

題目

摘要

摘要

关键字：关 键 字

一、问题重述与问题分析

1.1 问题重述

玻璃是早期丝绸之路的重要商品，从西亚和埃及地区传入我国后，古人因地制宜、就地取材制作出外观相似但化学成分不同的玻璃制品。玻璃炼制时需要助熔剂，不同的助熔剂会导致玻璃的化学成分不同，例如铅钡玻璃和钾玻璃分别以铅矿石和草木灰作为助熔剂。同时，古代玻璃易与外界环境进行元素交换而导致化学成分比例发生改变。

现有提供高钾玻璃和铅钡玻璃的相关数据，完成下列问题：

1. 分析玻璃文物表面风化和玻璃类型、纹饰和颜色的关系；对不同的玻璃类型分析表面风化与否和玻璃化学成分含量之间的统计规律；根据风化点的检测数据预测风化前的化学成分含量；
2. 根据附件数据分析铅钡玻璃和高钾玻璃的分类规律；对不同的玻璃类型，选择适当的化学成分划分亚类，并给出具体的方法和结果，并对分类结果进行合理性和敏感性分析；
3. 针对附件表单 3 的化学成分，鉴别其所属的类型，同样需要对结果进行敏感性分析；
4. 分析不同类别玻璃文物的化学成分之间的关系以及比较不同类别化学成分关系的差异性。

1.2 问题分析

1.2.1 问题一的分析

问题一要求分析附件表 1 中文物表面风化与玻璃类型、颜色和纹饰之间的关系，随后分析表 2 数据，得出风化与化学成分含量的统计规律，最后依据表 2 风化点的检测数据，预测风化前的化学成分含量。本文首先对附件表单 1 的数据进行可视化，作出以玻璃类型、纹饰和颜色为自变量，是否风化为因变量的三维气泡图，再对高钾玻璃和铅钡玻璃分别作风化与纹饰和颜色的二维散点图，得到四者之间的定性关系。接着，利用各个化学成分含量风化前后变换的散点图与 spearman 系数表，相互印证，得到对于某种玻璃类型最重要的几个化学成分指标，进一步分析得出化学成分与风化的统计规律。最后，利用前文挑选的重要的化学成分指标构建逻辑回归模型，并同时通过不等幅度地调节成分含量，使风化结果翻转，达到预测风化前化学成分含量的效果。

1.2.2 问题二的分析

问题二要求分析两类玻璃的分类规律，之后选取合适的化学成分划分亚类并给出方法和具体结果，并对亚类划分结果做合理性和敏感性分析。本文首先作出不同玻璃文物的不同化学成分含量的散点图，得出初步的高钾玻璃和铅钡玻璃的分类规律。随后通过计算所有玻璃文物的类别与化学成分的 *spearman* 系数、风化文物的类别与化学成分的 *spearman* 系数和未风化文物的类别与化学成分的 *spearman* 系数对分类规律做定量验证，最终得出最后的玻璃分类规律。通过对高钾玻璃和铅钡玻璃分别做化学成分含量的分布散点图，并结合方差分析选定最终亚类划分的化学成分，然后根据挑选的化学成分分类对高钾玻璃和铅钡玻璃做 K-means 聚类得到最后的亚类划分。

1.2.3 问题三的分析

问题三要求预测附件表 3 中文物的类型，并对结果做敏感度分析。通过观察附件表 3 的数据可以发现氧化纳、氧化锡、二氧化硫等化学成分由于空白信息过多而变得离散，可将其视为“标签”特征，因此引入决策树处理这些“离散的信息”。表三还提供了如二氧化硅、氧化铝、五氧化二磷等没有空白值的化学成分，SVM 更善于处理这样的化学成分特征，因此本文联合决策树和支持向量机两种分类方法，利用二者特点，训练不同特点的化学成分信息，得到两个分类模型，最后综合给出最后的类别预测。

1.2.4 问题四的分析

问题四要求分析同类玻璃化学成分的联系以及不同类别化学成分联系的差异性。未完待续

二、模型假设与符号说明

2.1 模型假设

- 假设各个化学成分含量仅受风化影响，不受人为或其他与奉化无关的自然因素的影响
- 假设在表单的检测中，若未检测到某一化学成分，则此化学成分的含量为 0；
- 假设表面风化的无风化检测点具有的无风化特性较多，而表面无风化的风化检测点具有的风化特性较多；

三、数据预处理

3.1 数据清洗

3.1.1 处理缺失值

对于表单 1 的数据，我们首先找出缺失值，缺失值数据如表1所示

表1 表单1 缺失值数据

文物编号	纹饰	类型	颜色	表面风化
19	A	铅钡	NaN	风化
40	C	铅钡	NaN	风化
48	A	铅钡	NaN	风化
58	C	铅钡	NaN	风化

可以观察到，表单 1 的数据仅存在颜色属性的缺失，缺失数据量为 4，本文中将其作为一种新的颜色特征进行分析，故对此数据保留。

3.1.2 处理含量空白值

对于表单 2 的数据，由于空白处表示未检测到相关化学成分，故对空白处使用“补 0”操作。

3.1.3 处理异常值

对于表单 2 的数据，针对每个文物样品进行成分比例数据的加和，求解出 15, 17 号文物的成分比例累加和低于 85%，属于异常数据，进行剔除操作。

3.1.4 处理奇异值

对于表单 2 的数据，部分数据存在文物编号为“未风化点”，而表面“风化”，对于此类数据进行分类与标记，在数据处理过程中需考虑其“未风化”的特性。

3.2 数据规约

3.2.1 归一化处理

由于表单 2 中不同的化学成分含量相差较大，故考虑首先对各列化学成分数据进行标准化处理，使处理后的数据更加能体现该化学成分含量的相对大小，使数据更加直观且具有可比性。此处使用了最小——最大规范化方法 (minmax-scale)，方法公式如下：

$$x_i = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3.2.2 字符数据数值化

由于表单中数据多为离散数据，且问题多为分类问题，故考虑将“字符变量”转换为“0-1 变量”。本文做出了如下定义：

- 风化为 1，无风化为 0；
- 铅钡玻璃类型是 1，高钾玻璃类型是 0

3.3 数据侧写

由于本题为经典的数据分析问题，且表单 2 数据较为重要，故先对表单 2 的数据进行总体刻画有利于把握整体问题。

- 对化学成分含量的观测如图1，为两种玻璃类型的化学成分含量组成的数据展示。

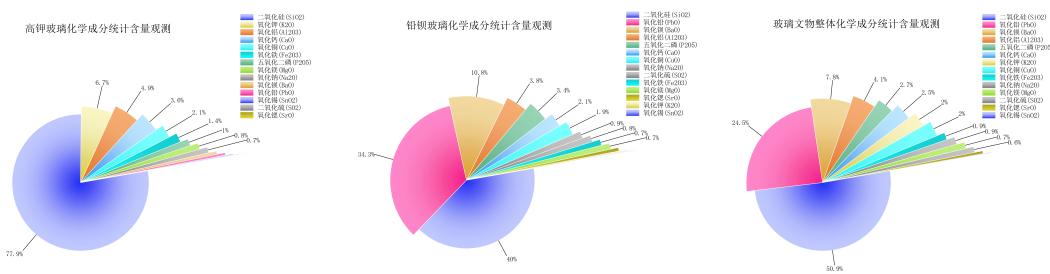


图 1 化学成分含量组成的数据观测

由图可得， SiO_2 的化学成分含量最多，高钾玻璃的 K_2O 、 CaO 含量较铅钡玻璃类多，铅钡玻璃的 PbO 、 BaO 含量较高钾玻璃类多，符合化学常识，具体问题分析时应该考虑不同成分含量所产生的影响。

