玻璃风化问题研究

摘要

本文基于单因素方差分析模型对题目中所给的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。首先我们对题目所给数据进行观察分析，表中有少量数据有缺失现象，应对表中数据进行预处理，补全缺失数据。

问题一中，最后以表面风化作为因素，以玻璃类型、纹饰和颜色的情况作为检验指标，根据模型原理，得到玻璃类型的显著性p值为0.002,则有显著性影响，而玻璃颜色的显著性p值为0.070和玻璃纹饰的显著性p值为0.078，表现为无差异显著性影响。

问题二中，首先建立玻璃类型与化学成分之间的线性回归模型，通过回归系数来解释玻璃种类的分类规律。而对于亚类的划分，我们综合了分类结果的敏感性和合理性问题，引入了“抗风化能力”这一概念，并构建了基于Topsis的古代玻璃抗风化能力的评价体系，将其与Kmeans++得到的聚类结果进行比对，发现聚类结果与Topsis得分高低有着极高的匹配度，可在玻璃种类大类下划分出抗风化能力强、抗风化能力中、抗风化能力弱三种亚类，增强了结果的合理性。

问题三中，我们首先基于表单2数据构建了玻璃种类与化学成分含量之间的回归模型，通过对回归系数的分析得到对玻璃种类的划分影响程度较大的若干种化学成分。然后基于决策树模型构建了这些成分与玻璃种类的分类模型，然后对决策树模型的最大深度、最小叶子节点数和最小训练样本分枝数等超参数进行敏感性分析，对模型进行调优，最终得到一个稳定的分类结果。

问题四中，我们认为化学成分之间的关系与玻璃本身的抗风化能力有关，基于灰色关联分析方法，同时结合第二问中构建的综合评价方法以分析同种玻璃化学成分之间的关联关系以及不同种玻璃化学成分关联关系的差异性。最终得到了不同化学成分对于同种玻璃抗风化能力强弱的影响程度占比，以及不同种玻璃之间化学成分差异性导致的抗风化能力差异性。

关键词：单因素方差分析 数学建模 K means聚类 决策树 灰色关联

# 问题重述

丝绸之路是我国古代连接中西方的交流通道，其中我国古代的玻璃成为了我国早期与各国贸易往来的宝贵物证。早期的古代玻璃在西亚和埃及地区常常被制作成珠形饰品传入我国内部，所以我国古代玻璃吸收其西方技术后在本土就地取材制作，但是不同的是我国古代玻璃与外来的玻璃制品外观相似，但是化学成分却大不相同。

玻璃的主要原料是石英砂，它的主要的化学成分是二氧化硅（SiO2 ），在炼制的过程中需要添加助熔剂和稳定剂。在我国古代有常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等。其中以添加石灰石作为稳定剂，石灰石煅烧以后转化为氧化钙（CaO）可以作为稳定剂。如果添加的助熔剂不同，其主要化学成分也会不同。古代玻璃极易受埋藏环境的影响而风化。并且在风化过程中，内部元素与环境元素进行着大量交换，导致其成分比例会发生变化，从而会影响对其类别的正确判断。文物标记可表现为表面风化和表面无风化两种。其中的文物标记为表面无风化的特征是从表面上可以明显看出其文物的颜色和纹饰，但是不排除局部有较浅的风化，而文物标记表现为表面风化的特征为表面大面积灰黄色区域为风化层，其表面有明显的风化区域，紫色部分是一般的风化表面，但是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。我们要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。

根据题目背景与附件数据需要解决以下问题：

（1）结合表单1数据，分析玻璃文物是否风化与其本身类型、纹饰和颜色之间的关系；对于不同种类的玻璃，分析表面是否风化对于其内部化学成分含量的影响；根据表单2中的检测数据，预测风化玻璃风化前化学成分含量的值。

（2）结合附件数据，分析玻璃的分类主要受到哪些化学成分或者玻璃本身特征的影响；针对不同的玻璃种类，选择合适的化学成分进行亚类划分，给出具体的划分方法以及划分结果，并分析结果的合理性和敏感性。

（3）分析表单3中未知类别玻璃文物的化学成分，进行类型鉴定，并分析结果的敏感性。

（4）对于不同种类的玻璃，分析每种玻璃内部化学成分之间的联系，并且比较不同种类玻璃内部化学成分之间联系的差异性。

# 问题分析

## 问题一的分析

针对问题一,目前已知玻璃类型,纹饰,颜色与是否风化和检测点各氧化物数据，题目要求对玻璃文物各特征和表面是否风化、风化成分进行分析，并且能根据此关系来预测风化前的成分含量。

