古代玻璃制品的成分分析与鉴别

摘要

丝绸之路是古代中西方文化交流的重要通道，玻璃是丝绸之路早期贸易往来的宝贵物证，研究其成分及其鉴别方法对于丝绸之路连绵深远的文化影响具有重要意义。

针对问题一，提取表面风化与玻璃纹饰、类型、颜色数据，做出直方图；利用卡方检验，得出表面风化与玻璃类型有较强的相关性。分别统计分析不同类型有无风化的化学成分含量数据，做描述性统计并检验正态性，画出四组折线图对比分析不同类型有无风化的化学成分含量变化规律；得到高钾风化数据均呈现正态分布，其余组数据均有大量元素不满足正态分布；高钾玻璃风化后，氧化钾含量衰减最为厉害；铅钡玻璃风化后，氧化铅、氧化钡占比上升明显。对于每类玻璃计算每个成分含量的平均值，结合风化前后的均值变化规律建立函数关系，针对风化点数据，往前推得到风化前的预测结果。

针对问题二，用随机森林模型对表单 2 的数据进行分类训练，并利用 OOB 数据估计输入指标的重要性，再计算重要指标在两类玻璃中平均含量，分析其差异得出当玻璃的 BaO 和 PbO 占比皆不超过 1%，K_2O 占比接近 6.4%，Al_2O_3 占比接近 5% 时，可分类为高钾玻璃，当 BaO 和 PbO 占比皆大于 9% 左右，K_2O 占比不超过 1%，Al_2O_3 占比接近 4.5% 时，可分类为铅钡玻璃。使用向后回归法分别筛选出适合高钾、铅钡玻璃进行亚分类的合适化学成分，高钾玻璃筛选出的成分是 CaO、Al_2O_3、SrO，铅钡的是 CaO、Fe_2O_3、SnO，将上述指标设为 X，Y，Z 分别建立基于 K-Means 算法的聚类亚类划分模型，利用手肘图确定了最优 k 值，将高钾玻璃中 18 个采样点数据分为两类，又将铅钡玻璃中的 49 个采样点分为三类，最终得到了高钾玻璃采样点的划分结果为高钙高铝型和低钙低铝低锶型，铅钡玻璃采样点的划分结果为低钙低铁低锡型、中钙低铁型和高钙中铁型。将数据降维，列出完备组合方式，重新聚类分析，分析轮廓图及其指标，认为原划分的结果具有高度合理性。设定两组以 CaO 为扰动对象，扰动量为 5% 以及-5% 的数据进行灵敏性测试，结果显示模型优良，有较强的鲁棒性。

针对问题三，首先采用决策树和逻辑回归模型对训练集进行分类训练，经过对比发现，决策树分类准确性高但其分支只有 2 个，不能保证预测的准确性，因此采用逻辑回归模型；接着根据问题二的分类指标重要性结果，对模型进行特征选择，得到当模型的自变量为 7 个最重要的指标时，AUC 值达到 1，即分类全部正确，对附件表单 3 进行预测得到结果为：高钾、铅钡、铅钡、铅钡、铅钡、高钾、高钾和铅钡；最后进行敏感性分析，将指标重要性最高的自变量 BaO 和 PbO 的含量在-5% 到 5% 波动，模拟外界的干扰，最终得到预测结果不变，模型具有较好的稳定性，有较强的鲁棒性。

针对问题四，采用皮尔逊相关系数、热力图分析各化学成分之间的关联关系；得到在高钾玻璃中二氧化硅与氧化钾、氧化钙、氧化铝的关系非常紧密等五组非常紧密关系；铅钡玻璃中两组非常紧密关系。各提取出 14 个化学元素两两之间，共计 91 个对应关系，用两组皮尔逊相关系数进行成对样本 T 检验。得到铅钡化学成分之间的关联关系明显弱于高钾，两种关联关系差异明显。

关键字：卡方检验　随机森林　向后回归法　K-Means 算法　决策树　逻辑回归

一、 问题重述

1.1 问题背景

丝绸之路是古代中西方文化交流的通道，是 21 世纪国家的重要战略工程之一，其影响广泛而深远。古代玻璃作为一种古代贸易往来的宝贵物证，早期的玻璃在西亚和埃及地区常被制作成珠形饰品传入我国，我国古代玻璃吸收其技术后在本土就地取材制作，因此与外来的玻璃制品外观相似，但化学成分却不相同。因此，研究其化学成分及鉴别方法，对于丝绸之路久远的历史文化传播具有重要意义。

玻璃的主要原料是石英砂，主要化学成分是二氧化硅（SiO2）。由于纯石英砂的熔点较高, 为了降低熔化温度，在炼制时需要添加助熔剂。古代常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等，并添加石灰石作为稳定剂，石灰石煅烧以后转化为氧化钙（CaO）。添加的助熔剂不同，其主要化学成分也不同。例如，铅钡玻璃在烧制过程中加入铅矿石作为助熔剂，其氧化铅（PbO）、氧化钡 BaO 的含量较高，通常被认为是我国自己发明的玻璃品种，楚文化的玻璃就是以铅钡玻璃为主。钾玻璃是以含钾量高的物质如草木灰作为助熔剂烧制而成的, 主要流行于我国岭南以及东南亚和印度等区域。

古代玻璃极易受埋藏环境的影响而风化。在风化过程中，内部元素与环境元素进行大量交换，导致其成分比例发生变化，从而影响对其类别的正确判断。在部分风化的文物中，其表面也有未风化的区域。

现有一批我国古代玻璃制品的相关数据, 考古工作者依据这些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。附件表单 1 给出了这些文物的分类信息，附件表单 2 给出了相应的主要成分所占比例（空白处表示未检测到该成分）。这些数据的特点是成分性，即各成分比例的累加和应为 100%，但因检测手段等原因可能导致其成分比例的累加和非 100% 的情况。本题中将成分比例累加和介于 85% ~ 105% 之间的数据视为有效数据。

1.2 问题提出

鉴于以上背景，本文需要建立数学模型解决以下问题：

1. 对这些玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系进行分析；结合玻璃的类型，分析文物样品表面有无风化化学成分含量的统计规律并根据风化点检测数据，预测其风化前的化学成分含量。
2. 依据附件数据分析高钾玻璃、铅钡玻璃的分类规律；对于每个类别选择合适的化学成分对其进行亚类划分，给出具体的划分方法及划分结果并对分类结果的合理性和敏感性进行分析。

3. 对附件表单 3 中未知类别玻璃文物的化学成分进行分析，鉴别其所属类型并对分类结果的敏感性进行分析。
4. 针对不同类别的玻璃文物样品，分析其化学成分之间的关联关系，并比较不同类别之间的化学成分关联关系的差异性。

二、 问题分析

2.1 研究现状分析

在考古界，对于出土玻璃制品的研究有着十分完善的系统方案。针对题目内容，查阅了众多相关度高的文献，对于一些有助于解题的重要文献中的有效内容做如下综述。

依据制品所使用的主要助熔剂（含量不低于 10%[2]），确定玻璃制品的成分体系。铅钡玻璃的成分体系为 $PbO-BaO-SiO_2$[4]; 钾玻璃的成分体系为 K_2O-SiO_2[1].

由于风化程度的不同，玻璃制品的 K_2O 含量呈现较大的变化。风化较弱的样品，K_2O 含量在 7%~15% 之间；中等风化的样品，K_2O 含量在 4%~6% 之间；风化严重的样品，K_2O 含量低于 4%[1] 甚至低于 1%[2].

根据钾玻璃中 CaO 和 Al_2O_3 的浓度，以往研究将亚洲各地发现的古代钾玻璃进一步划分为中等钙铝（$m-K-Ca-Al$）、低钙（$m-K-Al$）和低铝（$m-K-Ca$）3 个亚类，m 表示样品中的 K_2O 来自富钾矿物。在不同亚类钾玻璃中微量元素 Rb,Sr 的含量和比例页显示出不同的特征，根据这个明显的特征可以明显划分 2 个亚类[1]。

2.2 问题分析

首先对数据进行数据清洗，利用统计学知识和文献综述对数据中的异常值和缺失值进行合理地处理。

2.2.1 问题一分析

为探究数据之间的关系，先做出数据的直方图，然后提出假设使用卡方检验得到显著性，最后对数据进行分析，得到数据之间的相关性结论。针对不同类型有无风化的四组数据做描述性统计分析，计算偏度峰度 Z 得分检验其正态性；做出各组数据的折线图，对比分析不同类型风化前后的化学成分含量变化规律。计算各类别玻璃风化前和风化后的每个化学成分含量的平均值，对这两个平均含量求比值，根据其比值，用风化后的数据预测其风化前的化学成分含量。

2.2.2 问题二分析

首先对表单 2 的数据使用随机森林模型进行分类学习，自变量为 14 个化学成分以及 1 个表面风化指标，接着利用其可以估计输入指标重要性的特性，得出重要指标在两类玻璃中的平均占比，分析得到分类规律。先通过向后回归法筛选出主要的化学成分，

然后利用主要成分确定 X、Y、Z，在通过手肘图得到最优分类数，最后通过 K-Means 算法画出聚类散点图得到分类结果，为探究分类结果的合理性，先假设几组由其他指标或者少考虑一个指标组成的数据，做出他们的轮廓图、聚类散点图以及手肘图，分析这些分组结果，最终验证分类结果的合理性，在讨论灵敏性的时候，设置扰动程度，再重新使用模型得到分类结果的灵敏性结论。

2.2.3 问题三分析

问题三要鉴别附件表单 3 中未知类别玻璃文物的所属类型，首先对于 15 个特征变量建立起逻辑回归模型，计算 AUC 值观察分类效果。接着根据问题二的结果提取出一定数量的重要特征变量，同样地，对这些重要特征变量进行逻辑回归并计算 AUC 值，通过比较 AUC 值选择出分类效果最好的模型。最后对分类结果进行敏感性分析，将重要性最高的两个化学成分变量 BaO 和 PbO 进行一定范围内扰动，观察输出结果的变化情况。

2.2.4 问题四分析

针对两种玻璃制品的数据，借助皮尔逊相关系数分析各化学成分之间的关联关系，用皮尔逊相关系数反应，并画出热力图直观展示各元素之间的相互影响程度。针对两组关联关系的差异性探讨，选择配对样本 T 检验发现其关联关系之间的差异性。

三、 模型假设

假设 1：各样本风化的衰减程度相同；
假设 2：样本数据总体满足正态分布或近似正态分布；
假设 3：实验数据准确可靠，不存在粗大误差；
假设 4：扰动项服从独立的正态分布。

四、 符号说明

符号	描述
x_{1i}	高钾玻璃风化后第 i 个化学成分含量
x_{2i}	铅钡玻璃风化后第 i 个化学成分含量
y_{1i}	预测的高钾玻璃风化前第 i 个化学成分含量
y_{2i}	预测的铅钡玻璃风化前第 i 个化学成分含量

五、 数据准备

“附件.xlxs”中的表单 1 中给出了 57 组针对不同的古代玻璃观测数据，表单 2 中给出了针对 58 个文物的 69 个不同的采样点的化学成分数据，表单 3 中给出了未分类的文物化学成分数据。

