基于logistic回归对古代玻璃制品的成分分析和鉴别

摘　要

本文从逻辑回归，K-Means算法的角度，综合了敏感性分析，通过机器学习，实现了对古代玻璃制品的分类问题，结合线性回归实现了对未风化玻璃的成分预测。本文针对每个问题，抓住主要矛盾，给出了合理的数学模型。

针对问题一，经过分析，对定量分类数据进行了哑变量处理并将数据转化为**0-1变量**，提高了数据的维度，将数据从三维升到了**十三维**，其次采用logistic回归分析方法求解出了风化与否的统计规律，通过该模型分别得到预测的准确率为83.3%、95.92%、100%。考虑到高钾类型的数据量太少，只有18组，可能导致其离散程度比较高，预测结果才为100%。综上，预测结果良好。在预测风化前数据时利用一元线性回归预测了未风化玻璃的成分。

针对问题二，如一利用用logistic模型得出了玻璃分类的逻辑回归系数，根据**系数的正负**，来判断不同的化学成分增减对玻璃文物类别的判断影响，从而揭示统计规律。第二个亚分类问题，主要采用**K-Means算法**，得到两种数据的**k值聚类偏差图**，并且通过图表分别看出两种类型数据分类个数。其中高钾的分类为4类，铅钡的分类为3类。通过Matlab及python实现了对高钾及铅钡玻璃的亚分类，并给出了亚分类的分类准则.

针对问题三，利用问题二中逻辑回归模型求解出的玻璃文物类型的统计规律，沿用其函数关系，对表格三的数据进行预测，得到的结果为:A2、A3、A4、A5为铅钡类型，其他的都为高钾类型。模型的准确率达到了100%,分类效果良好。在进行灵敏性分析的时候，我们设置**原数据的5%为微小变化**，代入逻辑回归模型得到回归概率变化值的数量级小于10^{-4}。说明当原数据发生微小变化的时候，对最后的预测结果基本没影响。

针对问题四，利用了**相关系数**并通过假设检验，在给定的显著水平下，结合p值得出了不同种类化学成分样本总体的线性相关性，作为关联关系及其差异性的量化标准。

关键词:数据挖掘 logistic回归预测 K-Means聚类算法相关系数检验

1 问题的重述

1.1 问题背景

丝绸之路是古代中西方文化交流的通道，其中玻璃是早期贸易往来的宝贵物证。玻璃文物中最主要的成分的玻璃，而玻璃的主要原料是石英砂，主要的化学成分是二氧化硅(SiO_2)。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。所以玻璃制品中还含有很多的其他的化学元素。由此，添加助熔剂不同，其主要的化学成分也就不同。

现有两种玻璃制文物，一种是高钾玻璃，另一种是铅钡玻璃。为判断两种类型的玻璃文物的风化与否，以及类型的判断，提供了几个判别方式和角度。从直观上来看，可以观察其表面的纹路、颜色，甚至可以直接观察出是否发生风化，但是这里也不排除局部较浅的风化，从内部的元素来分析，可以通过对其内部的元素的含量分析来鉴别，同时还要考虑风化的影响，因为内部元素与环境元素进行大量交换，会导致其成分比例发生变化，给鉴别带来影响。

1.2 问题提出

题目中给出了58个数据样品，并且给出了这58个数据样本的化学成分含量，其中成分比例累加和介于 85% 105%之间的数据视为有效数据。此外还给出了8个待分类的数据样本，需要解决的问题如下：

1. 问题一：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. 问题二：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. 问题三：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. 问题四：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。

2 问题分析

2.1 对问题一的分析

第一个问题有三个小问题，第一个子问题是要求我们求解表面风化与其玻璃类型、纹饰、颜色的关系。首先我们将表面风化与否当成分类标准，而其他的特征当成判断指标。考虑到这些数据都是定类的，所以我们采用逻辑回归之前先对每个变量进行数据预测里，将数据升维，对每一个特征中的不同类型都设成一个单独的指标，比如花纹变成：A、B、C三种指标，用0-1表示这个变量是否含有这个指标，这样我们就得到了一个13维的数据。然后通过逻辑回归来对是否风化分类，最终得到分类的logistic函数。

然后第二个子问题考虑玻璃的类型，将数据划分成两类——铅钡和高钾。并将这两个单独的类别来判断表面有无风化和化学含量的统计规律，考虑到文件中很多空白处（未检测出这些成分），我们将这些空白处直接设置为0，然后继续通过逻辑回归，得到两者之间的关系。

第三个子问题根据风化点数据，我们需要考虑这其中的比例的类别，同时考虑是未风化的数据样本，寻找未风化的化学含量的统计规律从而预测出其风化前的数据，数据在文中给出

2.2 对问题二的分析

第二个问题有两个主要要求，第一个要求我们根据化学元素来确定玻璃类型的分类规律，这个问题同样可以使用逻辑回归来解决，将玻璃类型当成分类指标，建立玻璃类型和化学元素之间的关系，得到逻辑回归函数即可。这里不考虑文物的花纹和颜色，只对化学成分进行分析。

第二个要求我们在选择适当的化学成分，对玻璃的类型进行亚分类，在选择化学成分这一方面，计划通过使用k-Means聚类算法对数据进行划分，并确定分类的依据和方法，同时对分类的结果并进行合理性分析。在灵敏性分析方面，我们对数据样本多的化学成分进行微小变化，查看微小变化对轮廓系数的变化影响。

2.3 对问题三的分析

这个问题中涉及到用分类模型来预测未知文物的分类，我们可以使用前两问的逻辑回归，将表格三中的数据代入到我们求解到的逻辑回归方程中得到最终概率。通过计算的最后的概率将得到的文物进行。

敏感性分析采用当其中的某一个成分发生改变的时候，对应$ln\frac{p}{1-p}$的变化量。并且反推出概率p的变化的数量级。

2.4 对问题四的分析

第一个子问题要求我们分析不同类型玻璃的化学组成成分的关联关系，考虑最简单的相关关系——线性相关关系，计算两个成分组成之间的相关系数。考虑到化学组成复杂，计划舍弃对线性相关性弱的变量对的分析，着重分析相关关系强的成分组成对。鉴于相关系数并不一定能够对总体样本的相关性作出正确的估计，考虑采用假设验证的方法验证相关系数（对总体的估计效果）的显著性。

第二个子问题要求分析不同类型玻璃的化学组成成分关联的差异性，通过比对不同类型玻璃具有的显著线性相关关系的化学成分对来分析差异性。

3 问题假设

一、颜色、类型、条纹之间没有多重共线性

颜色的主要决定因素和化学含量由一定的关系，也和类型有一定的关系，这之间的概率假设很小，即没有高精确相关关系或高度相关关系，而条纹一般只决定于历史文化，与颜色和类型之间没有的关系[1]。

二、将文物的不同的部位测到的数据视作一个独立文物

在分析的时候将文物不同位置测到的数据视作类型相同的另一个独立文物，部分未风化的文物在风化的部位测定数据视作独立的以风化样本，部分已风化的文物在未风化部分测定的数据视作一个未风化独立样本

三、风化与否和类型均为化学组成单值函数且二者相互独立

考虑化学元素就能够决定其最终的类型，而且风化与否只是一个外在的标签，所以在与预测预测中舍去风化与否一列，因为通过化学元素的本质就可以得出最后的类别。

4 符号说明

符号	意义	说明
f_i	表示花纹的种类(i = 1，2，3)，分别对应三种花纹	0-1变量
c_i	表示的颜色的种类(i= 1，...，8)	0-1变量
s_i	表示的是玻璃文物的种类(i= 1，2)	0-1变量
p_i	表示的是某种文物是否表面风化的概率(i = 1,...,54)	-
$p_i^{'}$	表示化学元素为自变量的logistic模型的概率(铅钡)	-
$p_i^{''}$	表示化学元素为自变量的logistic模型的概率（高钾）	-
β_i	回归系数(i= 0,...,13)	-
$\beta_i^{'}$	化学元素为自变量的回归系数（铅钡）	因变量为是否风化
$\beta_i^{''}$	化学元素为自变量的回归系数（高钾）	因变量为是否风化
$\hat{\beta}_i$	化学元素和是否风化作为自变量回归系数	因变量为类型
x_{ij}	表示第j号文物的化学成分i的含量	顺序同表单二
$\hat{x}_{ij}$	表示的是第j号文物的化学成分i的归一化数据	-
Δx_{ij}	表示的是对表单三种的数据第j号文物的第i个化学成分的微小变化	-
σ_i	表示第i个化学成分的的标准差	顺序同表单二
ρ_{ij}	表示的是第i类化学成分与第j类化学成分的样本总体相关系数	-
r_{ij}	表示对ρ_{ij}的估计值	-
$\overline{x}_i$	表示的是第i个化学成分的均值	顺序同表单二