结合题目相关条件进行斯皮尔曼相关分析，建立风化玻璃与各化学成分所占比例的多元线性回归模型，分析玻璃是否风化与各氧化物比例的显著相关性，得出氧化物因风化导致的含量变化规律。

针对风化前的化学成分含量预测，首先对比同一文物有无风化点的各氧化物含量，得出风化后显著含量变化的氧化物种类， 以未风化数据为训练集，进行线性与非线性回归，输出显著变化氧化物的含量，最后整合预测出文物风化前的化学成分含量。

## 问题二的分析

通过对表单1数据的初步分析，不难发现用于玻璃种类分析的最有效信息为未风化时的颜色以及纹饰信息。而当玻璃风化后，外在信息不易被观察，所以可以通过玻璃内部化学成分含量进行分析，对于此我们决定构建玻璃14个化学成分和3个外在特征与玻璃种类的多元线性回归模型，通过回归系数，可以分析玻璃的分类规律。

而对于亚类的划分，我们首先考虑分类结果的合理性，观察到随着风化程度的不同玻璃表面的颜色呈现一种递进的变化。所以推测，不同的玻璃有着不同的风化程度，这代表着不同的抗风化能力，我们通过观察同一个样本不同风化点的数据发现，部分化学成分含量与风化程度显著相关。所以，我们选取这些风化前后发生显著变化的化学成分，进行Topsis分析，同时，针对不同大类的玻璃我们基于Kmeans++算法进行聚类。最后，将聚类结果与Topsis得分进行对比，观察两者之间是否具有一定的联系。

## 问题三的分析

观察表单3的数据，这里除了化学成分特征外，还包含了表面是否风化的特征。这里我们结合第一题多元线性回归分析的统计规律结果，从中挑选出对于玻璃种类划分影响程度较大的几个变量。然后构建这些变量与玻璃种类之间的决策树模型，并把表单2数据作为训练集，通过对决策树模型的超参数进行敏感性分析确定最优参数，最终将表单3的特征输入决策树模型，得到预测结果。

## 问题四的分析

结合前三题的分析我们知道，对于同一种类的玻璃文物样品，其化学成分含量与其表面是否风化有一定程度的联系。而在第二题中，我们引入了“抗风化能力”这一概念，所以我们决定通过分析化学成分对于玻璃文物样品抗风化能力的贡献程度，来进一步分析同种玻璃文物样品化学成分之间的关联性与不同种玻璃文物样品化学成分关联的差异性。

# 模型假设

（1）忽略各种玻璃类型在检测过程中可能受到的风化影响；

（2）检测玻璃化学成分含量时的环境一致；

（3）假定实验数据在进行处理的过程中没有人为的操作误差；

（4）假定在试验过程中除因素自身外其他影响指标的因素都保持不变；

（5）假设文物采样点未检测出化学成分的含量为0。

# 符号说明

|  |  |
| --- | --- |
| **符号** | **说明** |
| N | 样本数量 |
|  | 斯皮尔曼相关系数 |
|  | 多元线性回归系数 |
|  | 评价指标 |
|  | 归一化得分 |
| J | 总畸变程度 |
| R-squared | 拟合度 |
| a | 两极最小差 |

# 数据的预处理

（1）颜色补充

首先观察数据发现，在题目中所给的58个数据中，在纹饰A和C各有两个数据缺失，缺失项是颜色，其中缺失的四个文物，A纹饰的其他两个指标均为“铅钡类型、表明风化”，C纹饰的其他两个指标均为“铅钡类型、表明风化”，首先对A纹饰文物进行数据进行补充，我们可以通过众数补充法进行补充，观察数据，缺失位置应当补充的数据为浅蓝色。

对于C纹饰缺失的颜色，我们可以通过热卡填充方式进行，下面是对于C纹饰中，同为铅钡且表面风化的各个颜色统计分析：

表1 颜色分布（C相关）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 颜色 | 蓝绿 | 浅蓝 | 浅绿 | 深绿 | 紫 |
| 对应文物编号 | 56,57 | 11,25,43,51,52,54 | 41 | 34,36,38,39 | 08,26 |

我们观察表中数据，数据为未更改的确定性数据，剩余40以及58号文物颜色尚且未知通过热卡填充，我们得到，40号、58号文物颜色分别为深绿色，浅蓝色。由此我们对于文物颜色补充完成，数据预处理已经完成。