5.1 表单 1 数据清洗

针对表单 1 中的数据，首先进行基础数据清洗，利用 Excel 中的筛选功能识别缺失值，筛选出文物编号为 19、40、48 以及 58 这四组，上述四组颜色值都是空白值，将这些数据通过直接删除方式剔除，以免数据对后续的研究产生影响。且本文发现表单 1 中的数据均为定性数据，为了后面使用 SPSS 操作时能够正常运行，本文将表单 1 中出第一行数据，其他所有均用数字代替例如：风化：1，无风化：0

5.2 表单 2，表单 3 数据清洗

针对表单 2 中的数据，首先进行基础数据清洗。根据题干中“空白处表示未检测到该成分”，利用 Excel 表格中替换功能将空白处数据全部替换成 0，再根据题干中“成分比例累加和介于 85%-105% 之间的数据视为有效数据”的意思，利用 Excel 表格中的函数进行简单的计算发现两组异常数据，文物采样点为 15 和 17 的两组数据的成分比例均超小于 85%，为了不对下面的研究产生影响，本文剔除了这两组异常数据。

六、 模型建立与求解

6.1 问题一的模型建立

题目要求我们探讨表面风化与其玻璃纹饰、类型以及颜色的关系；对不同类型有无风化的化学成分含量进行统计规律探索；对风化点的数据进行预测，得到其风化前各成分含量的占比。

初步分析可知：由于关于表面风化与玻璃纹饰、类型以及颜色的相关数据为定性数据，且数据不满足正态分布，且根据表单一中给出的相关数据本文得到了相关数据的频数表表 1。

表 1　表单一相关数据的频数表

纹饰	类型	颜色	表面风化	频数	纹饰	类型	颜色	表面风化	频数
A	高钾	蓝绿	无风化	5	A	高钾	深蓝	无风化	1
A	铅钡	黑	风化	2	A	铅钡	蓝绿	风化	1
A	铅钡	浅蓝	风化	6	A	铅钡	蓝绿	无风化	4
A	铅钡	深蓝	无风化	1	B	高钾	蓝绿	风化	6
C	高钾	蓝绿	无风化	1	C	高钾	浅蓝	无风化	4
C	高钾	深绿	无风化	1	C	铅钡	蓝绿	风化	2
C	铅钡	绿	无风化	1	C	铅钡	浅蓝	风化	6
C	铅钡	浅绿	无风化	2	C	铅钡	浅绿	风化	1
C	铅钡	深绿	无风化	2	C	铅钡	深绿	风化	4
C	铅钡	紫	无风化	2	C	铅钡	紫	风化	2

基于以上分析，本文以表面风化分别与玻璃纹饰、类型以及颜色两两配对，分析配对数据是否满足卡方检验的前提条件，而且根据生活常识可知有无风化的同一指标的数据是符合卡方检验的条件的。

接着我们就可以对数据进行多组样本数据的卡方检验，从而对有无表面风化对玻璃纹饰、类型以及颜色的影响结果的显著性差异进行检验。

6.1.1 卡方检验简介

统计知识指出：

卡方检验是一种用途很广的计数资料的假设检验方法。属于非参数检验，主要是比较两个及两个以上样本率（构成比）以及两个分类变量的关联性分析。

卡方检验的前提条件为[3]：

（1）若 $n \geq 40$，且任意一个格子的理论频数 $T_{ij} \geq 5$，可以直接使用 χ^2 检验公式。

$$\chi^2 x^2 = \sum \frac{(A-T)^2}{T} \tag{1}$$

（2）若 $n \geq 40$，但出现 1 个格子的理论频数 $1 \geq T_{ij} < 5$ 时，则需对值按以下公式进行连续性校正。

$$\chi^2 = \sum \frac{(|A-T|-0.5)^2}{T} \tag{2}$$

3）若 $n < 40$ 或者任意一个格子的理论频数 $T_{ij} < 1$ 时，则检验不再适用，宜采用 Fisher 确切概率法进行处理。

其中，n 为样本量。

卡方检验的原理是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，卡方值越大，越不符合；卡方值越小，偏差越小，越趋于符合，若两个值完全相等时，卡方值就为 0，表明理论值完全符合。该检验相应的假设为：

$H_0 : p < 0.05$ 两者之间存在关系，$H_1 : p \geq 0.05$ 两者之间不存在关系。

6.1.2 化学成分含量的统计规律

基于统计学知识，将玻璃的高钾，铅钡类型分开考虑；再以是否风化为标准分别探讨玻璃制品化学成分含量的统计规律，分为高钾风化、高钾无风化、铅钡风化和铅钡无风化四类数据。考察其数据的范围、最小值、最大值、均值、标准偏差、方差、偏度、峰度和标准错误，并计算其 Z 评分来检验其正态性，以反映其化学成分含量的统计规律。其标准错误的计算公式如表3, 偏度 Z 评分、峰度 Z 评分的计算公式如4所示。

$$SEM = \frac{SD}{\sqrt{n}} \tag{3}$$

$$\text{偏度}Z_{score} = \frac{\text{偏度}}{\text{偏度}SEM} \qquad \text{峰度}Z_{score} = \frac{\text{峰度}}{\text{峰度}SEM} \tag{4}$$

SD 为标准差，SEM 为标准错误，n 为样本大小。Z 评分的值在-1.96~+1.96 之间，则认为数据满足正态分布，否者不然。

针对不同类型有无风化的四组数据，画出其折线图并进行对比，分析出其含量的占比及变化规律。

6.1.3 玻璃风化前后化学成分含量的函数关系

对每一类别的玻璃文物，计算出每个化学成分风化前后的平均含量，再把风化后的平均含量除以风化前的平均含量，得到风化前后平均含量的比值，用该比值便能得到风化前的化学成分含量。因此建立起如下的函数关系：

$$y_{1i} = x_{1i} \cdot \left(\frac{B_{1i}}{F_{1i}}\right)^{-1} \tag{5}$$

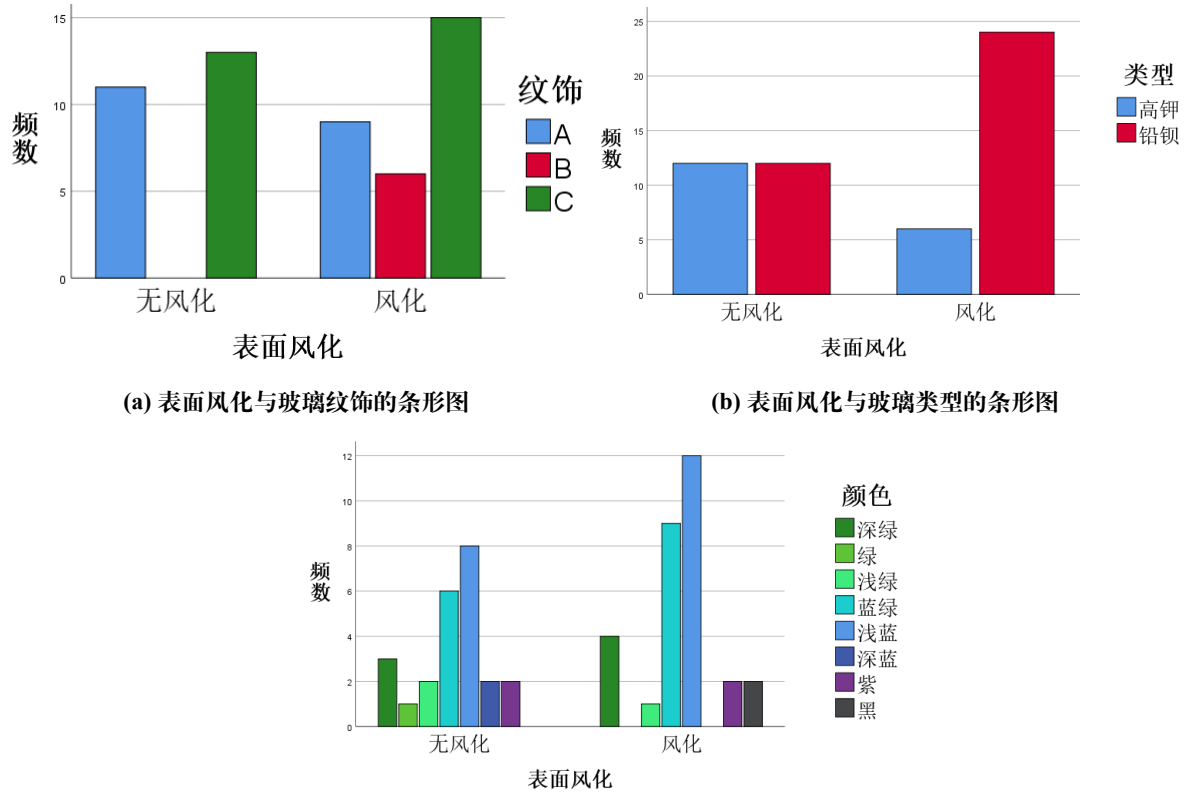
$$y_{2i} = x_{2i} \cdot \left(\frac{B_{2i}}{F_{2i}}\right)^{-1} \tag{6}$$

其中 F_{1i}、F_{2i} 分别为高钾玻璃、铅钡玻璃风化前的第 i 个化学成分的平均含量，B_{1i}、B_{2i} 分别为高钾玻璃、铅钡玻璃风化后的第 i 个化学成分的平均含量，x_{1i}、x_{2i} 分别为已知的高钾玻璃、铅钡玻璃风化后的第 i 个化学成分含量，y_{1i}、y_{2i} 分别为预测的高钾玻璃、铅钡玻璃风化前的第 i 个化学成分含量。

6.2 问题一的模型求解

6.2.1 表面风化与玻璃纹饰、类型、颜色的关系

根据表 1中的数据提取表面风化与玻璃纹饰、玻璃类型、颜色的相关数据将这三组数据代入 SPSS 中进行条形图以及卡方检验得到 p 值如下：



(a) 表面风化与玻璃纹饰的条形图

(b) 表面风化与玻璃类型的条形图

(c) 表面风化与玻璃颜色的条形图

图 1　工艺参数与结构变量的多项式函数拟合图

表 2　卡方检验 P 值

类别	卡方检验 p 值
表面风化与玻璃纹饰	0.056
表面风化与玻璃类型	0.02
表面风化与玻璃颜色	0.507

通过上表进行显著性分析：

（1）表面风化与玻璃纹饰的显著性分析问题附件中已给出表面风化与玻璃纹饰的相关数据表格，本文通过统计频数对相同类型利用加权实现数据的完整性。对于表面风化

结果构建卡方检验。得到显著性概率 $p = 0.056 > 0.05$。则拒绝原假设，即玻璃纹饰与表面风化无显著性相关。

（2）表面风化与玻璃类型的显著性分析问题附件中已给出表面风化与玻璃类型的相关数据表格，本文通过统计频数对相同类型利用加权实现数据的完整性。对于表面风化结果构建卡方检验。得到显著性概率 $p = 0.02 < 0.05$。则接受原假设，即玻璃类型与表面风化有显著性相关。

（3）表面风化与玻璃颜色的显著性分析问题附件中已给出表面风化与玻璃颜色的相关数据表格，本文通过统计频数对相同类型利用加权实现数据的完整性。对于表面风化结果构建卡方检验。得到显著性概率 $p = 0.507 > 0.05$。则拒绝原假设，即玻璃颜色与表面风化无显著性相关。

