2	27.00	24.01	0.4041	0.4200	CuO	1.2625	5.470	8.2084	14.3301
Na_2O	0.3667	0	2.0723	0.8796	PbO	44.70	34.38	4.6371	2.8919
K_2O	0.2467	0.2100	0.8808	0.4939	BaO	10.332	19.010	0.1656	0.1374
CaO	3.026	2.482	3.5717	1.904	P_2O_5	3.5292	6.010	0.3852	0
MgO	0.7133	0.4133	0.9022	0.8405	SrO	0.4500	0.350	4.6043	5.8748
Al_2O_3	3.357	2.523	1.6324	4.1742	SnO_2	0.1425	0	0	0
Fe_2O_3	0.925	0.5317	14.713	8.7418	SO_2	1.329	2.425	0	0

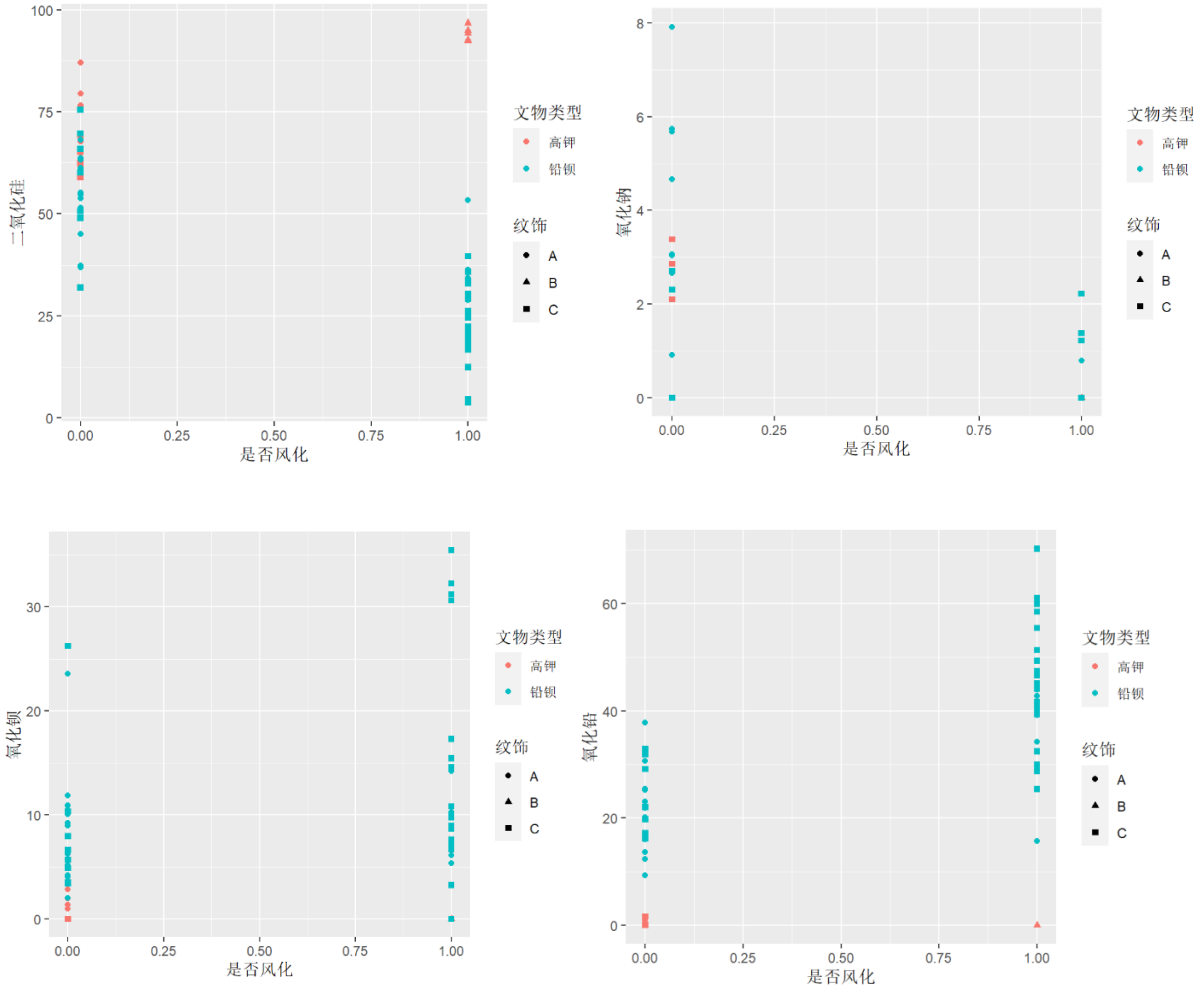


图 2 三维数据散点图

结合表 10 列出的数据以及不同玻璃类型与各个化学因素在风化前后的散点图（由于篇幅所限，仅放四张，其余置于附件 3 中）可得知，二氧化硅含量总体来说高钾玻璃文物均高于铅钡玻璃文物，且风化后高钾玻璃文物二氧化硅含量均高于风化前，但与之相反的是铅钡玻璃文物在风化后二氧化硅含量急剧下降；而氧化钠的分布含量与玻璃类型基本无关，且不风化下含量差距也较大；总体来说，铅钡的氧化钾含量在风化前后差别不大，但高钾玻璃文物的氧化钾含量在风化前较高，风化后迅速下降；铅钡文物的氧化钙、氧化镁、氧化铁含量在风化前后差别不大，高钾玻璃文物氧化钙、氧化铝含量在风化前要高于风化后，且氧化镁含量风化后略小于风化前含量；铅钡文物的氧化铝含量风化后含量略低于风化前；铅钡玻璃文物的氧化铜含量在风化后略微上涨，但高钾玻璃的氧化铜、氧化铅含量几乎不变；铅钡玻璃文物的氧化铝、氧化钡含量在风化后有所上升；铅钡文物风化后的五氧化二磷、氧化锶含量分布较为均匀，但风化前高钾和铅钡的五氧化二磷含量较低；高钾玻璃文物风化前氧化锶含量较低，风化后也无氧化锶含量，铅钡含量于风化前分布含量较为均匀；高钾和铅钡玻璃中含氧化锡、二氧化硫的玻璃较少。

针对二氧化硅含量在铅钡玻璃中，风化后含量降低的现象，王婕等（2014）[3]对此现象有过相关研究，研究表明铅钡类玻璃的硅元素会顺着风化层流失，导致硅的含量减少；类似的，钡元素在外层聚集，导致含量增加；铅元素同样也有增加的趋势。

5.2.2.2 基于弹性网正则化 Logistic 模型的统计规律分析

根据题意，即在玻璃类型一样的情况下，分析玻璃文物有无风化与化学成分含量之

间是否具有相关的统计规律。由于有无风化作为一种 0-1 变量，因此在分析化学成分和该变量的统计规律时，将采用 Logistic 回归，构建而二元离散选择模型。又有引入弹性网正则化对模型进行回归，能发挥 LASSO 和 Ridge 方法的优势，可以有效的排除弱相关变量或者增加模型的预测准确率，即通过查阅相关文献后，发现 Ridge 模型可能会带来多重共线性问题，LASSO 模型可能会排除有关的重要变量，即变量删减太多而降低了模型的预测效果，因此综合使用传统 Logistic 模型与 LASSO 和 Ridge 模型结合的弹性网正则化 Logistic 模型，在 Logistic 模型中引入两个模型的惩罚项，可以综合两模型优劣，能有效减轻两者模型存在的缺陷，提高模型的准确率。

在 Logistics 回归模型的基础上，利用弹性网正则化进行约束，其表达式为：

$$\hat{\beta}=\arg\min\left\{\|Y-X\beta\|^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^{\,2}\right\}$$

令 $\alpha=\dfrac{\lambda_1}{\lambda_1+\lambda_2}$，$\lambda=\lambda_1+\lambda_2$，则

$$\hat{\beta}=\arg\min\left\{\|Y-X\beta\|^2+\lambda\left[\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j=1}^{p}\beta_j^{\,2}\right]\right\}$$

据此得到弹性网正则化的惩罚项为：

$$\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j=1}^{p}\beta_j^{\,2}$$

该惩罚项恰好是岭回归函数和 LASSO 函数的线性组合，即$\alpha=0$时，该弹性网模型为岭回归；当$\alpha=1$时，该弹性网模型为 LASSO 回归。

基于此，通过设定解释变量

$$z=\begin{cases}1 & \text{有风化}\\ 0 & \text{无风化}\end{cases}$$

通过构建 7:3 训练集与测试集进行弹性网正则化 Logistic 模型，通过 R 语言循环语句，计算对 K-fold 交叉验证在 k 不同的时候使得该模型的均方误差最小。因此在不同玻璃文物类型（高钾、铅钡）下，计算可以得到：

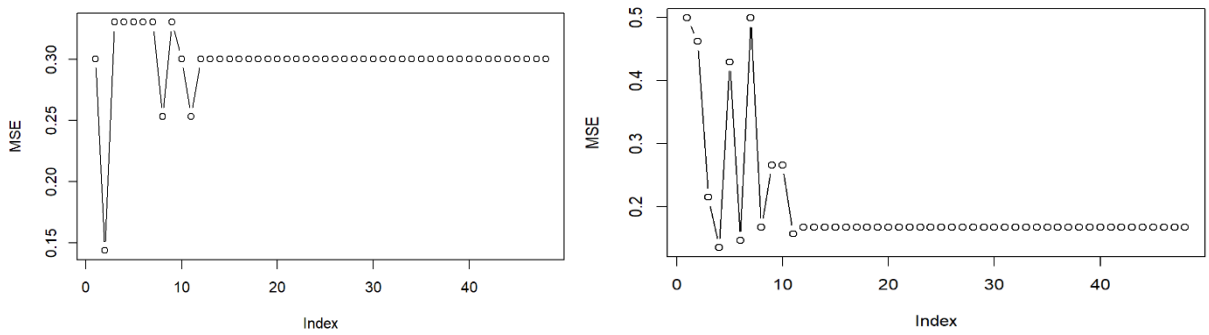


图 3 左边为高钾文物 MSE 右边为铅钡文物 MSE

最后可以发现，在$k=4,6$时，高钾和铅钡玻璃文物的 Logistic 模型的 MSE 模型取到最小值，最后，我们可以得到 Logistic 模型的系数如下所示（未在表格中写明的参数

的系数均为 0）：

表 12 高钾玻璃文物 Logistic 模型系数

Coefficients:	S1
(Intercept)	-6.06159377
x_{i1}	0.06814247

表 13 铅钡玻璃文物 Logistic 模型系数

Coefficients:	S1
(Intercept)	-0.20365911
x_{i1}	-0.04183614
x_{i4}	0.04130565
x_{i8}	0.04079594
x_{i10}	0.12998674

通过上述弹性网正则化 Logistic 模型能有效地表明不同的玻璃类型，文物样品表面化学成分含量和有无风化之间的函数关系，且通过测试集（如下）我们也可以看出，该函数关系拟合的很好。