5 问题一的logistic回归模型对风化与否分类

5.1 建模思路

这是一个很明显的分类模型，所以我们使用传统的**logistic回归**预测模型来得到逻辑回归函数，以此得到分类的结果。这些结果均可在SPSSAU中实现。

5.2 数据预处理

5.2.1 剔除异常值

1. 由于在表格一中有三组数据的颜色未给出，所以将这三组变量剔除，则直接在Excel表格中将这部分删除，对剩下的数据进行分析和挖掘
2. 此外在表格二中通过Excel对每一行的成分含量求和，并且对所有表单二中的数据依据总和逆序排列，发现有两组数据不满足题设中成分比例累加和介于 85% 105%之间。都是低于85%,分别是71.89以及79.47。结果如下

文物采样点	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al_2O_3)	氧化铁 (Fe_2O_3)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P_2O_5)	氧化锶 (SrO)	氧化锡 (SnO_2)	二氧化硫 (SO_2)	总和
17	60.71	2.12	5.71		0.85		1.04	1.09	0.19	0	0.18		0		71.89
15	61.87	3.21	7.44		1.02	3.15	1.04	1.29	0.19	0	0.26				79.47

Figure 1: 删除的两个数据的各化学成分和含量总和

5.2.2 数据升维

考虑到给出的数据全部是定类模型，无法直接作为逻辑回归输入，所以我们引入0-1变量来将每个属性都当成一个单独的自变量。这些变量满足如下规定

$$\sum_{i=1}^{3} f_i = 1, \sum_{i=1}^{8} c_i = 1, \sum_{i=1}^{2} s_i = 1 \tag{1}$$

其中s_i, c_i, f_i都是0-1二值变量

文物编号	颜色-1-蓝绿	颜色-2-浅蓝	颜色-3-紫	颜色-4-深绿	颜色-5-深蓝	颜色-6-浅绿	颜色-7-黑	颜色-8-绿	类型-1-高钾	类型-2-铅钡	表面风化
1	1	0	0	0	0	0	0	0	1	0	无风化
2	0	1	0	0	0	0	0	0	0	1	风化
3	1	0	0	0	0	0	0	0	1	0	无风化
4	1	0	0	0	0	0	0	0	1	0	无风化
5	1	0	0	0	0	0	0	0	1	0	无风化
6	1	0	0	0	0	0	0	0	1	0	无风化
7	1	0	0	0	0	0	0	0	1	0	风化
8	0	0	1	0	0	0	0	0	0	1	风化
9	1	0	0	0	0	0	0	0	1	0	风化
10	1	0	0	0	0	0	0	0	1	0	风化
11	0	1	0	0	0	0	0	0	0	1	风化
12	1	0	0	0	0	0	0	0	1	0	风化
13	0	1	0	0	0	0	0	0	1	0	无风化
14	0	0	0	1	0	0	0	0	1	0	无风化
15	0	1	0	0	0	0	0	0	1	0	无风化
16	0	1	0	0	0	0	0	0	1	0	无风化
17	0	1	0	0	0	0	0	0	1	0	无风化
18	0	0	0	0	1	0	0	0	1	0	无风化
20	0	1	0	0	0	0	0	0	0	1	无风化
21	1	0	0	0	0	0	0	0	1	0	无风化
22	1	0	0	0	0	0	0	0	1	0	风化
23	1	0	0	0	0	0	0	0	0	1	风化
24	0	0	1	0	0	0	0	0	0	1	无风化
25	0	1	0	0	0	0	0	0	0	1	风化

Figure 2: 前25个样本处理过后的结果图

经过这样的处理之后每个样本的数据的维度由三维变成了十三维，通过Matlab实现，然后这些数据都可以直接代入到逻辑回归方程中计算逻辑回归的参数。升维结果如Figure 2

5.2.3 数据补充

在表格二和三中的数据中有很多数据是空白的，根据题设这些空白的部分是未检测到的成分，所以可以设置为含量是0%,这一步对后续的求解十分重要，不然在数学处理工具中会显示为NAN，即问题数据，不能参与统计与计算。我们小组是通过Matlab对这里面的为空的数据设置为0。

5.3 模型建立及分析步骤

对于子问题一：通过相关的数据建立logistic回归模型，设表面风化的概率是p_i，则未风化的概率是$1-p_i$,其中i表示的第i号文物（i = 1,2,...,54），因为中间由剔除的原因，这里的第i号文物只是原剔除后的数据按照文物的编号的大小的一个顺序排列。通过指标拟合出回归方程中的回归系数。

$$ln\frac{p_i}{1-p_i} = \beta_0 + \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3 + \beta_4 c_1 + ... + \beta_{11} c_8 + \beta_{12} s_1 + \beta_{13} s_2 \tag{2}$$

其中$p_=\frac{1}{1+e^{-(\beta_0+\beta_1 f_1+\beta_2 f_2+\beta_3 f_3+\beta_4 c_1+...+\beta_{11} c_8+\beta_{12} s_1+\beta_{13} s_2)}}$

由此计算出每一个样本的表面风化的概率,并给出以下假设： 如果通过（1）式计

Table 1: 假设检验表

假设检验	假设	检验结果
表面风化假设检验	H_0:$p_i > 0.5$;H_1:$p_i < 0.5$	接受 H_0,拒绝H_1

算出来的概率值是大于0.5，那就接受原假设，表明此时表面以风化

对于子问题二：继续通过logistic回归模型，首先将**文物的类别分开**，两者分别考虑。计算的公式和（1）类似。

$$ln\frac{p_i^{'}}{1-p_i^{'}} = \sum_{i=1}^{14} \beta_i^{'} x_{ij}, j \in \{\text{铅钡类文物的标号}\} \tag{3}$$

$$ln\frac{p_i^{''}}{1-p_i^{''}} = \sum_{i=1}^{14} \beta_i^{''} x_{ij}, j \in \{\text{高钾类文物的标号}\} \tag{4}$$

然后通过对逻辑函数的进行参数检验，得到相应的系数。在预测风化点之前数据可以通过对未风化数据进行分析，然后由此预测风化点之前的数据,逻辑回归的分析步骤如下：

5.3.1 建立逻辑回归模型

Step 1：因变量的分布描述

由数据的频数可知，前两者数据量没有出现严重的不平衡，不需要对数据做特殊的处理,而到高钾的判断的时候，发现数据的比例是1：2，出现了不平衡，但是属于正常范围内，继续使用逻辑回归模型

Table 2: 因变量的频数分布（风化与花纹、颜色、类型）

选项	频数	百分比
风化	30	55.556%
无风化	24	44.444%
总计	54	100%

Table 3: 因变量的频数分布（铅钡中风化与化学元素）

选项	频数	百分比
风化	26	53.06%
无风化	23	46.94%
总计	49	100%

Table 4: 因变量的频数分布（高钾中风化与化学元素）

选项	频数	百分比
风化	6	33.3%
无风化	12	66.6%
总计	18	100%

Step 2：模型似然比卡方检验

通过SPSSPRO中的似然卡方值检验，同时计算p值。

对于第一个分类标准，即根据花纹、颜色、类型来得到对是否风化分类，这个用在logistic模型上得到的卡方值和p值分别为 30.23、0.000。对第二个分类标准，即根据化学元素得到受否风化的分类标准，这部分的数据分成了两类，分别应用到此模型上得到的铅钡数据计算出来的卡方值和p值分别是 55.132、0.000，高钾计算出的卡方值和p值是23.32 ，0.001。

这两个数据的显著性p值均小于0.05，说明这个模型是有效的。

Step 3：拟合参数值

将数据输入，计算得到的系数值代入方程（1）中，得到β_i、β_i'和β_i''的值

5.3.2 预测原理分析

提出假设：由于后面数据过少，考虑到样本量少带来的误差，我们只将SiO_2、Al_2O_3、CuO、P_2O_5、SrO 这几类样本量多的化学成分考虑进入。对同一个文物在风化和未风化的部位，出于简化考虑，利用线性回归，建立风化和未风化的线性关系，从而实现对其他样本未风化前化学成分的预测。

因为是线性回归，在预测的时候必然会出现比例超过100以及比例为负值的，比例为负值的产生原因是，原风化的文物这项的化学成分本来就很少，经过线性回归向前预测未风化的数据，由于这个项的系数是负值，减少的值比原来的结果更加大，所以产生复制，这样的结果统一通过转换为0来处理，而超过一百的数据当成100，然后再计算每个化学含量的百分比，得到最后的结果，由于数据预测量太多，我们就只给出部分数据的预测值

5.4 模型求解与检验

5.4.1 计算系数结果

通过 Matlab拟合最后的结果，如下表所示：

Table 5: β_i拟合的参数结果

β_0	β_1	β_2	β_3	β_4	β_5	β_6
−7.288	-34.602	38.674	-11.36	32.803	12.442	-11.205
β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}
−10.512	-22.23	-11.898	34.7	-31.387	-37.142	29.854