- 对有无风化的观测如图2，为两种玻璃类型的有无表面风化的数据展示。

由图可得，铅钡玻璃的风化概率更高，这提示我们风化概率可能与玻璃类型有较大关系。

四、问题一模型的建立与分析

4.1 风化结果与玻璃类型、纹饰与颜色的关系的描述性分析

首先对数据的整体进行刻画，绘制以玻璃类型、纹饰和颜色为自变量，是否风化为因变量的三维散点气泡图，如图3所示。

为了得到更加具体的关系特征，对图像进行降维处理，分别得到高钾玻璃和铅钡玻璃的纹饰、颜色与有无分化关系的二维散点图，如图4所示。

由图4，可以得出如下的规律：

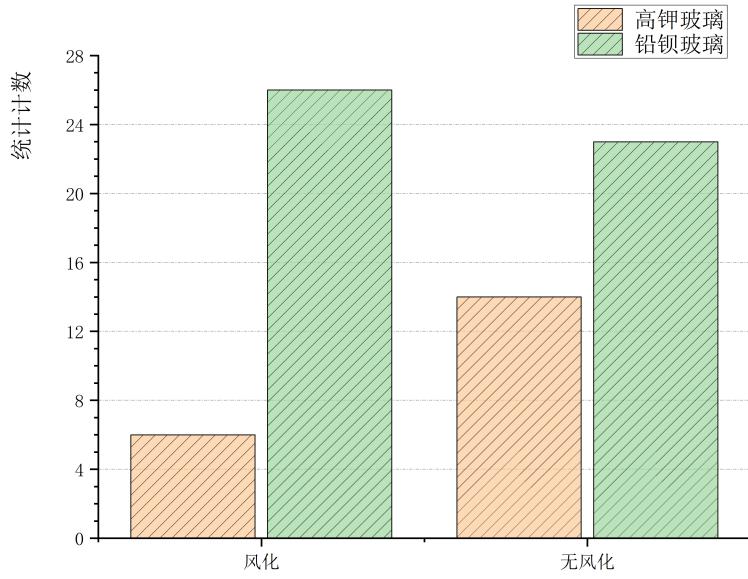


图 2 两种玻璃类型的有无表面风化的数据观测

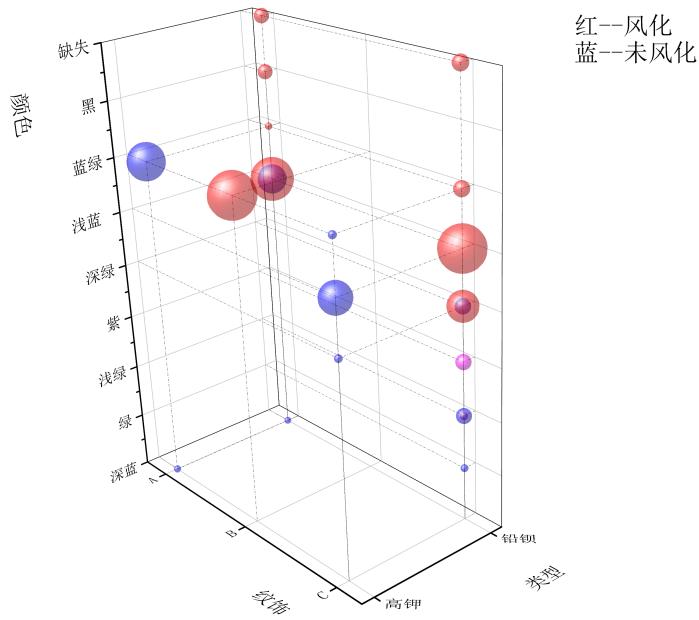


图 3 描述数据整体特征的三维散点气泡图

- 由高钾玻璃的二维图，可以观察到高钾玻璃中 B 纹饰——紫颜色的种类风化结果表现为风化，且数据样本较多，其他类型的纹饰与颜色均表现为未风化；
- 由铅钡玻璃的二维图，可以观察到 A 纹饰的浅蓝色同时表现出了未风化、风化两种结果，推测当 A 纹饰的铅钡玻璃介于风化与未风化的过渡地带时，往往呈现出浅蓝色，而 A 纹饰的其他颜色都具有较好的区分属性，如 A 纹饰——深蓝色表现为未风化，而 A 纹饰——蓝绿色/黑色/缺失颜色均表现为风化；C 纹饰同理，存在过渡地带。

基于上述的规律，本文做出了三个维度数据与风化结果的关系特征的推导：

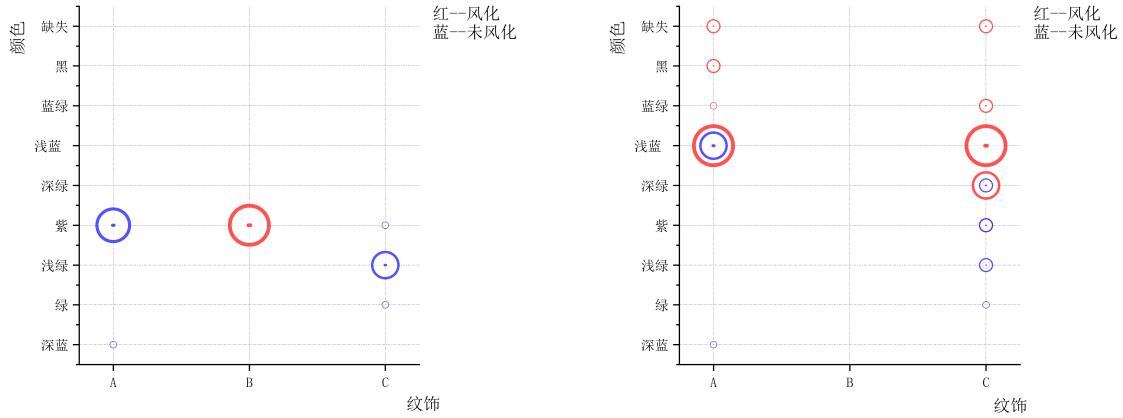


图4 描述高钾玻璃(左)和铅钡玻璃(右)数据特征的二维散点图

- 玻璃类型会对风化结果产生较显著的影响，铅钡玻璃风化比重比高钾玻璃大；
- 风化过程中会存在**颜色的渐变**，这种渐变过程会因玻璃类型和纹饰存在差异，如铅钡玻璃的C纹饰，初始为绿色，经过风化由于化学成分改变逐渐变成浅绿、深绿的过度颜色，最终被完全风化，变成蓝绿色。

4.2 基于 Spearman 相关系数的按含量比例分类的分类模型

本文设计了基于 Spearman 相关系数的按含量比例分类的模型，具体流程如下所示：

首先针对高钾玻璃做具体的流程分析：

- Step1: 绘制高钾玻璃风化前后的各个归一化后的化学成分含量的散点图如图5 (左) 所示，观察数据特征。

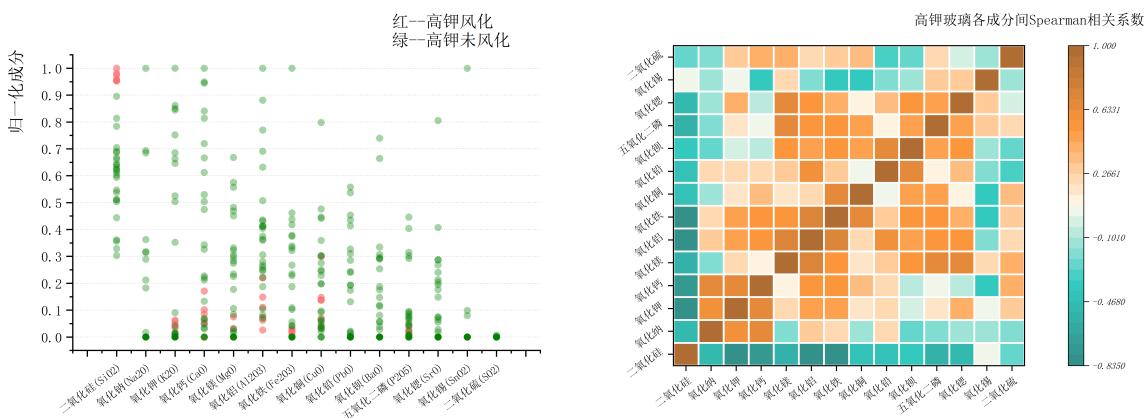


图5 高钾玻璃成分含量散点图(左)与各成分间的 Spearman 相关系数热力图(右)

由图5 (左) 可以初步得出，高钾玻璃风化前后对部分化学成分含量有较大影响，其中，对 SiO_2 的含量有正向影响、对 K_2O 、 CaO 、 MgO 、 Al_2O_3 、 Fe_2O_3 、 CuO 、 P_2O_5 的含量有负向影响。

- Step2: 计算高钾玻璃各个化学成分的 Spearman 相关系数。

定义: X 和 Y 为两组数据, 其斯皮尔曼 (等级) 相关系数为:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

其中, d_i 为 X_i 和 Y_i 之间的等级差。

若某两种 Spearman 相关性较大, 则可以推测在风化前后这两种化学成分极可能存在某种化学转换关系, 结合步骤 1, 由于众多成分中只有 SiO_2 含量明显增加, 故主要探究 SiO_2 与其他化学成分含量的 Spearman 系数, 将各成分间的 Spearman 相关系数按颜色划分, 得到热力图如图5 (右) 所示。

以 Spearman 相关系数 0.6 为分界, 选出与 SiO_2 相关性最强的五个化学成分, 分别为: K_2O 、 CaO 、 MgO 、 Al_2O_3 、 Fe_2O_3 , 此结果与步骤一的结果得到了相互印证。

- Step3: 以化学成分含量的比例作为是否风化的依据。

基于上述分析, 构造 SiO_2 与其他五个化学成分的比例关系 p_1, p_1 公式如下所示:

$$p_1 = \frac{n(SiO_2)}{n(K_2O) + n(CaO) + n(MgO) + n(Al_2O_3) + n(Fe_2O_3)}$$

各个高钾玻璃的文物样品的 p_1 如表2所示, 由于篇幅关系, 仅展示 2 个风化样品与 2 个未风化样品的 p_1 值, 完整表格见附录A。

表 2 高钾玻璃样品的 p_1 示意值

文物采样点	类型	表面风化	二氧化硅 (SiO_2)	氧化钾 (K_2O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al_2O_3)	氧化铁 (Fe_2O_3)	高钾的比例 p_1 (区分风化)
10	高钾	风化	96.77	0.92	0.21	0.00	0.81	0.26	43.99
.....									
22	高钾	风化	92.35	0.74	1.66	0.64	3.50	0.35	13.40
17	高钾	无风化	60.71	5.71	0.00	0.85	0.00	1.04	7.99
.....									
06 部位 2	高钾	无风化	59.81	7.68	5.41	1.73	10.05	6.04	1.93

依照表格信息, 可以选取 $p_1 = 10.7$ 作为区分高钾玻璃风化与否的依据。

- Step4: 总结统计规律。

当高钾玻璃中的 $p_1 = \frac{n(SiO_2)}{n(K_2O) + n(CaO) + n(MgO) + n(Al_2O_3) + n(Fe_2O_3)}$ 大于 10.7 时, 高钾玻璃出现风化, 小于 10.7 时, 高钾玻璃未风化。

接着对铅钡玻璃进行分析, 铅钡玻璃的分析流程与高钾玻璃类似, 首先得出铅钡玻璃的风化前后化学成分含量归一化后的散点图, 如图6(左) 所示

从图中除了能分析出风化后 SiO_2 含量减少之外, 较难分析铅钡玻璃中风化前后化学成分的变化, 故需要求解 Spearman 相关系数进行进一步观察, 可视化 Spearman 相关

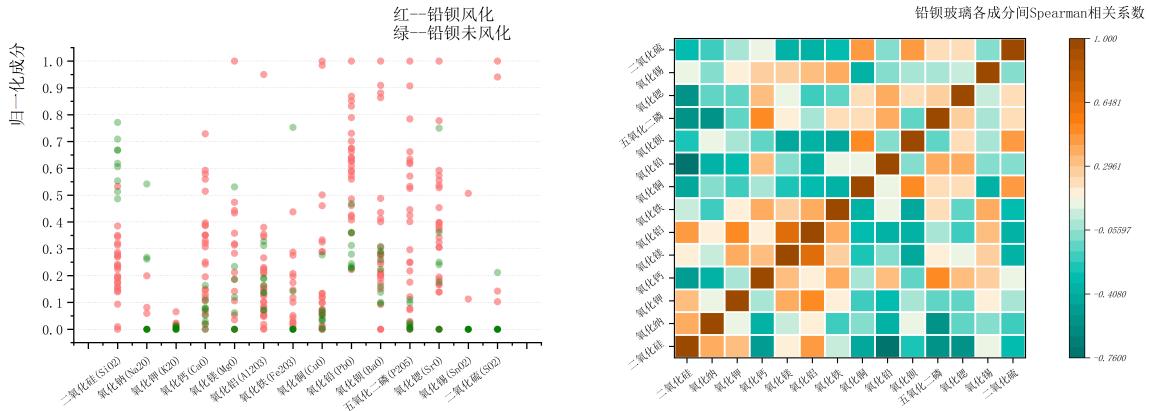


图6 铅钡玻璃成分含量散点图(左)与各成分间的 Spearman 相关系数热力图(右)