综上所示，表面风化与玻璃纹饰和颜色无显著性相关，当玻璃发生表面风化时，在玻璃表面的类型中铅钡的占比较高，纹饰 B 可能在玻璃发生风化后产生。

6.2.2 化学成分含量的描述性统计

按照分类的数据，借助 SPSS 软件做描述性统计分析分别得到四组统计数据，展示玻璃制品类型为高钾，表面风化为风化的一类数据的描述统计结果如表3所示。

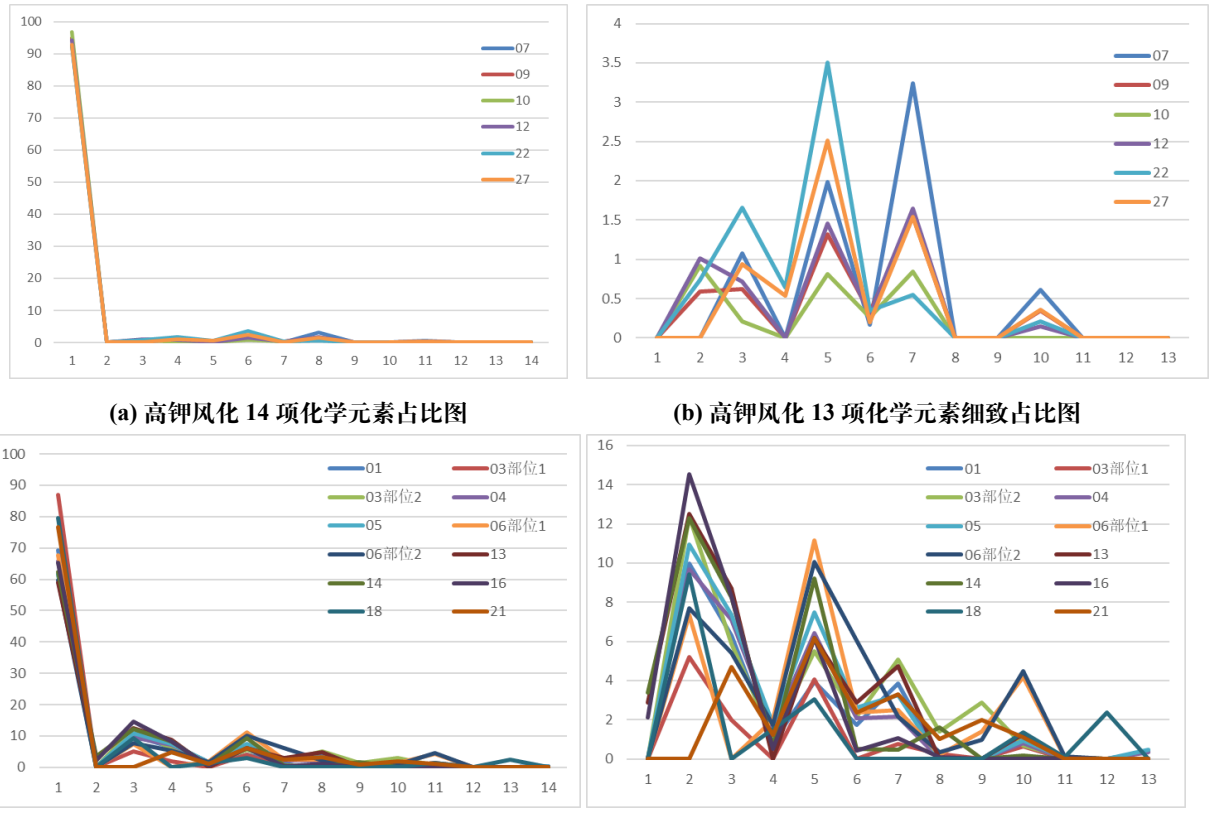
表 3　高钾风化描述性统计

统计指标	SiO_2	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO	P_2O_5
范围	4.42	1.01	1.45	0.64	2.69	0.18	2.69	0.61
最小值	92.35	0	0.21	0	0.81	0.17	0.55	0
最大值	96.77	1.01	1.66	0.64	3.5	0.35	3.24	0.61
均值	93.963	0.543	0.87	0.197	1.93	0.265	1.562	0.28
均值 SEM	0.708	0.182	0.199	0.125	0.394	0.028	0.382	0.086
标准偏差	1.734	0.445	0.488	0.306	0.964	0.069	0.935	0.21
方差	3.005	0.198	0.238	0.094	0.93	0.005	0.874	0.044
偏度	0.854	-0.537	0.504	1.014	0.779	-0.3	1.218	0.399
偏度 SEM	0.845	0.845	0.845	0.845	0.845	0.845	0.845	0.845
峰度	-0.388	-1.913	0.988	-1.598	0.181	-1.418	2.231	0.372
峰度 SEM	1.741	1.741	1.741	1.741	1.741	1.741	1.741	1.741

对 14 个化学元素均做了描述性统计，表格中未有的化学元素，其统计指标均为 0. 其余三组数据的统计指标见附录。利用偏度、峰度及其两者的标准错误，判定了这四组

数据的正态分布情况。高钾风化的的数据均满足正态分布；高钾无风化的数据有七个化学成分数据不满足正态分布；铅钡有无风化的数据中，均有九个化学成分数据不满足正态分布。

针对四组数据中的每一项文物样点做出其化学元素成分占比的折线图。玻璃制品类型为高钾，表面有无风化的折线如图2所示



(a) 高钾风化 14 项化学元素占比图

(b) 高钾风化 13 项化学元素细致占比图

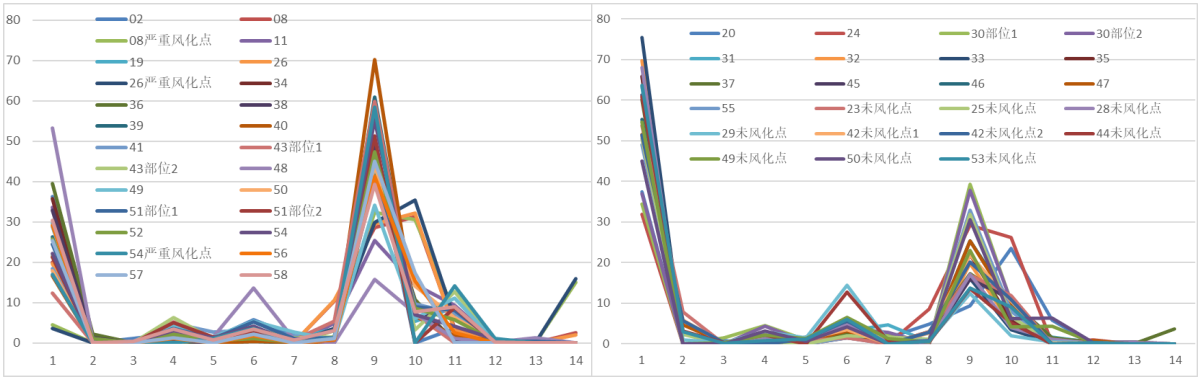
(c) 高钾无风化 14 项化学元素占比图

(d) 高钾无风化 13 项化学元素细致占比图

图 2　高钾优无风化化学元素占比图

图11a为元素整体占比规律图，图11b为图11a去除二氧化硅元素而放大的规律折线图，其更能反应其余十三种元素的变化规律；高钾无风化的元素占比图也是同样处理。从中可以看出，在众多文物样点中，各化学元素的成分占比具有高度的共线性，占比规律非常相似。其中二氧化硅的占比最大，不同文物样点在不同元素上有着不同但相似的占比份额。高钾玻璃在风化后，二氧化硅含量上升，其余成分含量下降，以氧化钾最为明显。

玻璃制品类型为铅钡，表面有无风化的折线如图3所示



(a) 铅钡风化 14 项化学元素占比图　　**(b) 铅钡无风化 13 项化学元素细致占比图**

图 3　铅钡有无风化化学元素占比图

从中可以看出，在众多文物样点中，各化学元素的占比呈现出强烈的共线性，二氧化硅的占比均为最大，不同文物样点在不同元素上有着相似但不同的占比份额。铅钡玻璃在风化后，二氧化硅的含量衰减严重；氧化铅、氧化钡的占比升高。

6.2.3 风化前的化学成分含量的预测结果

根据风化前后化学成分含量的函数关系，得出风化前的化学成分含量的部分预测结果如表 4和表 5所示，剩余结果见附录图 18和图 19。

表 4　高钾玻璃风化前化学含量的预测结果

文物编号	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO
07	67.02	0	0	6.56	0	6.79	1.24	5.09
09	68.75	0	10.13	3.80	0	4.53	2.33	2.43
10	70.01	0	15.80	1.29	0	2.78	1.90	1.32
12	68.22	0	17.35	4.41	0	5.01	2.11	2.59
22	66.82	0	12.71	10.17	3.51	12.01	2.55	0.86
27	67.08	0	0	5.76	2.96	8.61	1.46	2.42

表 5　铅钡玻璃风化前化学含量的预测结果

文物编号	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO
23	51.12	2.13	0	0.42	0.36	0.74	0	3.67
25	48.09	0.62	0	0.53	0	1.00	3.01	1.37
28	64.70	0	0.40	1.12	0.51	2.46	0.80	0.40
29	60.15	0.25	0.46	2.49	0.76	7.51	1.57	0.91
42	48.71	1.54	0.23	0.66	0.56	1.85	0	3.28
42	48.78	1.53	0.54	0.00	0.59	2.97	0	3.34
44	57.72	0.82	0.31	1.79	0	6.65	1.49	0.53
49	51.89	0	0.46	1.74	0.61	3.41	2.46	0.55
50	42.78	0	0	2.61	0.28	2.18	0	0.86
53	60.49	0.82	0.17	0.65	0.58	3.18	0	0.66

把预测出来的化学成分含量比例进行累加，得到的结果都在 85% ～ 105% 之间，是有效数据，因此预测具有一定的准确性。

6.3 问题二的模型建立

要求探索分析高钾、铅钡玻璃的分类规律；对每个类别选择合适的化学成分进行亚类划分，给出划分方法及结果，并检验其合理性与敏感性。

6.3.1 基于随机森林的指标重要性计算

为了探讨高钾玻璃和铅钡玻璃的分类规律，首先使用随机随林计算输入指标重要性，提取出分类中的重要指标，再计算重要指标在两种玻璃的平均占比，最终通过对比占比差异得到分类规律。随机森林是利用多个决策树对样本进行训练并预测的一种集成分类器，通常比个体学习器表现的效果更好，而且能够估计输入指标的重要性，使用随机森林模型进行分类的基本思想如图 4所示。