Table 6: $\beta_i^{'}$拟合的参数结果

$\beta_0^{'}$	$\beta_1^{'}$	$\beta_2^{'}$
4.164	-22.785	10.008
$\beta_3^{'}$	$\beta_4^{'}$	$\beta_5^{'}$
−6.989	-26.328	2.298
$\beta_6^{'}$	$\beta_7^{'}$	$\beta_8^{'}$
47.649	-21.854	8.755
$\beta_9^{'}$	$\beta_{10}^{'}$	$\beta_{11}^{'}$
11.927	-7.175	12.541
$\beta_{12}^{'}$	$\beta_{13}^{'}$	$\beta_{14}^{'}$
−3.018	5.617	7.426

Table 7: $\beta_i^{''}$拟合的参数结果

$\beta_0^{''}$	$\beta_1^{''}$	$\beta_2^{''}$
−143.870	1.691	8.494
$\beta_3^{''}$	$\beta_4^{''}$	$\beta_5^{''}$
2.098	-5.874	41.645
$\beta_6^{''}$	$\beta_7^{''}$	$\beta_8^{''}$
−5.116	10.976	8.697
$\beta_9^{''}$	$\beta_{10}^{''}$	$\beta_{11}^{''}$
12.635	-25.761	-12.601
$\beta_{12}^{''}$	$\beta_{13}^{''}$	$\beta_{14}^{''}$
57.749	-27.887	-89.551

将上述的数据代入逻辑回归函数中，即可得到的最后的概率值，通过对这些概率进行假设检验表中的检验即可进行分类，而最后的logistic回归函数就是这个分类的标准和关系，同时可以通过逻辑函数的系数来揭示文物表面有无风化与化学成分的统计规律。工具分析规律

对与铅钡类型，对应的是$\beta_i^{'}$，SiO_2、K_2O、 CaO 、Fe_2O_3、 BaO和 SrO的回归系数是负的，这些含量的上升会使逻辑回归得到的概率变小，得到风化的概率变小，是未风化的概率相对就变大了，在这几个化学成分之间 CaO和 SiO_2的回归系数绝对值最大，分别为26.328和22.785，说明这两个的含量越大，这个玻璃文物就越可能没有风化，相反其他的化学成分的相关系数是正的，那就说明，当那些成分变大的时候，这个文物更可能已经风化，其中Al_2O_3的相关系数最大，为47.649。综合说明，当Al_2O3化学成分多的时候，且CaO和SiO_2的含量少的时候，说明铅钡类玻璃文物最有可能是已经风化。

对高钾类型，对应的是$\beta_i^{''}$,这其中CaO 、Al_2O_3、 BaO、 SnO_2、 SO_2的回归系数是负的，类比上述分析，当这些含量增加的时候，其可能其未风化的概率会更大，其中SO_2的相关系数的绝对值是最大的，为89.551。其次是SnO_2,为27.887。但是这两个化学成分很少在文物中检测出来，同时又因为其起着重要的决定性，所以如果在文物中检测出少量的SnO_2、 SO_2，有很大的可能性这个文物还未风化；回归系数为正的主要

有SrO和MgO，分别为57.749和41.645，说明这些含量高的时候，高钾类的玻璃文物很可能已经风化

5.4.2 模型的代入检验

将原数据代入模型根据上述假设，得到分类结果如下：

已风化	3	27
未风化	18	6
	未风化	已风化

Figure 3: 是否风化与花纹、颜色、类型，回代分类结果

已风化	1	25
未风化	22	1
	未风化	已风化

Figure 4: 是否风化与化学元素回代分类结果

根据表格可以知道，在第一问的第一个子问题的分类中，已风化的分类结果更优，误判只有$\frac{1}{10}$,但是未风化的部分分类结果要差一些，误判有$\frac{1}{4}$,综合两个的数据的结果，得到模型的准确率见下表。

在第二个子问题中，我们的分类结果有明显的优越性，在对铅钡的分类中，误判的只有$\frac{1}{26}$以及$\frac{1}{23}$，而在对高钾类型的玻璃文物的判别中，误判为0，即召回率为100%（分析其原因，可能是高钾类数据量少，只有18组，导致数据样本点容易分离）综合上述数据，计算得到最终的准确率见下表。并且计算最后的准确率如下：

Table 8: 模型准确率

选项	准确率
判断风化与否	0.833
判断风化（铅钡）	0.9592
判断风化（高钾）	1.00

5.4.3 未风化前的预测结果

编号	SiO2	Al2O3	CuO	P2O5	SrO
2	83.377%	5.937%	0.000%	10.686%	0.000%
7	60.153%	2.308%	0.000%	37.539%	0.000%
8	63.024%	1.538%	6.246%	29.193%	0.000%
8	65.155%	2.078%	1.775%	30.993%	0.000%
9	65.837%	0.992%	1.821%	31.350%	0.000%
10	65.401%	3.001%	0.000%	31.598%	0.000%
11	65.143%	1.110%	2.077%	31.669%	0.000%
12	66.601%	0.523%	0.000%	32.876%	0.000%
19	49.478%	2.585%	1.061%	46.876%	0.000%

Figure 5: 风化前数据的预测

6　问题二的K均值聚类算法对类型的亚分类

6.1　建模思路

本问需要我们对文物的玻璃类型进行分类,可以继续使用逻辑回归，将数据代入，得到参数估计值。

而对每种的玻璃文物进行亚分类，那就要求我们在铅钡和高钾的类别基础上，进行再分类，得到更加细分的结果，我们采用**K-means**方法，得到每个大类下的亚类的数据，然后对这些数据的化学成分求解均值，作为这一类的代表，同时归纳出数据的统计规律，确定不同类别的属性。

6.2　K-Means模型建立

Step 1：数据归一化

通过Matlab计算出每个化学成分的平均值和标准差，如下计算得到归一化的数据

$$\hat{x}_{ij} = \frac{x_{ij} - \overline{x}_{ij}}{\sigma_i} \tag{5}$$

Step 2：确定k-means的k值

通过将重复多次迭代k-means包，其中选出SSE（簇类误差平方和）最小值作为k值,再取得了k值之后进行k-means分类[2]

Step 3：

根据kmeans代码求得idx分类表，并根据表格将原有数据分为3组，并绘成条形图，观察其特征进行分类总结

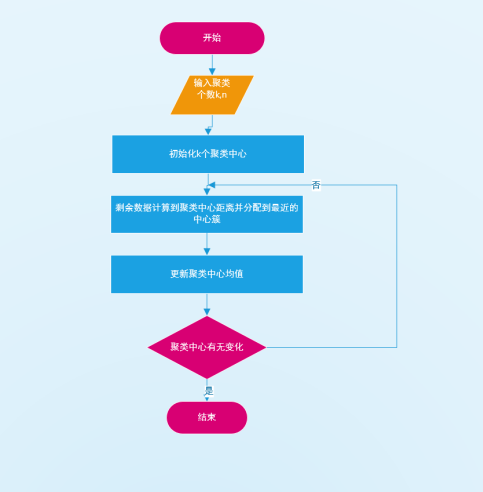


Figure 6: k-Means聚类算法流程

6.3 模型求解

6.3.1 玻璃类型分类规律

通过SPSSAU的二元逻辑回归分类求解到回归系数$\hat{\beta}_i$的值如下所示：

Table 9: β_i拟合的参数结果

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$
38.259	-7.372	-0.267	0.119	1.533	0.181	-6.025	-0.512
$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$
0.726	6.822	-0.924	-2.803	-0.580	-17.498	7.033	-0.490

6.3.2 k-means分类求解

我们编写了Matlab代码画出高钾和铅钡两种类型不同k值的聚类偏差图，结果如下所示： 通过观察折线图的走势，从k = 2开始，当k逐步增大的过程中，k的聚类偏差整

Figure 7: 两种数据的k值聚类偏差图

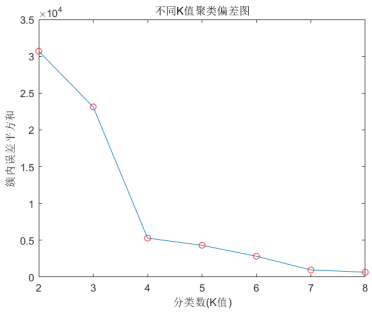


Figure 8: 高钾

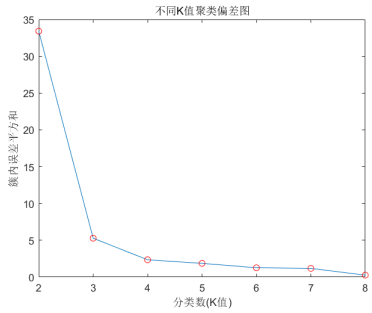


Figure 9: 铅钡

体是减少的，在k = 2到3之间下降的最大。我们需要寻找曲线的拐点，通过观察得到高钾曲线的拐点是k = 4，铅钡的曲线拐点是k = 3。

然后，在分类得到的不同类别中，求每个类别的化学成分的平均值作为总体的均值估计以及后续的分类依据，绘制的直方图如下：

6.3.3 分类标准确定

我们聚类的数据的处理后的数据，所以在我们分类的时候，以及将来对样本数据之外的数据分类是，首先就要进行零均值归一化的处理，归一化之后再通过以下的准则判断类别。以下讨论的数据都是归一化之后的数据。下述是对图表的分析结果。