系数热力图如图6(右)所示。

从中提取出与 SiO_2 相关性较强的 PbO 、 CuO 、 BaO 三个化学成分，同样地，构造 SiO_2 与这三个化学成分含量的比例关系值 p_2 ，即

$$p_2 = \frac{n(SiO_2)}{n(PbO) + n(CuO) + n(BaO)}$$

各个铅钡玻璃的文物样品的 p_2 如表3所示，由于篇幅关系，仅展示 2 个风化样品与 2 个未风化样品的 p_2 值，完整表格见附录B。

表3 铅钡玻璃样品的 p_2 示意值

文物采样点	类型	表面风化	二氧化硅 (SiO ₂)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	铅钡的比例 p_2 (区分风化)
26 严重风化点	铅钡	风化	3.72	3.60	29.92	35.45	0.05
.....							
11	铅钡	风化	33.59	4.93	25.39	14.61	0.75
50 未风化点	铅钡	风化 (实际无风化)	45.02	0.70	30.61	6.22	1.20
.....							
29 未风化点	铅钡	无风化	63.30	0.74	12.31	2.03	4.20

依照表3信息，选取 $p_2 = 0.70$ 作为分解规律，得到如下结论：

当铅钡玻璃中的 $p_2 = \frac{n(SiO_2)}{n(PbO)+n(CuO)+n(BaO)}$ 大于 0.7 时，高钾玻璃出现风化，小于 0.7 时，高钾玻璃不风化。此分类结果不完全准确，有 3 个样本不符合此规律，分别是 11、36 和 48。

4.3 基于反向传播变换的线性逻辑回归的预测模型

预测风化前化学成分含量之前，先将其中的化学成分分为三类，如图7所示。



图 7 三类化学成分

模型算法流程如图8所示。

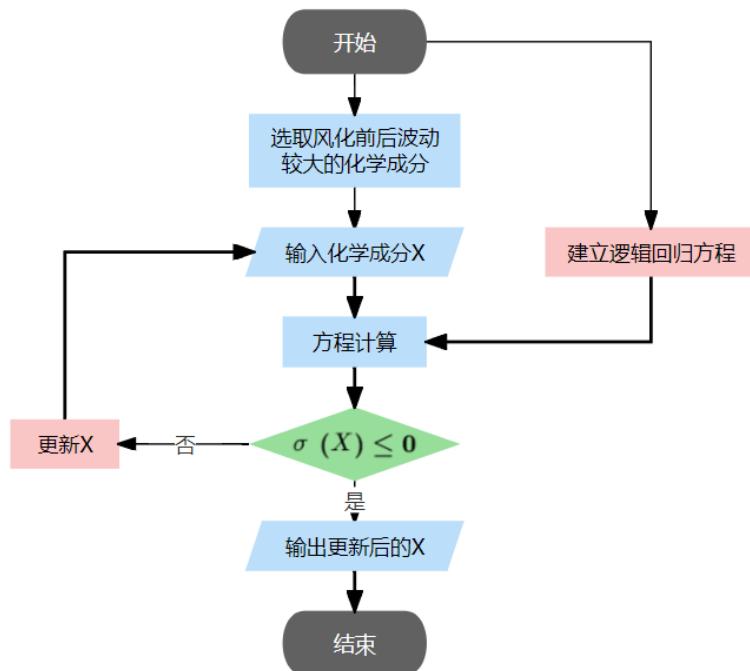


图 8 模型算法流程图

本节首先对高钾玻璃风化前的化学成分进行预测，预测过程如下：

4.3.1 选取特征（化学成分含量）

特征量的选取主要基于图5(左)，由上述分析可知，风化前后只对 SiO_2 、 K_2O 、 CaO 、 MgO 、 Al_2O_3 、 Fe_2O_3 、 CuO 的含量有显著影响，故将这七个化学成分含量视为系统输入的特征值。为了方便表达，选取的七个特征值依次用 x_1 、 x_2 …… x_7 表示。对于其他含量，风化前后波动较平稳的 P_2O_5 使用原值近似预测值，风化后消失的元素使用风化前元素的平均值近似预测值。

4.3.2 建立线性逻辑回归方程

- step1: 构建一般方程。

引入 sigmoid 函数，sigmoid 有两个特性，第一个是其函数值在趋于正无穷或负无穷时，函数趋近平滑状态，第二个是 sigmoid 函数输出范围为 $(0, 1)$ ，因此，sigmoid 函数经常用于二分类问题，其函数表达式如下所示：

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

为了建立输出与特征的关系，考虑使用 x_i 的线性表达来替代 z ，则 z 可以被表达为

$$z = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_7 \cdot x_7 = \vec{w}^T \cdot \vec{x}$$

具体到本问题中， \vec{w} 、 \vec{x} 的维度是 8，且 \vec{x} 的第一个元素是 1，将与 w_0 相乘，形成偏置。此时，线性逻辑回归的表达式就可以表达为：

$$\sigma_w(x) = \frac{1}{1 + e^{-\vec{w}^T \cdot \vec{x}}}$$

- step2: 构建逻辑回归的损失函数。

单个样本点的损失值记为 $cost$ ，构造如下所示的 $cost$ 的表达式。

$$cost(\sigma_w(x), y) = \begin{cases} -\log(\sigma_w(x)) & \text{if } y = 1 \\ -\log(1 - \sigma_w(x)) & \text{if } y = 0 \end{cases}$$

构建 $cost$ 表达式后，就可以构建出逻辑回归的损失函数

$$\begin{aligned} J(\theta) &= \frac{1}{m} \sum_{i=1}^m cost\left(\sigma_w\left(x^{(i)}\right), y^{(i)}\right) \\ &= -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log \sigma_w\left(x^{(i)}\right) + (1 - y^{(i)}) \log (1 - \sigma_w\left(x^{(i)}\right)) \right] \end{aligned}$$

- step3: 代入数据，训练模型。

代入高钾玻璃的数据，标记风化为 1，无风化为 0，进行训练，通过对损失函数求偏导更新参数，最终建立逻辑回归方程。

- step4: 方程准确度的验证。

如图9，可视化方程预测值与实际值的差距

方程的具体参数如表4所示。

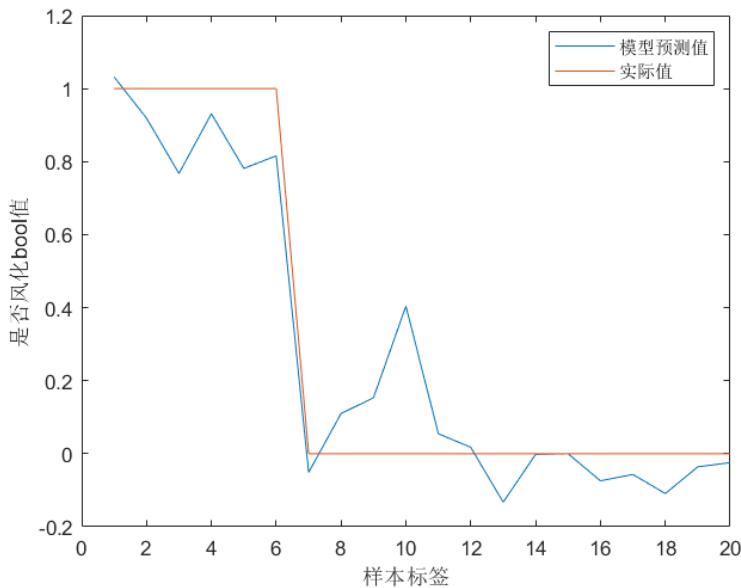


图 9 方程预测值与实际值的风化 bool 值

表 4 模型方程的具体参数

权重系数 (coef)	0.03059	0.00604	-0.043613	-0.115602	0.03004	0.17706	-0.18683
偏置 (intercept)	-1.99501						
准确率	0.82063						

4.3.3 化学成分含量的反向传播

本文通过按照一定比例依次调整风化后样本的化学成分含量，直到逻辑回归的方程判定由“1”变为“0”为止，达到预测风化前化学成分含量的效果。由于是对多组化学变量同时调整，故需要设定不同的调整幅度，此幅度由不同化学成分含量风化前后的均值的变化量之比决定。幅度比如下所示：

$$\begin{aligned} \Delta\mu_{\text{CuO}} : \Delta\mu_{\text{Fe}_2\text{O}_3} : \Delta\mu_{\text{Al}_2\text{O}_3} : \Delta\mu_{\text{MgO}} : \Delta\mu_{\text{CaO}} : \Delta\mu_{\text{K}_2\text{O}} : \Delta\mu_{\text{SiO}_2} \\ = 1 : 2.166 : 5.587 : 1.213 : 5.209 : 11.814 : -37.913 \end{aligned}$$

本算法的伪代码如下所示：

```
Init mu, alpha      #初始化上述比例系数行向量, alpha为幅值系数, 本例设为0.005
while :
    x = x*(1+alpha*mu) #更新x, 每个x的更新幅度由mu决定
    if (f(x)==0):
        break           #若经过逻辑回归判定为0, 未风化, 则跳出循环
```

更新后，得到六组风化后的高钾玻璃的化学成分的预测量，如表5所示，结果同样将展示在附录C中。

对于铅钡玻璃的化学成分的预测量，采用同样的预测方式，由于铅钡玻璃各个成分

表5 风化后高钾玻璃化学成分的预测量

文物 采样点	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P ₂ O ₅)	氧化锶 (SrO)	氧化锡 (SnO ₂)	二氧化硫 (SO ₂)
12	67.549	0.976	9.951	4.472	0.862	5.431	1.830	2.360	0.380	0.513	0.150	0.036	0.169	0.087
22	71.648	0.976	9.312	5.361	1.523	7.470	1.889	1.236	0.380	0.513	0.210	0.036	0.169	0.087
9	72.596	0.976	8.994	4.321	0.868	5.289	1.862	2.267	0.380	0.513	0.350	0.036	0.169	0.087
27	70.284	0.976	8.373	4.657	0.540	2.510	1.740	2.254	0.380	0.513	0.360	0.036	0.169	0.087
7	70.183	0.976	8.394	4.783	0.864	5.947	1.709	3.954	0.380	0.513	0.610	0.036	0.169	0.087
10	75.371	0.976	9.343	3.912	0.862	4.779	1.800	1.550	0.380	0.513	0.000	0.036	0.169	0.087