表 14 弹性网 Logistic 模型下高钾测试集结果

Predicted \ TRUE	**0**	**1**
0	2	0
1	0	4

表 15 弹性网 Logistic 模型下铅钡测试集结果

Predicted \ TRUE	**0**	**1**
0	37	1
1	9	1

通过计算可以得出准确率分别达到 1，0.93，可见该函数关系较为科学。

5. 2. 3 风化前后预测模型

5.2.3.1 预测方法及其过程

通过观察数据特点，发现不同类别风化前后不同化学含量存在细微差别，因此我们通过计算不同类别风化前后的均值差，结合风化点的检测数据和均值差，预测出该风化点风化前的数据。考虑到结果可能出现负值，结合现实实际情况，出现负值说明该化学成分含量在风化过程中是增加的，因此我们将负值统一替换为 0 值，最后再通过表单二中文物采样点为 49、49 未风化点、50、50 未风化点的数据，验证预测方法的合理性。

①预测过程：

利用 Excel 按不同类别分别求出未风化各化学成分含量的平均值、风化后各化学成分含量的平均值以及二者的差值，最后用风化前的数据加上差值。限于篇幅，最终预测结果见附录。

②验证：

以二氧化硅为例，验证预测结果是否合理

表 16 预测结果验证

	49 风化点预测	49 未风化点	50 风化点预测	50 未风化点
二氧化硅含量	59.11132997	54.61	48.30132997	45.02
相对误差	-0.082426844		-0.07288605	

通过相对误差值可以看出，预测结果较好。

5.3 问题二模型的建立与求解

5.3.1 问题二数据预处理

为了体现风化文物或风化采样点对采样点化学成分含量的影响，通过结合表单一和表单二信息，在表单中添加一列新的变量风化，风化值设定规则如下：

①根据表单一，对有风化文物的风化值设为 1，对无风化文物的风化值设为 0；

②如果检测部位是风化文物的未风化点，则将其风化值修改为 0；

③如果检测部位是风化文物的严重风化点，则将其风化值修改为 2。

基于此，我们的数据顾及到了风化作用对文物化学成分含量检测的影响，减小了由于风化对文物分类的影响，是一组层次分明且特征完备的数据组。

表 17 化学元素的训练集和测试集均值

分子描述符	Mean		分子描述符	Mean	
	Train	Test		Train	Test
二氧化硅(SiO2)	48.70717	51.90143	氧化铜(CuO)	1.776667	1.232857
氧化钠(Na2O)	0.7435	1.175714	氧化铅(PbO)	25.41033	26.29571
氧化钾(K2O)	1.917833	1.962857	氧化钡(BaO)	7.217333	5.018571
氧化钙(CaO)	2.623667	3.06	五氧化二磷(P2O5)	2.748667	1.59
氧化镁(MgO)	0.6805	0.15	氧化锶(SrO)	0.263667	0.264286
氧化铝(Al2O3)	4.043	4.382857	氧化锡(SnO2)	0.086833	0.062857
氧化铁(Fe2O3)	0.8375	1.202857	二氧化硫(SO2)	0.374833	0

5.3.2 基于随机森林的玻璃文物类型分类模型建立

5.3.2.1 随机森林模型框架

随机森林模型[4]的训练分为两个阶段（如下图）：

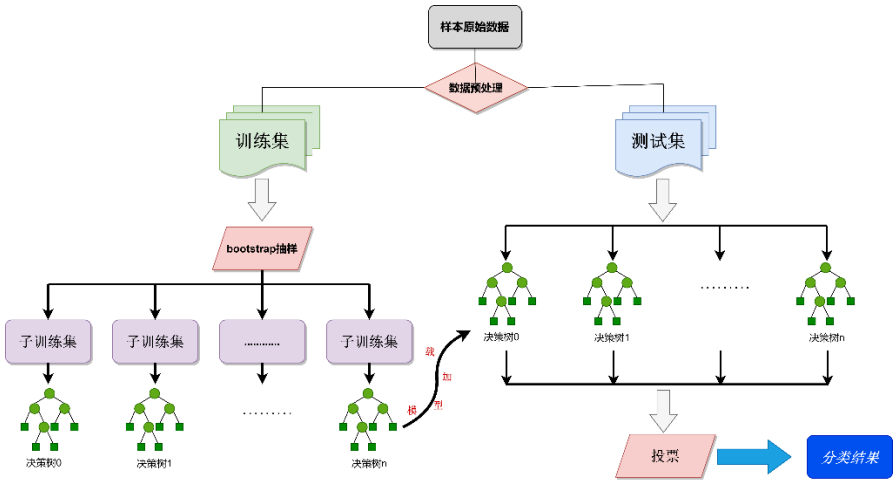


图 4 随机森林流程图

①训练集准备阶段：随机森林的多棵决策树所需要的数据采用 Bootstrap 抽样(有放

回的随机抽样)，利用有放回的采样方式使得部分数据重复采样，最终导致采样后样本是原样本的子集，但各个子集样本互相交叉，使树的评判并不片面，有利于最后评判投票选择正确的分类；利用随机的取样方式使得每棵树的训练集不同，从而形成更加全面的评判标准。

②决策树建立阶段：决策树的建立，先是基于信息论关于信息熵、信息增益的理论[5]，得到不同特征对对结果分类的信息增益效果，将重要的判别依据作为树根，能够更高效地划分类别。

5.3.2.2 随机森林参数调优——网格搜索

对随机森林算法模型参数进行优化，使用 GridSearchCV（网格搜索）方法进行最佳参数搜索。在此之前，我们采用了控制单一变量的方式依次调节了参数：criterion、n_estimators、max_depth、min_samples_split、max_features，最终得到的范围如下图时，模型接近最优：

```
param_grid = {
    'criterion':['entropy','gini'],
    'n_estimators':[11,13,15,17],
    'max_depth':[5, 6, 7],
    'min_samples_split':[4,8,12,16],
    'max_features':[0.3,0.4,0.5]
}
```

(注：实验得到在上述范围内的分类准确率都比较高)

表 18 test 数据的均值和标准差

序号	Mean_test	序号	Std_test_score
0	1	0	0
1	1	1	0
⋮	⋮	⋮	⋮
286	0.9667	286	0.05774
287	0.9667	287	0.05774

①上述参数网格是结合参数特点和问题特点进行多次网格范围收缩实验的结果；

②参数范围的选择还参考了最后可视化的结果，防止树太高产生过拟合或者树太宽出现数据利用不充分的状况。

5.2.2.3 结果

实验可能会由于 bootstrap 抽样出现一定变动，但分类效果始终良好。

下面是某次实验的测试结果：

表 19 测试集结果总结结果

RFC 最优模型参数	测试集中文物真实类型	测试集中文物预测类型
{'criterion': 'entropy', 'max_depth': 5, 'max_features': 0.3, 'min_samples_split': 4, 'n_estimators': 11}	['铅钡'], ['铅钡'], ['铅钡'], ['铅钡'], ['高钾'], ['铅钡'], ['铅钡']	['铅钡' '铅钡' '铅钡' '铅钡' '铅钡' '高钾' '铅钡' '铅钡']

最终得到随机森林结果如下图所示，具体请见附录：

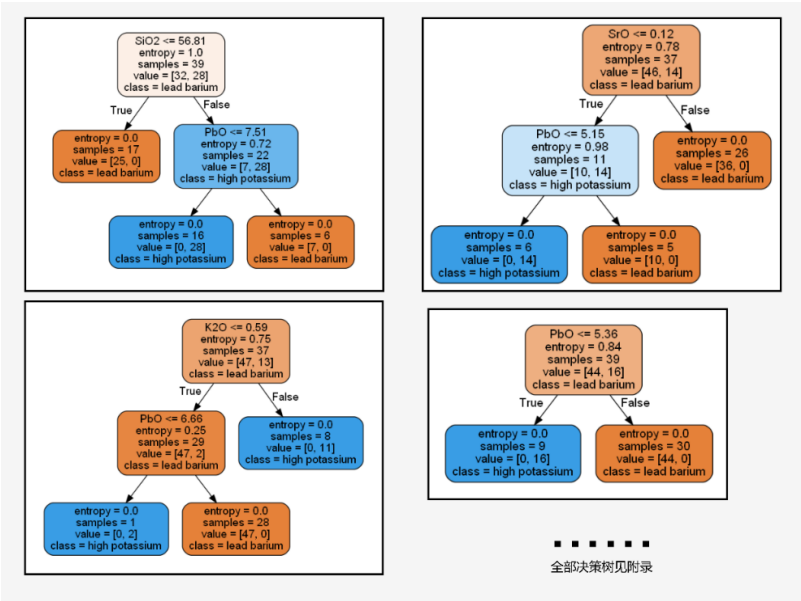


图 5 随机森林树图

5.3.2.3 分类规律

随机森林中，我们基于信息增益的概念，再加上决策树投票方法的加权，最终得出了不同化学成分的对分类结果的信息增益强度，也就是对分类结果的重要性百分比，如下图：

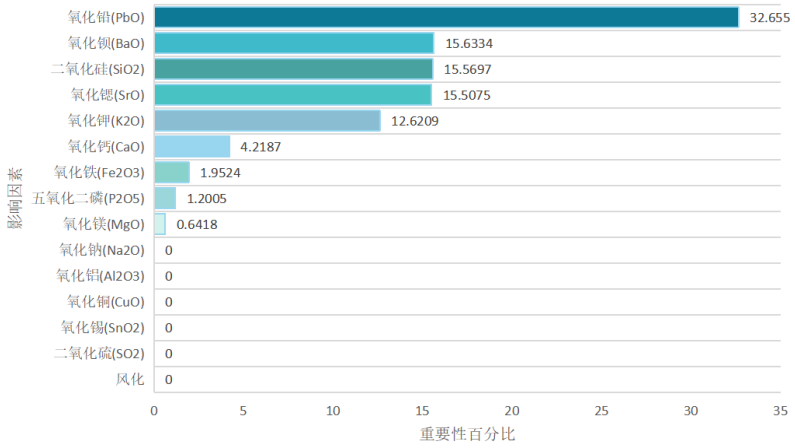


图 6 不同化学成分影响分类结果的重要性占比

从上图重要性百分比可以看出，高钾玻璃和铅钡玻璃的分类氧化铅起主导作用，另外氧化钡、二氧化硅、氧化锶、氧化钾的作用对高钾玻璃和铅钡玻璃的分类起到了一定的作用，但强度弱于氧化铅，影响程度大概为氧化铅的一半，其余化学物质只有微弱或者无影响。

5. 3. 3 基于熵权法选择合适的化学成分指标

对于指标选取类问题，通常采用两种方法：

①分析出指标对结果的影响大小，选择影响程度大的指标，但是前提是要寻找到因变量；

②利用数据本身的特点，客观地得出数据指标的权重。

本题并未给出亚类划分的依据，故选择第二种方法。熵权法就是一种利用数据进行赋权重的方法。熵权法[6]评价方法步骤如下：

①利用原始数据矩阵，计算第 i 个元素的第 j 个指标值所占比重：$p_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n}x_{ij}}$；