对于**高钾类型**的文物,根据图示，高钾A类的显示它所有化学含量数据的均值都是在0之间波动，且波动距离不超过1，距离0最大的均值的化学成分是CaO和Al_2O_3，最主要的一点是其中大部分的数据都是负的，可以通过这一类来划分高钾A类；高钾B类则考虑Na_2O、K_2O、CaO、Al_2O_3、PbO,这五类数据，如果这些某文物的这几项指标

Figure 10: 各化学含量均值的直方图（归一化后）高钾

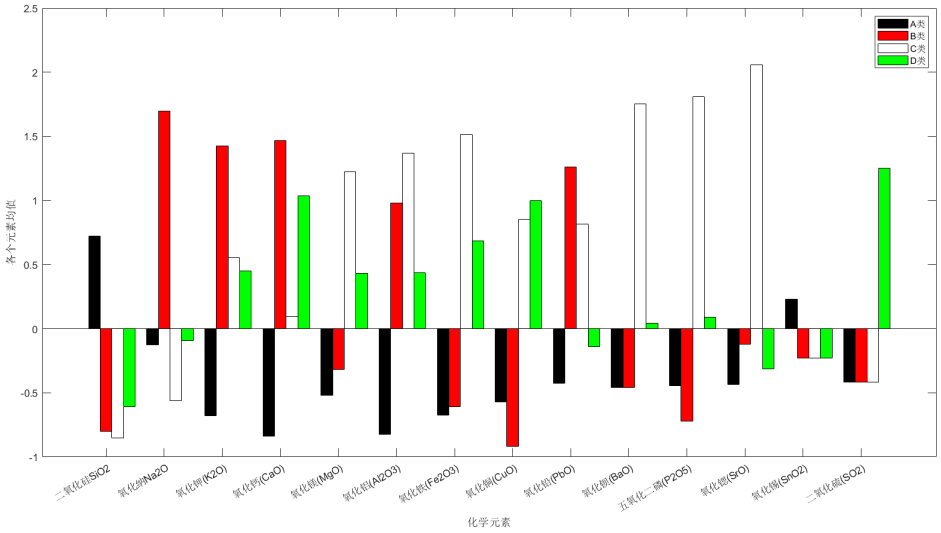
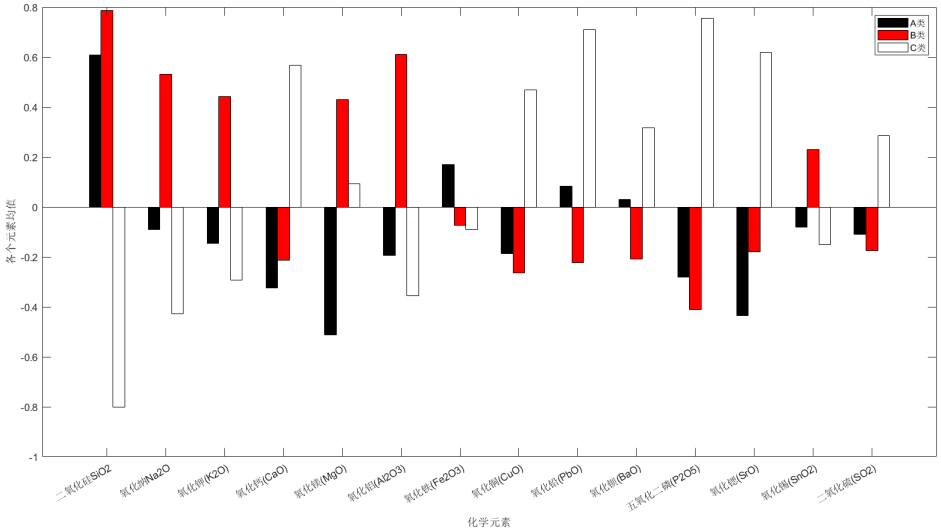


Figure 11: 各化学含量均值的直方图（归一化后）铅钡



的数据都为正值，并且到0的距离都能在1左右，甚至超过，那就判断为高钾B类；对于高钾C类，考虑SrO、P_2O_5、BaO、Fe_2O_3、Al_2O_3,这几类数据，如果数据值是都为正且都能够在1左右及以上，那么就判断为高钾C类；高钾D类则考虑SO_2、CuO、CaO,这三项数据能够为正，且在1左右下限为0.5，由此判断为高钾D类。

对于**铅钡类型**的文物，首先观察铅钡A类，发现SiO_2的数据和MgO的归一化数据到0的距离分别为正的最大值和负的最大值，可通过这两类这个指标来进行铅钡A的分类；对于铅钡B类，考虑SiO_2、Na_2O、K_2O、MgO、Al_2O_3这几个化学成分，当这几个数据的结果都是正值并且到0的距离都在0.4左右，那就判定这类为铅钡B类；铅钡C类考虑Si_2O、P_2O_5、PbO，这几个化学成分，如果SiO_2数据为负的，且到0的距离大概在0.6左右，其他两个数据都为正的并且距离在0.6附近，满足这样的数据就分类为铅钡C类。

大类	亚类	评价标准(归一化数据)
高钾	A类	考虑归一化的化学成分平均值在(-1,1)之间波动，化学成分中CaO和Al_2O_3的归一化数值绝对值最大，且其中大部分的数据都是负的
	B类	考虑Na_2O、K_2O、CaO、Al_2O_3、PbO：数值在1左右
	C类	考虑SrO、P_2O_5、BaO、Fe_2O_3、Al_2O_3：数值大于1 且都能够在由左右及以上
铅钡	D类	考虑SO2、CuO、CaO：数值在0.5以上，且三项数据均值大概为1
	A类	考虑SiO2和MgO：数值分别在所有化学组成中表现为最大值与最小值
	B类	考虑SiO_2、Na_2O、K_2O、MgO、Al_2O_3：数值处于区间(0,0.4)
	C类	考虑Si2O、P2O5、P bO:分别处于区间(-0.6,0)(0,0.6),(0,0.6)

Figure 12: 亚分类方法汇总表

合理性说明：

1. 数据演算的合理性，我们进行k-means聚类算法得到不同的类型，同时在选取成分估计值时采用了平均值，进一步减小误差
2. 指标选择的合理性，我们选择的分类的参考指标不少单一的，考虑了多个化学成分数据，使得对样本的数据能够彻底的划分成不同的类别，具有参考价值

6.3.4 灵敏性分析

首先引入**轮廓系数**：其适用于实际类别信息未知的情况。对于单个样本，设a_i是与它同类别中其他样本的平均距离，b_i是与它距离最近不同类别中样本的平均距离，其轮廓系数为：

$$\hat{s}_i = \frac{b_i - a_i}{max(a_i, b_i)}, 其中, i = 1, ..., 6 \tag{6}$$

对于一个样本集合，它的轮廓系数是所有样本轮廓系数的平均值。廓系数的取值范围是$[-1, 1]$，同类别样本距离越相近不同类别样本距离越远，分数越高。

研究当每个样本的某几个化学成分发生微小变化的时候，轮廓系数（Silhouette Coefficient）的变化量。首先，我们选择了样本数多的化学成分Al_2O_3、BaO、CaO、CuO、PbO、SiO_2进行分析。此时微小变化依旧取值为$0.05x_{ij}$,然后将其代入k-Means模型中计算轮廓系数的变化率。

$$\frac{\Delta \hat{s}_i}{\hat{s}_i} \times 100\% \tag{7}$$

得到的结果如上表所示，易知当化学成分发生微小变化的时候，其轮廓系数的变化率很小,对模型的判断效果影响很微小

Table 10: 各化学成分轮廓系数

化学成分	Al_2O_3	BaO	CaO	CuO	PbO	SiO_2
铅钡（k = 3）	-1.023	-4.247	0.486	-0.699	0.693	-3.127
高钾（k = 4）	2.571	0.410	3.333	0.756	6.861	6.743

7 问题三基于逻辑回归方程对未知文物类型的分类

7.1 数据预处理

1. 将表单三中的数据中空白的部分进行全部设置为0。
2. 将表单三中的定类数据转换成0-1二值，为后续的求解作准备，设置未风化为1，风化为0。

7.2 模型求解

通过问题二建立的对文物的类型分类的logistic模型，来对问题三中的类型进行分类预测。

Table 11: 模型的预测结果

文物编号	表面风化	判断结果
A1	无风化	高钾
A2	风化	铅钡
A3	无风化	铅钡
A4	无风化	铅钡
A5	风化	铅钡
A6	风化	高钾
A7	风化	高钾
A8	未风化	高钾