变化均较大，故选取铅钡的除了二氧化硫的十三个化学成分进行训练，构建铅钡玻璃的逻辑回归模型，最终反向更新化学成分含量，得到铅钡玻璃风化前化学成分的预测量，具体预测值见附录D。

五、问题二模型的建立与求解

5.1 高钾、铅钡玻璃分类规律的描述性统计分析

本文首先将预处理好的附件表二中玻璃类别与14种化学组成成分的关系用散点图进行可视化分析，得到“所有玻璃文物类别与成分关系图”、“风化玻璃文物类别与成分关系图”和“未风化玻璃文物类别与成分关系图”对类别与成分之间的关系进行直观的定性分析；随后计算类别与单个化学成分之间的*spearman* 相关系数对类别与成分之间的关系做定量分析。

5.1.1 统计规律的定性分析

所有玻璃文物类别与成分关系散点图如图10所示。

从图中可以看出：高钾玻璃的氧化钾含量基本上高于比铅钡玻璃，且有较明显的分界线；而高钾玻璃的氧化铅含量普遍低于铅钡玻璃。在二氧化硅含量方面，高钾玻璃的含量大多偏高，而铅钡玻璃的含量则相对中等，整体趋势高钾玻璃的二氧化硅含量更高，但二者含量的交集较明显。

将预处理后的表二数据按照风化与否分别画出玻璃类别与化学成分含量的散点图如图11所示。

从风化玻璃的类别与成分关系散点图可以看出：风化后的高钾玻璃的二氧化硅含量明显高于风化后的铅钡玻璃，且可以看出风化后的高钾玻璃化学成分含量集中在二氧化硅上；风化后的高钾玻璃不含氧化铅与氧化钡，而风化后的铅钡玻璃的氧化铅和氧化钡的含量很高；

从未风化玻璃的类别与成分关系散点图可以看出：未风化的高钾玻璃在氧化纳、氧化钾和氧化钙含量上明显高于铅钡玻璃且有较明显的分界；而未风化的高钾玻璃在氧化铅和氧化钡上含量明显低于铅钡玻璃。

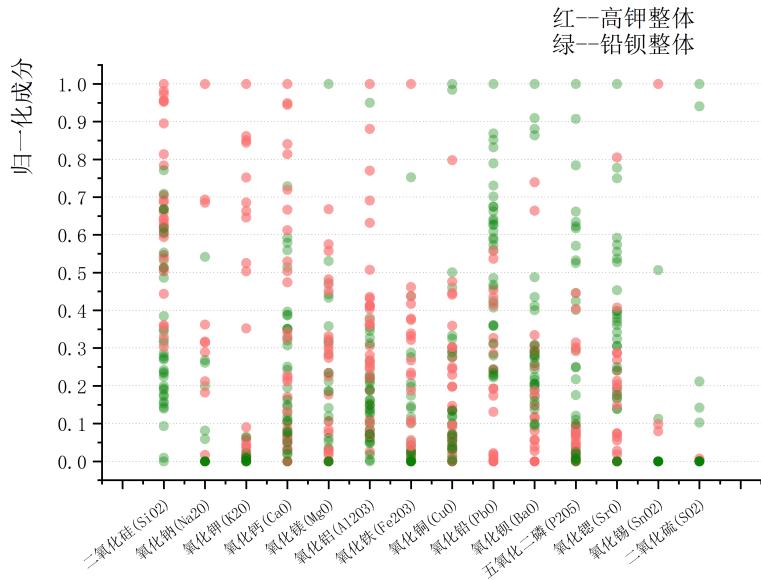


图 10 玻璃文物整体的类别与成分关系散点图

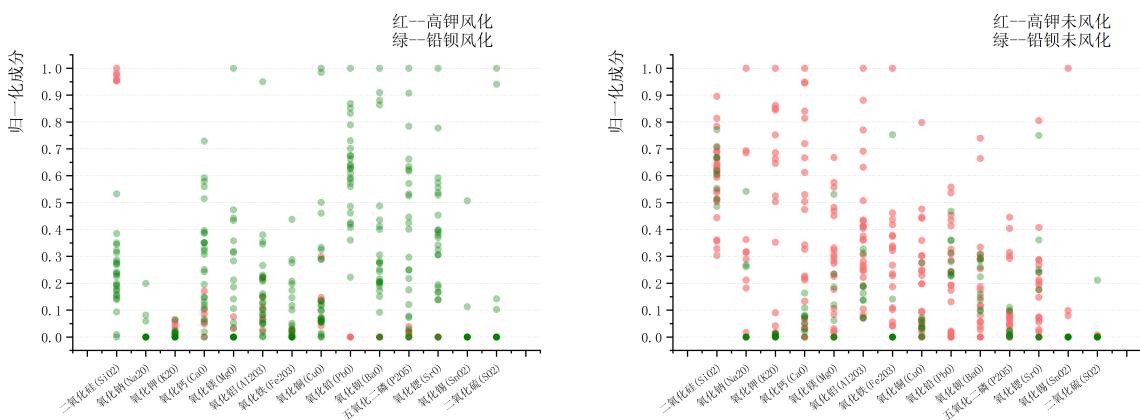


图 11 风化玻璃(左)与未风化(右)玻璃的类别与成分关系散点图

5.1.2 统计规律的定量分析

在 5.1.1 节通过散点图进行了定性分析后，接下来通过计算类别与各化学成分之间的 *spearman* 相关系数，进一步分析玻璃分类的统计规律。

首先，所有玻璃文物的类别与化学成分之间的 *spearman* 相关系数如表6所示：

表 6 所有玻璃文物的类别与成分的 spearman 相关系数表

化学成分	二氧化硅	氧化纳	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
Spearman 相关系数	-.670**	0.117289	-.583**	-0.19334	-0.09478	-.241*	-.289*	-0.22728	.770**	.714**	0.156965	.588**	0.063272	-0.06492

从表中可以看出：在 P 值为 0.01 的条件下，玻璃类别与氧化铅、氧化钡有较强的相关性：氧化铅和氧化钡的含量越高，越有可能是铅钡玻璃；其他相关性并不明显。

接着分别分析风化与否的玻璃类别与化学成分之间的 *spearman* 相关系数，结果如表7所示：

表 7 未风化玻璃文物的类别与成分的 spearman 相关系数表

化学成分	二氧化硅	氧化纳	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
Spearman 相关系数	-.442*	0.026336	-.750**	-.567**	-.397*	-.586**	-.037817	-.038615	.867**	.866**	-.28823	.417*	0.074139	-.22109

从未风化玻璃文物的类别与成分的 *spearman* 相关系数表中可以看出：在 P 值为 0.01 的条件下，玻璃类别与氧化铅、氧化钡和氧化钾有极强的相关性：氧化铅和氧化钡的含量越高，越有可能是铅钡玻璃；而氧化钾的含量越高越有可能是高钾玻璃；

从风化玻璃文物的类别与成分的 *spearman* 相关系数表中可以看出，在 P 值为 0.01 的条件下，玻璃类别与二氧化硅、氧化钡、氧化铅和氧化锶的含量有着一般的相关性：风化玻璃的二氧化硅含量越高，越有可能是高钾玻璃类型；氧化钡、氧化铅和氧化锶的含量越高，越有可能是铅钡玻璃类型。

5.1.3 分类规律总述

综合玻璃类别与化学成分散点图和 *spearman* 相关系数的分析可得：高钾玻璃的氧化钾含量明显高于铅钡玻璃；高钾玻璃的氧化铅和氧化钡含量之和明显低于铅钡玻璃的氧化铅和氧化钡含量之和；高钾玻璃的二氧化硅含量整体高于铅钡玻璃，但是二者的二氧化硅含量有明显的交集。基于以上观测，我们推测通过 $\frac{PbO+BaO}{K_2O}$ 这一比值可初步确定玻璃文物是高钾玻璃还是铅钡玻璃。

5.2 基于方差分析的 K-means 聚类模型的建立与求解

5.2.1 选择合适的化学成分

由于 K-means 聚类是对特征比较敏感的算法，噪声数据会对最终分类结果带来不利影响，因此在进行 K-means 聚类之前，需要进行特征筛选。

本文首先对附件表单 2 的高钾玻璃和铅钡玻璃分别做化学成分含量的分布散点图，观察不同类的化学成分的分布规律，再分别计算两种玻璃成分含量的方差，选出最终的化学成分。高钾玻璃和铅钡玻璃的化学成分含量分布图如图12

结合图12和 4.2 节的图6可以发现：氧化铅的含量分布可以近似将高钾玻璃分为风化与未风化，氧化锶、氧化镁和氧化铝可以进一步划分风化的铅钡玻璃；同理可以分析得出选取氧化铅、氧化纳、氧化钾、氧化钙和氧化铜作为高钾玻璃的聚类特征。

再对高钾玻璃内部的原始数据做归一化后计算各化学成分的方差，可以发现所选化学成分的方差较大，印证了此前的想法。

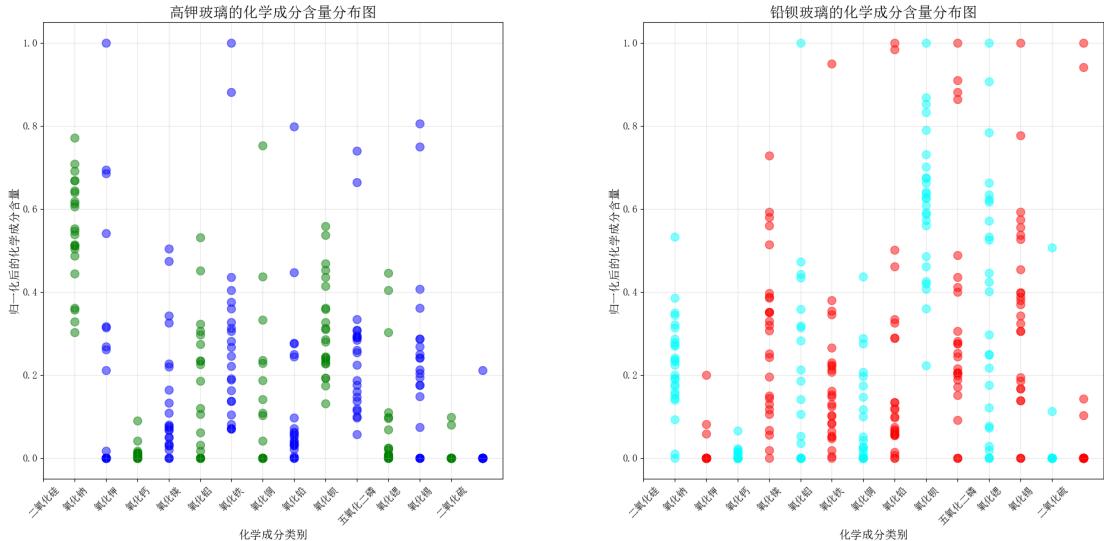


图 12 高钾玻璃和铅钡玻璃的化学成分含量分布图

5.2.2 确定最佳聚类数

K-means 聚类分析存在超参数最佳聚类数 K ，本文使用了肘部法则和平均轮廓法实现 K 值的选取。

- 肘部法则引入指标误差平方和 SSE，随着 K 值升高，SSE 值减小，当 SSE 与 K 值的关系到达拐点时，具有最好的分类效果，此时即可以挑选出不同特征的类别，又避免了过度分类。