②计算第 j 项指标的熵值：$z_j=-\frac{1}{\ln n}\sum_{i=1}^{n}p_{ij}\ln np_{ij},j=1,\cdots,m.$；

③计算第 j 项指标的变异系数：$y_j=1-z_j,j=1,\cdots m.$即对于第 j 项指标，z_j越大，指标值的变异程度越小；

④计算第 j 项指标的权重：$\omega_j=\frac{y_j}{\sum_{j=1}^{m}y_j}\ j=1,2,\cdots m.$

由第一问得出的结论，样品是否风化，其化学成分会发生很大变化，考虑到后续聚类易受到风化的影响，导致聚类结果仍体现为样品是否风化，因此为了排除风化对样品本身化学性质的影响，我们将数据分成四组，如下图所示：

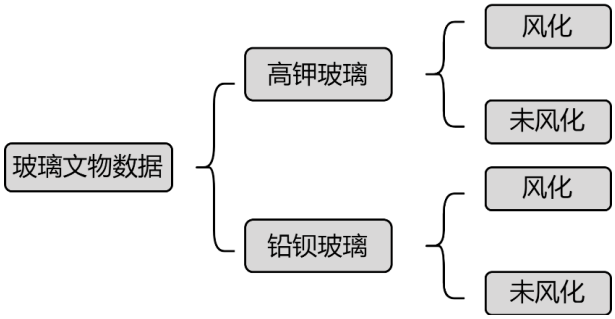


图 7 玻璃文物分组

接下来分别对四组数据分别进行熵权法赋值，找出每一组的化学成分指标，各化学成分的权重大小如下图所示（限于篇幅，只显示前 6 个）：

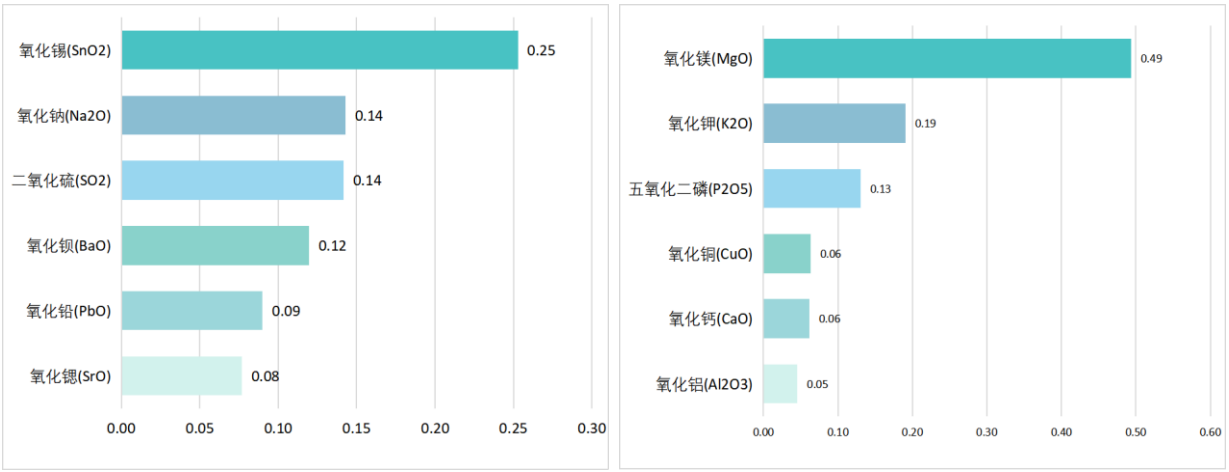


图 8 高钾未风化（左）风化（右）权重图

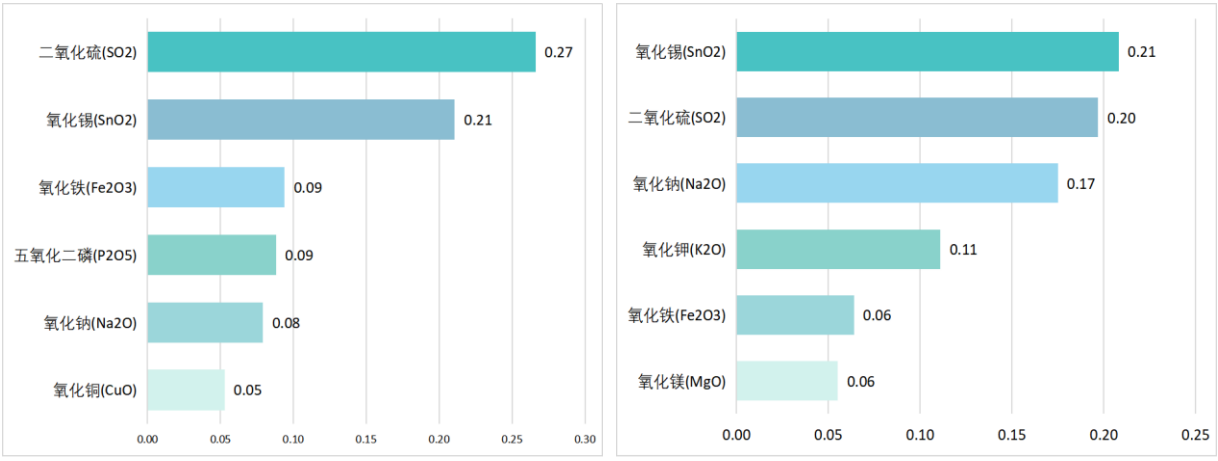


图 9 铅钡 未风化（左）风化（右）权重图

因而在下文中，我们选择权重排名前 4 的化学成分指标。选择出合适的化学成分后，下一步就是对 4 组数据进行亚类划分

5. 3. 4 层次聚类得出聚类结果

层次聚类是一种自下而上的聚类方法，是根据样本点的距离进行聚类并不断迭代，以最后类的个数是否为 1 作为迭代停止的条件，其具体算法流程图可见下图：

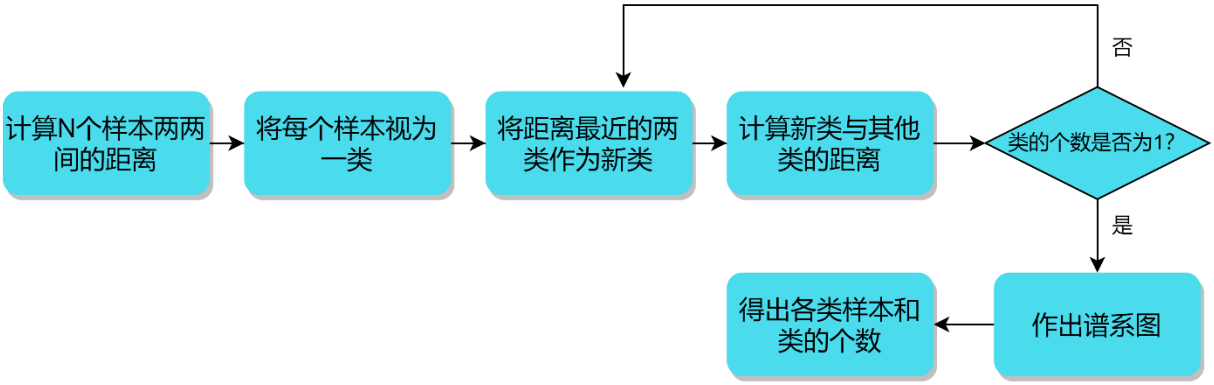


图 10 聚类算法流程图

使用 SPSS 软件实现层次聚类，得出四组数据的谱系图如下：

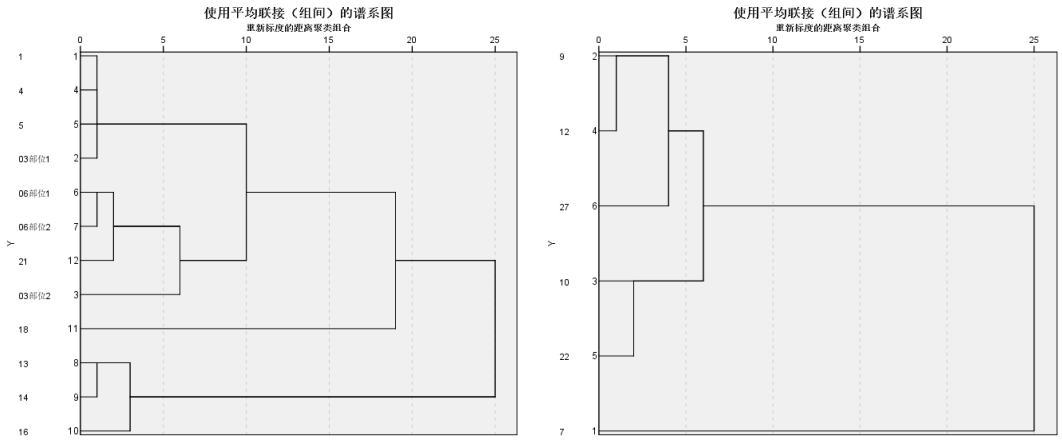


图 11 高钾未风化（左）风化（右）聚类谱系图

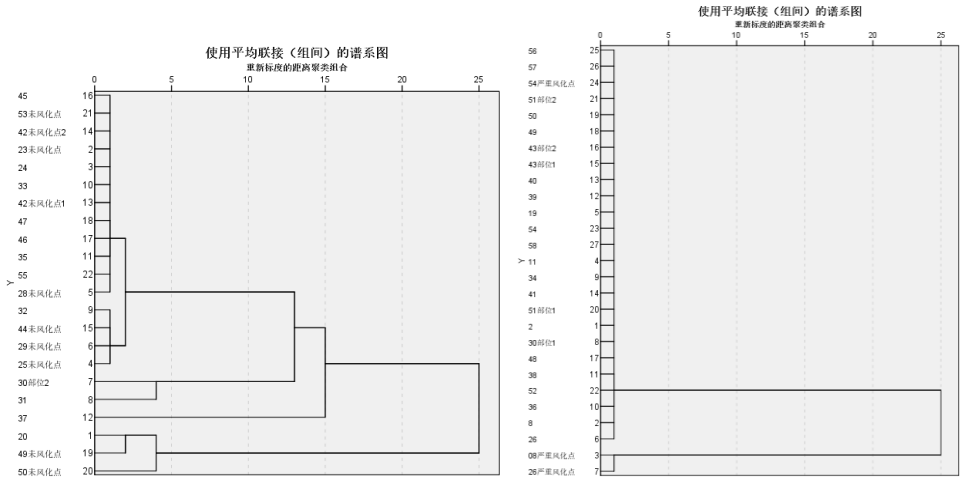


图 12 高钾未风化（左）风化（右）聚类谱系图

通过肘部法则，得出类别数 K 与聚合系数的折线图如下：

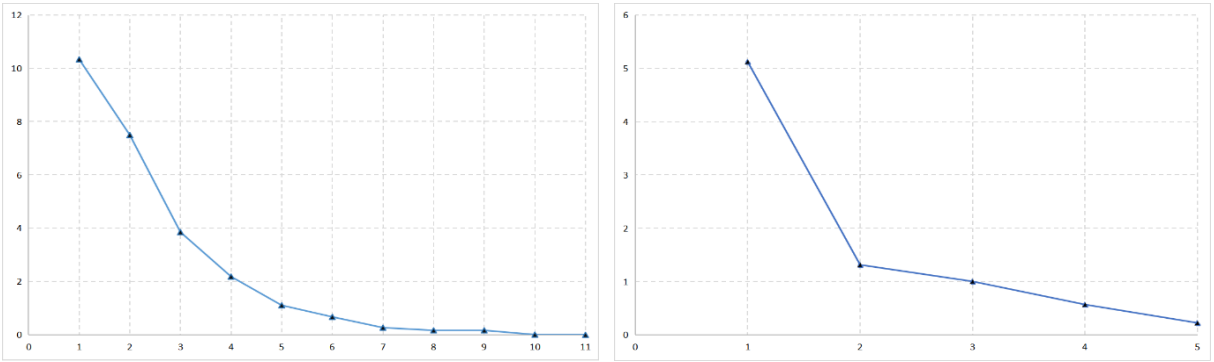


图 13 高钾未风化（左）风化（右）聚合系数折线图

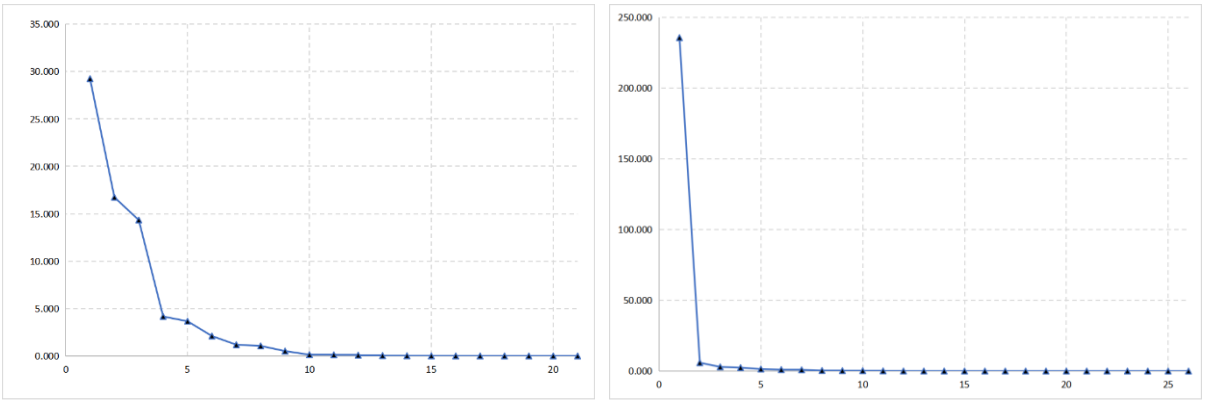


图 14 铅钡未风化（左）风化（右）聚合系数折线图

依据肘部法则，选择出最佳的聚类类别，聚类结果如下、高钾风化 2 类，铅钡未风化 4 类、铅钡风化 2 类。

表 20 高钾未风化，共分为 4 类

文物采样点	氧化铁	五氧化二磷	氧化锡	二氧化硫	文物采样点	氧化铁	五氧化二磷	氧化锡	二氧化硫
		类别 1			06 部位 1	0	1.38	0	0
13	2.86	0	0	0	06 部位 2	0	0.97	0	0
14	3.38	0	0	0			类别 4		
16	2.1	0	0	0	1	0	0	0	0.39

类别 2					4	0	0	0	0.36
18	0	0	2.36	0	5	0	0	0	0.47
类别 3					03 部位 1	0	0	0	0
03 部位 2	0	2.86	0	0					

表 21 高钾风化，共分为 2 类

文物采样点	氧化钾	氧化镁	氧化铜	五氧化二磷	文物采样点	氧化钾	氧化镁	氧化铜	五氧化二磷
类别 1					10	0.92	0	0.84	0
7	0	0	3.24	0.61	12	1.01	0	1.65	0.15
类别 2					22	0.74	0.64	0.55	0.21
9	0.59	0	1.55	0.35	27	0	0.54	1.54	0.36

表 22 铅钡未风化，共分为 4 类

文物采样点	氧化铁	五氧化二磷	氧化锡	二氧化硫	文物采样点	氧化铁	五氧化二磷	氧化锡	二氧化硫
类别 1					28 未风化点	0.41	1.04	0.23	0
37	0	1.46	0	3.66	29 未风化点	0.81	0.41	0	0
类别 2					32	1	0.17	0	0
30 部位 2	2.74	1.41	0.44	0	33	0	0.13	0	0
31	4.59	1.62	0	0	35	0.17	0.42	0	0
类别 3					42 未风化点 1	0	0.08	0	0
20	1.51	5.75	0	0	42 未风化点 2	0	0	0	0
49 未风化点	1.27	4.32	0	0	44 未风化点	0.77	0	0	0
50 未风化点	0	6.34	0	0	45	0	0	0	0
类别 4					46	0	0.2	0	0
23 未风化点	0	0	0	0	47	0	0.1	0	0
24	0	0.14	0	0	53 未风化点	0	0	0	0
25 未风化点	1.55	0.19	0	0	55	0	0.35	0	0

表 23 铅钡风化，共分为 2 类

文物采样点	氧化钠	氧化钾	氧化锡	二氧化硫	文物采样点	氧化钠	氧化钾	氧化锡	二氧化硫
