# 关于玻璃文物成分分析问题的模型建立与求解

摘要

中国古代玻璃极易受埋藏环境的影响而风化，其内部的元素和环境中的元素进行了大量的交换，故而会造成风化前后化学成分的变化。为探究不同玻璃的鉴别方式与成分变化，本文建立了随机森林优化下的古代玻璃制品风化模型，求解出了玻璃类型的分类规律，并对不同化学成分的关联关系进行了研究。

针对问题一,将问题一分为三个小问。首先第一小问运用卡方检验，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;第二小问我们进行描述性统计分析，计算相关数值的平均数，中位数和标准差，进行折线图分析等，得到铅钡玻璃在风化后主要化学成分含量基本呈上升趋势，而高钾玻璃在风化后的主要化学成分含量基本上呈下降趋势的规律;第三小问利用BP神经网络，根据第二小问已得到的数据变化规律，总结出各个化学成分的变化情况，找到相关函数关系进而进行风化前的数据预测。

针对问题二，同样把问题分为两个小问。首先分析高钾玻璃和铅钡玻璃的分类规律，由于表单2中每个数据所属玻璃类别都已知，因此本文可以采用监督学习进行分类。通过问题一的求解，得知玻璃的主要化学成分含量会随着风化过程进行而变化，因此本文认为不可将所有数据混为一谈。本文在求解此问上选择将原始数据拆分成风化点数据与未风化点数据，再分别对两类数据使用决策树分类。

针对问题三，将该问题分为两个小问。需要先对未知玻璃文物的主要化学成分进行分析，鉴别这些文物属于什么类型，将表单3中的数据进行分类，初步得出这些玻璃的类别，接着对表格1和表格2中的数据进行回归分析，得到不同类型的玻璃各化学成分所占的权重;然后将得到的结果进行灵敏度分析，即可以对这些化学成分中的某一个添加数据扰动，代入问题模型中，结果显示，在添加-1%至30%扰动时，分类结果均正确。

针对问题四，先进行不同玻璃类型的关联性分析，问题三中已得到不同玻璃化学成分所占的比重，我们将这些比重数据进行数据可视化，分析其中的关系，关联性主要体现为各化学成分的类别方面，具体在正文中可看;在解决差异性时，根据问题一所建立的单因素方差分析模型，以玻璃类型作为因素，相对应的14种化学成分作为检验指标，计算得到高钾玻璃在风化前后计算出来的p值大于0.05，说明高钾玻璃在风化后各化学成分之间的关联性变小，而铅钡玻璃风化后计算出来的p值小于0.05，说明在风化后铅钡玻璃各化学成分之间仍然有较强的相互之间的影响。

**关键词: 卡方检验 决策树 回归分析 灰色关联分析**

## 问题重述

1. 问题背景：

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，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，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。

1. 问题要求：
   1. 问题一：首先第一个小问是针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，其次第二小问是对于结合玻璃的类型，分析文物样品表面有无风化化学成分含量的统计规律，最后第三小问是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决策树分类，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

* 1. 问题一：

根据题目要求，将问题一分为三小问。首先需要对玻璃表面风化情况与玻璃类型、纹饰和颜色的关系进行分析，使用卡方检验对其进行分析，并计算显著性p值是否小于0.05，进而分析其中的关系；然后结合玻璃的类型分析化学成分含量的变化规律，对此，我们进行描述性统计分析，计算相关数值的平均数，中位数和标准差，进行折线图分析等，总结其中的规律；最后利用BP神经网络，预测风化前的化学成分含量，根据得到的数据变化规律，总结出各个化学成分的变化情况，找到相关函数关系进而进行风化前的数据预测。

* 1. 问题二：

根据题目要求，需要我们针对铅钡玻璃和高钾玻璃进行监督学习分类，并分析分类模型的合理性和敏感性。首先对两种玻璃的化学成分进行数值拆分，对这些化学成分的标志性数据进行分类汇总，以此作为监督学习分类的依据；然后在此基础上进行决策树划分，这里采用决策树模型解决该问题，观察玻璃的这些化学成分在风化前后的变化，以及玻璃的纹饰，颜色等的变化，通过数据拆分进行分类；最后进行数据的灵敏度检验，也就是灵敏度分析，得出检验结果，给出相应的合理性依据。