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

- 轮廓系数法轮廓系数法的核心思想是通过调整 K ，以达到分类结果可以簇内紧密，簇外远离的效果，为了刻画这一效果，引入了指标 $s(i)$ ，其计算公式如下所示：

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} = \begin{cases} 1 - \frac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases}$$

其中， $a(i)$ 表示该样本点 i 到簇内其他数据的平均距离， $b(i)$ 表示到临近簇数据的平均距离；如果 $s(i)$ 接近 1，说明聚类效果好； $s(i)$ 接近 -1，说明聚类存在失真； $s(i)$ 接近 0，说明聚类效果效果差。

分别使用两种方法得到的效果曲线如图13所示：

由图可得，应该选取的 K 值为。

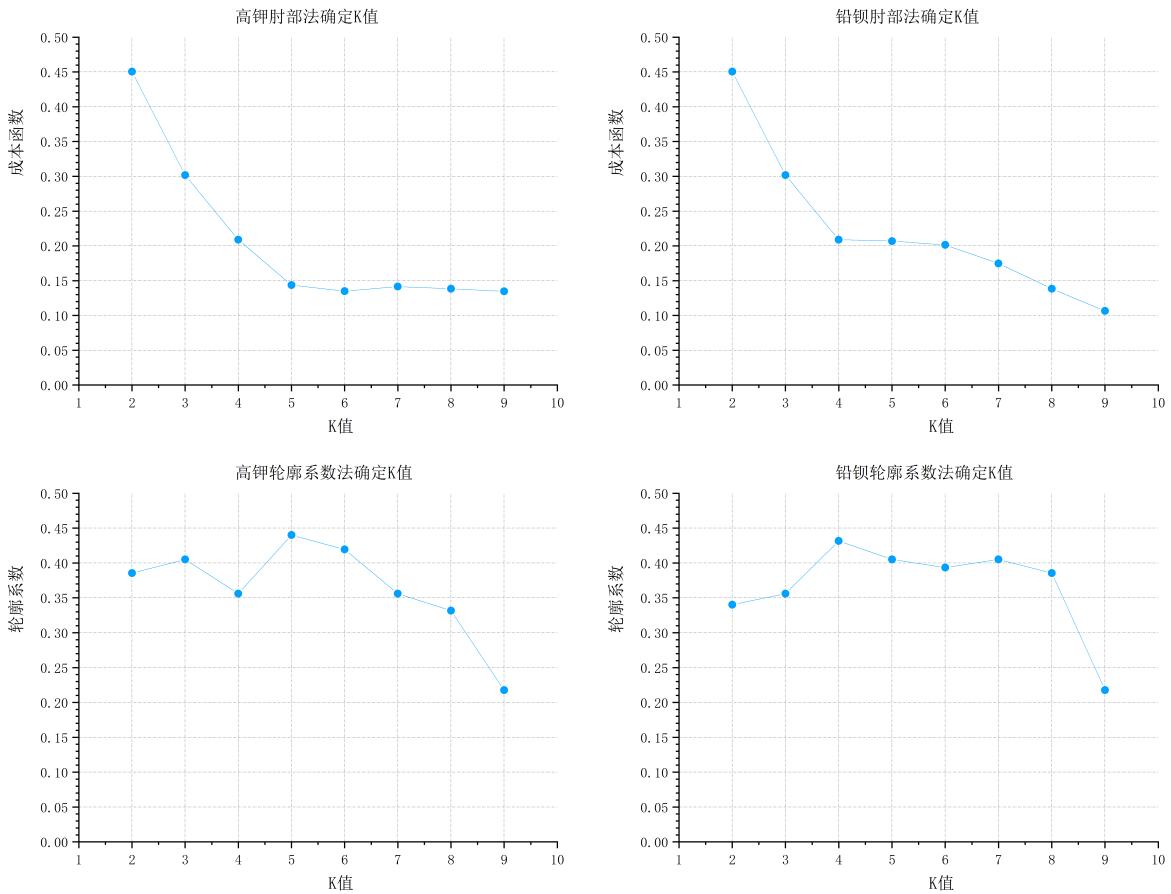


图 13 肘部法和轮廓系数法获取 K 值

5.2.3 K-means 聚类及结果

在确定高钾玻璃和铅钡玻璃的 K 值后，进行 K-means 聚类分析得到亚类结果如下：

5.3 聚类模型的合理性与敏感度分析

5.3.1 合理性分析

我们将不筛选化学成分，将所有成分作为输入进行 K-means 聚类得到的结果如下：对比可知，进行化学成分筛选后的 K-means 聚类效果更好。

六、问题三模型的建立与求解

观察附件表单三的信息，发现给出的化学成分信息依旧存在空白值，空白值的大量存在会使得本应是连续的化学成分信息变得离散，例如：氧化钠、氧化锡、二氧化硫等化学成分由于空白信息过多而变得离散。决策树是处理这些“离散特征”的好工具，但除了存在大量空白含量的“离散”化学成分外，表三还提供了如二氧化硅、氧化铝、五氧化二磷等没有空白值的化学成分，SVM 更善于处理这样的化学成分特征，因此本文

联合决策树和支持向量机两种分类方法，利用二者特点，训练不同特点的化学成分信息，得到两个分类模型，最后综合给出最终的类别预测。

6.1 支持向量机模型的构建

6.1.1 支持向量机简介

支持向量机是经典的线性分类模型，其基本思想是寻找一个合适的超平面将不同类型的样本点分隔开来，特别适用于二分类。支持向量机采用 hinge 损失作为自己的损失函数具体表达式如下：

$$L(y \cdot (w \cdot x + b)) = [1 - y(w \cdot x + b)]_+$$

其中 $[z]_+ = \begin{cases} z, & z > 0 \\ 0, & z \leq 0 \end{cases}$ 引入正则化项后，最后的损失函数如下：

$$\sum_i^N [1 - y_i (w \cdot x_i + b)]_+ + \lambda \|w\|^2$$

虽然支持向量机是一个线性分类模型，但通过引入适当的核函数，就能实现在高维空间上线性可分从而实现当前空间下非线性分类，本文便采用使用核技巧的核支持向量机。

6.1.2 化学成分特征选择

因为支持向量机是通过寻找一个超平面实现分类的，因此选择良好的化学成分特征尤为重要。根据 5.1 节的分析可以得出：氧化钾、氧化铅和氧化钡与化学成分的类别具有较强的相关性，但是考虑到附件表三的氧化钡和氧化铅存在较多未检测到的空白成分，且整个表单二的数据量极少，为了防止训练时模型过拟合，本文不直接使用氧化铅和氧化钡作为支持向量机模型的输入特征，而是通过 $\frac{PbO + BaO}{K_2O}$ 这一比值的方式将氧化铅和氧化钡引入到模型训练中；同时，考虑到二氧化硅百分比分数远大于除氧化铅、氧化钡以外的其他化学成分的百分比，因此在训练时避免直接将二氧化硅作为输入，而是通过 $\frac{K_2O}{SiO_2}$ 这一比值的方式引进模型训练；最后联合没有空白信息或仅有唯一空白信息的氧化钙、氧化铝、氧化铁和氧化铜组成最终的输入特征，如表8所示。

6.1.3 模型超参数的选择及模型训练

由于使用的是核支持向量机，使用何种核函数是我们首先要确定的。文章将 3 种常见的核函数：线性核函数（不使用核函数）、多项式核函数和高斯核函数均尝试了一遍，

表 8 输入特征表

化学成分	表面	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
支持向量机	类型	是否风化	(SiO ₂)	(Na ₂ O)	(K ₂ O)	(CaO)	(MgO)	(Al ₂ O ₃)	(Fe ₂ O ₃) (CuO)
						√		√	√
	决策树	√	√	√	√				
支持向量机	化学成分	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫 (SO ₂)	$\frac{n_K}{n_{Pb}+n_{Ba}}$	$\frac{n_K}{n_{Si}}$
	类型	(PbO)	(BaO)	(P ₂ O ₅)	(SrO)	(SnO ₂)			
	决策树	√	√			√	√	√	