类别 1					41	0	0.44	0	0
08 严重风化点	0	0	0	15.03	43 部位 1	0	0	0	0
26 严重风化点	0	0.4	0	15.95	43 部位 2	0	0	0	0
类别 2					48	0.8	0.32	1.31	0
2	0	1.05	0	0	49	0	0	0	0
8	0	0	0	2.58	50	0	0	0	0
11	0	0.21	0	0	51 部位 1	0	0	0.47	0
19	0	0	0	0	51 部位 2	0	0	0	0
26	0	0	0	1.96	52	1.22	0	0	0
30 部位 1	0	1.41	0.4	0	54	0	0.32	0	0
34	0	0.25	0	0	54 严重风化点	0	0	0	0
36	2.22	0.14	0	0	56	0	0	0	0
38	1.38	0	0	0	57	0	0	0	0
39	0	0	0	0	58	0	0.34	0	0
40	0	0	0	0					

5.3.4.1 对聚类结果进行分析

①对于高钾未风化的类别 1，观察聚类结果和化学成分含量可发现，仅类别 1 含有氧化钠，故**类别 1** 中的文物具有**含钠元素的特性**；同理**类别 2** 中的文物具有**含锡元素的特性**；**类别 3** 具有**含钡元素的特性**；**类别 4** 具有**含硫元素的特性**。

②对于高钾风化后的类别 1，观察聚类结果可发现，类别 1 的氧化铜含量高于类别 2，且不含氧化钾，故**类别 1** 的特性为**不含钾元素、铜元素含量相对更高**；**类别 2** 的特性为**具有钾元素且铜含量相对更低**。

③对于铅钡未风化的类别 1，观察聚类结果可发现，仅类别 1 含有二氧化硫，故**类别 1** 中的文物具有**含硫元素的特性**；类别 2 的氧化铁含量高于其他类别，故**类别 2** 的特征为**铁元素含量相对更高**；类别 3 的五氧化二磷含量高于其他类别，故**类别 3** 的特征为**磷元素含量相对更高**；同样的，类别 4 的五氧化二磷含量低于其他类别，故**类别 4** 的特征为**磷元素含量相对更低**。

④对于铅钡风化后的类别 1，观察聚类结果可发现类别 1 的二氧化硫含量明显高于类别 2，故**类别 1** 的特性为**硫元素含量相对更高**；同理，**类别 4** 的特征为**硫元素含量相对更低**。

通过上述分析我们不难发现，同样类型风化情况一致的文物，具有不同的化学成分的特征，这也成为了本模型划分亚类的依据，究其原因，本文认为：对于同类型未风化文物，文物的制造方法的差异性造成了文物具有不同的化学成分特征；对于同类型风化文物，不仅要考虑到制造方法的不同，文物所处的环境也具有很大影响，如铅钡风化后类别 1 和类别 2 二氧化硫含量的差别，可能是类别 1 文物所处环境空气污染较大，大气中含二氧化硫成分多造成的，因此环境气氛对玻璃的风化具有显著影响[1]。

5.3.5 分析分类结果的敏感性

层次聚类中，影响聚类结果的重要因素有计算距离的方法和聚类的类别数量，计算类与类距离常用的计算方法有：组间联接、组内联接、最近距离法、最远距离法、重心聚类、瓦尔德法等。利用 SPSS 使用不同方法对数据进行聚类，发现聚类结果基本一致，证明聚类结果对聚类方法的敏感性不高。

聚类类别的选择能直接地影响到聚类结果，本文使用肘部法则确定 K 值，通常使用各类别的样本点到该类中心的距离来度量聚类结果的好坏程度，即聚类系数，由聚类系数折线图可以明显发现当类别越多，聚类效果越好，但此时每一个样本即为一类，达不到聚类的效果。因此 K 值的选择需兼顾聚类效果和聚类类别。

因此将折线图斜率减少处作为聚类的 K 值。因此聚类结果对聚类类别数较为敏感，利用肘部法则也很好地解决了模型的敏感性问题。

5.4 问题三模型的建立与求解

5.4.1 随机森林模型的预测

通过问题二设定的模型，带入相关因素，最终得出结果。

5.4.2 用于论证的 Logistic 模型

逻辑回归模型是一种从统计学领域借鉴来的二分类方法，具有更强的数学原理。其次，逻辑回归模型适合该题的场景——连续数据的二分类。

5.4.2.1 数据预处理和分析

逻辑回归的数据本应该与随机森林保持一致，但其相对与随机森林模型而言，在反向传播时如果不进行归一化会导致损失函数进行梯度下降时梯度太小，降低了逻辑回归的算法效率。

我们使用 MinMaxScaler 的归一化方法，其原理是将数据点按其比例映射到[0,1]

区间内，数据处理的数学原理公式如下：

$$X_{std} = \frac{X - X.\min(axis = 0)}{X.\max(axis = 0) - X.\min(axis = 0)}$$

$$X_{scaled} = X_{std} \times (\max - \min) + \min$$

5.4.2.2 逻辑回归模型的数学原理

逻辑回归前半部分依旧是将特征值函数与其对应的权重参数相乘形成多元回归，这样最大程度地反映了模型对数据的贴合；逻辑回归后半部分加上一个逻辑函数，需要对前半部分的结果寻找一个阈值形成二分类