（1）无效数据删除

题目要求成分比例累加和介于 85%~105%之间的数据视为有效数据，根据计算采样点15和17不符合要求，故在后续建模中删去。

表2 无效数据

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **文物采样点** | **二氧化硅(SiO2)** | **氧化钠(Na2O)** | **氧化钾(K2O)** | **氧化镁(MgO)** | **氧化铝(Al2O3)** | **氧化铁(Fe2O3)** | **氧化铜(CuO)** | **氧化铅(PbO)** | **五氧化二磷(P2O5)** | **合计** |
| 15 | 61.87 | 3.21 | 7.44 | 1.02 | 3.15 | 1.04 | 1.29 | 0.19 | 0.26 | 79.5 |
| 17 | 60.71 | 2.12 | 5.71 | 0.85 |  | 1.04 | 1.09 | 0.19 | 0.18 | 71.9 |

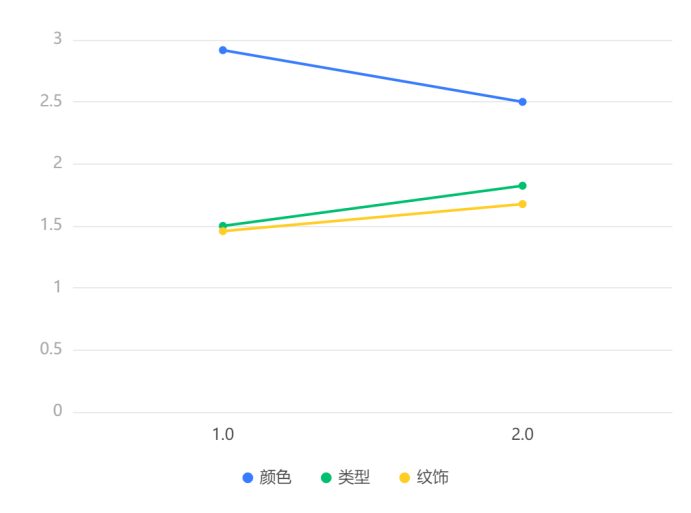
# 模型的建立与求解

## 问题一模型的建立与求解

### 问题一第一小问模型的建立与求解

我们将数据导入SPSSPRO中，利用网站中所含的单因素方差分析进行相关性检验，

**单因素方差分析对比图**



**方差分析结果表**

| 变量名 | 变量值 | 样本量 | 平均值 | 标准差 | 方差检验 | Welch's方差检验 |
| --- | --- | --- | --- | --- | --- | --- |
| 颜色 | 1.0 | 24 | 2.917 | 1.932 | F=0.785  P=0.380 | F=0.741  P=0.394 |
| 2.0 | 34 | 2.5 | 1.638 |
|  | 总计 | 58 | 2.672 | 1.761 |  |  |
| 类型 | 1.0 | 24 | 1.5 | 0.511 | F=7.537  P=0.008\*\*\* | F=6.853  P=0.012\*\* |
| 2.0 | 34 | 1.824 | 0.387 |
|  | 总计 | 58 | 1.69 | 0.467 |  |  |
| 纹饰 | 1.0 | 24 | 1.458 | 0.509 | F=1.476  P=0.230 | F=1.692  P=0.199 |
| 2.0 | 34 | 1.676 | 0.768 |
|  | 总计 | 58 | 1.586 | 0.676 |  |  |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | |

上表展示了方差分析的结果，包括均值±标准差的结果、F检验结果、显著性P值。  
1. 分析每个分析项的P值是否显著(P<0.05)。  
2. 若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据均值±标准差的方式对差异进行分析，反之则表明数据不呈现差异性。

1.0与2.0在颜色上的均值分别为：2.917/2.5；由于满足方差齐性，采用单样本方差检验，方差分析结果P值为0.380>0.05，因此统计结果不显著，说明不同的表面风化在颜色上不存在显著差异。1.0与2.0在类型上的均值分别为：1.5/1.824；由于不满足方差齐性，采用Welch's 方差检验，方差分析结果P值为0.012\*\*≤0.05，因此统计结果显著，说明不同的表面风化在类型上存在显著差异。1.0与2.0在纹饰上的均值分别为：1.458/1.676；由于不满足方差齐性，采用Welch's 方差检验，方差分析结果P值为0.199>0.05，因此统计结果不显著，说明不同的表面风化在纹饰上不存在显著差异。

### 问题一第二小问模型的建立与求解

根据spearman相关性分析的结果，建立风化玻璃与各化学成分所占比例的多元线性回归模型，分析玻璃是否风化与各氧化物比例的显著相关性，得出化学成分的统计规律，模型如下。