最终选择结果最优的核函数。

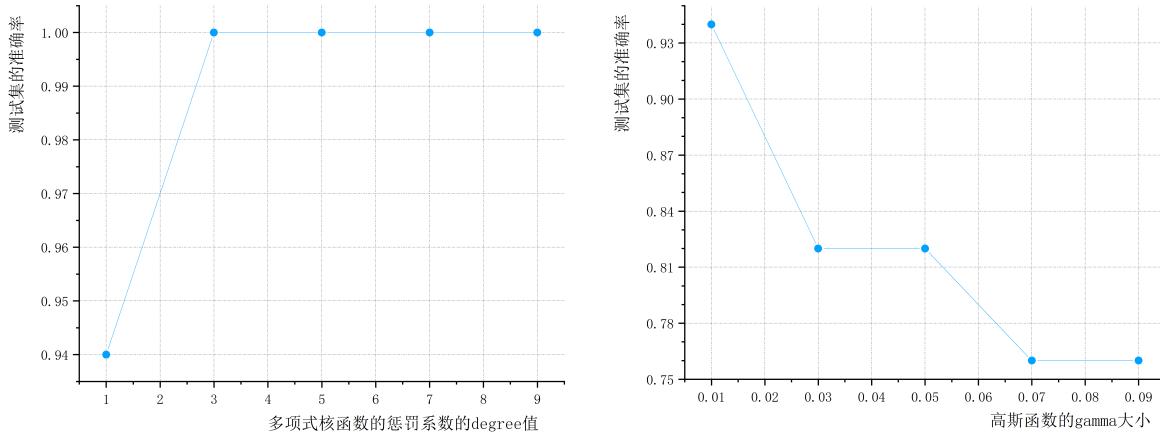


图 14 不同核函数下的测试集的准确度

* 线性核函数默认参数下准确率为 94%，作对照用

通过图14可知，最优的核函数是多项式核函数，且 degree 选择 3。

6.2 决策树模型的构建

6.2.1 CART 决策树模型简介

常见的决策树有许多种，例如：ID3、C4.5、CART 分类回归树等等。结合附件中的数据特征：既有连续量又有如“是否风化”这样的标签量以及因空白信息过多导致的呈现“标签性质”的连续量，本文最终选择 CART 分类回归树，一种可以同时处理连续量和标签量的分类算法。同时，考虑到附件表单二的数据集较小，使用基尼系数相较于使用信息熵能够抑制模型的过拟合。CART 分类回归树采用基尼系数作为衡量一个分割点决策好坏的指标，其数学表达式如下：

$$\text{Gini}(P) = \sum_{k=1}^K P_k (1 - P_k) = 1 - \sum_{k=1}^K P_k^2$$

其中 K 代表将分类的总类别数, P_k 是第 k 个类别的概率。本题是进行二分类, 可将表达式进一步化简如下:

$$\text{Gini}(P) = 2P(1 - P)$$

在模型训练之前, 需要设置一个样本个数阈值和基尼系数阈值作为模型的训练结束条件。通过不断迭代, 调整分支使得最后的决策树满足阈值条件便得到最后的分类模型。

6.2.2 化学成分特征选取

数据的输入特征在决策树中充当节点的角色, 因此选取合适的输入特征, 更能构建一颗优良的决策树。由上文分析可知, 高钾玻璃和铅钡玻璃的二氧化硅、氧化钾、氧化钡和氧化铅含量差异明显, 甚至具有肉眼可见的分界面, 因此这些属性是良好的决策属性。除此之外: 玻璃文物表面风化与否也是很好的分类特征。观察附件表三所给的数据: 二氧化硫、氧化锡和氧化纳这些化学成分由于空白过多, 导致其在预测阶段丧失了连续属性, 可以看作标签信号加入进决策树的模型训练中, 最终决策树的输入化学成分如 6.1.2 的表8所示。

6.2.3 模型超参数的选择及模型训练

6.3 结果预测

将附件表单三的未知类别的风化信息以及化学成分信息输入多项式核的支持向量机以及决策树得到的对应结果如表9所示。

表 9 预测结果输出

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
支持向量机鉴别类别	高钾玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃	高钾玻璃	高钾玻璃	铅钡玻璃
决策树鉴别类别	高钾玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃	高钾玻璃	高钾玻璃	铅钡玻璃
最终预测结果	高钾玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃	高钾玻璃	高钾玻璃	铅钡玻璃

6.4 敏感性分析

对于得到的分类模型, 本文通过让输入的化学成分的值上下浮动, 观察分类结果是否发生反转从而分析模型对输入化学成分的敏感性。

具体而言: 对任意输入的化学成分 E_i , 使其从 $0.5E_i$ 增长到 $2E_i$, 以 $0.1E_i$ 为步长, 观察分类结果是否反转, 统计使成分降低时的下界和成分上升时的上界, 得到统计表如表10所示。

表 10 成分降低时的下界和成分上升时的上界统计

	氧化钙 (CaO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	五氧化二磷 (P ₂ O ₅)	$\frac{n_K}{n_{Pb}+n_{Ba}}$	$\frac{n_K}{n_{Si}}$
A1	1.8	1.7	1.5	1.5	1.6	1.1	1.1
A1	0.6	0.5	0.6	0.6	0.4	0.8	0.9
A2	1.6	1.6	1.6	1.4	1.3	1.1	1.2
A2	0.4	0.6	0.4	0.4	0.5	0.9	0.8
A3	1.7	1.4	1.8	1.6	1.9	1.2	1.3
A3	0.6	0.6	0.4	0.4	0.7	0.7	0.9
A4	1.4	1.8	1.6	1.5	1.8	1.4	1.2
A4	0.6	0.6	0.4	0.4	0.3	0.8	0.8
A5	1.3	1.6	1.8	1.6	1.4	1.3	1.3
A5	0.4	0.7	0.4	0.6	0.6	0.4	0.8
A6	1.5	1.3	1.6	1.4	1.5	1.6	1.1
A6	0.6	0.5	0.5	0.8	0.6	0.5	0.9
A7	1.7	1.5	1.7	1.6	1.7	1.3	1.2
A7	0.8	0.4	0.4	0.9	0.6	0.9	0.8
A8	1.6	1.6	1.6	1.5	1.2	1.1	1.1
A8	0.4	0.7	0.6	0.6	0.6	0.8	0.8

从表10中可以看出氧化钙、氧化铝、氧化铁、氧化铜和五氧化二磷有较大的浮动空间，模型对这些成分的变化不敏感，而 $\frac{n_K}{n_{Pb}+n_{Ba}}$ 和 $\frac{n_K}{n_{Si}}$ 浮动空间较小，表明模型对氧化钾、氧化铝、二氧化硅和氧化钡这些化学成分较敏感。

七、问题四的分析

7.1 化学成分关联关系的描述性分析

对于高钾玻璃，化学成分有如下的关联关系

- 在风化进程中 SiO_2 含量明显的增多，而 K_2O 、 CaO 、 MgO 、 Al_2O_3 、 Fe_2O_3 、 CuO 等的含量明显的减小。
- 上述几种化学成分的 spearman 相关系数较高

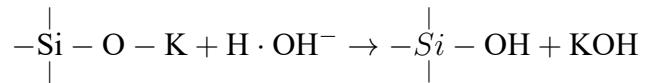
对于铅钡玻璃，化学成分有如下的关联关系

- 在风化进程中 SiO_2 含量明显的减少，而 PbO 、 CaO 、 BaO 等化学成分含量明显增多，较多的化学成分含量风化前后无明显增多
- 铅钡玻璃的化学成分的 spearman 相关系数较低

7.2 不同玻璃类型间化学成分关系的差异性分析

本文从化学原理入手，分析不同玻璃类型将化学成分关系的差异性。

- 水解氢化反应的差异¹ 水解氢化反应对风化过程起着非常重要的作用。在高钾玻璃中，存在下面的反应：



而在铅钡玻璃，在 $\text{Si} - \text{O}$ 键破坏后，可能有 $\text{Pb} - \text{O}$ 或 Pb^{++} 或 Ba^{++} 出现；² 推论：在水解氢化反应中，铅钡玻璃 $\text{Si} - \text{O}$ 键被破坏，导致 SiO_2 含量减少，而 PbO 、 BaO 含量则可能增加。

- 与玻璃内在结晶水、外界空气反应的差异³ 在高钾玻璃中，存在下面的化学转换关系： $\text{CaO} + \text{CO}_2 \rightarrow \text{CaCO}_3$ $\text{Na}_2\text{SiO}_3 + \text{CO}_2 \rightarrow \text{Na}_2\text{CO}_3 + \text{SiO}_2$ $\text{K}_2\text{SiO}_3 + X\text{H}_2\text{O} \rightarrow 2\text{KOH} + (X - 1)\text{H}_2\text{O} \cdot \text{SiO}_2$ ⁴ 等而在铅钡玻璃中，由于 $\text{PbO}/\text{K}_2\text{O}$ 比例较高，形成网络，网络中的桥氧 (O_6) 结合能较高，形式稳定，使表面的氧化物不易发生反应。推论：在与内在结晶水、外界 CO_2 反应的过程中，铅钡玻璃由于网络的生成其活性较差，反应较少，这也是铅钡玻璃中化学成分含量 spearman 相关系数较低的原因；而高钾玻璃频繁地发生金属氧化物生成碳酸金属化合物与二氧化硅的生成反应，导致其中 SiO_2 含量增多， CaO 、 Al_2O_3 等含量减少，大量化学关系的转换也是高钾玻璃中化学成分含量 spearman 相关系数较大的原因。

参考文献

- ¹吴宗道;周福征;史美光.几个古玻璃的显微形貌、成分及其风化的初步研究 [J].电子显微学报,1986,(04):65-71.
- ²Chengyu, W., Ying, T., Min, C., & Ming, H. (1991). Weathering of Soda-Lime and Lead Glasses. Transactions of the Indian Ceramic Society, 50(6), 171–177.
- ³陈敏.影响铅玻璃风化的因素 [D].大连工业大学,1987.
- ⁴刘壮飞.玻璃制品的风化及其预防措施 [J].商业科技,1987,(05):30-31.

附录 A 高钾玻璃文物样品的 p_1 值

文物采样点	类型	表面风化	二氧化硅 (SiO ₂)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	高钾的比例 (区分风化)
10	高钾	风化	96.77	0.92	0.21	0.00	0.81	0.26	43.99
09	高钾	风化	95.02	0.59	0.62	0.00	1.32	0.32	33.34
07	高钾	风化	92.63	0.00	1.07	0.00	1.98	0.17	28.77
12	高钾	风化	94.29	1.01	0.72	0.00	1.46	0.29	27.09
27	高钾	风化	92.72	0.00	0.94	0.54	2.51	0.20	22.13
22	高钾	风化	92.35	0.74	1.66	0.64	3.50	0.35	13.40
17	高钾	无风化	60.71	5.71	0.00	0.85	0.00	1.04	7.99
03 部位 1	高钾	无风化	87.05	5.19	2.01	0.00	4.06	0.00	7.73
18	高钾	无风化	79.46	9.42	0.00	1.53	3.05	0.00	5.68
21	高钾	无风化	76.68	0.00	4.71	1.22	6.19	2.37	5.29
15	高钾	无风化	61.87	7.44	0.00	1.02	3.15	1.04	4.89
01	高钾	无风化	69.33	9.99	6.32	0.87	3.93	1.74	3.03
06 部位 1	高钾	无风化	67.65	7.37	0.00	1.98	11.15	2.39	2.96
04	高钾	无风化	65.88	9.67	7.12	1.56	6.44	2.06	2.45
03 部位 2	高钾	无风化	61.71	12.37	5.87	1.11	5.50	2.16	2.28
16	高钾	无风化	65.18	14.52	8.27	0.52	6.18	0.42	2.18
05	高钾	无风化	61.58	10.95	7.35	1.77	7.50	2.62	2.04
14	高钾	无风化	62.47	12.28	8.23	0.66	9.23	0.50	2.02
13	高钾	无风化	59.01	12.53	8.70	0.00	6.16	2.88	1.95
06 部位 2	高钾	无风化	59.81	7.68	5.41	1.73	10.05	6.04	1.93