逻辑回归是基于线性回归$z = \varepsilon^T x + b$梯度下降使得损失最小化后，再将z通过"Sigmoid 函数"$y = \frac{1}{1+e^{-z}}$次方在阈值的比较下转化为 0 和 1 两类。总体来讲，是对多元回归、逻辑函数、阈值分类的综合，如公式所示：$y = \frac{1}{1+e^{-(\varepsilon^T x + b)}}$。整个过程是求解是的损失最小的权重参数和阈值的过程。

5. 4. 3 最终结果

通过随机森林模型，并通过 Logistic 模型进行检验，最终可以求出结果如下：

表 24 化学成分分析结果

文物编号	所属类型	文物编号	所属类型
A1	高钾	A5	铅钡
A2	铅钡	A6	高钾
A3	铅钡	A7	高钾
A4	铅钡	A8	铅钡

5. 4. 4 分类结果的敏感性分析

ROC 曲线是通过不断改变二分类模型的阈值生成的，常用于评价一个模型的好坏，也可以用于探究分类结果对分类设定阈值的敏感性。ROC 曲线的横纵坐标分别为 FPR 值和 TPR 值，分别表示错误的预测为正的概率和正确的预测为正的概率。

AUC 是 ROC 曲线与 X 轴围成的面积，当 AUC 值越接近于 1 时，分类器的性能越好。

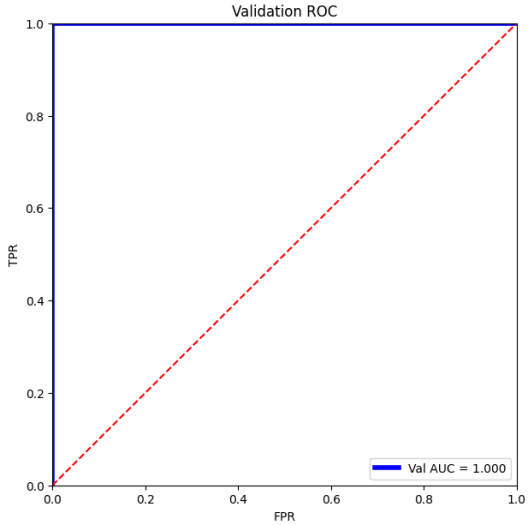


图 15 ROC 曲线

通过上述 ROC 曲线可以看出，当分类器的阈值发生变化时，TPR 值均为 1，说明无论阈值如何变化均不影响分类结果。说明本文建立的模型的敏感性好。

5.5 问题四模型的建立与求解

问题四分为两个小问，先针对不同类型的玻璃文物样品，分别探究样品中化学成分之间的关系，得出两两指标间的相关系数；再对两个类型的相关系数进行方差分析，分析其差异性。

5.5.1 基于 spearman 相关系数分析化学成分之间的关系

由于化学成分含量之间不存在线性关系，因而在分析不同类别的玻璃文物样品，其成分之间的关联关系，不能使用前提严格的皮尔逊相关系数进行分析，因为皮尔逊相关系数衡量的是连续数据之间的关联性大小，且只能衡量数据之间的线性关系。因而在本题中，我们采用斯皮尔曼相关系数，分析不同化学成分之间的关联关系，斯皮尔曼相关系数具体公式如下：

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^{2}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^{2}}}$$

其中n表示样本容量的大小，x,y表示两因素具体数据的排序，本题中表示化学成分两两之间对应元素的含量排序的数据。

通过 Matlab 软件中 corr 命令，可以分别得出高钾玻璃和铅钡玻璃两种类型的相关系数热力图如下（显著性水平见附录）：

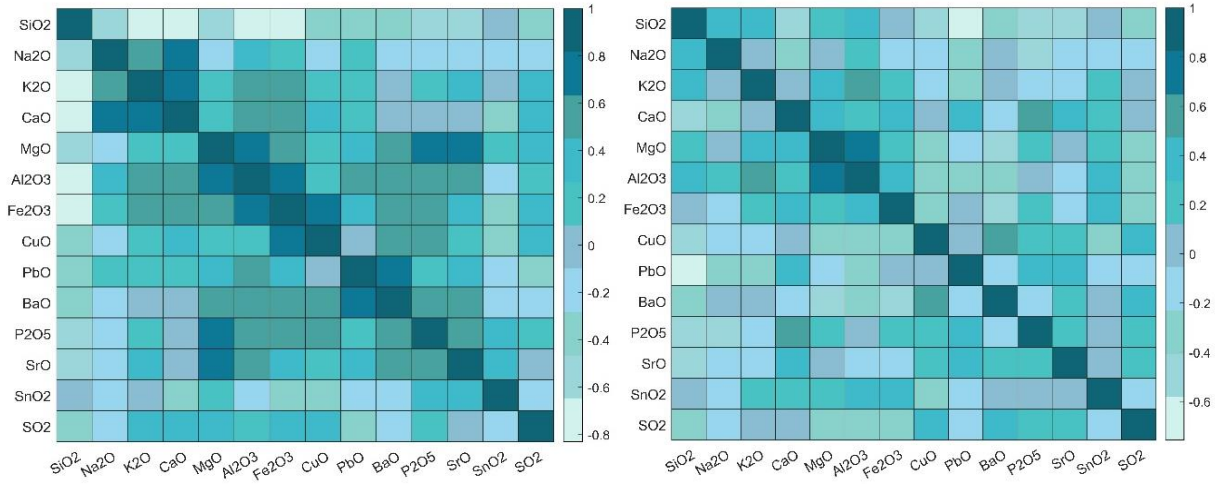


图 16 高钾（图左）铅钡（图右）相关系数热力图

得出相关系数后，为确定化学元素之间是否存在相关关系，设定原假设和备择假设如下：

$$H_0{:}R=0;H_0{:}R\neq 0$$

构造检验统计量：

$$u=R\sqrt{n-1}\sim N(0,1)$$

最后，计算得出相应的 P 值，与显著性水平$\alpha=0.10$进行比较，最终结果限于篇幅，

请见附录，通过上述计算，我们可以得出结论如下：

①高钾玻璃中二氧化硅与其他化学含量均为负相关，其中与氧化铝的相关性最为显著、其次为氧化钾，氧化钙等；这是因为高钾玻璃含二氧化硅成分少，含其他化学成分较多，经过风化和外界交换元素，其他化学成分的含量降低，导致二氧化硅含量增加；而铅钡玻璃中二氧化硅与其他化学含量不全为负相关；

②高钾玻璃中相关性最强的两个化学成分为二氧化硅和氧化铝，呈极强的负相关关系；铅钡玻璃中相关性最强的两个化学成分为二氧化硅和氧化铅，呈强负相关关系。究其原因，郭国林等[7]（2006）曾有过相关研究，在风化过程中，钾长石会出现粘土化，相关反应方程式为

$$4K[AlSiO_8]+4H_2O+2CO_2\rightarrow 2K_2CO_3+8SiO_2+Al_4[Si_4O10](OH)_8$$

其中$Al_4[Si_4O10](OH)_8$是一种铝盐，呈白色软泥状，易从玻璃表面脱离，反应同时生成了二氧化硅，所以二氧化硅和氧化铝的变化呈现很强的负相关关系;

③高钾玻璃中除二氧化硅外，其他化学成分的相关性基本为正相关，这表明除二氧化硅外，其他化学元素的变化具有趋同性。

5.5.2 基于单因素方差分析比较不同类别之间化学成分关联关系的差异性

由前面的分析，我们求出了不同类别的化学成分的相关大小，属于数值型变量，而不同类别属于分类型变量，因此我们使用方差分析来研究不同类别之间化学成分相关系数有无差异。

方差分析可以考虑所有样本，增加分析的可靠性，提高检验的效率[8]。但是方差分析的使用前提是样本需服从正态分布，因此对高钾玻璃和铅钡玻璃两种类型的相关系数的绝对值进行正态分布检验。

5.5.2.1 利用 Jarque-Bera 正态性检验

构造统计量：

$$J=\frac{n}{6}\left[S^2+\frac{(K-3)^2}{4}\right]$$

其中 S 为偏度，K 为峰度，n 为样本容量。

若样本服从正态分布，则该统计量应服从自由度为 2 的卡方分布。基于此，提出原假设和备择假设：

H_0:该变量服从正态分布 H_1:该变量不服从正态分布

利用 Matlab 的 jbtest 命令，计算出对应 p 值：

$$P_{高钾}=0.1182$$

$$P_{铅钡}=0.1819$$

均大于$\alpha=0.05$，故不能拒绝原假设，认为两组数据均服从正态分布

5.5.2.2 方差分析

具体步骤如下：

①数据预处理

根据相关系数的定义，相关系数的绝对值越大，表明关联程度越高。因此需对所有相关系数取绝对值后再进行方差分析。

②提出原假设和备择假设如下：

H_0:$u_1 = u_2$ 不同类别之间的化学成分关联关系没有差异

H_1:$u_1 \neq u_2$ 不同类别之间的化学成分关联关系有显著差异

构造检验统计量 F：

$$F = \frac{SSA/(k-1)}{SSE/(n-k)} = \frac{MSA}{MSE}$$

其中，SSA 表示组间误差，SSE 表示组内误差，MSA 表示组间均方，MSE 表示组内均方，n 为样本容量，k 为因素水平的个数。

③利用 Excel“数据分析”功能实现方差分析

Excel 进行单因素方差分析的结果如下表所示：

表 25 数据的描述性统计

SUMMARY				
组	观测数	求和	平均	方差
高钾玻璃	91	33.35139	0.366499	0.043043
铅钡玻璃	91	23.20767	0.255029	0.027929

表 26 单因素方差分析结果

差异源	SS	Degree of freedom	MS	F	P-value	F crit
组间	0.565357	1	0.565357	15.93188	9.5493E-05	3.89364
组内	6.387462	180	0.035486			
总计	6.952819	181				

④得出结论

$P < 0.10$，故拒绝原假设，高钾玻璃和铅钡玻璃之间的化学成分关联关系有显著差异，并且由表 11 可知，高钾玻璃的平均值要高于铅钡玻璃，故可得出：不同类别之间的化学成分关联关系是有显著差异的，且高钾玻璃内部化学成分之间的关联度要显著大于铅钡玻璃。

六、 模型的评价、改进与推广

6.1 模型的评价

①对于问题 2 文物类型的分类预测模型，对缺失数据和噪声数据有很好的容忍度；恰好符合本题数据测量有误和风化影响成分含量的场景；另外该模型在较小数据集上具有一定的优势；同时从决策树集成而来，具有更好的可解释性；而且我们可以先用决策树进行参数范围收缩，再利用小范围进行网格搜索，调参便捷；网格搜索的动态模型建立，使得模型具有很好的泛化能力；

②对于问题 3 未知类型文物的论证模型，从另一种角度——Logistic 回归得到了相同的结果，回归的方式使得模型与数据更加贴切，且训练速度很快；

③使用弹性网正则化 Logistic 模型进行离散选择变量的拟合，可以有效的防止多重共线性问题，又有效的提高了模型的拟合准确率；

④层次聚类法对是否给出聚类数没有要求，能够很好地发现类与类的关系。

6.2 模型的不足

①随机森林分类预测模型虽然在分类效果上具有明显的优势，但是由于模型建立的过程需要网格搜索选择合适超参数，内存和速度上都不及其他模型。

②Logistic 的论证模型较为简单，如果数据复杂一点，很难保证能拟合出数据的真实分布。

6.3 模型的改进与推广

①随机森林作为集成学习方法，基分类器并不一定只用决策树，如果使用更加高效的模型，或许在泛化能力和训练速度上又会有一定的提升。

②可以通过使用神经网络论证，但巨大的参数量不适合进行实际预测，并且也不好解释其中分类的原理，还要防止风化带来的过拟合。

③对这四个问题的分析建模，本文得到了玻璃文物在风化过程中的变化规律，这对考古工作者关于玻璃文物的成分分析和鉴别具有一定的参考性意义。

参考文献

[1]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.