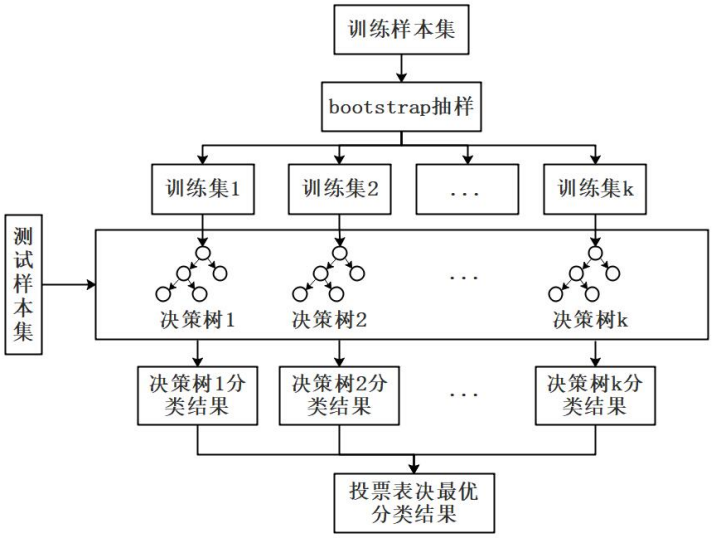


图 4　随机森林算法流程图

依据随机森林模型的基本原理，下面针对高钾玻璃与铅钡玻璃的分类设计算法流程。

步骤 1：确定训练集。训练集选用表单 2 中的 14 个化学成分，此外，在问题一的分析中可知，不同类型的玻璃有着不同的风化程度，因此将该指标加入训练集，一共构成 15 个指标。

步骤 2：对训练集进行 bootstrap 抽样，即从训练集中抽取 k 个样本集，并且每个样本容量与训练集的相同[5]。

步骤 3：对每个样本集分别建立各自的决策树模型，得到组合分类器，得到 k 种分类结果。

步骤 4：对每个样本进行投票表决得到最优的分类结果，即有

$$S(x) = \arg\max_{Y} \sum_{i=1}^{k} I\left(s_i(x) = Y\right) \tag{7}$$

其中 $S(x)$ 是组合分类模型的结果，$s_i(x)$ 是单个决策树的分类结果，Y 是因变量，$I(\cdot)$ 是示性函数。

步骤 5：利用袋外数据（OOB）计算 15 个特征变量的重要性。首先使用 OOB 数据计算每一颗决策树的袋外数据误差，记为 $err1$，接着随机对袋外数据的某个特征变量加入噪声值，再次计算袋外数据误差，记为 $err2$，则某个特征变量的重要性为：

$$\frac{\sum(err2 - err1)}{k} \tag{8}$$

从公式可以看出，当对某个特征变量加入随机噪声后，袋外数据的准确率会下降，则 $err2$ 的值会增大，说明这个特征对于样本的预测分类结果有较大影响，进而说明其重要程度比较高。因此利用随机森林拥有这样的特性，便能够对特征变量进行重要性排序。

6.3.2 基于向后回归法筛选合适的化学成分

首先建立起一个全因素的回归方程；再去掉其中一个因素，建立起各回归方程；剔除使方程残差平方和减少最少的自变量，逐一剔除；直到剔除自变量不会使残差平方和显著减小为止。残差平方和计算如公式9。

$$\left(pv^2\right) = p_1v_1^2 + p_2v_2^2 + \ldots + p_nv_n^2 \tag{9}$$

式中 v_i 是测量数据 l_i 的残差，p_i 为相应的权。对玻璃的类型进行亚类划分，应当选择偏离成分回归方程远的的自变量来区分，以足以有明确的区分度。

6.3.3 基于 K-Means 算法的亚类划分模型

在上述模型中已经探讨出了高钾玻璃以及铅钡玻璃中的主要指标。考虑到这些化学成分之间的关系以及题意的，且数据已经完成了数据的预处理，因此考虑 K-Means 算法对数据进行划分，使用 K-Means 算法对样本点进行分析的基本思想如图 5所示。

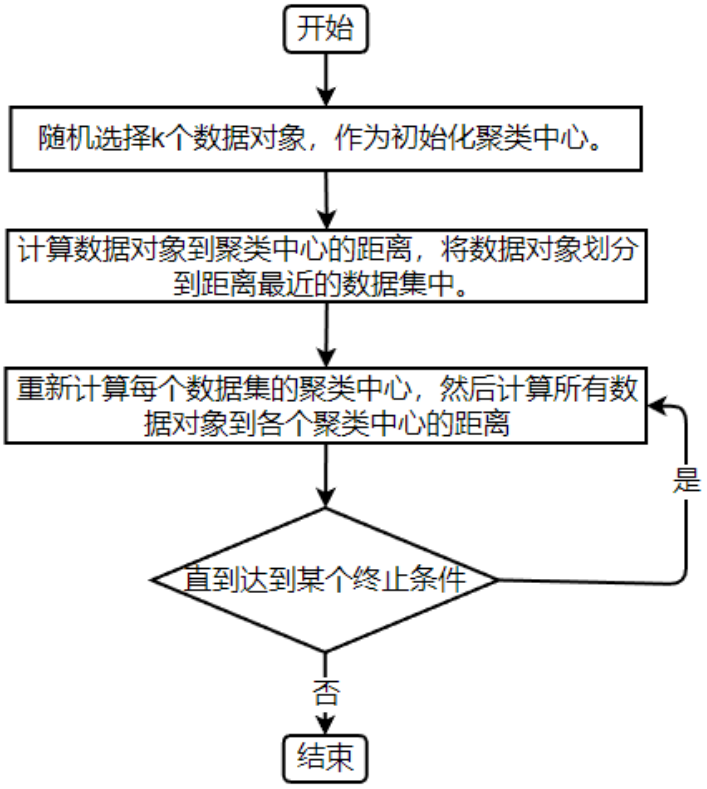


图 5　K-Means 算法流程图

根据 K-Means 算法的基本思想，下面是分别针对高钾玻璃和铅钡玻璃的亚类划分的 K-Mean 算法流程：

步骤 1：step1: 选择初始化的 k 个采样点样本作为初始聚类中心 $a = a_1, a_2, ...a_k$；

步骤 2：针对数据集中每个样本 x_i，计算它到 k 个聚类中心的距离并将其分到距离最小的聚类中心所对的类中；

步骤 3：针对每个类别 a_j，重新计算它的聚类中心 $a_j = \frac{1}{|c_i|}\sum_{x \in c_i} x$(即属于该类的所有样本的质心)；

步骤 4：重复上面步骤 2 和步骤 3 两步操作，直到达到某个终止条件（迭代次数、最小误差变化等）；

上面已经讨论了基于 K-Means 算法的聚类算法基本流程，下面讨论 K-Means 算法中的最优 k 值确定方法-手肘法。

假设将数据样本划分为 k 类，随着聚类数 k 的逐渐增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和 SSE 自然会逐渐变小。SSE 的计算公式如下 (10)：

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{10}$$

其中，C_i 是第 i 个簇，p 是 C_i 中的样本点,m_i 是 C_i 的质心

6.4 问题二的模型求解

6.4.1 高钾玻璃和铅钡玻璃的分类规律

利用 MATLABx 工具箱进行随机森林模型的求解，确定随机森林的学习器数量为 30，最大分裂数为 66，交叉验证为 5 次，计算得到模型的混淆矩阵如图 6所示。

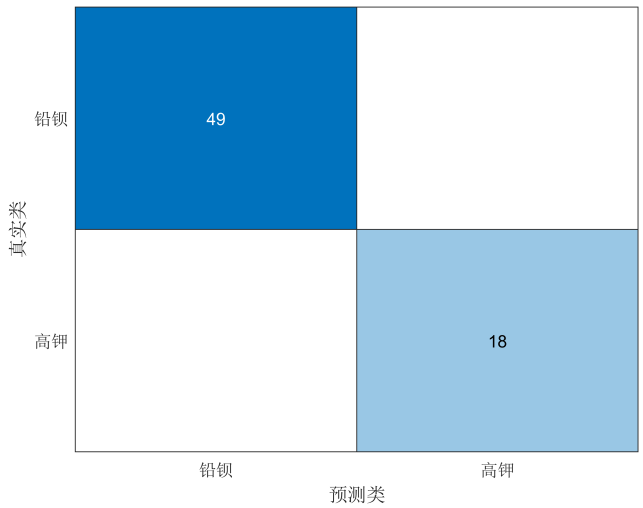


图 6　随机森林模型混淆矩阵图

从混淆矩阵可以看出，随机森林模型对表单 2 数据的分类正确，因此对该模型进行指标重要性的计算，结果如图 7所示。

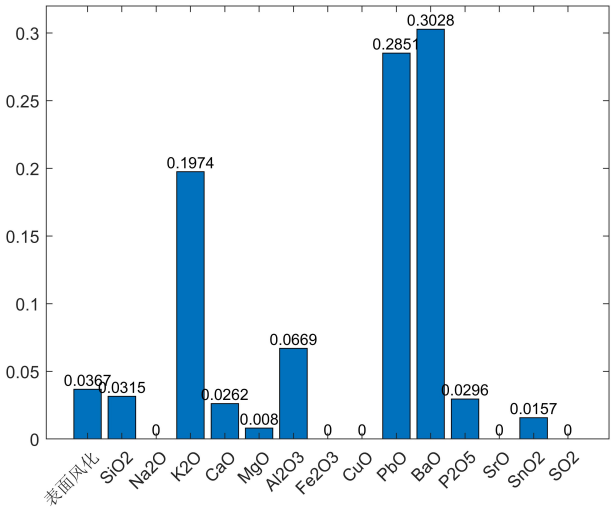


图 7　玻璃分类的指标重要性柱状图

从图 7可以看出，BaO 和 PbO 的重要性都很高，其次是 K_2O、Al_2O_3 等，按照重要性排序来计算这些重要的化学成分的平均占比，得到结果如表 7所示。

表 6　重要分类指标的平均占比值

	BaO	PbO	K_2O	Al_2O_3
高钾玻璃	0.40	0.27	6.40	5.06
铅钡玻璃	9.00	22.08	0.22	4.46

根据上述结果得到，当玻璃的 BaO 和 PbO 占比皆不超过 1%，K_2O 占比接近 6.4%，Al_2O_3 占比接近 5% 时，可分类为高钾玻璃，当 BaO 和 PbO 占比皆大于 9% 左右，K_2O 占比不超过 1%，Al_2O_3 占比接近 4.5% 时，可分类为铅钡玻璃。

6.4.2 亚类划分化学成分提取

利用 SPSS 统计软件，导入各类型化学成分数据，利用回归分析，使用向后法，输出结果如表7所示。

表 7　提取成分结果

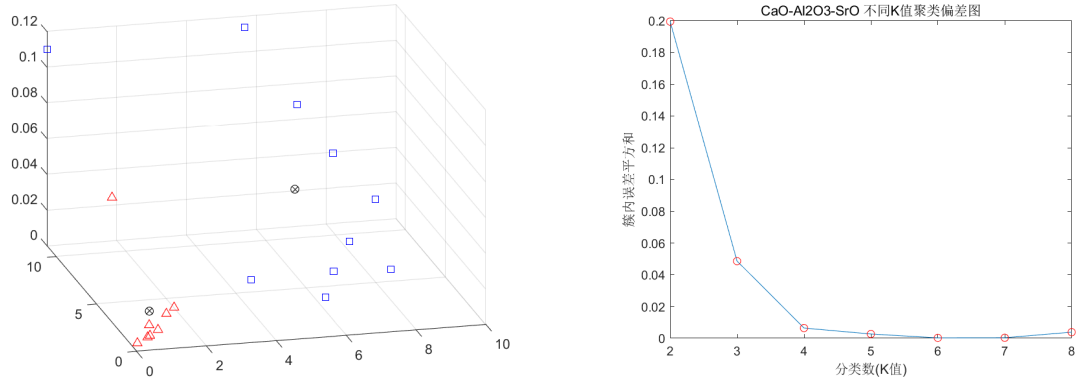
类型	成分	类型	成分
高钾玻璃	氧化钙	铅钡玻璃	氧化钙
高钾玻璃	氧化锶	铅钡玻璃	氧化铁
高钾玻璃	氧化铝	铅钡玻璃	氧化锡