7.3 灵敏性分析

研究玻璃文物的表面的化学元素的成分含量的微小对$ln\frac{p_i^{''}}{1-p_i^{''}}$值的影响，这里将$f(p_i^{''}) = ln\frac{p_i^{''}}{1-p_i^{''}}$将其最后的计算结果和未变化之前的数据进行对比，这里采用数值分析，具体的计算步骤如下：

Step 1：数据进行微小变化处理

将每个化学含量的值通过matlab乘上1.05。这样得到了$x_{ij}+\Delta x_{ij}$,此时Δx_{ij}为 $0.05x_{ij}$然后将每一行的数据单独复制到原数据中，得到只有一个化学成分数据变化14*8的表格[3]。

Step 2：因变量变化量的计算

通过如下的计算式，计算到函数值的差值

$$\Delta f(p_i^{''}) = \sum_{i=1}^{14} \beta_i^{''} \Delta x_{ij}, j \in \{\text{A1，A2,...,A8}\} \tag{8}$$

具体的结果是：通过Matlab将数据循环读取，分别计算，计算结果一打印到表格中。表格的数据如下 由上图的结果可以推知，当每种的元素的结果变化

文物编号	二氧化硅(SiO_2)	氧化钠(Na_2O)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)	氧化锶(SrO)	氧化锡(SnO_2)	二氧化硫(SO_2)
A1	-0.037895	1.009413	1.009413	1.064437	0.449087	0.824324	1.087458	1.729134	1.009413	1.009413	0.978673	0.983166	1.009413	0.996917
A2	1.989087	2.493049	2.493049	2.562101	2.493049	2.433401	2.493049	2.493049	0.908389	2.493049	2.079219	2.493049	2.493049	2.493049
A3	2.840366	3.266898	3.371142	3.331968	3.022886	3.19189	3.523176	3.338529	1.438302	2.609595	3.189178	2.81195	3.266898	3.266898
A4	2.62541	3.098935	3.159488	3.125089	2.782622	2.917943	3.33307	3.426391	1.977199	1.934288	2.853885	2.853963	3.098935	3.098935
A5	1.515626	2.381037	2.402258	2.38874	1.668972	2.047498	2.4033	2.694531	1.808872	2.071174	2.368388	2.190169	2.546206	2.373897
A6	-0.600909	0.64291	0.746388	0.648702	0.579648	0.603998	0.652711	1.233014	0.64291	0.64291	0.63682	0.64291	0.64291	0.64291
A7	-0.357051	0.855529	0.930647	0.865666	0.855529	0.725994	0.864242	1.254617	0.855529	0.855529	0.85176	0.855529	0.855529	0.852835
A8	-0.105205	0.577247	0.594877	0.585302	0.577247	0.522975	0.577247	3.650558	-0.404041	-1.012054	0.534907	0.306028	0.577247	0.521877

Figure 13: 每个类型的灵敏性变化结果

Step 3：结果分析

观察上述的数据，发现总体的正负性没有发生改变，同时整体的数据变化不大，推测可能是采用了取ln的缘故。通过整体的数据分析发现，分类的结果还是和之前一样，说明微小变化不会对模型的预测产生影响，说明这个分类模型的稳定性很高，同时说明这两类（文物的玻璃类型）是线性可分的。

对具体的数据进行分析，由于数据偏离0的位置基本没有变化，所以转换到p的变化是十分微小的，不超过e^{-9},这个结果大概是10^{-4},所以得出结论，数据的化学成分的微小变化基本不会带来概率的变化。

8 问题四的对化学组成成分关联性和差异性分析

8.1 建模思路

本问有两个主要要求，第一个要求就是要我们将数据分成铅钡和高钾两类，然后分别对每一类分析化学成分之间的联系，所以我们使用相关系数分析来解释两两变量之间的是否存在线性相关关系。第二个要求是让我们比较不同类别之间的化学成分关联关系的差异性，这个差异性通过对要求一得到的关联性进行分析，定性得出差异性规律

8.2 模型建立与求解

相关分析中最简单的相关关系就是两个变量之间是否存在线性相关关系我们用样本的相关系数r_{ij}估计总体的相关系数ρ_{ij}

Step1：相关系数的计算

分别计算两两样本之间的简单相关系数r_{ij},其构成矩阵,其中n = 14

$$\begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{bmatrix} \tag{9}$$

$$r_{ij} = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}} \tag{10}$$

引入**相关性判断标准：**

- r绝对值大于0.8即具有强相关关系。
- r绝对值介于0.5至0.8即具有中等相关关系。
- r绝对值小于0.5时相关关系不明显

计算的结果如下下图所示，分析图表13的结果可以得到，在铅钡的这个文物类别中 SiO_2与Al_2O_3、CuO、PbO、BaO呈强正相关; Al_2O_3与CuO、PbO、BaO呈强正相关; CuO与PbO、BaO呈强正相关; PbO与BaO呈强正相关;

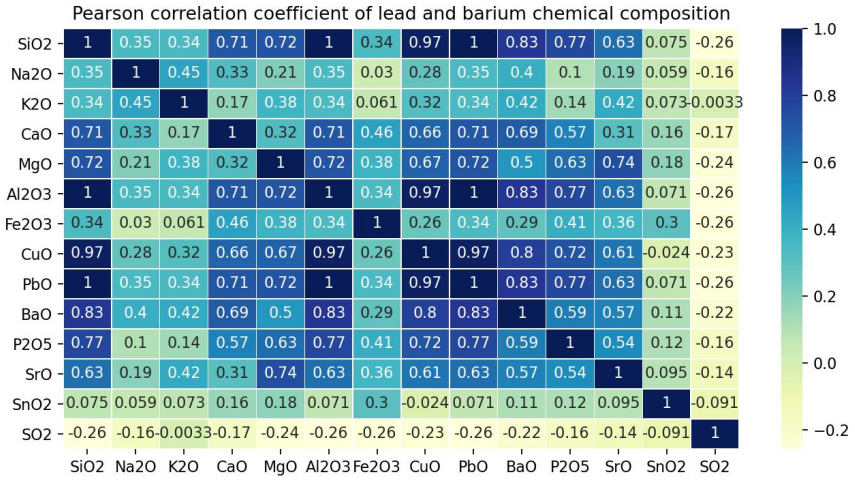


Figure 14: 铅钡玻璃各组成成分的Pearson相关系数表

同样根据相关性判断准则，结合图表14可知，在高钾这个文物类别中 $SiO2$与Al_2O_3、CuO、PbO、BaO呈强正相关 Al_2O_3与Fe_2O_3、CuO、P_2O_5呈强正相关 Fe_2O_3与CuO、P_2O_5呈强正相关 CuO与P_2O_5呈强正相关

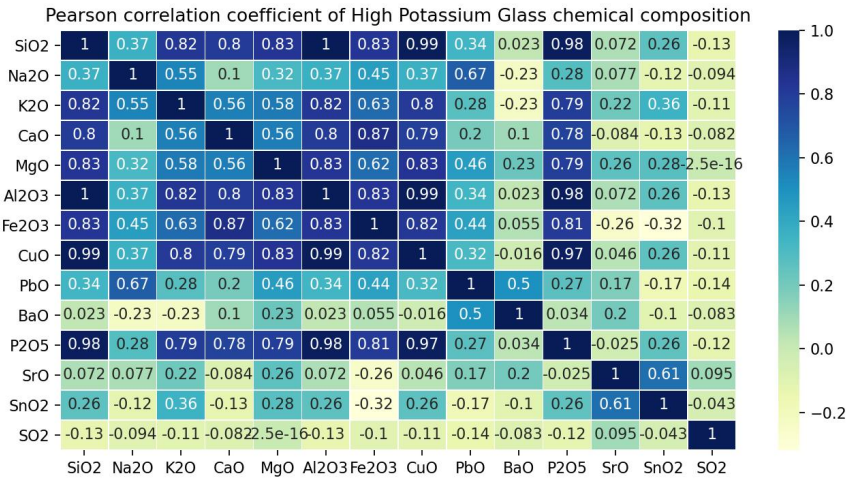


Figure 15: 高钾玻璃各组成成分的Pearson相关系数表

Step 2：假设检验流程

第一步、建立假设检验 第二步、构造t分布作为检验统计量:

Table 12: 假设检验表

假设检验	假设	检验结果
样本都是从$\rho = 0$的总体抽取的样本	H_0:$\rho = 0$;H_1:$\rho \neq 0$	接受 H_0,拒绝H_1

$$t_{ij} = r_{ij}\sqrt{\frac{n-2}{1-r_{ij}^2}}, \text{其中n为样本数量,} t_{ij} \text{为} r_{ij} \text{对应的t值} \tag{11}$$