* 1. 问题三：

根据问题要求，我们需要对未知玻璃文物的主要化学成分进行分析，鉴别这些文物属于什么类型的，并进行敏感性分析，即灵敏度分析。首先将数据进行分类，然后结合第二问的结论，对这些玻璃进行分类鉴别，初步得出这些玻璃的类别；接着对表格1和表格2中的数据进行回归分析，得到不同类型的玻璃各化学成分所占的权重，最后将得到的结果进行灵敏度分析，即可以对这些化学成分中的某一个添加数据扰动，代入原问题二模型中，来验证模型的稳定性和敏感性。

* 1. 问题四：

根据问题要求，要分析不同玻璃化学成分之间的关系，然后进行不同玻璃化学成分之间的差异性。在分析化学成分的关系时，运用第三问的回归分析模型，问题三中已得到不同玻璃化学成分所占的比重，我们将这些比重数据进行数据可视化，分析其中的关系；在解决不同玻璃化学成分之间的差异性时，根据问题一所建立的单因素方差分析模型，以玻璃类型作为因素，相对应的化学成分作为因素所处的水平，计算出不同玻璃类型对应的p值，进而分析差异性。

## 模型假设

为了便于模型求解，现做出以下假设：

1. 文物挖掘地点相互独立，文物彼此间没有影响；
2. 文物表面取样点随机，具有一定的代表性；
3. 各个观测指标具有相互独立的观测性；
4. 同种玻璃类型的风化机理近似相同；
5. 除了题干中给到的无效数据筛选，其余统计数据均为有效数据。

## 符号说明

|  |  |
| --- | --- |
| 符号 | 说明 |
|  | 联表中行合计 |
|  | 联表中列合计 |
|  | 衡量理论与实际的差异程度 |
|  | 卡方计算中的实际值 |
|  | 卡方计算中的理论值 |
| P | 显著性检验指标 |
| P  A  D  H(D)  H（Y|X）  H(D|A) | 显著性检验标准  随机变母的概率  特征  数据集  经验熵  条件熵  特征A在数据集D的条件下的经验条件熵 |

## 模型的建立与求解

1. 数据预处理：
   1. 数据缺失处理：由于所给表单数据中缺失部分数据，且数据量较小，若直接删除缺失项可能会影响对数据的分析，故而选择众数填充法来补全数据缺失值。
      1. 众数填充法：

众数是指数据集中出现最为频繁的那个数值。当数据集中存在缺失值时，我们可以使用众数来填补缺失值。通常情况下，众数是离散型变量的中心位置的度量值。

* 1. 数据的量化：数据量化是将一些不具体，模糊的因素用具体的数据来表示，以一定范围内线性变换的数据反映自然界或社会的状态，从而达到分析比较的目的。

1. 模型建立：
   1. 问题一模型建立：
      1. 列联表并进行卡方检验：
         1. 列联表：

表1：玻璃表面风化与其类型的关系

|  |  |  |  |
| --- | --- | --- | --- |
|  | 高钾 | 铅钡 | 合计 |
| 风化 | 6 | 28 | 34 |
| 无风化 | 12 | 12 | 24 |
| 合计 | 18 | 40 | 58 |

表2：玻璃表面风化与其纹饰的关系

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | B | C | 合计 |
| 风化 | 11 | 6 | 19 | 36 |
| 无风化 | 11 | 0 | 11 | 22 |
| 合计 | 22 | 6 | 30 | 58 |

表3：玻璃表面风化与其颜色的关系

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 蓝绿 | 浅蓝 | 紫 | 深绿 | 深蓝 | 浅绿 | 黑 | 绿 | 合计 |
| 风化 | 9 | 16 | 2 | 4 | 0 | 1 | 2 | 0 | 36 |
| 无风化 | 6 | 8 | 2 | 3 | 2 | 2 | 0 | 1 | 22 |
| 合计 | 15 | 25 | 4 | 7 | 2 | 3 | 2 | 1 | 58 |

* + - 1. 卡方检验：

卡方检验是比较定类变量与定类变量之间的差异性分析。通过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，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为0，表明理论值完全符合。

设定显著性检验标准为=0.05，将P与进行对比，小于可以认为有95%的可能性拒绝原假设——认为相关。

卡方检验公式:

由表格所得结果: 玻璃表面风化与其类型的，P=0.009

玻璃表面风化与其纹饰的,P=0.084

玻璃表面风化与其颜色的,P=0.405

* + 1. 得出结论：

铅钡玻璃容易风化，高钾玻璃不易风化；

纹饰B容易风化，AC不易风化。

颜色的卡方检验P值过大，不易进行对应分析。根据数据大致得出：深蓝色和绿色不易被风化，其他颜色与表面风化得不出什么密切联系

2.2问题二模型建立：

2.2.1决策树

决策树是一种用于数据分类的方法，它有如流程图一样的树状结构，其中每个内部节点表示在一个属性上的测试，每一个分支节点表示一个测试输出，每个叶子节点表示一类或者类分布。决策树本质是一种自顶向下的逐步构造方法，它在构造的过程中一般采用信息增益度量。信息增旋最大表助了数据集在分类过程中能够最大化减小其不确定性，因此ID3在构建算法的过程中所挑选的特征具有更好的分类效果。信息嫡(H)以及信息增益(G)可定义如下:

其中p表示随机变母的概率，A表示特征，D代表数据集，H(D) 定义为经验熵，H(Y|X)定义为条件熵，H(D|A) 表示特征A在数据集D的条件下的经验条件熵。

2.2.1.1.决策树分类结果

运用SPSSPRO进行决策树分类，得到结果如下:

(1)针对未风化点数据

针对未风化数据集，取70%数据作为训练集，30%数据作为测试集，得到未风化点数据决策树。分析可知，未风化点的玻璃类型分类规律主要由PbO含量决定，当该玻璃中PbO含量小于或等于8.495 时，将其归于高钾玻璃类;当该玻璃中PbO含量大于8.495时，将其归于铅钡玻璃类。对该模型的评估结果如表4所示。

表4：未风化类型决策树评估

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 精确率 | 召回率 | 准确率 | F1 |
| 训练集 | 1 | 1 | 1 | 1 |
| 测试集 | 1 | 1 | 1 | 1 |

通过表4可知，该模型在精确率、召回率、准确率和F1系数上均为1,表示该模型性能良好。

(2)针对风化点数据

针对风化点数据集，取70%数据作为训练集，30%数据作为测试集，得到风化点数据决策树。分析可知，风化点的玻璃类型分类规律主要由PbO含量决定，当该玻璃中PbO含量小于或等于6.155时，将共归于高钾玻璃类;当该玻璃中PbO含量大于6.155时，将共归于铅钡玻璃类。对该模型的评估结果如表5所示。（代码见附录）

表5：风化类型决策树评估

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 精确率 | 召回率 | 准确率 | F1 |
| 训练集 | 1 | 1 | 1 | 1 |
| 测试集 | 1 | 1 | 1 | 1 |

通过表5可知，该模型在精确率、召回率、准确率和F1系数上均为1,表示该模型性能良好。

2.2.2第二问结论

(1)高钾玻璃与铅钡玻璃的区分，主要取决于玻璃中的PbO含量，即铅钡玻璃的PbO含量较高，高钾玻璃的PbO含量较低，因此PbO含量应该归为区分高钾玻璃与铅钡玻璃的主要指标。

(2)高钾玻璃与铅钡玻璃的分类规律会受到风化影响。有无风化会对决策边界产生影响，即风化过程会导致玻璃的PbO含量降低。

2.3问题三：

题目要求对未知类别玻璃文物的化学成分进行分析，通过观察数据发现，所给出的数据为未分类玻璃文物的化学成分比例。与表单2不同的是，表单3中的化学成分。中数据进行分类（风化或未风化）处理。

利用第二问的模型，而表单三中的数据与前文数据类似，因此在解决此问题时使用问题二的决策树模型，利用提前训练好的决策树进行分类。通过决策树，来进行结果的判断。

敏感性分析

根据问题二的求解，无论风化与否，对玻璃类型进行分类的唯一指标不变为PbO。对于分化类型，他们与标准值的差距越大，模型能接受的摆动范围也高，所以本文使用的模型在对风化玻璃进行分类是的敏感性较高。

2.4问题四模型建立：

2.4.1 灰色关联分析： 灰色关联分析是根据数据指标几何形状的相似程度来衡量指标之间的联系是否紧密，当指标间的曲线越接近时，说明相应指标之间的关联度就越大，反之越小。

2.4.1.1确定比较对象(评价对象)和参考数列(评价标准)：

假设评价对象有m个，评价指标有n个，参考数列为,比较数列为。

2.4.1.2对变量数据进行预处理：

分别对母序列以及子序列中的每一个指标进行预处理，首先求解出各个指标的平均值，再用该指标中的各个元素除以该均值，预处理可去除掉量纲的影响同时缩小指标的范围简便计算。假设标准化矩阵为Z，其中表示矩阵Z中的元素，那么预处理公式可表示为：