在高钾类型中，当提取出三个因素时，模型无法再进行提取；在铅钡类型中，当提取出三个对残差平方和减小最少的三各因素时，任然可以继续提取，使回归方程残差平方和有显著减少。铅钡类型后续提取出的因素依次是二氧化硅、氧化钡、氧化镁。

根据上述模型，使用 MATLAB 来进行 K-Means 算法的实现以及对高钾玻璃和铅钡玻璃的亚类划分，由上述结论可知，高钾玻璃中的 14 个化学成分指标以氧化钙（CaO）、氧化铝（Al_2O_3）以及氧化锶（SrO）这三个成分为主要成分，铅钡玻璃中的 14 个化学成分指标以氧化钙（CaO）、氧化铁（Fe_2O_3）以及氧化锡（SnO_2）这三个成分为主要成分，以这些指标为 X、Y、Z，进行高钾玻璃采样点的亚类划分以及铅钡玻璃采样点的亚类划分如下。

6.4.3 高钾玻璃采样点的亚类划分

利用清洗后的表单 2 中的数据，以高钾玻璃采样点中的氧化钙（CaO）、氧化铝（Al_2O_3）以及氧化锶（SrO）这三个指标的数据作为 X、Y、Z 做出散点图，通过手肘图知最优分组为 2，使用 K-Means 算法对高钾玻璃采样点进行聚类，结果如图 8a以及图 8b所示



(a) 经过 K-Means 算法聚类后的高钾玻璃采样点散点图　　**(b) 高钾玻璃样本基于 K-Means 算法的手肘图**

图 8　高钾玻璃 K-Means 聚类以及手肘图

划分结果如下表 8所示

表 8　高钾玻璃采样点的划分结果

文物采样点	CaO	Al_2O_3	SrO	分类	文物采样点	CaO	Al_2O_3	SrO	分类
01	6.32	3.93	0.00	1	03 部位 1	2.01	4.06	0.00	2
03 部位 2	5.87	5.50	0.10	1	04	7.12	6.44	0.00	1
05	7.35	7.50	0.06	1	06 部位 1	0.00	11.15	0.11	1
06 部位 2	5.41	10.05	0.12	1	07	1.07	1.98	0.00	2
09	0.62	1.32	0.00	2	10	0.21	0.81	0.00	2
12	0.72	1.46	0.00	2	13	8.72	6.16	0.04	1
14	8.23	9.23	0.00	1	16	8.27	6.18	0.04	1
18	0.00	3.05	0.07	2	21	4.71	6.19	0.00	1
22	1.66	3.50	0.00	2	27	0.94	2.51	0.00	2

其中 1 代表高钙高铝型，2 代表低钙低铝低锶型

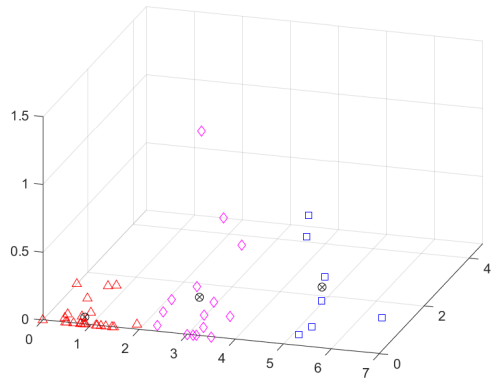
根据上述划分结果表 8以及图 8a, 可知当氧化钙（CaO）、氧化铝（Al_2O_3）以及氧化锶（SrO）都较小时有明显的聚类现象且都离质心较近，当氧化钙（CaO）、氧化铝（Al_2O_3）较高时，虽然点有些稀疏但还是可以认为这些之间存在一个中心点使得这些点能够划为一类。在考虑 CaO、Al_2O_3 这些含量较高的基础上，还需结合微量元素锶 Sr 的含量进行亚类划分。

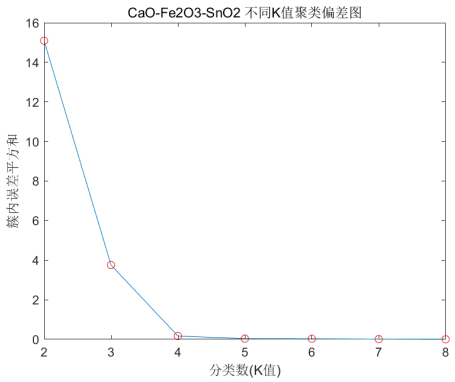
表 9　高钾玻璃亚分类标准

亚类类型	CaO/%	Al_2O_3/%	SrO%
低钙低铝低锶型	0~3	0~5	<0.08
高钙高铝型	4~10	4~12	

6.4.4 铅钡玻璃采样点的亚类划分

利用清洗后的表单 2 中的数据，以铅钡玻璃采样点中的氧化钙（CaO）、氧化铁（Fe_2O_3）以及氧化锡（SnO_2）这三个指标的数据作为 X、Y、Z，通过手肘图得知最优分组为 3，最后做出散点图并使用 K-Means 算法对高钾玻璃采样点进行聚类，结果如图。图 9a以及图 9b所示。





(a) 经过 K-Means 算法聚类后的铅钡玻璃采样点散点图　　**(b) 铅钡玻璃采样点基于 K-Means 算法的手肘图**

图 9　铅钡玻璃 K-Means 聚类以及手肘图

划分结果如下表 10所示（详见附录 F表 19）。

表 10　铅钡玻璃采样点的划分结果

文物采样点	分类	文物采样点	分类	文物采样点	分类
02	2	08	1	08 严重风化点	2
11	2	19	2	20	1
…	…	…	…	…	…
54 严重风化点	1	55	1	56	1
57	1	58	2		

注：其中 1 代表低钙低铁低锡型，2 代表中钙低铁型，3 代表高钙中铁型

根据上述划分结果表 10以及图 9a, 通过观察图 9a, 可明显看出数据有以 X 轴为标准，采样点的类有明显的分层的现象，其中红色点的集中程度很好，所有样本点基本都在质心附近，其他的类虽然没有明显的集中情况，但大多数点都比较紧凑，比较直观。

表 11　铅钡玻璃亚分类标准

亚类类型	CaO/%	Fe_2O_3/%	SnO%
低钙低铁低锡型	0~2	0~2	<0.3
中钙低铁型	2~4	0~2	
高钙中铁型	4~7	0~3	

6.4.5 分类结果合理性、敏感性分析

下面对上述模型的分类结果进行合理性分析，对于高钾玻璃的分类结果，在提取化学成分进行亚分类时，对比了多中提取方法，包括主成分分析，Lasso 回归，逐步回归等等。最终通过巧妙的算法分析提取高钾玻璃中的 14 个化学成分中 3 个最适合亚分类的成分，且本文的分类标准表 9与参考文献[2] 中的“钾玻璃亚类类型划分标准”类似。

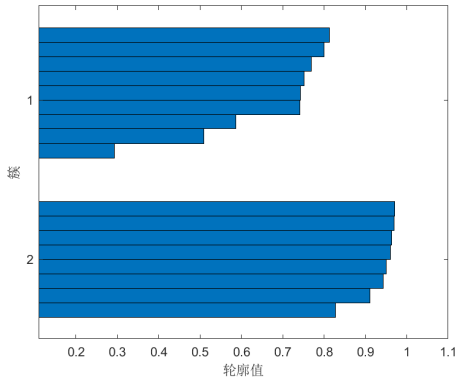
从已选出的三个主要成分中两两组合再挑选出三组成分为：

（1）氧化钙（CaO）- 氧化铝（Al_2O_3）

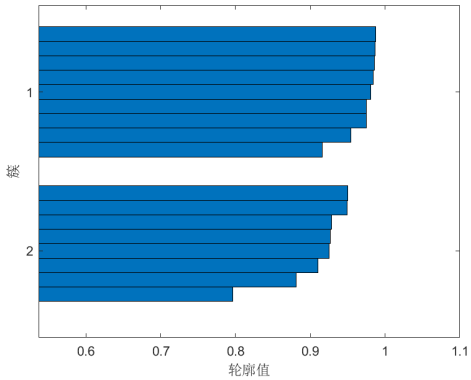
（2）氧化钙（CaO）- 氧化锶（SrO）

（2）氧化锶（SrO）- 氧化铝（Al_2O_3）

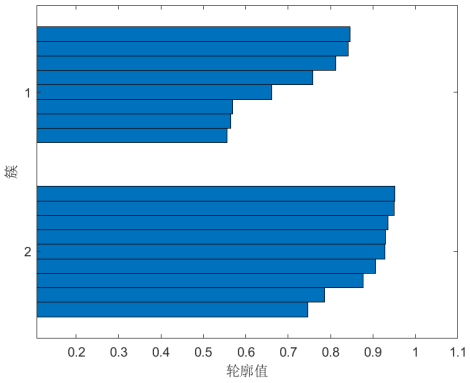
分别对上述三组成分的采样点样本进行 K-Means 聚类，探讨二维的聚类效果，进行对比并说明上述分类的合理程度。



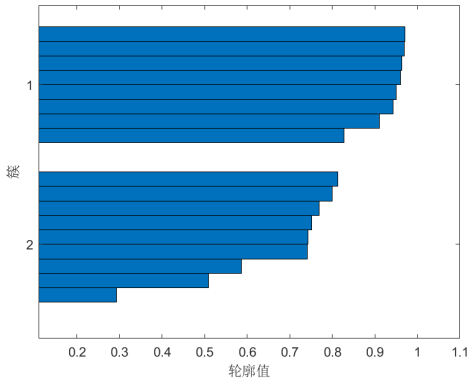
(a) CaO-Al2O3 为指标的 K-Means 聚类轮廓图



(b) CaO-SrO 为指标的 K-Means 聚类轮廓图



(c) SrO-Al2O3 为指标的 K-Means 聚类轮廓图



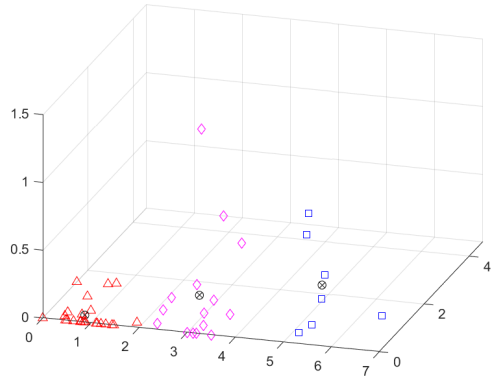
(d) CaO-Al2O3-SrO 为指标的 K-Means 聚类轮廓图

图 10　各种情况的 K-Means 聚类轮廓图

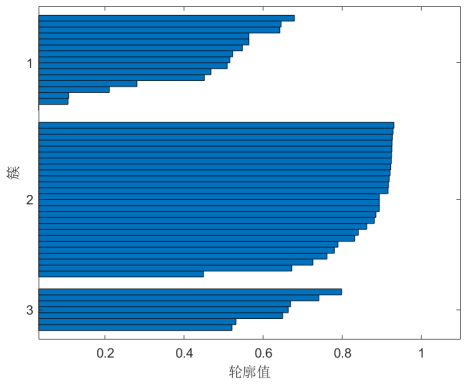
根据轮廓图概念以及观察可知，3 个主要成分（氧化钙（CaO）、氧化铝（Al_2O_3））、氧化锶（SrO）任取两组的聚类效果并不好，图中的数据的个别轮廓值不仅不能接近 1，且和其他轮廓值差距较大，聚类效果不好，分类不明显。与此相比，本文前面提到的模型分类结果就很好，比这三组假设模型聚类效果更好，更明显。