代入r值计算相应的检验值t，并且在给定置信水平 $\alpha = 0.05$,利用matlab计算t值的对应的p值

Step 3：计算结果与分析

在显著水平为0.05的情况下，将p值超过0.05的视为无效数据，在图中标红。结果表明（结合r值对p值解读）：

1. 铅钡玻璃中:SiO_2与Al_2O_3, CuO, PbO, BaO呈显著线性正相关；Al_2O_3与CuO, PbO, BaO呈显著线性正相关；CuO与BaO呈显著线性正相关；(而上述相关系数绝对值较大的几组化学成分$CuO - PbO$及$PbO - BaO$之间没有显著线性相关性)

2. 高钾玻璃中:SiO_2与Al_2O_3呈显著线性正相关； Al_2O_3与Fe_2O_3呈显著线性正相关；Fe_2O_3与CuO， P_2O_5呈显著线性正相关；(而上述相关系数绝对值较大的几组化学成分：SiO_2-CuO， SiO_2-PbO， SiO_2-BaO， Al_2O_3-CuO,CuO-P_2O_5之间没有显著线性相关性)

Step 4：关联性的差异性分析

按上述结果：

p值(p-value)

铅钡玻璃	高钾玻璃
p(SiO2-Al2O3)=0.00429	p(SiO2-Al2O3)=0.00401
p(SiO2-CuO)=0.0125	p(SiO2-CuO)=0.116
p(SiO2-PbO)=0.00000000149	p(SiO2-PbO)=0.121
p(SiO2-BaO)=0.00173	p(SiO2-BaO)=0.245
p(Al2O3-CuO)=0.0499	p(Al2O3-Fe2O3)=0.00152
p(Al2O3-PbO)=0.00151	p(Al2O3-CuO)=0.261
p(Al2O3-BaO)=0.017	p(Al2O3-P2O5)=0.000894
p(CuO-PbO)=0.562	p(Fe2O3-CuO)=0.0169
p(CuO-BaO)=6.03E-09	p(Fe2O3-P2O5)=9.56E-05
p(PbO-BaO)=0.337	p(CuO-P2O5)=0.469

Figure 16: 两两样本之间的p值

1. 按SiO_2 含量分析差异性：
   - 铅钡玻璃中：SiO_2与Al_2O_3,CuO,PbO,BaO呈显著线性正相关；Al_2O_3与CuO,PbO,BaO显著线性正相关；
   - 而高钾玻璃中:SiO_2仅与Al_2O_3呈显著线性正相关；
2. 按Al_2O_3 含量分析差异性：
   - 铅钡玻璃中:Al_2O_3与CuO, PbO, BaO均呈现显著线性正相关性；
   - 而高钾玻璃中:Al_2O_3仅与Fe_2O_3出现较为显著的线性正相关性；
3. 其他化学组成成分关联关系的差异:铅钡玻璃中:CuO还呈现出与BaO的显著线性正相关性；

9 模型评价

9.1 优点分析

(一)、通过0-1变量对数据转化

在遇到第一个问题的时候，我们将数据升维的方法，把三个指标上升到十三维度，每个维度仅对应0-1两个变量，这样就成功的将定类数据转化成了定量数据，这样通过逻辑回归得到回归系数之后，进行数据回代，发现其分类的精确度有0.833，属于效果比较好的分类。而如果保留之前的三维数据，是得不到这样的结果，甚至是无从下手。但是升维就完美的解决了这个问题。我们小队对这个方案一开始也是反对的，一般数据大部分是进行降维，结果我们这次的逆向思维带来了不错的效果。我们小组探讨的时候，觉得这样的升维更加完备的展示了数据属性与待分类类别内部关系。使得数据分化更明显。

(二)、分类规则明确且准确

采用逻辑回归，在拟合数据的时候，可以直接给出回归系数，得到回归方程。这就得到了一个明确的划分界限，当其拟合结果确定之后，用于划分的逻辑回归方程也就确定了，十分的清楚明白。同时，召回率很高，准确率也很高，四个分类结果分别是0.833、0.9592、1、1。如果不考虑过拟合的情况，这个准确率很高了。并且在第三问中的灵敏性分析也很好的反映了模型的健壮性，其数据的微小波动不会带来最终分类结果的判断。同时这也反映了logistic回归分类在解决这个问题的优越性。

(三)、分类方法多样

在使用逻辑回归之外，我们在亚分类的问题中，使用了K-Means聚类算法，两种数据的k值聚类偏差图都存在明显的拐点，说明这个算法的合理性，并且在每个大类中的亚类在化学成分方面有明显的区分度，引入轮廓系数来分析这个聚类算法的灵敏性，也发现微小变化对聚类算法评分影响不大。

9.2 缺点分析

(一)、在部分问题上舍弃了数据

当进行风化点数据预测其未风化是化学含量的时候，我们选择只选择了一部分数据，没有采用全部的数据进行分析，这样会导致最终的预测结果有很大的偏差，其次就是在灵敏性分析的时候只对数据样本的量够大化学元素进行微小变化设置，并没有对全部数据进行灵敏性分析。

(二)、可能存在过拟合的情况

由于在对高钾的数据判断风化与否的逻辑回归的时候，其两类的数据并没有接近1：1，而且在后续的对玻璃文物的分类的时候，两类数据也没有接近1：1。但召回率达到了100%，虽然两个数据比例并没有巨大偏差，但是也有可能存在过拟合的现象。

(三)、数据预测采用线性

风化前成分预测模型过于简单，只考虑了最简单的线性关系，由于数据样本稀疏，无法得出有效的参数估计值，具有一定的误差，模型的有效性不能够满足预测需求，即模型欠拟合,后续可以采用非线性回归等方式提高模型的复杂度，数据处理方面，可以考虑收集数据特征，使用重采样方法提高模型的拟合效果

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] 王森,刘琛,邢帅杰.K-means聚类算法研究综述[J/OL].华东交通大学学报:1-8[2022-09-15 10:22].https://doi.org/10.16749/j.cnki.jecjtu.20220914.001.DOI:10.16749/j.cnki.jecjtu.202209

[3] 王凌妍,张鑫雨,许胜楠,王禹力,甄志龙.逻辑回归的敏感性分析及在特征选择中的应用[J].信息记录材料,2022,23(07):30-33.DOI:10.16009/j.cnki.cn13-1295/tq.2022.07.051

附录1：python代码

```
# python的代码，计算轮廓系数的代码
from sklearn.cluster import KMeans
import warnings
import pandas as pd
warnings.filterwarnings("ignore") # 忽略警告
from sklearn.metrics import silhouette_score, silhouette_samples

# silhouette_score 返回所有样本的轮廓系数的均值
# silhouette_samples 返回每个样本的轮廓系数

'''数据集导入，若需要进行复现，请注意不同测试文件相对路径（或绝对路径）并更改文件名字
以下gj_yuan指高钾类型玻璃化学成分的原始数据，即，我们计算出来的轮廓系数是原始数据对应的值
可以通过改变导入的文件计算不同化学成分在增加5%的情况下轮廓系数的数值。
'''
X = pd.read_excel('gj_yuan.xlsx')
'''注意：文件放在了data文件夹下'''

sil_scores = []
for n in range(3,4):
    km = KMeans(n_clusters=n).fit(X)
    # 轮廓系数接收的参数中，第二个参数至少有两个分类
    sc = silhouette_score(X, km.labels_)
    sil_scores.append(sc)
    print("n_clusters: {}\t silhoutte_score: {}".format(n,  sc))
```

附录2：matlab代码

```
%灵敏性分析的微小变化的处理代码
clc,clear;
A = readmatrix('sheet3.xlsx');
data = A(:,3:16);
data = data*1.05;
writematrix(data,'sheet3.xlsx','range','C2:P9');
```

```
xiangguanxishu=regress(y,x);
LinearModel.fit(data(:,i),data(:,j))$\%$计算p值
data=xlsread('8-14 zhengfu.xlsx');
a=zeros(length(data),1);
i=1;
while i<=length(data)
a(i,1)=38.259-7.372*data(i,1)-0.267*data(i,2)+0.119*data(i,3)+1.533*data(i,4)+0.181*d
i=i+1;
end
```

4.1求p值判断类别

```
%读取数据，生成容纳预测结果的空矩阵,便于快速完整的记录结果
data=xlsread('predict');
a=zeros(length(data),1);
%因为公式为ln(p/(1-p))=a0+a1*x1+a2*x2+.... ,为了计算简便，我们将等式左端先
用a记录出来。
i=1;
while i<=length(data)
a(i,1)=38.259-7.372*data(i,1)-0.267*data(i,2)+0.119*data(i,3)+1.533*data(i,4)+0.181*d
i=i+1;
end
%通过公式变化，得出来a到p的映射关系，带入a值，运算得出p值。
p=exp(a)./(1+exp(a));
%p代表是否为高钾的概率，因为a值算出结果为正负，所以p值要么是1，要么为0。将
预测结果p值为1的数据总结为高钾，预测结果p值为0的数据总结为铅钡。
```