Step1.建立多元线性回归模型，涉及m个自变量的多元线性回归模型可表示为

我们通过n组实际数据而引入矩阵记号：

其中成为模型设计矩阵，与是随机向量，是回归系数构成的常数向量，未知待定，且

（为阶单位阵）

Step2.回归系数的最小二乘估计

选取的一个估计值，计为，使随机误差的平方和达到最小，即

由最小二乘法的要求，由多元函数取得极值的必要条件可求解回归参数的标准方程如下：

虽然当X不满秩时，解不唯一，但对任意一组解都能使残差平方和最小，当X秩时，正规方程组的解为，即回归系数的估计值。

Step3.逐步回归分析

由于并非每一个因子都对y影响程度大，可以通过逐步回归的方法来对因子进行筛选,通过对回归系数显著性检验，取t值对应最大概率值,判断是否≤0.05，若满足则可接受，再次建立多元线性回归方程，反之拒绝输出方程。

将（Ⅲ）式中表示为方程的回归系数，表示二氧化硅，氧化钠等化学成分所占的比例。考虑到文物表面化学成分重复性较高，将显著性低于0.1的自变量视作与因变量关系密切，将标准化系数绝对值较大的视作对自变量影响较大的因变量，结合图2与表3可得如下结论：

氧化铅的成分含量对表面风化有较大的正相关影响，氧化钾对表面风化有较大的负相关影响，其中氧化铁，氧化铜，氧化硫的含量与样品表面是否风化相关性较高，二氧化硅无论是否风化，都是文物样品表面主要成分。结合玻璃类型，则认为高钾玻璃不易风化，氧化钾，氧化铁含量较高，铅钡玻璃易风化，氧化铅，氧化铜含量较高。

表3 回归系数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 未标准化系数 | | 标准化系数 | t | 显著性 |
| B | 标准错误 | Beta |
| (常量) | 0.297 | 2.655 | 0 | 0.112 | 0.911 |
| 二氧化硅 | 0.008 | 0.027 | 0.396 | 0.296 | 0.768 |
| 氧化钠 | 0.052 | 0.043 | 0.177 | 1.205 | 0.234 |
| 氧化钾 | 0.061 | 0.037 | -0.488 | 1.663 | 0.102 |
| 氧化钙 | 0.028 | 0.046 | 0.131 | 0.596 | 0.554 |
| 氧化镁 | 0.006 | 0.107 | -0.008 | 0.059 | 0.953 |
| 氧化铝 | 0.034 | 0.030 | 0.215 | 1.136 | 0.261 |
| 氧化铁 | 0.114 | 0.057 | -0.277 | 1.997 | 0.051 |
| 氧化铜 | 0.069 | 0.035 | 0.317 | 1.940 | 0.058 |
| 氧化铅 | 0.018 | 0.027 | 0.734 | 0.668 | 0.507 |
| 氧化钡 | 0.015 | 0.036 | -0.258 | 0.417 | 0.679 |
| 五氧化二磷 | 0.040 | 0.039 | 0.288 | 1.012 | 0.316 |
| 氧化锶 | 0.413 | 0.261 | -0.227 | 1.584 | 0.119 |
| 氧化锡 | 0.071 | 0.159 | 0.049 | 0.448 | 0.656 |
| 二氧化硫 | 0.048 | 0.029 | 0.268 | 1.688 | 0.097 |

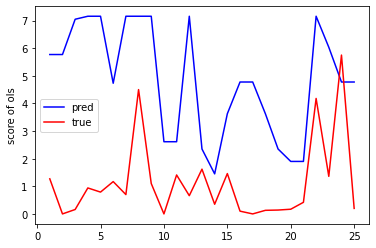
根据上述结论和风化前后各氧化物比例对比，可知玻璃风化前后化学含量主要变化的有二氧化硅，氧化钾，氧化铁，氧化铜，氧化铅，氧化锶，二氧化硫，故建立未风化玻璃类型、颜色、纹饰为自变量，样本表面氧化物比例为因变量的逐步多元线性回归方程。

物含量进行多元线性回归，模型的显著性检验为

表4 二氧化硅(SiO2)回归结果

|  |  |  |  |
| --- | --- | --- | --- |
| Dep. Variable: | 二氧化硅(SiO2) | R-squared : | 0.731 |
| Model: | GLS | Adj. R-squared : | 0.424 |
| Method: | Least Squares | F-statistic: | 2.38 |
| Date: | Sun, 18 Sep 2022 | Prob (F-statistic): | 0.135 |

得到的预测结果为



## 问题二模型的建立与求解

### 基于多元线性回归的玻璃分类模型

基于对数据的分析，本文认为附录2提供的数据表明化学成分对玻璃分类有关，并建立多元回归模型：

在（Ⅲ）式中表示二氧化硅，氧化钠等化学成分所占的比例， 表示第i个样本为铅钡玻璃，为高钾玻璃，利用 matlab 统计工具箱建立多元线性回归方程，对结果统计量进行分析得到R-squared为0.927， 表明预测值和实际值接近92.7％，回归方程显著。