对于铅钡玻璃的分类结果，和高钾玻璃一样首先是通过详细而又周全的统计分析从铅钡玻璃 14 个化学成分中经过优化筛选得到三个主要成分。

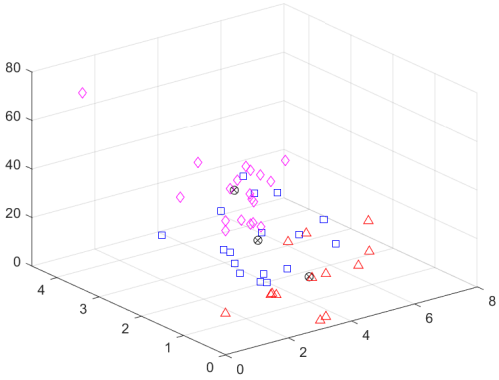
选取二氧化硅（SiO_2）代替氧化锡（SnO_2）组成一组新的指标与原本铅钡玻璃聚类模型进行比较，下面做出上述铅钡玻璃聚类组和（SiO_2-CaO-Fe_2O_3）的轮廓图以及三维聚类图如所示：



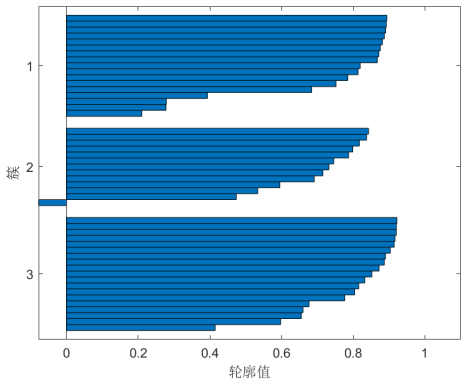
(a) CaO-Fe2O3-SnO2 为指标的 K-Means 算法的三维聚类图



(b) CaO-Fe2O3-SnO2 为指标的 K-Means 聚类轮廓图



(c) CaO-Fe2O3-SiO2 为指标的 K-Means 算法的三维聚类图



(d) CaO-Fe2O3-SiO2 为指标的 K-Means 聚类轮廓图

图 11　多角度比较图

由上图 11可知，以 CaO-Fe2O3-SnO2 为指标的点比较有规律且比较密集，则 CaO-Fe2O3-SnO2 为指标的 K-Means 聚类效果要比 CaO-Fe2O3-SiO2 为指标的 K-Means 聚类效果好，所以在目前现有的 14 个指标下，铅钡玻璃采样点的亚类划分的模型挑选出的三个主要成分较好，以此为基础的分类结果较好。

对于高钾玻璃亚分类模型，为了分析分类效果的敏感度，将模型中三个指标重要性最高的一个化学成分 CaO 加入扰动，即将数据增加或降低一定范围内的百分比，这里将范围设定在 [-5% 5%], 为了更直观看到效果，选取 2 个值：-5%、5%。

表 12　扰动 CaO 后的分类结果

文物采样点	原数据	-5%	5%	文物采样点	原数据	-5%	5%
01	1	1	1	03 部位 1	2	2	2
03 部位 2	1	1	1	04	1	1	1
05	1	1	1	06 部位 1	1	1	1
06 部位 2	1	1	1	07	2	2	2
09	2	2	2	10	2	2	2
12	2	2	2	13	1	1	1
14	1	1	1	16	1	1	1
18	2	2	2	21	1	1	1
22	2	2	2	27	2	2	2

从表 12可以看出，将 CaO 在 [-5% 5%] 这个范围内进行扰动后，结果仍然不变，因此在这个范围内，分类结果受变化参数的影响很小，模型具有较好的稳定性，敏感程度较低。

6.5 问题三的模型建立

要求对未知类别的文物样点进行鉴别，并对分类结果的敏感性进行分析。

6.5.1 基于逻辑回归模型的玻璃类型预测

为了鉴别附件表单 3 中未知类别玻璃文物的所属类型，需要使用预测模型，常用有决策树和逻辑回归模型等。若使用决策树并且调参至最优时，得到的结果只有两个分支，如图 12所示，经过分析认为，这可能是训练集数据太少的缘故，因此即使当前训练集的分类结果是准确的，预测效果的准确性仍不能保证，故最终选择逻辑回归模型作为预测模型。

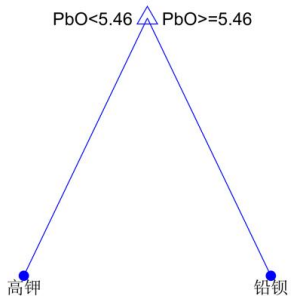


图 12　决策树分支

逻辑回归模型一般用于解决二分类问题，能够将一个因变量和多个自变量构成多元回归关系，从而预测某一事件的发生概率，下面是关于逻辑回归预测玻璃类型的具体算法步骤。

步骤 1：建立预测函数。令因变量为 Y，取值为 Y=0 和 Y=1，分别代表高钾玻璃和铅钡玻璃。自变量为 14 个化学成分及 1 个表面风化指标，一共 15 个，分别表示为 X_1,X_2,...,X_{15}，在 15 个自变量作用下，分类结果为铅钡玻璃的条件概率为 $P = P(Y = 1 \mid X_1, X_2, \ldots, X_{15})$, 则 logistic 回归模型可表示为:

$$z = a_0 + a_1X_1 + a_2X_2 + \cdots + a_{15}X_{15} \tag{11}$$

$$P = \frac{1}{1 + e^{(-z)}} \tag{12}$$

其中 z 是中间变量参数，a_0 是回归常数，a_i 是第 i 个自变量的回归系数，X_i 是第 i 个自变量的取值，高钾玻璃取 0，铅钡玻璃取 1，P 是铅钡玻璃发生概率的回归预测值[6]。

步骤 2：在极大似然法的基础上，得出逻辑回归的损失函数：

$$Cost(P, Y) = \begin{cases} -log(P) & Y = 1 \\ -log(1 - P) & Y = 0 \end{cases} \tag{13}$$

则最终的代价函数形式为:

$$J(a) = -\frac{1}{m}\left[\sum_{i=1}^{m} Y_i \log(P_i) + (1 - Y_i)\log(1 - P_i)\right] \tag{14}$$

其中 m 代表样本数据的总量，i 表示第几个样本数据。

步骤 3：使用梯度下降法求解代价函数最小值和参数。

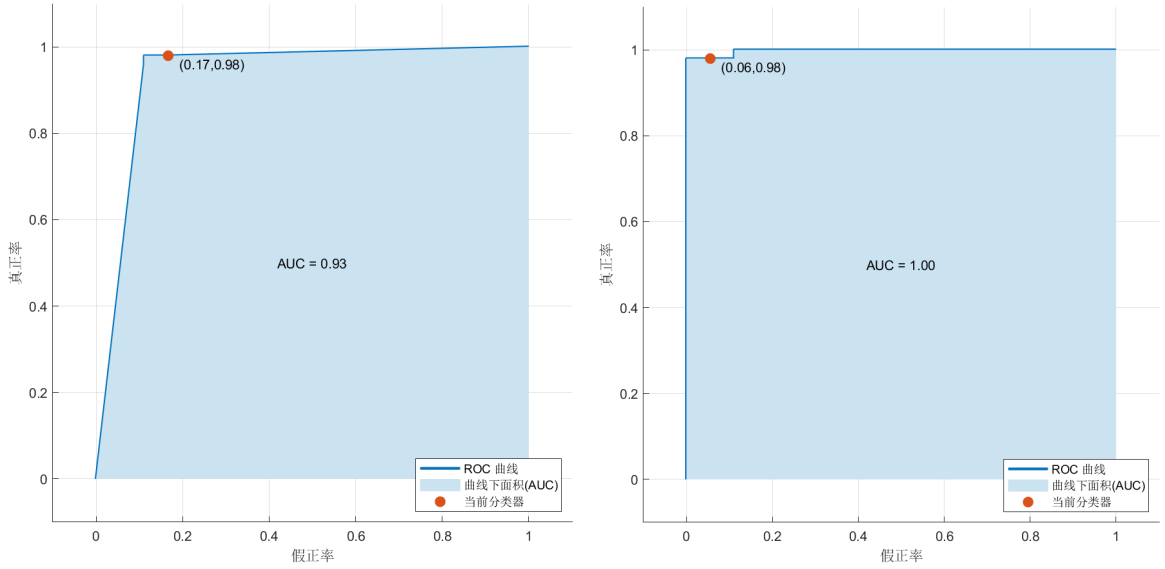
步骤 4：计算逻辑回归模型的 AUC 值，观察模型的分类效果。若分类不够准确，根据问题二的指标重要性结果进行特征提取，仿照步骤 1 重新建立预测函数，再次计算 AUC 值，与之前的结果进行对比，直到当前得到的值最大。

步骤 5：将附件表单 3 的数据代入逻辑回归模型，得到预测值。

6.6 问题三的模型求解

6.6.1 未知类别玻璃文物的预测结果

使用 MATLAB 的 Classification Learner 工具箱对 15 个自变量进行逻辑回归，得到的 ROC 曲线和 AUC 值如图 13a所示，同时根据问题二中的指标重要性结果提取出 7 个最重要的特征变量，以此作为自变量进行逻辑回归，得到的 ROC 曲线和 AUC 值如图 13b所示。



(a) 15 个特征变量的结果　　**(b) 7 个特征变量的结果**

图 13　逻辑回归的 ROC 曲线和 AUC 值

观察这两个图可知，当特征变量为 7 个时，AUC 值为 1，比特征变量为 15 个时的 AUC 值大，说明当特征变量为 7 个时的模型效果更好，同时因为样本量较少，模型的分类正确性显得较为重要，故从图 7 中选择 7 个最重要的特征变量来进行逻辑回归。最终预测结果如表 13所示。

表 13　附件表单 3 的预测结果

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

6.6.2 敏感性分析

为了分析分类结果的敏感性，将表单 3 中指标重要性最高的两个化学成分 Ba0 和 Pb0 加入扰动，即将数据增加或降低一定范围内的百分比，这里将范围设定在 [-5% 5%]，为了更直观看到效果，选取 6 个值：-5%、-3%、-1%、1%、3%、5%。分类结果如表 14和表 15所示。

表 14　扰动 $Ba0$ 后的预测结果

	-5%	-3%	-1%	1%	3%	5%
A1	高钾	高钾	高钾	高钾	高钾	高钾
A2	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A3	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A4	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A5	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A6	高钾	高钾	高钾	高钾	高钾	高钾
A7	高钾	高钾	高钾	高钾	高钾	高钾
A8	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡

表 15　扰动 Pb0 后的预测结果

	-5%	-3%	-1%	1%	3%	5%
A1	高钾	高钾	高钾	高钾	高钾	高钾
A2	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A3	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A4	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A5	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A6	高钾	高钾	高钾	高钾	高钾	高钾
A7	高钾	高钾	高钾	高钾	高钾	高钾
A8	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡

从图可以看出，将 BaO 和 PbO 在 [-5% 5%] 这个范围内进行扰动后，结果仍然不变，因此在这个范围内，模型输出受变化参数的影响很小，模型具有较好的稳定性，具有较强的鲁棒性。

6.7 问题四模型建立

题目要求针对不同类别的玻璃文物样品，分析其化学成分之间的关联关系，并比较不同类别之间的化学成分关联关系的差异性。

6.7.1 皮尔逊相关系数分析模型

建立皮尔逊相关系数[7] 分析模型，分析不同类别玻璃样品化学成分之间的关联关系。

针对不同类别的玻璃制品，分别探讨其化学成分之间的相关关系，建立起皮尔逊相关系数模型。反映皮尔逊相关系数的一个重要参数是协方差。协方差反映了两个随机变量相关的层度，如公式15：

$$cov\left(X,Y\right)=\frac{\sum_{n}^{i=1}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{n-1} \tag{15}$$

皮尔逊相关系数由此协方差除以两个变量的标准差之积而得到，如公式16：

$$\rho_{X,Y}=corr\left(X,Y\right)=\frac{cov\left(X,Y\right)}{\sigma_X\sigma_Y} \tag{16}$$

一般认为皮尔逊相关系数大于 0.7，认为两者关系非常紧密；系数在 0.4 到 0.7 之间，认为两者关系紧密；系数在 0.2 到 0.4 之间，认为两者关系一般。

6.7.2 配对样本 T 检验模型

建立配对样本 T 检验[8] 模型，比较不同类别之间化学成分关联关系的差异性。

玻璃制品分为高钾和铅钡，将其各 14 个化学成分之间的相关系数进行配对。假设两样本所表示的总体服从正态分布或近似正态分布，两总体方差相等，则合并其方差记为 σ_c^2，如公式17，n_1,n_2 为样本量。

$$\sigma_c^2=\frac{\sum x_1^2-\frac{\sum\left(x_1\right)^2}{n_1}+\sum x_2^2-\frac{\sum\left(x_2\right)^2}{n_2}}{n_1+n_2-2} \tag{17}$$

建立检验假设，确定检验水准，μ_1,μ_2 为两样本均值

假设 $H_0:\mu_1=\mu_2$，化学成分之间的关联关系对玻璃制品类型的反应无差别

假设 $H_1:\mu_1\neq\mu_2$，化学成分之间的关联关系对玻璃制品类型的反应有差别

检验统计量 t 值计算如公式18,$\bar{X}_1$,$\bar{X}_2$ 分别表示两样本的均数。

$$t=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\sigma_c^2\left(\frac{1}{n_1}-\frac{1}{n_2}\right)}} \tag{18}$$

确定 P 值，根据统计学推断是否拒绝原假设 H_0

6.8 问题四模型求解

6.8.1 化学成分关联关系求解

将两类玻璃的十四种化学元素分别代入分析。利用 SPSS 数据分析软件，分析出各类别化学元素中两两之间的皮尔逊相关系数，如表1617所示。其中 * 指在 0.05 级别（双尾），相关性显著；** 指在 0.01 级别（双尾），相关性显著。

表 16　高钾化学成分之间的皮尔逊相关系数

化学成分	二氧化硅	氧化纳	……	氧化锡	二氧化硫
二氧化硅	1	-0.457	……	0.049	-0.357
氧化纳	-0.457	1	……	-0.083	-0.194
……	……	……	……	……	……
氧化锡	0.049	-0.083	……	1	-0.108
二氧化硫	-0.357	-0.194	……	-0.108	1

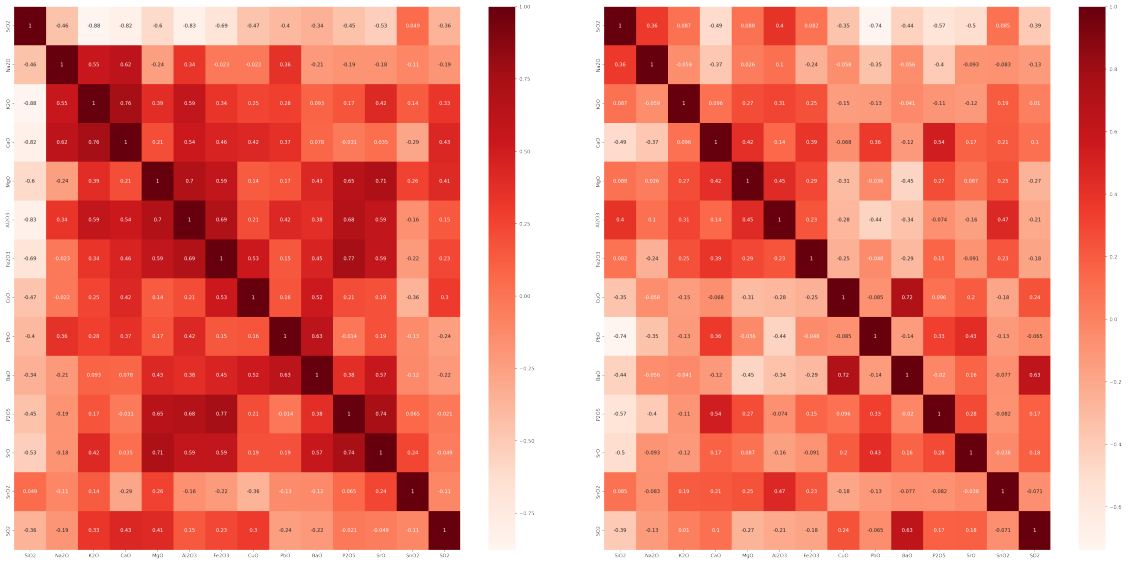
在高钾化学成分之间，二氧化硅与氧化钾、氧化钙、氧化铝的关系非常紧密；氧化钾与二氧化硅、氧化钙的关系非常紧密；氧化镁与氧化锶的关系非常紧密；氧化铁与五氧化二磷的关系非常紧密；五氧化二磷与氧化锶的关系也非常紧密。上述关系的显著性均在 0.01 级别相关性显著。

表 17　铅钡化学成分之间的皮尔逊相关系数

化学成分	二氧化硅	氧化纳	……	氧化锡	二氧化硫
二氧化硅	1	.362*	……	0.085	-.386**
氧化纳	.362*	1	……	-0.083	-0.13
……	……	……	……	……	……
氧化锡	0.085	-0.083	……	1	-0.13
二氧化硫	-.386**	-0.13	……	-0.13	1

在铅钡化学成分之间，二氧化硅与氧化铅的关系非常紧密，且在 0.01 级别相关性显著；与氧化钙、氧化铝、氧化钡、五氧化二磷、氧化锶的关系紧密；与氧化纳、氧化铜、二氧化硫关系一般，与其余化学成分几乎无关。氧化铜与氧化钡关系非常紧密，且在 0.01 级别相关性显著。

利用 python 做出高钾、铅钡化学成分之间关系的热力图，直观反应其关联关系如图14a14b所示。



(a) 高钾热力图　　(b) 铅钡热力图

图 14　两种类型化学成分热力图

6.8.2 关联关系差异性求解

上文得到化学成分之间的相关系数，共计 392 个数据。分析提取出 14 个化学元素两两之间，共计 91 个对应关系，带入皮尔逊相关系数，结合两种类型，构建起分析数据。借助 SPSS 统计软件，做配对样本 T 检验，得到结果如表18所示。

表 18　关联关系 T 检验结果

类型	$\bar{X} \pm S$	t	p	差值 95%CI
铅钡	0.0021±0.2841	-2.864	0.005	-0.2264~-0.0409
高钾	0.1358±0.3987			

S 为标准偏差,CI 为置信区间。结果显示，铅钡化学成分之间的平均相关系数远低于高钾；高钾玻璃与铅钡玻璃化学成分之间的关联关系具有显著的差异性。

七、 模型评价与推广

7.1 模型优点

1. 问题一使用了大量的统计学方法对数据进行分析和检验，在此基础上得出的数据结果可信度较高。
2. 问题二在探讨两种玻璃类型时使用了随机森林估计指标重要性的特性，选择重要性高的指标进行化学成分上的差异，较有说服力的函数关系。

通过散点图观察后建立四次多元拟合函数，经过检验，拟合效果良好，根据上述拟合函数，找到了工艺参数和结构变量之间的关系。预测时使用机器学习进行预测，预测的结果跟函数拟合的预测结果基本一致，比较具有科学性。

3. 问题三的模型使用逻辑回归模型时进行了特征选择，通过比较模型的分类效果得到最好的模型，考虑相对全面，预测结果合理性较强，
4. 问题四提出了做了很多相关性分析且还使用 t 检验，对于探讨化学成分之间的关联关系非常全面，可以很好的探究出各个化学成分之间的关系。

7.2 模型改进

1. 对于数据的预处理，对未检测到的数据进行填补，以达到成分累计和为百分百；对表单 1 中缺失四组颜色进行大胆探索，探索其应当出现的颜色。
2. 对于数据中文物采样点再进行细致的探讨，针对严重风化点、一般风化点等分别纳入讨论，分析其可能具有的不同特性。

参考文献

[1] 付强, 邝桂荣, 吕良波, 莫慧旋, 李青会, 干福熹. 广州出土汉代玻璃制品的无损分析[J]. 硅酸盐学报,2013,41(07):994-1003.

[2] 刘松, 吕良波, 李青会, 熊昭明. 岭南汉墓出土玻璃珠饰与汉代海上丝绸之路中外交流[J]. 文物保护与考古科学,2019,31(04):18-29.DOI:10.16334/j.cnki.cn31-1652/k.2019.04.004

[3] 卡方检验的应用条件[J]. 临床肝胆病杂志,2022,38(06):1292.

[4] 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

[5] 萧超武, 蔡文学, 黄晓宇, 陈康. 基于随机森林的个人信用评估模型研究及实证分析[J]. 管理现代化,2014,34(06):111-113.

[6] 王彤, 栗金晶, 刘欢, 张浩祥, 韦彪, 朱多林, 刘嘉祥. 基于逻辑回归的管道健康状态评价模型及应用 [J]. 水电能源科学,2021,39(05):13

[7] Sedgwick P. Pearson’s correlation coefficient[J].Bmj, 2012, 345.

[8] De Winter J C F. Using the Student’s t-test with extremely small sample sizes[J]. Practical Assessment, Research, and Evaluation, 2013, 18(1): 10.