1.3风化前预测

```
%%本题（1.3）我们选择利用线性回归进行预测
%因为同一个文物未分化与分化点都有的数据过少，文物某些元素含量过低导致变化值
不易处理等问题，我们对此题进行了以下假设
%1.认为风化前后对其他元素无影响，仅对SiO2，Al2O3，CuO，P2O5，SrO含量产生影
响
%2.认为风化前后与风化到严重风化的影响相似，可以利用这些数据进行线性分析
Data=xlsread('bian.xlsx');
```

```
%将风化前后的数据分离开，作为x，y值
bianqian=[Data(1,:);Data(3,:);Data(5,:);Data(7,:);Data(9,:)];
bianhou=[Data(2,:);Data(4,:);Data(6,:);Data(8,:);Data(10,:)];
%%所变化的化学元素SiO2，Al2O3，CuO，P2O5，SrO分别为1，6，8，11，12列，并以此作为元素编号
%%x为分化前的数据，y为分化后的数据，角标为元素编号，利用regress函数直接求得线性回归方程中的b值
x1=bianqian(:,1)';y1=bianhou(:,1)';X1=[ones(size(x1')) x1'];Y1=y1';
[b1]=regress(Y1,X1);
x6=bianqian(:,6)';y6=bianhou(:,6)';X6=[ones(size(x6')) x6'];Y6=y6';
[b6]=regress(Y6,X6);
x8=bianqian(:,8)';y8=bianhou(:,8)';X8=[ones(size(x8')) x8'];Y8=y8';
[b8]=regress(Y8,X8);
x11=bianqian(:,11)';y11=bianhou(:,11)';X11=[ones(size(x11')) x11'];Y11=y11';
[b11]=regress(Y11,X11);
x12=bianqian(:,12)';y12=bianhou(:,12)';X12=[ones(size(x12')) x12'];Y12=y12';
[b12]=regress(Y12,X12);
%%导入所有数据，找到分化了的数据，导入矩阵fenhua
suoyou=xlsread('数据.xlsx');
i=1,j=1;
while i<=69
if suoyou(i,6)==1
fenhua(j,:)=suoyou(i,:)
j=j+1;
end
i=i+1;
end
%%处理分化数据，仅保留化学元素的7-20列
fenhua(:,1:6)=[];
%开始预测，根据已求的b值带入，处理方程为分化后到分化前的映射，将分化后的数据带入方程，求得预测值
fenhuaqian1=(fenhua(:,1)-b1(1))./b1(2);
fenhuaqian6=(fenhua(:,6)-b6(1))./b6(2);
fenhuaqian8=(fenhua(:,8)-b8(1))./b8(2);
fenhuaqian11=(fenhua(:,11)-b11(1))./b11(2);
fenhuaqian12=(fenhua(:,12)-b12(1))./b12(2);
fenhuaqian=[fenhuaqian1,fenhuaqian6,fenhuaqian8,fenhuaqian11,fenhuaqian12]
```

```
%因为化学含量百分比不可能超过100，也不可能低于0，所以我们对这些异常的预测进行处理（大于100的记为100，小于0的记为0）
i=1,j=1;
[h,z]=size(fenhuaqian);
while i<=h
j=1;
while j<=z
if fenhuaqian(i,j)>100
fenhuaqian(i,j)=100
end
if fenhuaqian(i,j)<0
fenhuaqian(i,j)=0
end
j=j+1;
end
i=i+1;
end
```

2.2高钾求k值

```
% % 选择聚类算法，利用k-means进行分类
% %
% % 因为k-means算法的结果对k值依赖很重，所以我们需要更精确的方法求解k值。
% % 这里采用"肘部法"，观察SSE误差平方和的变化。其核心思想是分类数k越大，样本划分会更加精细，每个类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
% % 通过matlab绘图，画出相应的SSE随k值变化的折线图，最终我们选择图像拐点，例如此题高钾问题中k=4，这点之前SSE下降迅速，之后SSE下降缓慢也就是说4类确实是效果最好。
%提取数据，并利用零均值对已知数据进行标准化
clear,clc;
Gj=xlsread('gj.xlsx');
Gjjun=mean(Gj,1);
Gjstd=std(Gj,1);
h=1;
while h<=14 %%共14种元素
gj(:,h)=(Gj(:,h)-Gjjun(1,h))./Gjstd(1,h);
h=h+1;
```

```
end
[n,p]=size(gj);
%计算1~8的K值情况下SSE图
K=8;D=zeros(K,2);
for k=2:K
[lable,c,sumd]=kmeans(gj,k,'dist','sqeuclidean');
% lable: n维列向量，得出聚类标签结果;
% c: k×p向量，最终k个聚类质心的位置
% sumd: k维列向量，该类质心点与类间所有点距离之和
sse1=sum(sumd.^2);%对SSE求和
D(k,1)=k;
D(k,2)=sse1;
end
%绘图
plot(D(2:end,1),D(2:end,2))
hold on;
plot(D(2:end,1),D(2:end,2),'or');
%给图片添加标题，标签，便于观察
title('不同K值聚类偏差图')
xlabel('分类数(K值)')
ylabel('簇内误差平方和')
```

2.2高钾已知k后亚分类

```
%%通过肘部法绘制SSE图，确认好k值之后，对已有数据进行分类编号，并绘制出条形图，直观的反应各类比中各化学成分的含量，便于亚分类
%%(注：由于电脑分类后每类得取名不同，所以图会有变化)
k=4;
%%直接利用matlab自带的函数kmeans，输入已求得的k值，进行分类，输出结果保存于gjfenlei
gjfenlei = kmeans(gj,k)
%Ai,Bi,Ci,Di分别为各个类比的个数，用于分类时将数据分为4组。
i=1,Ai=1,Bi=1,Ci=1,Di=1;
%n是数据个数（上部分代码已导出）
%遍历各组数据，将类别不同的数据分别导入4个矩阵中
while i<=n
if gjfenlei(i,1)==1
A(Ai,:)=gj(i,:);
```

```
Ai=Ai+1;
end
if gjfenlei(i,1)==2
B(Bi,:)=gj(i,:);
Bi=Bi+1;
end
if gjfenlei(i,1)==3
C(Ci,:)=gj(i,:);
Ci=Ci+1;
end
if gjfenlei(i,1)==4
D(Di,:)=gj(i,:);
Di=Di+1;
end
i=i+1;
end
%%对每类中每个元素求均值，作为y轴
Ajun=mean(A,1);
Bjun=mean(B,1);
Cjun=mean(C,1);
Djun=mean(D,1);
%%把4类数据分别绘入同一张条形图,并对条形图宽，图间距，总长，颜色，x轴标签，
y轴标签，注释栏等进行设置，便于观察
bar1=bar([2:5:67],Ajun,'BarWidth',0.2,'FaceColor','k');
hold on;
bar2=bar([3:5:68],Bjun,'BarWidth',0.2,'FaceColor','r');
hold on;
bar3=bar([4:5:69],Cjun,'BarWidth',0.2,'FaceColor','w');
hold on;
bar4=bar([5:5:70],Djun,'BarWidth',0.2,'FaceColor','g');
ylabel('各个元素均值')
xlabel('化学元素')
legend('A类','B类','C类','D类');
labelID={'二氧化硅SiO2','氧化纳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)'}
set(gca,'XTick',3:5:71);
```

```
set(gca,'XTickLabel',labelID)
```

2.2铅钡求k值

```
% % 选择聚类算法，利用k-means进行分类
% %
% % 因为k-means算法的结果对k值依赖很重，所以我们需要更精确的方法求解k值。
% % 这里采用"肘部法"，观察SSE误差平方和的变化。其核心思想是分类数k越大，样本划分会更加精细，每个类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
% % 通过matlab绘图，画出相应的SSE随k值变化的折线图，最终我们选择图像拐点，例如此题中铅钡问题中k=3，这点之前SSE下降迅速，之后SSE下降缓慢也就是说分3）类确实是效果最好。
%提取数据，并利用零均值对已知数据进行标准化
clear,clc;
Qb=xlsread('铅钡.xlsx');
Qbjun=mean(Qb,1);
Qbstd=std(Qb,1);
h=1;
while h<=14 %%共14种元素
qb(:,h)=(Qb(:,h)-Qbjun(1,h))./Qbstd(1,h);
h=h+1;
end
[n,p]=size(qb);
%计算1~8的K值情况下SSE图
K=8;D=zeros(K,2);
for k=2:K
[lable,c,sumd]=kmeans(qb,k,'dist','sqeuclidean');
% lable: n维列向量，得出聚类标签结果;
% c: k×p向量，最终k个聚类质心的位置
% sumd: k维列向量，该类质心点与类间所有点距离之和
sse2=sum(sumd.^2);%对SSE求和
D(k,1)=k;
D(k,2)=sse2;
end
%绘图
plot(D(2:end,1),D(2:end,2))
hold on;
```

```
plot(D(2:end,1),D(2:end,2),'or');
%给图片添加标题，标签，便于观察
title('不同K值聚类偏差图')
xlabel('分类数(K值)')
ylabel('簇内误差平方和')
```