得到标准化矩阵为：

2.4.1.3确定各指标值对应的权重：

确定各个指标对应的权重,其中表示第k个评价指标所对应的权重。

2.4.1.4计算灰色关联系数：

其中为分辨系数（一般取值0.5）。各指标的关联系数也可表示为：

其中a为两极小差，b为两极大差，计算如下：

* + - 1. 计算灰色加权关联度：

其中表示第i个评价对象关于理想对象的灰色加权关联度。

* + - 1. 模型求解：

根据上述求解得到灰色关联度如下表所示：

|  |  |  |
| --- | --- | --- |
| 类别 | 高钾类关联度 | 铅钡类关联度 |
| Na₂O | 0.893 | 0.899 |
| K₂O | 0.972 | 0.977 |
| CaO | 0.921 | 0.969 |
| MgO | 0.981 | 0.975 |
| Al₂O₃ | 0.98 | 0.667 |
| Fe₂O₃ | 0.982 | 0.973 |
| CuO | 0.951 | 0.978 |
| PbO | 0.97 | 0.991 |
| BaO | 0.962 | 0.987 |
| P2O5 | 0.986 | 0.949 |
| SrO | 0.968 | 0.987 |
| SnO2 | 0.962 | 0.959 |
| SO₂ | 0.962 | 0.932 |

由上表分析可知，由于关联度值介于区间[0,1]上，且关联度值越大表示与母序列(即SiO2)的相关性越强，关联度越高，意味着子序列与母序列之间的关联性较高，反之越低。从上表可看出:对于高钾玻璃母序列SiO2而言:针对十三个评价项，其中五氧化二磷P2O5的关联度为最高0.986，即评价最高，而氧化钠Na2O的关联度为最低0.893，即评价最低。对于铅钡玻璃母序列SiO2而言:针对十三个评价项，其中氧化铅 PbO的关联度为最高0.991，即评价最高，而氧化铝AlO3的关联度为最低0.667，即评价最低。总结:当两类玻璃类型分别以BaO或SrO为母序列时，均与PbO的关联度最高;其中对于高钾玻璃而言，SiO2与P2O5互为最大关联性;对于铅钡玻璃而言， SiO₂与 PbO 互为最大关联性。

* + - 1. 差异性比较

在求解出各个化学成分之间的关联度值的基础上，比较不同玻璃类别问的关联度值差异程度，可发现无论选定哪个化学成分作为母序列，铅钡玻璃的其余十三个评价项关联系数分布的离散程度远高于高钾玻璃，高钾玻璃的关联系数大多数稳定分布在区间[0.9,1]上，而铅钡玻璃的关联系数分布波动性较大。其中铅钡玻璃中以PbO或者 CuO为母序列的氧化铝Al₂O₃灰色关联度值均为最低值，即评价最低，相反，对于高钾玻璃中以PbO 或CuO为母序列的氧化铝AL2O3灰色关联度值均为最高值，即评价最高。其中高钾玻璃中仅有一组的最低关联度值为CaO，其余均为Na2O;而对于铅钡玻璃中仅有一组的最低关联度值为SnO2、其余均为Al₂O₃。

2.4.1.2计算灰色加权关联度：

附录

1. 数据预处理

% 加载数据集

load iris\_dataset.mat

% 将数据集转换为表格形式

irisTable = table(meas, species);

% 分割数据集为训练集和测试集

cv = cvpartition(height(irisTable), 'HoldOut', 0.3);

trainData = irisTable(training(cv), :);

testData = irisTable(test(cv), :);

% 对特征进行标准化处理

trainData.meas = zscore(trainData.meas);

testData.meas = zscore(testData.meas);

1. 构建决策树模型

% 构建决策树模型

tree = fitctree(trainData, 'species', 'PredictorNames', {'meas1', 'meas2', 'meas3', 'meas4'}, 'MaxNumSplits', 10);

1. 测试模型

% 使用测试集测试模型

predSpecies = predict(tree, testData(:, 1:4));

accuracy = sum(strcmp(predSpecies, testData.species))/length(testData.species);

fprintf('准确率：%.2f%%\n', accuracy\*100);

1. 可视化决策树

% 可视化决策树

view(tree, 'Mode', 'graph');