由t分布的p值可知，除氧化锶以外，其他13种化学成分均与类别有显著相关性，根据回归系数绝对值大小的升序排列，可知氧化钾对玻璃类型的分类的决定作用最大；同时其标准误差偏小，结果更具解释性。同时在对回归系数和标准误差的综合考量下，氧化钡也对玻璃分类起重要作用。

表5 回归系数

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| 二氧化硅**(SiO2)** | 0.0086 | 0.001 | 12.207 | 0.000 | 0.007 | 0.010 |
| 氧化钠**(Na20)** | -0.0386 | 0.014 | -2.775 | 0.008 | -0.067 | -0.010 |
| 氧化钾**(K20)** | 0.0425 | 0.011 | 3.975 | 0.000 | 0.021 | 0.064 |
| 氧化钙**(CaO)** | 0.0115 | 0.019 | 0.618 | 0.540 | -0.026 | 0.049 |
| 氧化镁**(MgO)** | -0.0103 | 0.047 | -0.220 | 0.827 | -0.105 | 0.084 |
| 氧化铝**(AI2O3)** | -0.0324 | 0.009 | -3.507 | 0.001 | -0.051 | -0.014 |
| 氧化铁**(Fe2O3)** | 0.0276 | 0.026 | 1.066 | 0.293 | -0.025 | 0.080 |
| 氧化铜**(CuO)** | 0.0354 | 0.015 | 2.423 | 0.020 | 0.006 | 0.065 |
| 氧化铅**(PbO)** | -0.0029 | 0.002 | -1.635 | 0.110 | -0.007 | 0.001 |
| 氧化锲**(BaO)** | -0.0174 | 0.005 | -3.232 | 0.002 | -0.028 | -0.007 |
| 五氧化二磷**(P2O5)** | -0.0008 | 0.008 | -0.098 | 0.923 | -0.017 | 0.015 |
| 氧化锂**(SrO)** | 0.0812 | 0.110 | 0.739 | 0.464 | -0.141 | 0.303 |
| 氧化揚**(SnO2)** | 0.0268 | 0.067 | 0.399 | 0.692 | -0.109 | 0.162 |
| 二氧化硫**(S02)** | 0.0300 | 0.012 | 2.574 | 0.014 | 0.006 | 0.054 |

### 基于keans++聚类玻璃亚分类模型

先将玻璃分成高钾和铅钡两大类，基于对数据的分析，对于同一样本，风化点与未风化的化学成分具有一定的变化趋势，我们观察多组这样的数据发现，部分化学成分的变化趋势相同。本文推测随着风化程度的不同，样本的部分化学成分有着往一个趋势增加或者减少的特性。

将高钾玻璃的亚分类聚成两类，对比原数据的分化类型与聚类结果的一致性；将铅钡玻璃的亚分类聚成三类，对比原数据的分化类型与聚类结果的一致性。

对每类进行kmeans++聚类，具体算法流程如下图所示：

Step1.输入聚类数量k和迭代次数n;

Step2.初始化聚类中心,随机选取一个样本点作为第一个聚类中心，计算每个样本与当前已有聚类中心的最短距离，用轮盘法选出下一个聚类中心，直到聚类中心数量达到k。

Step3.计算其余的各个数据对象到这K个初始聚类中心的距离，把数据对象划归到距离它最近的那个中心所处在的簇类中。

Step4.调整新类并且重新计算出新类的中心;

Step5:Step3和Step4直到达到最大迭代次数n或者收敛时输出聚类结果。

根据化学成分与风化程度的相关性，将高钾玻璃划分为两类，以风化程度的高低作为划分依据；铅钡玻璃划分为三类，以风化程度强、中、弱为划分依据。

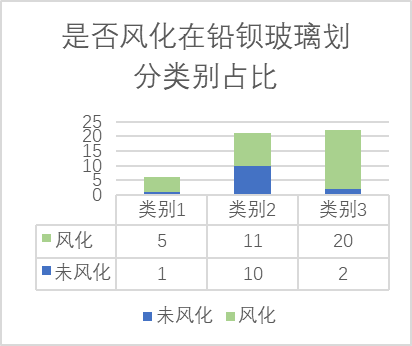
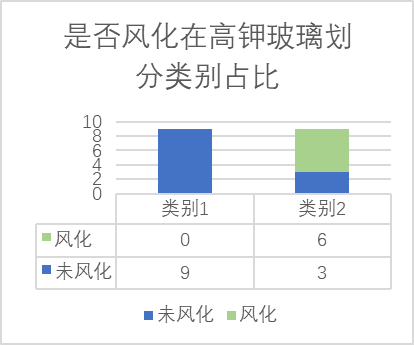