2.2铅钡已知k后亚分类

```
%%通过肘部法绘制SSE图，确认好k值之后，对已有数据进行分类编号，并绘制出条形图，直观的反应各类比中各化学成分的含量，便于亚分类
%%(注：由于电脑分类后每类得取名不同，所以图会有变化)
k=3;
%%直接利用matlab自带的函数kmeans，输入已求得的k值，进行分类，输出结果保存于qbfenlei
qbfenlei = kmeans(qb,k)
%Ai,Bi,Ci,Di分别为各个类比的个数，用于分类时将数据分为4组。
i=1,Ai=1,Bi=1,Ci=1;
%n是数据个数（上部分代码已导出）
%遍历各组数据，将类别不同的数据分别导入4个矩阵中
while i<=n
if qbfenlei(i,1)==1
A(Ai,:)=qb(i,:);
Ai=Ai+1;
end
if qbfenlei(i,1)==2
B(Bi,:)=qb(i,:);
Bi=Bi+1;
end
if qbfenlei(i,1)==3
C(Ci,:)=qb(i,:);
Ci=Ci+1;
end
i=i+1;
end
%%对每类中每个元素求均值，作为y轴
Ajun=mean(A,1);
Bjun=mean(B,1);
Cjun=mean(C,1);
```

```
Djun=mean(D,1);
%%把4类数据分别绘入同一张条形图,并对条形图宽，图间距，总长，颜色，x轴标签，
y轴标签，注释栏等进行设置，便于观察
bar1=bar([2:5:67],Ajun,'BarWidth',0.2,'FaceColor','k');
hold on;
bar2=bar([3:5:68],Bjun,'BarWidth',0.2,'FaceColor','r');
hold on;
bar3=bar([4:5:69],Cjun,'BarWidth',0.2,'FaceColor','w');
ylabel('各个元素均值')
xlabel('化学元素')
legend('A类','B类','C类');
labelID={'二氧化硅SiO2','氧化纳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)'}
set(gca,'XTick',3:5:71);
set(gca,'XTickLabel',labelID)
```

附录3：数据图片

文物编号	表面风化	二氧化硅 (SiO_2)	氧化钠 (Na_2O)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al_2O_3)	氧化铁 (Fe_2O_3)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P_2O_5)	氧化锶 (SrO)	氧化锡 (SnO_2)	二氧化硫 (SO_2)	判断结果
A1	无风化	78.45	0	0	6.08	1.86	7.23	2.15	2.11	0	0	1.06	0.03	0	0.51	高钾
A2	风化	37.75	0	0	7.63	0	2.33	0	0	34.3	0	14.27	0	0	0	铅钡
A3	无风化	31.95	0	1.36	7.19	0.81	2.93	7.06	0.21	39.58	4.69	2.68	0.52	0	0	铅钡
A4	无风化	35.47	0	0.79	2.89	1.05	7.07	6.45	0.96	24.28	8.31	8.45	0.28	0	0	铅钡
A5	风化	64.29	1.2	0.37	1.64	2.34	12.75	0.81	0.94	12.23	2.16	0.19	0.21	0.49	0	铅钡
A6	风化	93.17	0	1.35	0.64	0.21	1.52	0.27	1.73	0	0	0.21	0	0	0	高钾
A7	风化	90.83	0	0.98	1.12		5.06	0.24	1.17	0	0	0.13	0	0	0.11	高钾
A8	无风化	51.12	0.00	0.23	0.89	0.00	2.12	0.00	9.01	21.24	11.34	1.46	0.31	0.00	2.26	高钾

Figure 17: 问题三预测结果全数据表

纹饰-A	纹饰-B	纹饰-C	颜色-1-青绿	颜色-2-浅青	颜色-3-紫	颜色-4-深绿	颜色-5-深青	颜色-6-浅绿	颜色-7-黑	颜色-8-绿	类型-1-高钾	表面风化
0	0	1	1	0	0	0	0	0	0	0	1	0
1	0	0	0	1	0	0	0	0	0	0	0	1
1	0	0	1	0	0	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	0	0	1	0
0	1	0	1	0	0	0	0	0	0	0	1	1
0	0	1	0	0	1	0	0	0	0	0	0	1
0	1	0	1	0	0	0	0	0	0	0	1	1
0	1	0	1	0	0	0	0	0	0	0	1	1
0	0	1	0	1	0	0	0	0	0	0	0	1
0	1	0	1	0	0	0	0	0	0	0	1	1
0	0	1	0	1	0	0	0	0	0	0	1	0
0	0	1	0	0	0	1	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	1	0	0	0	1	0
1	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	1	0
0	1	0	1	0	0	0	0	0	0	0	1	1
1	0	0	1	0	0	0	0	0	0	0	0	1
0	0	1	0	0	1	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	0	0	0	1
0	0	1	0	0	1	0	0	0	0	0	0	1
0	1	0	1	0	0	0	0	0	0	0	1	1
1	0	0	0	1	0	0	0	0	0	0	0	1
1	0	0	0	1	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1	0	0	0	0	0
0	0	1	0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	1	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0	0	1
0	0	1	0	0	0	0	0	1	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0	0	1
0	0	1	0	0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0	0	1
0	0	1	0	0	0	1	0	0	0	0	0	1
0	0	1	0	0	0	0	0	1	0	0	0	1
1	0	0	0	1	0	0	0	0	0	0	0	1
0	0	1	0	1	0	0	0	0	0	0	0	1
1	0	0	0	1	0	0	0	0	0	0	0	1
1	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	1
1	0	0	0	0	0	0	0	0	1	0	0	1
0	0	1	0	1	0	0	0	0	0	0	0	1
0	0	1	0	1	0	0	0	0	0	0	0	1
1	0	0	0	1	0	0	0	0	0	0	0	1
0	0	1	0	1	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0	1	0	0
0	0	1	1	0	0	0	0	0	0	0	0	1
0	0	1	1	0	0	0	0	0	0	0	0	1

Figure 18: 问题一的十三维数据

编号	SiO2	Al2O3	CuO	P2O5	SrO
2	83.377%	5.937%	0.000%	10.686%	0.000%
7	60.153%	2.308%	0.000%	37.539%	0.000%
8	63.024%	1.538%	6.246%	29.193%	0.000%
8	65.155%	2.078%	1.775%	30.993%	0.000%
9	65.837%	0.992%	1.821%	31.350%	0.000%
10	65.401%	3.001%	0.000%	31.598%	0.000%
11	65.143%	1.110%	2.077%	31.669%	0.000%
12	66.601%	0.523%	0.000%	32.876%	0.000%
19	49.478%	2.585%	1.061%	46.876%	0.000%
22	73.901%	5.804%	13.683%	6.613%	0.000%
26	64.567%	6.978%	0.000%	28.454%	0.000%
26	57.537%	1.571%	2.113%	38.779%	0.000%
27	60.362%	1.481%	0.000%	38.156%	0.000%
34	76.639%	5.797%	11.859%	5.705%	0.000%
36	61.419%	11.631%	0.000%	26.950%	0.000%
38	74.215%	3.826%	18.951%	2.888%	0.120%
39	36.753%	1.324%	34.691%	27.139%	0.093%
40	73.588%	10.262%	2.913%	13.048%	0.188%
41	56.555%	2.822%	0.000%	40.492%	0.131%
43	70.201%	1.707%	0.000%	27.847%	0.246%
43	35.741%	0.537%	34.817%	28.740%	0.166%
48	88.641%	11.044%	0.000%	0.000%	0.315%
49	86.192%	8.467%	4.928%	0.000%	0.413%
50	70.287%	7.720%	0.000%	21.608%	0.386%
51	39.345%	3.240%	25.403%	31.254%	0.759%
51	54.748%	0.355%	0.000%	44.518%	0.379%
52	28.799%	2.879%	25.143%	42.339%	0.841%
54	28.677%	2.818%	18.804%	49.297%	0.404%
54	65.426%	3.859%	0.000%	29.995%	0.720%
56	46.900%	0.356%	0.000%	52.167%	0.577%
57	57.456%	6.811%	0.000%	34.906%	0.826%
58	83.003%	10.722%	4.195%	0.000%	2.081%

Figure 19: 风化前预测结果