[2]王承遇,陶瑛.硅酸盐玻璃的风化[J].硅酸盐学报,2003,(01):78-85.

[3] 王婕,李沫,马清林,张治国,章梅芳,王菊琳.一件战国时期八棱柱状铅钡玻璃器的风化研究[J].玻璃与搪瓷,2014,42(02):6-13.DOI:10.13588/j.cnki.g.e.1000-2871.2014.02.002.

[4] 董红瑶,王弈丹,李丽红.随机森林优化算法综述[J].信息与电脑(理论版),2021,33(17):34-37.

[5] Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

[6] 司守奎,孙玺菁.数学建模算法与应用[M].三版.北京：国防工业出版社,2021.4.

[7] 郭国林, 郭福生, 刘晓东, 等. 丹霞地貌砂岩的微观化学风化作用电子探针研究[D]. , 2006.

[8] 贾俊平. 统计学[M].7 版.北京:中国人民大学出版社,2018.

附录

附录 1

R 问题一（Fisher 精确检验、卡方检验、正态性检验、弹性网正则化 Logistic 回归）

```
getwd()
setwd("C:\\Users\\HUAWEI\\Desktop") #可以掉换成任何你存放文件的路径
library(openxlsx)
a<-read.xlsx("附件.xlsx","表单 1")#读取文件中数据,这里 x，y 可以被替换
a
fh<-as.vector(unlist(a$表面风化))  #转化为向量
ws<-as.vector(unlist(a$纹饰))#转化为向量
lx<-as.vector(unlist(a$类型))#转化为向量
ys<-as.vector(unlist(a$颜色))#转化为向量
table(a$纹饰,a$表面风化)
table(a$类型,a$表面风化)
table(a$颜色,a$表面风化)
table(a$纹饰,a$表面风化)/54
table(a$类型,a$表面风化)/54
table(a$颜色,a$表面风化)/54
#Fisher 精确检验
fisher.test(fh,lx)
fisher.test(fh,ys)
fisher.test(fh,ws)
#卡方检验
chisq.test(fh,lx)
chisq.test(fh,ys)
chisq.test(fh,ws)
#正态性检验
shapiro.test(a$二氧化硅.SiO2.)
shapiro.test(a$氧化钠.Na2O.)
shapiro.test(a$氧化钾.K2O.)
shapiro.test(a$氧化钙.CaO.)
shapiro.test(a$氧化镁.MgO.)
shapiro.test(a$氧化铝.Al2O3.)
shapiro.test(a$氧化铁.Fe2O3.)
shapiro.test(a$氧化铜.CuO.)
shapiro.test(a$氧化铅.PbO.)
shapiro.test(a$氧化钡.BaO.)
shapiro.test(a$五氧化二磷.P2O5.)
shapiro.test(a$氧化锶.SrO.)
shapiro.test(a$氧化锡.SnO2.)
shapiro.test(a$二氧化硫.SO2.)
```

```
getwd()
setwd("C:\\Users\\HUAWEI\\Desktop")
bd2<-read.csv("附件 表单 22.csv",encoding = "UTF-8")
names(bd2)<-c("文物采样点","文物类型","纹饰","是否风化","二氧化硅","氧化钠
","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")
library(ggplot2)
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=二氧化硅,color=
文物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化钠,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化钾,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化钙,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化镁,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化铝,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化铁,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化铜,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化铅,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化钡,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=五氧化二
磷,color=文物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化锶,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=氧化锡,color=文
物类型,shape=纹饰))
ggplot(data=bd2)+geom_point(mapping=aes(x=是否风化,y=二氧化硫,color=
文物类型,shape=纹饰))
    getwd()
setwd("C:\\Users\\HUAWEI\\Desktop")
gj2<-read.csv("第四问高钾数据.csv",encoding = "UTF-8")
names(gj2)<-c("文物采样点","是否风化","二氧化硅","氧化钠","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")
summary(gj2[1:6,])
summary(gj2[7:18,])
c<-c()
```

```
for(i in 5:18){
  m<-sd(gj2[1:6,i])
  n<-sd(gj2[7:18,i])
  c<-c(c,m,n)
}
c
getwd()
setwd("C:\\Users\\HUAWEI\\Desktop")
gj2<-read.csv("第四问铅钡数据.csv",encoding = "UTF-8")
names(gj2)<-c("文物采样点","是否风化","二氧化硅","氧化钠","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")
summary(gj2[1:6,])
summary(gj2[7:18,])
c<-c()
for(i in 5:18){
  m<-sd(gj2[1:6,i])
  n<-sd(gj2[7:18,i])
  c<-c(c,m,n)
}
   c
#高钾数据 Logistic 模型
library(corrplot)  #掉用包
getwd()
setwd("C:\\Users\\HUAWEI\\Desktop")
a<-read.csv("第四问高钾数据.csv",encoding = "UTF-8")
x<-a[1:18,3:16]  #化学元素
y<-a[1:18,2]     #0-1 变量
library(glmnet)
set.seed(205)
select<-sample(1:nrow(a),length(a$X.U.FEFF.文物采样点)*0.7)
train=a[select,]
test=a[-select,]
cc<-c()
for(i in 3:50){
     cvfit <- cv.glmnet(data.matrix(x), y,nfolds = i ,family =
"binomial", type.measure = "class")
     cvfit$lambda.min
     summary(cvfit)
     plot(cvfit)
     coef(cvfit, s =cvfit$lambda.min)
     ms<-
assess.glmnet(cvfit,newx=as.matrix(test[,3:16]),newy=test[,2],fami
ly = "binomial")
```

```
      cc<-c(cc,ms$mse)
    }  ##循环
plot(cc,type="b",ylab="MSE")
which.min(cc)
cvfit <- cv.glmnet(data.matrix(x), y,nfolds = 4,family =
"binomial", type.measure = "class")
cvfit$lambda.min
summary(cvfit)
plot(cvfit)
coef(cvfit, s =cvfit$lambda.min)
ms<-
assess.glmnet(cvfit,newx=as.matrix(test[,3:16]),newy=test[,2],fami
ly = "binomial")
b<-predict(cvfit,newx=as.matrix(test[,3:16]),type="class")
library(Hmisc)
somers2(as.numeric(b),test[,2])
cnf <- confusion.glmnet(cvfit,
newx=as.matrix(test[,3:16]),newy=test[,2])
cnf
#铅钡数据
    library(corrplot)
    getwd()
    setwd("C:\\Users\\HUAWEI\\Desktop")
    a<-read.csv("第四问铅钡数据.csv",encoding = "UTF-8")
    x<-a[1:18,3:16]  #相关化学元素
    y<-a[1:18,2]     #0-1 变量
    library(glmnet)
    set.seed(205)
    select<-sample(1:nrow(a),length(a$X.U.FEFF.文物采样点)*0.7)
    #训练集测试集
    train=a[select,]
    test=a[-select,]
    cc<-c()
    for(i in 3:50){
      cvfit <- cv.glmnet(data.matrix(x), y,nfolds = i ,family =
"binomial", type.measure = "class")
      cvfit$lambda.min
      summary(cvfit)
      plot(cvfit)
      coef(cvfit, s =cvfit$lambda.min)
      ms<-
assess.glmnet(cvfit,newx=as.matrix(test[,3:16]),newy=test[,2],fami
ly = "binomial")
      cc<-c(cc,ms$mse)
```

```
    }  #循环
    plot(cc,type="b",ylab="MSE")
    which.min(cc)
    cvfit <- cv.glmnet(data.matrix(x), y,nfolds = 6,family =
"binomial", type.measure = "class")
    cvfit$lambda.min
    summary(cvfit)
    plot(cvfit)
    coef(cvfit, s =cvfit$lambda.min)
    b<-predict(cvfit,newx=as.matrix(test[,3:16]),type="class")
    library(Hmisc)
    somers2(as.numeric(b),test[,2])
    cnf <- confusion.glmnet(cvfit,
newx=as.matrix(test[,3:16]),newy=test[,2])
    cnf
```

附录 2

Fisher 精确检验以及卡方检验步骤

Fisher 精确检验是基于超几何分布进行的检验，是交叉分类数据，即列联表独立性检验中的一个重要的方法，具体步骤如下：

①根据所给数据分类，进行假设检验，设定原假设和备择假设，并设定显著性水平α；

②对数据预处理，列出交叉分类数据的列联表；

表 27 2×2 列联表

x_2 \ x_1	x_{11}	x_{12}	Row Total
x_{21}	a	b	$a+b$
x_{22}	c	d	$c+d$
Column Total	$a+c$	$b+d$	$a+b+c+d(=n)$

③基于 Fisher 精确检验公式，计算出两类定性数据的 P 值，与显著性水平比较，得出结论：

$$p=\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}=\frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}}$$