图6 是否分化在玻璃划分类别占比

由是否风化在高钾、铅钡玻璃中划分类别占比可得划分依据为风化程度，通过横向纵向交叉对比，按照所分类别，从类别1到类别2、类别3，风化样品占比逐步增加，所以本文所亚分类划分的依据为风化程度。

合理性分析：

通过猜想化学成分含量一定程度上反映玻璃的抗风化能力，本文使用Topsis模型对玻璃抗风化能力进行评估，由于化学成分类别过多，采用前文中使用的基于多元线性回归的玻璃分类模型中对回归系数和标误的综合分析，提取出影响玻璃分类最大的前五个指标。

这五个指标依次为二氧化硅,氧化钾,氧化钙,氧化铝,五氧化二磷，且他们分别为极大型、极小型、极小型、极小型、极小型指标。最终得分高低为玻璃抗风化能力的强弱，对比划分的亚类结果发现一致性，增强了模型的合理性。

### 指标计算

对不同化指标进行归一化处理，处理公式如下：

效益性指标： 成本性指标：

Step 1：将数据导入matlab，得到二氧化硅,氧化钾,氧化钙,氧化铝五氧化二磷共五个指标的权重；

Step2：对五个指标数据进行指标属性同向化处理并构造归一化初始矩阵

(Ⅶ)

Step 3：计算得分并进行归一化；

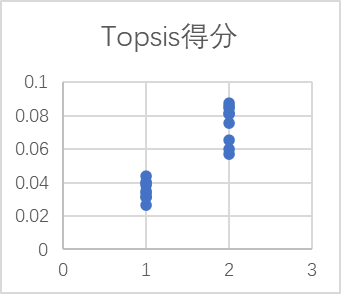
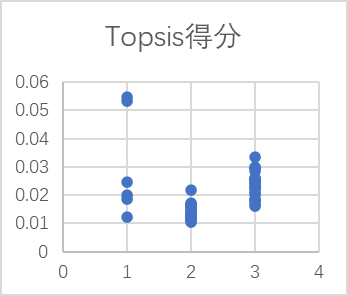
定义最大值：

定义最小值：

定义第 个评价对象与最大值的距离

定义第 个评价对象与最小值的距离   
那么, 我们可以计算得出第 个评价对象末归一化的得分: ，

通过对用Topsis得分对玻璃划分结果进行分析，在高钾玻璃中采用Topsis模型进行评价，得分越高，属于类别2的样本数量越集中；而在铅钡玻璃中，尽管未能发现得分的高低与类别有显著区分关系，但纵向观察，类别一致的样本Topsis得分呈现集中性。

高钾玻璃 铅钡玻璃

图7 topsis得分与玻璃亚类结果比对图

同时用系统聚类模型进行进一步验证，高钾玻璃的系谱图中，以重新标度的距离组合为15划分发现与前文所选用的初始K值相同，说明使用的keans聚类数的选取符合合理划分值,谱系图见附件1。

敏感性分析：

对于敏感性的分析使用总畸变程度。各个类的畸变程度等于该类重心与其内部成员位置距离的平方和；设一共将n个样本划分到K个类中（K≤n-1，即至少有一类中有两个元素）用表示第k个类（k=1，2，…，K），且该类重心的位置记为,那么第k个类的畸变程度为

定义所有类的总畸变程度

以横坐标为玻璃亚类类的类别数K，纵坐标为畸变总程度J。由图可直观感受到高钾玻璃在聚类类别数大于2时曲线趋于平缓，敏感度降低；铅钡玻璃在聚类类别数大于4时曲线趋于平缓，敏感度降低.