附录 B 铅钡玻璃文物样品的 p_2 值

文物 采样点	类型	表面 风化	二氧化硅 (SiO ₂)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	铅钡的比例 (区分风化)
26 严重风化点	铅钡	风化	3.72	3.60	29.92	35.45	0.05
08 严重风化点	铅钡	风化	4.61	3.14	32.45	30.62	0.07
43 部位 1	铅钡	风化	12.41	5.35	59.85	7.29	0.17
40	铅钡	风化	16.71	0.00	70.21	6.69	0.22
54 严重风化点	铅钡	风化	17.11	1.34	58.46	0.00	0.29
54	铅钡	风化	22.28	0.83	55.46	7.04	0.35
50	铅钡	风化	17.98	1.13	44.00	14.20	0.30
51 部位 2	铅钡	风化	21.35	0.75	51.34	0.00	0.41
41	铅钡	风化	18.46	0.19	44.12	9.76	0.34
39	铅钡	风化	26.25	0.88	61.03	7.22	0.38
43 部位 2	铅钡	风化	21.70	1.51	44.75	3.26	0.44
52	铅钡	风化	25.74	0.70	47.42	8.64	0.45
57	铅钡	风化	25.42	1.16	45.10	17.30	0.40
51 部位 1	铅钡	风化	24.61	1.37	40.24	8.94	0.49
38	铅钡	风化	32.93	0.73	49.31	9.79	0.55
26	铅钡	风化	19.79	10.57	29.53	32.25	0.27
19	铅钡	风化	29.64	3.51	42.82	5.35	0.57
08	铅钡	风化	20.14	10.41	28.68	31.23	0.29
56	铅钡	风化	29.15	0.79	41.25	15.45	0.51
02	铅钡	风化	36.28	0.26	47.43	0.00	0.76
34	铅钡	风化	35.78	1.51	46.55	10.00	0.62
58	铅钡	风化	30.39	3.13	39.35	7.66	0.61
49	铅钡	风化	28.79	0.70	34.18	6.10	0.70
30 部位 1	铅钡	无风化	34.34	0.00	39.22	10.29	0.69
36	铅钡	风化	39.57	0.68	41.61	10.83	0.74
30 部位 2	铅钡	无风化	36.93	0.00	37.74	10.35	0.77
24	铅钡	无风化	31.94	8.46	29.14	26.23	0.50
11	铅钡	风化	33.59	4.93	25.39	14.61	0.75
50 未风化点	铅钡	风化 (实际无风化)	45.02	0.70	30.61	6.22	1.20
55	铅钡	无风化	49.01	0.86	32.92	7.95	1.17
25 未风化点	铅钡	风化 (实际无风化)	50.61	1.12	31.90	6.65	1.28
47	铅钡	无风化	51.54	0.65	25.40	9.23	1.46
46	铅钡	无风化	55.21	0.77	25.25	10.06	1.53
42 未风化点 1	铅钡	风化 (实际无风化)	51.26	2.67	21.88	10.47	1.46

文物 采样点	类型	表面 风化	二氧化硅 (SiO ₂)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	铅钡的比例 (区分风化)
49 未风化点	铅钡	风化 (实际无风化)	54.61	0.45	23.02	4.19	1.97
42 未风化点 2	铅钡	风化 (实际无风化)	51.33	2.72	20.12	10.88	1.52
35	铅钡	无风化	65.91	0.16	22.05	5.68	2.36
23 未风化点	铅钡	风化 (实际无风化)	53.79	2.99	16.98	11.86	1.69
48	铅钡	风化	53.33	0.00	15.71	7.31	2.32
37	铅钡	无风化	60.12	3.01	17.24	10.34	1.97
32	铅钡	无风化	69.71	0.11	19.76	4.88	2.82
45	铅钡	无风化	61.28	0.53	15.99	10.96	2.23
28 未风化点	铅钡	风化 (实际无风化)	68.08	0.33	17.14	4.04	3.17
31	铅钡	无风化	65.91	0.44	16.55	3.42	3.23
20	铅钡	无风化	37.36	4.78	9.30	23.55	0.99
44 未风化点	铅钡	风化 (实际无风化)	60.74	0.43	13.61	5.22	3.15
53 未风化点	铅钡	风化 (实际无风化)	63.66	0.54	13.66	8.99	2.75
33	铅钡	无风化	75.51	0.47	16.16	3.55	3.74
29 未风化点	铅钡	无风化	63.30	0.74	12.31	2.03	4.20

附录 C 风化后高钾玻璃化学成分的预测

文物采样点	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P ₂ O ₅)	氧化锶 (SrO)	氧化锡 (SnO ₂)	二氧化硫 (SO ₂)
12	67.549	0.976	9.951	4.472	0.862	5.431	1.830	2.360	0.380	0.513	0.150	0.036	0.169	0.087
22	71.648	0.976	9.312	5.361	1.523	7.470	1.889	1.236	0.380	0.513	0.210	0.036	0.169	0.087
9	72.596	0.976	8.994	4.321	0.868	5.289	1.862	2.267	0.380	0.513	0.350	0.036	0.169	0.087
27	70.284	0.976	8.373	4.657	0.540	2.510	1.740	2.254	0.380	0.513	0.360	0.036	0.169	0.087
7	70.183	0.976	8.394	4.783	0.864	5.947	1.709	3.954	0.380	0.513	0.610	0.036	0.169	0.087
10	75.371	0.976	9.343	3.912	0.862	4.779	1.800	1.550	0.380	0.513	0.000	0.036	0.169	0.087

附录 D 风化后铅钡玻璃化学成分的预测

文物采样点	二氧化硅 (SiO ₂)	氧化钠 (Na ₂ O)	氧化钾 (K ₂ O)	氧化钙 (CaO)	氧化镁 (MgO)	氧化铝 (Al ₂ O ₃)	氧化铁 (Fe ₂ O ₃)	氧化铜 (CuO)	氧化铅 (PbO)	氧化钡 (BaO)	五氧化二磷 (P ₂ O ₅)	氧化锶 (SrO)	氧化锡 (SnO ₂)	二氧化硫 (SO ₂)
48	75.17	2.31	0.42	1.47	1.58	15.38	1.23	-0.86	-5.61	4.54	-3.27	0.10	1.31	0.00
02	66.20	1.50	1.16	0.97	1.20	7.45	2.02	-0.59	26.62	-2.86	-0.67	0.04	-0.02	0.00
11	66.29	1.47	0.31	2.17	0.71	4.33	0.15	4.22	4.21	11.90	5.21	0.22	-0.02	0.00
36	71.84	3.72	0.23	-1.01	-0.01	3.18	0.48	-0.17	20.64	8.33	4.32	0.07	-0.02	0.00
49	60.70	1.52	0.09	3.29	1.52	7.12	2.97	-0.15	13.53	3.44	7.13	0.32	-0.02	0.00
34	65.73	1.49	0.35	-0.62	-0.01	3.14	0.62	0.69	26.21	7.32	-4.03	0.07	-0.02	0.00
58	62.72	1.53	0.44	2.18	0.80	5.19	1.03	2.29	18.62	5.02	4.94	0.09	-0.02	0.00
19	62.16	1.48	0.09	1.56	0.60	5.30	1.49	2.71	21.79	2.58	4.63	0.04	-0.02	0.00
38	65.76	2.99	0.09	-0.70	-0.01	4.23	0.44	-0.12	28.38	7.17	-3.88	0.26	-0.02	0.00
56	61.43	1.51	0.09	-0.17	-0.01	3.35	0.16	-0.06	20.10	13.17	-1.73	-0.16	-0.02	0.00
51 部位 1	56.49	1.52	0.09	2.30	1.21	6.86	1.36	0.53	19.88	6.32	3.91	0.25	0.45	0.00
52	56.91	2.82	0.09	0.92	0.55	2.69	0.39	-0.14	27.12	5.93	1.49	0.30	-0.02	0.00
43 部位 2	51.90	1.49	0.09	5.05	0.98	5.06	1.62	0.67	24.18	0.46	8.96	0.32	-0.02	0.00
51 部位 2	52.12	1.50	0.09	3.91	1.50	4.12	0.58	-0.09	30.58	-2.87	4.56	-0.16	-0.02	0.00
57	55.54	1.50	0.09	-0.07	-0.01	3.81	0.16	0.32	24.07	14.80	-4.26	-0.16	-0.02	0.00
39	56.08	1.52	0.09	-0.27	-0.01	2.02	0.16	0.04	41.04	4.49	-3.17	0.47	-0.02	0.00
54	54.47	1.53	0.41	1.88	1.28	5.69	0.16	-0.01	35.92	4.35	0.01	0.75	-0.02	0.00
41	48.93	1.47	0.53	3.59	2.73	4.84	2.01	-0.67	23.09	7.21	3.32	0.32	-0.02	0.00
50	48.43	1.48	0.09	1.82	0.48	3.49	0.48	0.29	23.06	11.64	2.19	0.53	-0.02	0.00
08	50.72	1.49	0.09	0.11	-0.01	2.86	0.16	9.87	7.60	29.03	-0.64	0.22	-0.02	0.00
54 重风化点	47.95	1.47	0.09	-1.42	1.14	5.24	0.15	0.50	37.37	-2.82	10.37	0.97	-0.02	0.00
26	51.14	1.50	0.09	0.07	-0.01	2.25	0.15	9.79	8.58	29.48	-1.13	0.31	-0.02	0.00
40	46.52	1.49	0.09	0.51	-0.01	1.96	0.36	-0.85	49.97	3.94	-2.54	0.55	-0.02	0.00
43 部位 1	43.93	1.48	0.09	4.00	0.89	3.86	0.93	4.54	40.52	4.56	-4.24	0.51	-0.02	0.00
08 重风化点	35.32	1.48	0.09	1.89	-0.01	2.66	0.16	2.32	11.45	28.72	3.47	0.39	-0.02	0.00
26 重风化点	34.90	1.53	0.49	1.66	-0.01	2.69	0.15	2.78	8.96	34.21	1.88	0.48	-0.02	0.00