①根据所给数据分类，进行假设检验，设定原假设和备择假设，并设定显著性水平α；

②对数据预处理，列出交叉分类数据的频率分布表：

表 28 频率分布表

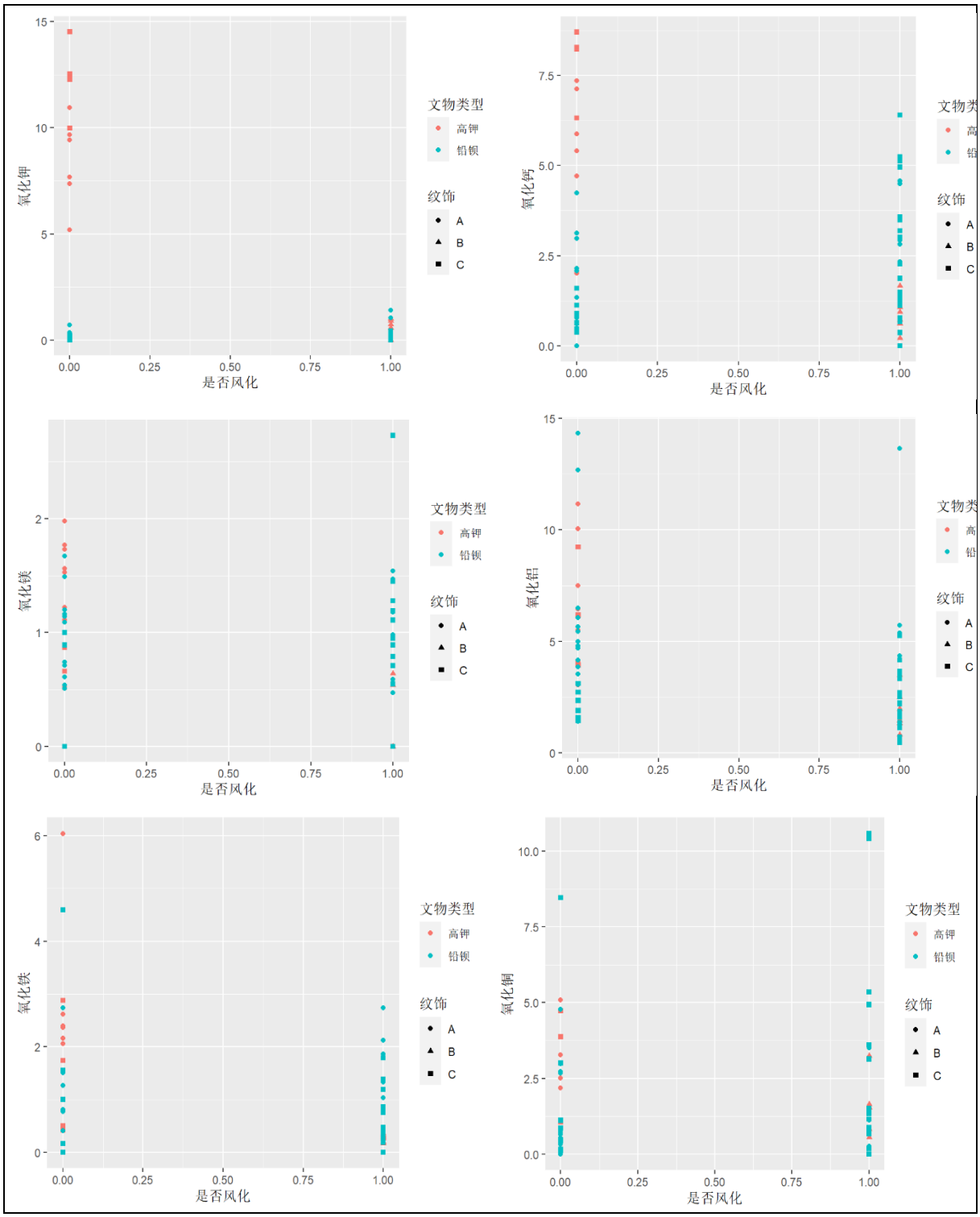
A	**B**					**行和**
	1	…	j	…	c	
1	p_{11}	…	p_{1j}	…	p_{1c}	$p_{1\cdot}$
⋮	⋮		⋮		⋮	⋮
i	p_{i1}	…	p_{ij}	…	p_{ic}	⋮
⋮	⋮		⋮		⋮	⋮
r	p_{r1}	…	p_{rj}	…	p_{rc}	$p_{r\cdot}$
列和	$p_{\cdot 1}$	…	$p_{\cdot j}$	…	$p_{\cdot c}$	1

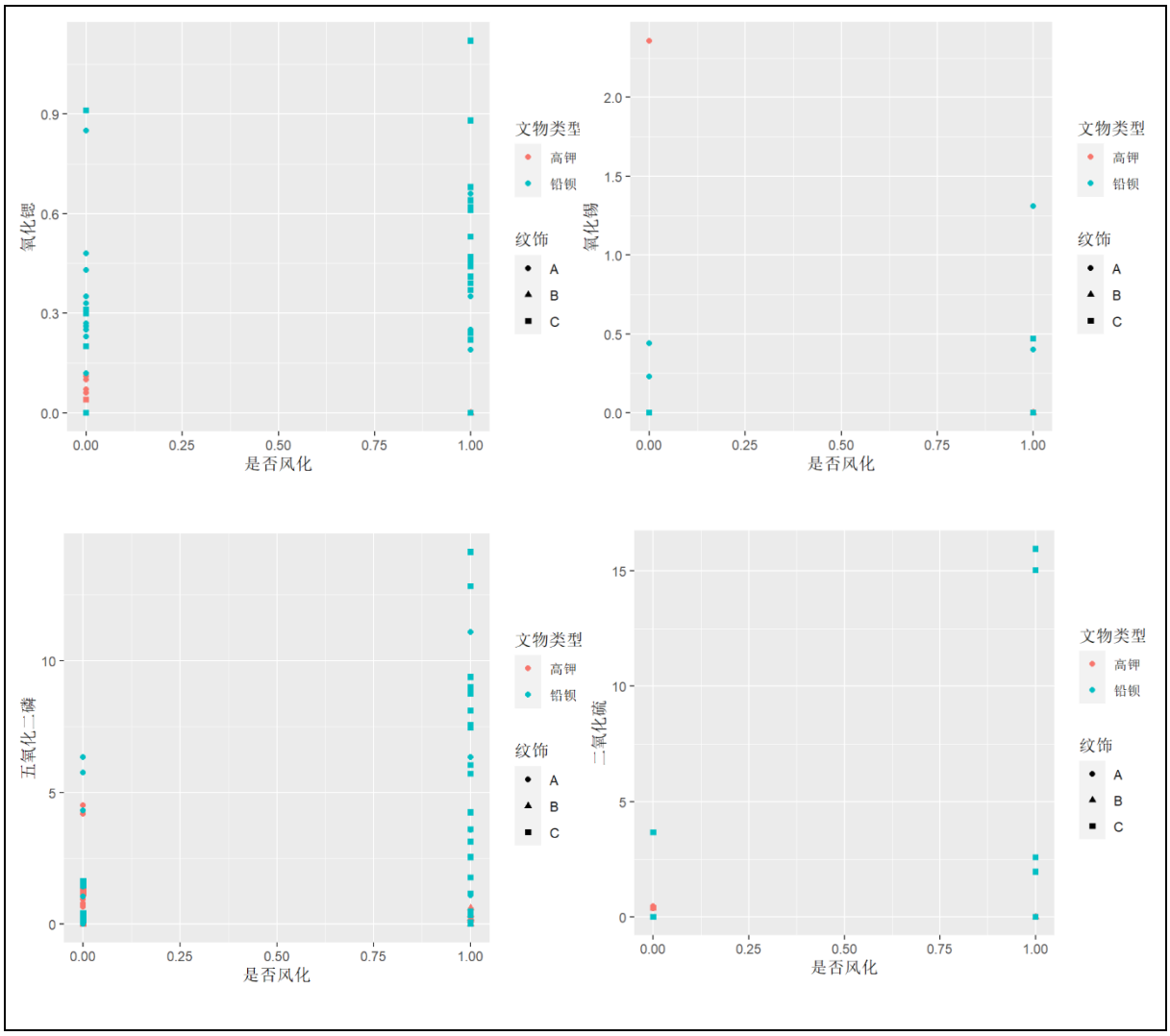
③基于卡方检验公式，计算出两类定性数据的检验统计量以及 P 值，与显著性水平比较，得出结论：

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}}\sim\chi^2\left((r-1)(c-1)\right)$$

检验拒绝域：$W=\{\chi^2\geq\chi^2{}_{1-\alpha}\left((r-1)(c-1)\right)\}$

附录 3
介绍：问题一统计规律分析散点图


氧化钾
是否风化
氧化钙
是否风化
氧化镁
是否风化
氧化铝
是否风化
氧化铁
是否风化
氧化铜
是否风化
文物类型
高钾
铅钡
纹饰
A
B
C



附录 4

介绍：决策树模型实验及其相关代码

```
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import KFold
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
################################################
# 加载数据集
data = pd.read_excel('data.xlsx',sheet_name=1)
labels = list(data.columns.values)
x_data = data.iloc[:,1:16]
y_data = data.iloc[:,16:17]
```

```
x = x_data.values
y = y_data.values
##############################################
def select_best_tree(x,y):
    k_fold = KFold(n_splits=10)
    tree_model = DecisionTreeClassifier(random_state=20)
    params =
{'max_depth':range(1,16),'criterion':np.array(['entropy','gini'])}
    grid = GridSearchCV(tree_model,
param_grid=params,scoring='neg_mean_squared_error',cv=k_fold)
    grid = grid.fit(x, y)
    return grid.best_estimator_
tree_model=select_best_tree(x,y)
tree_model.fit(x,y)
scores=cross_val_score(tree_model, x, y, cv=5,scoring =
"precision_weighted")
print("out:{} mean: {:.3f} (std:
{:.3f})".format(scores,scores.mean(),scores.std()),end="\n" )
##############################################
fn=labels[1:16]
cn=['铅钡','高钾']
##############################################
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus'] = False
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (4,4),
dpi=300)
tree.plot_tree(tree_model,
               feature_names = fn,
               class_names=cn,
               filled = True);
#fig.savefig('imagename.png')

pre_data = pd.read_excel('data.xlsx',sheet_name=2)
pre_x = pre_data.iloc[:,2:17].values
pre_res=tree_model.predict(pre_x)
print(pre_res)
```