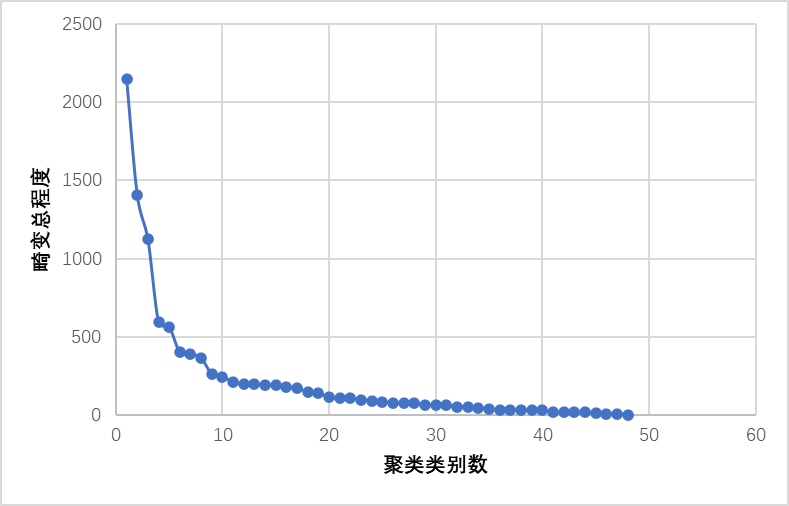


图8 铅钡敏感度

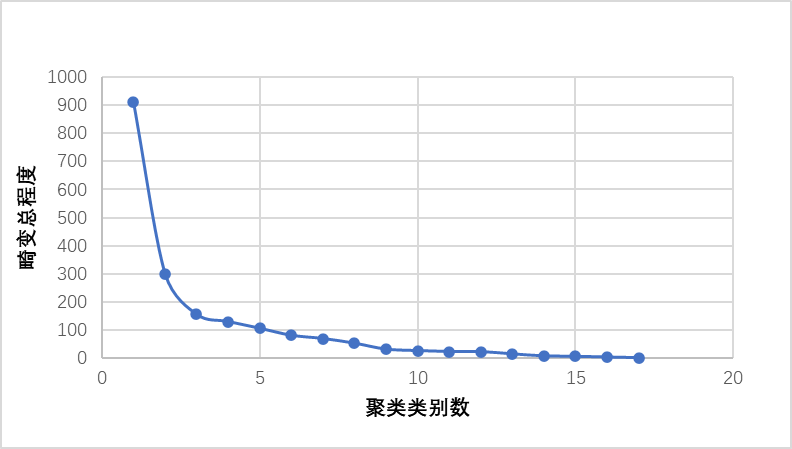


图9 高钾敏感度

## 问题三模型的建立与求解

为了对未知类别玻璃文物的分类有更好的合理性和敏感度，基于问题一所求各氧化物含量比例与玻璃风化的显著性，得到各属性对分类的重要程度，使用ID3决策树的算法建立以显著性高的化学成分为因变量的玻璃文物分类模型，对各化学成分进行风化贡献评级分析分类的合理性，通过选择最敏感或最不敏感的属性作为当前拓展属性进行分支，来判断分类模型是否稳健。

### 基于ID3决策树算法分类模型

在数据预处理中，基于问题一，将氧化铅变量去除。信息熵是由香农提出，表示对随机变量X不确定性的度量。熵越高，信息越多;熵越低，信息越少。信息熵如公式( 1)所示:

其中s表示随机变量,c表示随机变量的取值数量，pi表示每个取值时的概率。

信息增益:假设S是一组实例,A是一个特征，s v是S的子集，其中A = v, Values (A)是A 的所有可能值的集合，则信息增益为

当使用决策树中的节点将训练实例划分为更小的子集时，熵会发生变化。某个特征的信息增益是熵变化程度的量度。该算法具体的步骤如下：

Step1.特征选择

这一步主要根据特征选择算法来选取合适的特征来对当前样本集进行划分。在特征选择中本文选用使用信息增益为准则筛选出跟分类标签最相关性的特征。

Step2.决策树生成

选好特征后，就从当前树的根节点出发，遍历所有特征,找到那个使分信息增益最大的特征，将其设置为根节点，并且将这feature删除掉；根据该特征的不同取值建立子节点;对每个子节点使用相同的方式生成新的子节点，直到信息增益很小（满足一定的阈值）或者当前样本集不需要再划分为止；采用信息增益来选取最优分裂特征。

Step3.决策树剪枝

决策树剪枝的主要防止过拟合，通过提前去掉部分分支来降低过拟合的风险。剪枝的阈值设置是一个关键。

Step4.调整超参数分析敏感性

影响决策树敏感性的有超参数max\_depth，min\_samples\_leaf， min\_samples\_split。如果max\_depth设定的太低，那么决策树不能够很好的学习数据，如果设置的太大则极易造成模型的过拟合，min\_samples\_leaf超参数设置得太小会引起过拟合，设置得太大就会阻止模型学习数据，min\_samples\_split决定模型是否可以分支，还可以有效预防过拟合，本文对这三个超参数进行测试来分析模型的敏感性。