附录 A　高钾无风化描述统计

	范围	最小值	最大值	均值	均值SEM	标准偏差	方差	偏度	偏度SEM	峰度	峰度SEM
SiO2	28.04	59.01	87.05	68	2.325	8.382	70.264	1.191	0.616	0.831	1.191
Na2O	3.38	0	3.38	0.7	0.372	1.287	1.656	1.497	0.637	0.559	1.232
K2O	14.52	0	14.52	9.33	1.132	3.92	15.369	-1.203	0.637	1.9	1.232
CaO	8.7	0	8.7	5.33	0.893	3.092	9.563	-0.875	0.637	-0.518	1.232
MgO	1.98	0	1.98	1.08	0.195	0.676	0.457	-0.434	0.637	-1.015	1.232
Al2O3	8.1	3.05	11.15	6.62	0.719	2.492	6.208	0.482	0.637	-0.492	1.232
Fe2O3	6.04	0	6.04	1.93	0.481	1.667	2.778	1.176	0.637	2.568	1.232
CuO	5.09	0	5.09	2.45	0.479	1.66	2.756	0.101	0.637	-1.058	1.232
PbO	1.62	0	1.62	0.41	0.17	0.589	0.347	1.374	0.637	0.418	1.232
BaO	2.86	0	2.86	0.6	0.284	0.982	0.965	1.493	0.637	1.238	1.232
P2O5	4.5	0	4.5	1.4	0.414	1.434	2.056	1.678	0.637	1.876	1.232
SrO	0.12	0	0.12	0.04	0.014	0.048	0.002	0.571	0.637	-1.452	1.232
SnO2	2.36	0	2.36	0.2	0.197	0.681	0.464	3.464	0.637	12	1.232
SO2	0.47	0	0.47	0.1	0.054	0.186	0.034	1.396	0.637	0.055	1.232

图 15　高钾无风化描述统计

附录 B　铅钡风化描述统计

	范围	最小值	最大值	均值	均值SEM	标准偏差	方差	偏度	偏度SEM	峰度	峰度SEM
SiO2	49.61	3.72	53.33	24.913	2.001	10.4	108.15	0.318	0.448	1.387	0.872
Na2O	2.22	0	2.22	0.216	0.105	0.546	0.298	2.71	0.448	7.068	0.872
K2O	1.05	0	1.05	0.133	0.045	0.235	0.055	2.554	0.448	8.15	0.872
CaO	6.4	0	6.4	2.695	0.313	1.628	2.649	0.39	0.448	-0.355	0.872
MgO	2.73	0	2.73	0.65	0.133	0.693	0.48	1.054	0.448	1.366	0.872
Al2O3	13.2	0.45	13.65	2.97	0.497	2.583	6.672	2.876	0.448	11.082	0.872
Fe2O3	2.74	0	2.74	0.585	0.139	0.722	0.522	1.427	0.448	1.692	0.872
CuO	10.57	0	10.57	2.276	0.532	2.766	7.649	2.136	0.448	4.4	0.872
PbO	54.5	15.71	70.21	43.314	2.308	11.993	143.825	-0.033	0.448	0.36	0.872
BaO	35.45	0	35.45	11.807	1.883	9.784	95.736	1.299	0.448	0.984	0.872
P2O5	14.13	0	14.13	5.277	0.792	4.115	16.935	0.407	0.448	-0.729	0.872
SrO	1.12	0	1.12	0.418	0.05	0.26	0.067	0.554	0.448	0.92	0.872
SnO2	1.31	0	1.31	0.068	0.051	0.264	0.07	4.438	0.448	20.508	0.872
SO2	15.95	0	15.95	1.366	0.794	4.124	17.011	3.315	0.448	10.047	0.872

图 16　铅钡风化描述统计

附录 C　铅钡无风化描述统计

	范围	最小值	最大值	均值	均值SEM	标准偏差	方差	偏度	偏度SEM	峰度	峰度SEM
SiO2	49.61	3.72	53.33	24.913	2.001	10.4	108.15	0.318	0.448	1.387	0.872
Na2O	2.22	0	2.22	0.216	0.105	0.546	0.298	2.71	0.448	7.068	0.872
K2O	1.05	0	1.05	0.133	0.045	0.235	0.055	2.554	0.448	8.15	0.872
CaO	6.4	0	6.4	2.695	0.313	1.628	2.649	0.39	0.448	-0.355	0.872
MgO	2.73	0	2.73	0.65	0.133	0.693	0.48	1.054	0.448	1.366	0.872
Al2O3	13.2	0.45	13.65	2.97	0.497	2.583	6.672	2.876	0.448	11.082	0.872
Fe2O3	2.74	0	2.74	0.585	0.139	0.722	0.522	1.427	0.448	1.692	0.872
CuO	10.57	0	10.57	2.276	0.532	2.766	7.649	2.136	0.448	4.4	0.872
PbO	54.5	15.71	70.21	43.314	2.308	11.993	143.83	-0.033	0.448	0.36	0.872
BaO	35.45	0	35.45	11.807	1.883	9.784	95.736	1.299	0.448	0.984	0.872
P2O5	14.13	0	14.13	5.277	0.792	4.115	16.935	0.407	0.448	-0.729	0.872
SrO	1.12	0	1.12	0.418	0.05	0.26	0.067	0.554	0.448	0.92	0.872
SnO2	1.31	0	1.31	0.068	0.051	0.264	0.07	4.438	0.448	20.508	0.872
SO2	15.95	0	15.95	1.366	0.794	4.124	17.011	3.315	0.448	10.047	0.872

图 17　铅钡无风化描述统计

附录 D　高钾玻璃风化前的化学成分含量预测结果

文物编号	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)	成分比例累加和
07	67.02	0	0	6.56	0	6.79	1.24	5.09	0	0	2.78	0	0	0	89.47
09	68.75	0	10.13	3.80	0	4.53	2.33	2.43	0	0	1.59	0	0	0	93.57
10	70.01	0	15.80	1.29	0	2.78	1.90	1.32	0	0	0	0	0	0	93.09
12	68.22	0	17.35	4.41	0	5.01	2.11	2.59	0	0	0.68	0	0	0	100.37
22	66.82	0	12.71	10.17	3.51	12.01	2.55	0.86	0	0	0.96	0	0	0	109.59
27	67.08	0	0	5.76	2.96	8.61	1.46	2.42	0	0	1.64	0	0	0	89.93

图 18　高钾玻璃风化前的化学成分含量预测结果

附录 E　铅钡玻璃风化前的化学成分含量预测结果

文物编号	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)	成分比例累加和
23	51.12	2.13	0	0.42	0.36	0.74	0	3.67	19.91	13.91	0	0.42	0	0	92.68
25	48.09	0.62	0	0.53	0	1.00	3.01	1.37	37.40	7.80	0.14	0.26	0	0	100.21
28	64.70	0	0.40	1.12	0.51	2.46	0.80	0.40	20.10	4.74	0.76	0.15	0.65	0	96.79
29	60.15	0.25	0.46	2.49	0.76	7.51	1.57	0.91	14.43	2.38	0.30	0.32	0	0	91.55
42	48.71	1.54	0.23	0.66	0.56	1.85	0	3.28	25.65	12.28	0.06	0.45	0	0	95.27
42	48.78	1.53	0.54	0.00	0.59	2.97	0	3.34	23.59	12.76	0	0	0	0	94.09
44	57.72	0.82	0.31	1.79	0	6.65	1.49	0.53	15.96	6.12	0	0.33	0	0	91.73
49	51.89	0	0.46	1.74	0.61	3.41	2.46	0.55	26.99	4.91	3.15	0.39	0	0	96.58
50	42.78	0	0	2.61	0.28	2.18	0	0.86	35.89	7.29	4.63	0.30	0	0	96.81
53	60.49	0.82	0.17	0.65	0.58	3.18	0	0.66	16.02	10.54	0	0.35	0	0	93.46

图 19　铅钡玻璃风化前的化学成分含量预测结果

附录 F　铅钡玻璃采样点的划分结果（完整）

表 19　铅钡玻璃采样点的划分结果

文物采样点	分类	文物采样点	分类	文物采样点	分类
02	2	08	1	08 严重风化点	2
11	2	19	2	20	1
23 未风化点	1	24	1	24	1
25 未风化点	1	26	1	26 严重风化点	2
28 未风化点	1	29 未风化点	2	30 部位 1	1
30 部位 2	3	31	2	32	1
33	1	34	1	35	1
36	1	37	1	38	1
39	1	40	1	41	3
42 未风化点 1	1	42 未风化点 2	1	43 部位 1	3
43 部位 2	3	44 未风化点	2	45	1
46	1	47	1	48	2
49	3	49 未风化点	2	50	2
50 未风化点	2	51 部位 1	2	51 部位 2	2
52	2	53 未风化点	1	54	2
54 严重风化点	1	55	1	56	1
57	1	58	2		

注：其中 1 代表低钙低铝低锡型，2 代表中钙低铝型，3 代表高钙中铝型

附录 G　第二问基于 K-Means 算法的聚类代码

```
clc
clear
load gaojia.mat
[idx,cmeans3,cen]=kmeans(gaojia,2,'Replicates',1000);
figure(1)
```

```
silhouette(gaojia,idx)
color=['r','g','b'];
ptsymb = {'bs','r^','md','go','c+'};
figure(2)
for i = 1:3
    clust = find(idx==i);
    plot3(gaojia(clust,1),gaojia(clust,2),gaojia(clust,3),ptsymb{i});
    hold on
end
syms x
plot3(cmeans3(:,1),cmeans3(:,2),cmeans3(:,3),'ko');
syms y
plot3(cmeans3(:,1),cmeans3(:,2),cmeans3(:,3),'kx');
grid on
hold off
data = gaojia;
syms j
data=mapminmax(gaojia,0,1);
[n,p]=size(data);
figure(3)
syms o
K=8;D=zeros(K,2);
for k=1:K
[lable,c,sumd,d]=kmeans(data,k,'dist','sqeuclidean');
sse1 = sum(sumd.^2);
D(k,1) = k;
D(k,2) = sse1;
end
plot(D(2:end,1),D(2:end,2))
syms p
hold on;
plot(D(2:end,1),D(2:end,2),'or');
title('CaO-Al2O3-SrO 不同K值聚类偏差图')
xlabel('分类数(K值)')
ylabel('簇内误差平方和')
```

附录 H　第三问的 matlab 代码

```
clc;clear
load('DATA2.mat');
load('LGModel.mat');

load('BaO\BaO1.mat');load('BaO\BaO3.mat');load('BaO\BaO5.mat');load('BaO\BaO_1.mat');
load('BaO\BaO_3.mat');load('BaO\BaO_5.mat');
load('PbO\PbO1.mat');load('PbO\PbO3.mat');load('PbO\PbO5.mat');load('PbO\PbO_1.mat');
load('PbO\PbO_3.mat');load('PbO\PbO_5.mat');
```

```
y = LGModel.predictFcn(DATA2)

B(:,1)=LGModel.predictFcn(BaO_1);
B(:,2)=LGModel.predictFcn(BaO_3);
B(:,3)=LGModel.predictFcn(BaO_5);
B(:,4)=LGModel.predictFcn(BaO1);
B(:,5)=LGModel.predictFcn(BaO3);
B(:,6)=LGModel.predictFcn(BaO5);
B

P(:,1)=LGModel.predictFcn(PbO_1);
P(:,2)=LGModel.predictFcn(PbO_3);
P(:,3)=LGModel.predictFcn(PbO_5);
P(:,4)=LGModel.predictFcn(PbO1);
P(:,5)=LGModel.predictFcn(PbO3);
P(:,6)=LGModel.predictFcn(PbO5);
P
```