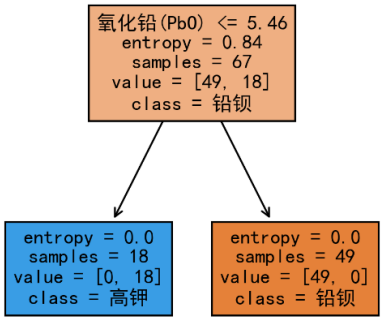


图 17 决策树模型

附录 5

介绍：随机森林模型

```
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import KFold
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
################################################
# 加载数据集
data = pd.read_excel('data.xlsx',sheet_name=1)
labels = list(data.columns.values)
x_data = data.iloc[:,1:16]
y_data = data.iloc[:,16:17]
x = x_data.values
y = y_data.values
################################################
def select_best_tree(x,y):
    k_fold = KFold(n_splits=10)
    tree_model = DecisionTreeClassifier(random_state=20)
    params =
{'max_depth':range(1,16),'criterion':np.array(['entropy','gini'])}
    grid = GridSearchCV(tree_model,
param_grid=params,scoring='neg_mean_squared_error',cv=k_fold)
    grid = grid.fit(x, y)
    return grid.best_estimator_
```

```
##############################################
tree_model=select_best_tree(x,y)

tree_model.fit(x,y)
scores=cross_val_score(tree_model, x, y, cv=5,scoring =
"precision_weighted")
print("out:{} mean: {:.3f} (std:
{:.3f})".format(scores,scores.mean(),scores.std()),end="\n" )

fn=labels[1:16]
cn=['铅钡','高钾']

plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus'] = False
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (4,4),
dpi=300)
tree.plot_tree(tree_model,
               feature_names = fn,
               class_names=cn,
               filled = True);
#fig.savefig('imagename.png')

pre_data = pd.read_excel('data.xlsx',sheet_name=2)
pre_x = pre_data.iloc[:,2:17].values
pre_res=tree_model.predict(pre_x)
print(pre_res)'''
训练需要一定时间但不长,主要时间用于网格搜索部分，接近 300 种模型组合
'''
##############################################
# 导入必要库(pip install numpy,pandas,matplotlib,sklearn,...)
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split,GridSearchCV
from sklearn import metrics
from sklearn.tree import export_graphviz
import os
import pickle
import warnings
warnings.filterwarnings('ignore')
# 预处理已经完成，加载已处理的数据集并划分
data = pd.read_excel('data.xlsx',sheet_name=1)
labels = list(data.columns.values)
```

```
x_data = data.iloc[:,1:16]
y_data = data.iloc[:,16:]
x_train,x_test,y_train,y_test =
train_test_split(x_data,y_data,test_size=0.1,random_state=6)
x_train = x_train.values
x_test = x_test.values
y_train= y_train.values
y_test = y_test.values
# 超参数（已实验收缩）--网格搜索
'''
下述参数网格为经过逐步缩小网格范围并结合参数特点和问题特点的多次实验的结果
参数范围的选择还参考了最后可视化的结果,防止树太高产生过拟合或者树太宽出现数据利
用不充分的状况
采用单一变量的原则，先后缩小：'criterion'、'n_estimators'、'max_depth'、
'min_samples_split'、'max_features'
'''
param_grid = {
    'criterion':['entropy','gini'],
    'n_estimators':[11,13,15,17],
    'max_depth':[5, 6, 7],
    'min_samples_split':[4,8,12,16],
    'max_features':[0.3,0.4,0.5]
}
Rfc_Basic = RandomForestClassifier(random_state = 66)
Rfc_GS = GridSearchCV(estimator=Rfc_Basic, param_grid=param_grid,
                      scoring='accuracy', cv=4)
Rfc_GS.fit(x_train, y_train)

# Rfc_GS.cv_results_
print(Rfc_GS.cv_results_[ 'mean_test_score'])
print(Rfc_GS.cv_results_[ 'std_test_score'])
print('RFC 最优模型参数：',Rfc_GS.best_params_)
Rfc_Best=Rfc_GS.best_estimator_
with open('RFC.pickle','wb') as f:
    pickle.dump(Rfc_Best,f)
print('\n-----------------------------------------------------')
print('  化学成分\t\t\t\t重要性百分比')
# 最优参数模型各特征对结果的重要性
importances = Rfc_Best.feature_importances_
indices = np.argsort(importances)[::-1]
for i in range(15):
    print("%2d)%-*s\t%f" % (i + 1, 30, labels[indices[i]+1],
importances[indices[i]]))
print('-----------------------------------------------------\n')
```

```
pred = Rfc_Best.predict(x_test)
print('测试集中文物真实类型：\t',y_test.tolist())
print('测试集中文物预测类型：\t',pred,'\n')

# 随机森林可视化,可视化代码部分有与 graphviz 的pip install 路径统一的要求,
否则可能无法正常运行
fn=['SiO2','Na2O','K2O','CaO','MgO','Al2O3','Fe2O3','CuO','PbO','B
aO','P2O5','SrO','SnO2','SO2','weathering']
cn=['lead barium','high potassium']
for index, Tree_estimator in enumerate(Rfc_Best):

    export_graphviz(Tree_estimator,
                    out_file='tree{}.dot'.format(index),
                    feature_names=fn,
                    class_names=cn,
                    rounded=True,
                    proportion=False,
                    precision=2,
                    filled=True)
    os.system('dot -Tpng tree{}.dot -o tree{}.png'.format(index,
index))
```

附录 6

介绍：问题三 随机森林模型分类

```
import pickle
import pandas as pd
#导入模型（来自问题 2 的随机森林）
with open('RFC.pickle','rb') as f:
Rfc_Best = pickle.load(f)
# 问题 3 预测
data_pred = pd.read_excel('data.xlsx',sheet_name=2)
x_pred = data_pred.iloc[:,2:17].values
result = Rfc_Best.predict(x_pred)
print('RandomForestClassifier 问题 3 文物类型预测结果：',result)
```

附录 7

介绍：问题四 Matlab

```
clear;clc
load data1.mat;
load data2.mat;
[R1,P1] = corr(data1,'type','Spearman');
[R2,P2] = corr(data2,'type','Spearman');
%figure_youwant
```

```
x1=tril(R1);%对高钾的相关系数进行正态分布检验
x2=tril(R2);
X1=x1(:);
X2=x2(:);
X1(find(X1==0))=[];
X1(find(X1==1))=[];
X2(find(X2==0))=[];
X2(find(X2==1))=[];
qqplot(X1);
qqplot(X2);

X1_excel=abs(X1);
X2_excel=abs(X2);
```

附录 7

介绍：问题一 风化前高钾玻璃文物化学成分含量预测

文物采样点	是否风化	二氧化硅	氧化钠	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
							预测值								
文物采样点															
7	66.65	0.70	8.79	5.53	0.88	6.67	1.84	4.13	0.41	0.60	1.73	0.04	0.20	0.10	
9	69.04	0.70	9.38	5.08	0.88	6.01	1.99	2.44	0.41	0.60	1.47	0.04	0.20	0.10	
10	70.79	0.70	9.71	4.67	0.88	5.50	1.93	1.73	0.41	0.60	1.12	0.04	0.20	0.10	
12	68.31	0.70	9.80	5.18	0.88	6.15	1.96	2.54	0.41	0.60	1.27	0.04	0.20	0.10	
22	66.37	0.70	9.53	6.12	1.52	8.19	2.02	1.44	0.41	0.60	1.33	0.04	0.20	0.10	
27	66.74	0.70	8.79	5.40	1.42	7.20	1.87	2.43	0.41	0.60	1.48	0.04	0.20	0.10	

附录 8

介绍：问题一 风化前铅钡玻璃文物化学成分含量预测

文物采样点	是否风化	二氧化硅	氧化钠)	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
预测值															
文物采样点															
2	66.60	1.55	1.03	0.75	1.14	7.17	1.89	0.00	25.57	0.00	0.00	0.04	0.00	0.00	
8	50.46	1.55	0.00	0.00	0.00	2.78	0.03	9.72	6.82	28.42	0.00	0.22	0.00	1.43	
08严重风化点	34.93	1.55	0.00	1.60	0.00	2.55	0.03	2.45	10.59	27.81	3.57	0.38	0.00	13.88	
11	63.91	1.55	0.19	1.92	0.67	4.13	0.03	4.24	3.53	11.80	5.39	0.22	0.00	0.00	
19	59.96	1.55	0.00	1.34	0.55	5.01	1.36	2.82	20.96	2.54	4.84	0.04	0.00	0.00	
26	50.11	1.55	0.00	0.00	0.00	2.14	0.03	9.88	7.67	29.44	0.00	0.30	0.00	0.81	
26严重风化点	34.04	1.55	0.38	1.42	0.00	2.62	0.03	2.91	8.06	32.64	2.05	0.47	0.00	14.80	
30部位	64.66	1.55	1.39	2.90	0.94	5.79	2.15	0.00	17.36	7.48	0.00	0.20	0.35	0.00	
34	66.10	1.55	0.23	0.00	0.00	3.06	0.50	0.82	24.69	7.19	0.00	0.07	0.00	0.00	

36	69.89	3.77	0.12	0.00	0.00	3.04	0.35	0.00	19.75	8.02	0.00	0.07	0.00	0.00
38	63.25	2.93	0.00	0.00	0.00	4.01	0.32	0.04	27.45	6.98	0.00	0.26	0.00	0.00
39	56.57	1.55	0.00	0.00	0.00	1.94	0.03	0.19	39.17	4.41	0.00	0.46	0.00	0.00
40	47.03	1.55	0.00	0.28	0.00	1.89	0.22	0.00	48.35	3.88	0.00	0.53	0.00	0.00
41	48.78	1.55	0.42	3.37	2.69	4.77	1.82	0.00	22.26	6.95	3.47	0.32	0.00	0.00
43部位	42.73	1.55	0.00	3.65	0.85	3.69	0.79	4.66	37.99	4.48	0.00	0.49	0.00	0.00
43部位	52.02	1.55	0.00	4.81	0.91	4.85	1.42	0.82	22.89	0.45	8.84	0.32	0.00	0.00
48	83.65	2.35	0.30	1.23	1.50	15.09	1.06	0.00	0.00	4.50	0.00	0.10	1.26	0.00
49	59.11	1.55	0.00	2.99	1.43	6.82	2.77	0.01	12.32	3.29	7.11	0.31	0.00	0.00
50	48.30	1.55	0.00	1.60	0.43	3.31	0.36	0.44	22.14	11.39	2.35	0.51	0.00	0.00
51部位	54.93	1.55	0.00	1.99	1.15	6.69	1.22	0.68	18.38	6.13	4.11	0.24	0.42	0.00
51部位	51.67	1.55	0.00	3.54	1.41	3.95	0.45	0.06	29.48	0.00	4.76	0.00	0.00	0.00
52	56.06	2.77	0.00	0.68	0.51	2.60	0.26	0.01	25.56	5.83	1.72	0.29	0.00	0.00
54	52.60	1.55	0.30	1.60	1.24	5.59	0.03	0.14	33.60	4.23	0.25	0.73	0.00	0.00
54严重	47.43	1.55	0.00	0.00	1.07	5.09	0.03	0.65	36.60	0.00	10.14	0.97	0.00	0.00

风化点														
56	59.47	1.55	0.00	0.00	0.00	3.29	0.03	0.10	19.39	12.64	0.00	0.00	0.00	0.00
57	55.74	1.55	0.00	0.00	0.00	3.62	0.03	0.47	23.24	14.49	0.00	0.00	0.00	0.00
58	60.71	1.55	0.32	1.90	0.75	4.96	0.89	2.44	17.49	4.85	5.00	0.09	0.00	0.00

附录 9

介绍：熵权法

```
    [a,b] = size(X);%一共有四个 mat 文件，其中的变量名均为 X，故运行前请依次导入四个 mat 文件，mat 文件请见支撑材料
%分别为 Pb_Ba_without_weathering.mat，Pb_Ba_weathering.mat，High_K_without_weathering.mat，High_K_weathering.mat
H = zeros(1,b);
for i = 1:b
    v = X(:,i);
    p = v / sum(v);
    e = -sum(p .* log(p)) / log(a);%e 为信息熵
    H(i) = 1- e; %信息的效用值
end
    W = H ./ sum(H);  % W 为最后各化学成分指标的权重
```