|  |  |
| --- | --- |
| 模型的求解  1. 模型的分类结果   class = 0 表示铅钡玻璃  class = 1 表示高钾玻璃  决策树  图10 基于决策树模型的对表单三化学成分预测  表单三化学成分预测具体结果位于附录中。   1. 敏感性检验结果   image-20220918005441176  图11 超参数max\_depth  首先，我们对决策树的最大深度进行敏感性分析。超参数max\_depth限制树的最大深度，超过设定深度的树枝将被全部剪掉用于限制树的最大深度。考虑到题目数据量较少，树的最大深度不宜过高，所以我们对该参数进行敏感性分析。如图所示，当max\_depth取5时，模型在验证集和训练集上的效果达到最佳，因此最优参数为max\_depth=5。  image-20220918005757811  图12 超参数min\_samples\_leaf  其次，对决策树超参数min\_samples\_leaf进行敏感性分析。min\_samples\_leaf表示一个节点在分枝后的每个子节点都必须包含至少min\_samples\_leaf个训练样本，否则分枝就不会发生，或者分枝会朝着满足每个子节点都包含min\_samples\_leaf个样本的方向去发生。min\_samples\_leaf一般和max\_depth参数配合使用，可以让模型变得更加平滑，同时，min\_samples\_leaf可以保证每个叶子的最小尺寸，可以在回归问题中避免低方差，过拟合的叶子节点出现。如图所示，随着min\_samples\_leaf的增加，train和test的准确率都有所下降，综合考虑模型的泛化能力以及准确度，我们认为min\_samples\_leaf的最优值为2.  image-20220918005824667  图13 超参数min\_samples\_split  最后，对超参数min\_samples\_split进行敏感性分析 min\_samples\_split限制决策树一个节点必须要包含至少min\_samples\_split个训练样本，这个节点才允许被分枝，否则分枝就不会发生，min\_samples\_split可以有效防止过拟合。但是由于玻璃样本的数据比较少，并且结合敏感性曲线分析，我们认为min\_samples\_split最优参数为4. 问题四模型的建立与求解 在第二题中，我们引入了“抗风化能力”这一概念，并且用Topsis得分去表征它，所以在这一题中，认为化学成分之间的关系与玻璃本身的抗风化能力有关，本文基于灰色关联模型构建了针对化学成分之间的关系与不同类别的玻璃文物抗风化能力的分析模型。  化学成分之间的灰色关联模型  Step1.将第二题所求得的铅钡与高钾玻璃Topsis评分当做母序列，其他化学成分当做子序列，以此来分析其他化学成分对于玻璃抗风化能力的影响程度，以及不同化学成分之间的联系。  Step2.对正向化后的矩阵进行预处理，得到矩阵Znxm=(Zij)nxm，具体矩阵数值见附件。  Step3.计算子序列中各个指标与母序列的关联系数。  母序列： （Ⅷ）  子序列：    两极最小差 （Ⅸ）  两极最大差 （Ⅹ）  (ρ为分辨系数，取0.05 ) （XI）  和的灰色关联度 （XII）  模型的求解  根据python计算，设定化学成分对应特征顺序为：二氧化硅，氧化钠，氧化钾，氧化钙，氧化镁，氧化铝，氧化铁，氧化铜，氧化铅，五氧化二磷，氧化锶，氧化锡，二氧化硫。  高钾的灰色关联度系数为：  [0.97835483, 0.84242264, 0.88733777, 0.884876  , 0.89560761,   0.91239705, 0.88283529, 0.9178258 , 0.86610521, 0.8480089 ,0.90449882, 0.85303659, 0.86511522, 0.84456953]  铅钡的灰色关联度系数为：  [0.94105583, 0.88245123, 0.91247829, 0.95657924, 0.92914516,  0.94592981, 0.91610475, 0.92718159, 0.97269554, 0.9536275,0.93650384, 0.9554802 , 0.88464863, 0.89391412]  图14 化学成分关联关系的差异性  高钾与铅钡玻璃其抗风化能力关联度由高到低如图所示：  图15 玻璃抗风化能力关联度  对于高钾玻璃而言，对抗风化能力贡献程度较大的化学成分主要为二氧化硅、氧化铜、氧化铝、五氧化二磷等成分。  而对于铅钡玻璃而言，对抗风化能力贡献程度较大的化学成分主要为氧化铅、氧化锶、氧化钡、氧化铝等成分。    图16 表面风化在类型的数量  如图所示，铅钡玻璃的风化样本明显多于高钾玻璃。而在高钾玻璃抗风化能力成分中占主要地位的二氧化硅，其含量在铅钡玻璃中相对较少。所以我们认为对于不同种类玻璃，高钾玻璃含有更多的“抗风化能力”成分，而相应成分在铅钡玻璃中比重则较小，这就是不同种玻璃化学成分关联关系的差异性，而这种差异性进一步影响了两种玻璃的抗风化能力。 |  |