附录 E 绘制逻辑回归真实值与预测值的 matlab 代码

```
% 绘制逻辑回归预测值、真实值
% y_pre提前通过拖拽excel表单载入数据
y_fact = [1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0];
x= linspace(1,20,20);
plot(x,y_pre)
hold on
plot(x,y_fact)

legend('模型预测值','实际值')
```

附录 F 处理含量空白值的 python 代码

```
import pandas as pd
import numpy as np

#读取excel
excel=pd.read_excel(r"E:\G\exercise\github\2022_model\excel\表单2处理.xlsx",sheet_name=1)
#处理缺失值
excel=pd.DataFrame(excel)
excel=excel.fillna(value=0)

excel.to_excel(r"E:\G\exercise\github\2022_model\excel\补零的表二数据.xlsx")
```

附录 G 用于编码的 python 代码

```
import pandas as pd
import numpy as np

# 打开excel表格
excel=pd.read_excel(r"E:\G\exercise\github\2022_model\题目\附件.xlsx",sheet_name=0)
#转化为numpy数组
arr=np.array(excel)
print(arr)
encoder=[]

for row in arr:
    tmp=[1,2,3,4]
    #对纹饰进行编码
    if row[1]=='A':
        tmp[0]=1
    elif row[1]=='B':
        tmp[0]=2
    else:
        tmp[0]=3
    #类型编码
    if row[2]=="高钾":
        tmp[1]=1
    else:
        tmp[1]=2
    if row[3]=='深蓝':
        tmp[2]=1
    elif row[3]=='绿':
        tmp[2] =2
    elif row[3]=='浅绿':
        tmp[2] =3
    elif row[3]=='紫':
        tmp[2] =4
    elif row[3]=='深绿':
        tmp[2] =5
    elif row[3]=='浅蓝':
        tmp[2] =6
    elif row[3]=='蓝绿':
        tmp[2] =7
    encoder.append(tmp)
```

```

    elif row[3]=='黑':
        tmp[2] =8
    else:
        tmp[2]=9
    #风化编码
    if row[4]=="风化":
        tmp[3]=1
    else:
        tmp[3]=2
    encoder.append(tmp)
pd_encoder=pd.DataFrame(encoder)
pd_encoder.to_excel(r'C:\Users\16033\Desktop\表一编码结果.xlsx')

```

附录 H 用于训练线性逻辑回归方程的 python 代码

```

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import math
# data=pd.read_csv('Dry_Bean_Dataset.csv')
data =pd.read_excel('C:\\\\Users\\\\lenovo\\\\Desktop\\\\myworkspace\\\\mydata.xlsx',sheet_name="Sheet1")
df=pd.DataFrame(data)
Y_train=[]
for i in df['Class'][0:20]:
    if i=='无风化':
        Y_train.append(0)
    else:
        Y_train.append(1)
si=df['si'][0:20]
k=df['k'][0:20]
ca=df['ca'][0:20]
mg=df['mg'][0:20]
al=df['al'][0:20]
fe=df['fe'][0:20]
p=df['p'][0:20]
X_train=list(zip(si,k,ca,mg,al,fe,p))

from sklearn.linear_model import LinearRegression # 导入sklearn库中的线性回归模块
lr =LinearRegression()                      # 定义一个线性回归模型
lr.fit(X_train, Y_train)                    # 将模型拟合到数据上

# y_pred = lr.predict(X_train)
# print(y_pred[:19])

def sigmoid_function(z):
    fz = []
    for num in z:
        fz.append(1/(1 +math.exp(-num)))
    return fz

# 输出w权值
print(lr.coef_)
print(lr.intercept_)

```

```

# 载出预测值
y_pred = lr.predict(X_train)
out_y = pd.DataFrame(y_pred)
out_y.to_excel("y_k_pred.xlsx")
#print(y_pred[:])
# x1 , x2 , x3 , x4 , x5 , x6 = X_train[0],X_train[1],X_train[2],X_train[3],X_train[4],X_train[5]

# 验证准确度
score = lr.score(X_train,Y_train)
print(score)

```

附录 I 肘部法获取聚类 K 值的 python 代码

```

import matplotlib.pyplot as plt
from scipy.spatial.distance import cdist
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans

# 读入数据
plt.plot()
X = pd.read_excel(r"element.xlsx")
colors = ['b', 'g', 'r']
markers = ['o', 'v', 's']

# 肘部法确定K值
distortions = []
K = range(2, 10)
for k in K:
    kmeanModel = KMeans(n_clusters=k).fit(X)
    kmeanModel.fit(X)
    distortions.append(sum(np.min(cdist(X, kmeanModel.cluster_centers_, 'euclidean'), axis=1)) / X.shape[0])

# 绘图
#设置中文格式宋体
plt.rcParams['font.sans-serif'] = ['SimSun']
plt.plot(K, distortions, 'bx-', color="blue", marker="o", linestyle="--")
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.grid(True)
plt.xlabel('K 值', fontsize=14)
plt.ylabel('成本函数', fontsize=14)
plt.title('肘部法确定 K 值', fontsize=14)
plt.show()

```

附录 J 轮廓分析获取聚类 K 值的 python 代码

```

from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import pandas as pd
import matplotlib.pyplot as plt

# 读入数据
data = pd.read_excel(r"element.xlsx")
data = data[['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)', '氧化钡(BaO)',

'五氧化二磷(P2O5)', '氧化锶(SrO)', '氧化锡(SnO2)', '二氧化硫(SO2)']]

]]] #选择需要的分类特征值
data_Array = data.values #获取数组，便于聚类

# 轮廓系数确定K值
def Silhouette_ALL(n):
    data_Cluster = KMeans(n_clusters=n)
    data_Cluster.fit(data_Array)
    label = data_Cluster.labels_
    Silhouette_Coefficient = silhouette_score(data_Array, label)
    return Silhouette_Coefficient

y = []
k = []
# 遍历不同k值下轮廓系数
for n in range(2, 10):
    k.append(n)
    data_data_Silhouette_mean = Silhouette_ALL(n)
    y.append(data_data_Silhouette_mean)
print(y)

# 绘图
#设置中文格式宋体
plt.rcParams['font.sans-serif'] = ['SimSun']
plt.plot(k, y, 'bx-', color="blue", marker="o", linestyle="--")
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.grid(True)
plt.xlabel('K 值', fontsize=14)
plt.ylabel('轮廓系数', fontsize=14)
plt.title('轮廓系数确定 K 值', fontsize=14)
plt.show()

```

附录 K K-means 的 matlab 代码

```

clc; close all; clear;
[num]=xlsread('element.xlsx')
data=num;
%聚类中心个数，即K
N=4;
[m,n]=size(data);

```

```

pattern=zeros(m,n+1);
%聚类中心的初始化
center=zeros(N,n);
pattern(:,1:n)=data(:,,:);
for x=1:N
%产生中心
center(x,:)=data( randi(49,1),:);
end
figure('name','聚类过程');
while 1
distence=zeros(1,N);
num=zeros(1,N);
new_center=zeros(N,n);

%loop draw figure

drawfigure(pattern,center)
%循环对各点计算距离，并按照最近的中心进行归类。
for x=1:m
for y=1:N
%计算到每个类的距离，寻找最近的中心
distence(y)=norm(data(x,:)-center(y,:));
end
%取第二个参数即最小值所在的索引
[~, temp]=min(distence);
pattern(x,n+1)=temp;
end
k=0;
for y=1:N
for x=1:m
if pattern(x,n+1)==y
%对第y个中心 (x,y) 求和相加
new_center(y,:)=new_center(y,:)+pattern(x,1:n);
%对第y个中心自增
num(y)=num(y)+1;
end
end
%求取中心的均值
new_center(y,:)=new_center(y,:)/num(y);
%一个中心变化不大时，说明该点就是中心点
%当聚集了N个中心没有点时，就可以结束运算
if norm(new_center(y,:)-center(y,:))<0.1
k=k+1;
end
end
if k==N
break;

```

```

else
center=new_center;
end
end
[m, n]=size(pattern);
%绘图，显示聚类后数据
figure;
hold on
for i=1:m
if pattern(i,n)==1
plot(pattern(i,1),pattern(i,2), 'r*');
plot(center(1,1),center(1,2), 'ko');
elseif pattern(i,n)==2
plot(pattern(i,1),pattern(i,2), 'g*');
plot(center(2,1),center(2,2), 'ko');
elseif pattern(i,n)==3
plot(pattern(i,1),pattern(i,2), 'b*');
plot(center(3,1),center(3,2), 'ko');
elseif pattern(i,n)==4
plot(pattern(i,1),pattern(i,2), 'y*');
plot(center(4,1),center(4,2), 'ko');
% elseif pattern(i,n)==5
% plot(pattern(i,1),pattern(i,2), 'g*');
% plot(center(5,1),center(5,2), 'ko');
% elseif pattern(i,n)==6
% plot(pattern(i,1),pattern(i,2), 'b*');
% plot(center(6,1),center(6,2), 'ko');
else
plot(pattern(i,1),pattern(i,2), 'm*');
plot(center(4,1),center(4,2), 'ko');
end
end
grid on;

```

```

function drawfigure(pattern,center)
[m, n]=size(pattern);

hold on;
for i=1:m
if pattern(i,n)==1
plot(pattern(i,1),pattern(i,2), 'r*');
plot(center(1,1),center(1,2), 'ko');
elseif pattern(i,n)==2
plot(pattern(i,1),pattern(i,2), 'g*');
plot(center(2,1),center(2,2), 'ko');
elseif pattern(i,n)==3

```

```
plot(pattern(i,1),pattern(i,2), 'b*');
plot(center(3,1),center(3,2), 'ko');
elseif pattern(i,n)==4
plot(pattern(i,1),pattern(i,2), 'y*');
plot(center(4,1),center(4,2), 'ko');
% elseif pattern(i,n)==5
% plot(pattern(i,1),pattern(i,2), 'g*');
% plot(center(5,1),center(5,2), 'ko');
% elseif pattern(i,n)==6
% plot(pattern(i,1),pattern(i,2), 'b*');
% plot(center(6,1),center(6,2), 'ko');
else
plot(pattern(i,1),pattern(i,2), 'm*');
plot(center(4,1),center(4,2), 'ko');
end
end
hold off;
pause(0.7);
clf
end
```