玻璃文物的化学成分分析与鉴别

郭佳欣
（北京建筑大学，北京 102616）

摘 要 玻璃是早期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，我国古代玻璃吸收西亚和埃及地区的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且极易受埋藏环境的影响而风化。本文根据所给玻璃文物的相关数据，对玻璃文物的表面风化情况、类型、化学成分建立数学模型，进行相应的分析、检测、预测及鉴别，并对结果进行敏感性分析。

关键词 Spearman 系数 Logistic 回归 Fisher 判别 K-means 聚类 假设检验
中图分类号：TQ171 **文献标识码**：A **文章编号**：1007-0745(2022)11-0049-03

1 问题重述

1.1 问题背景

玻璃文物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。

玻璃的主要原料是石英砂，主要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。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。

1.2 具体问题

1. 对玻璃文物的化学成分含量数据进行分析，将成分比例累加和介于 85%~105% 之间的数据视为有效数据。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。

2 问题分析

2.1 化学成分含量分析

针对问题一，该问要求需要对玻璃表面风化情况与玻璃类型，纹饰和颜色的相关性进行分析，并结合玻璃的类型分析化学成分含量的变化规律以及预测风化前的化学成分含量。

首先，对文本数据进行预处理，通过求解 Spearman 系数来对玻璃类型、颜色、纹饰、表面风化四个定类变量进行相关性分析，从而得到变量间的相关性。

其次，对玻璃文物的化学成分含量进行预处理，将成分比例累加和介于 85%~105% 之间的数据视为有效数据，从而删掉 15 号和 17 号并将表结合。结合玻璃的类型，我们分别对高钾类，和铅钡类的风化前后变化差异进行描述性统计分析，以及有无风化各成分的均值是什么，假设检验，从而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，通过给定的数据，判断出二氧化硅为主要成分，所以针对风化的玻璃文物数据做 Logistic 回归模型，令高钾为“0”，铅钡为“1”，预测出其类别，从而得出风化前的化学成分含量[1]。

2.2 化学成分亚类划分及敏感性分析

针对问题二，依据附件数据分析高钾玻璃、铅钡玻璃的分类规律，该问要求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。

通过数据分析高钾玻璃、铅钡玻璃的分类规律，我们采用聚类分析模型中的 K-means 算法，对于每个类别所有的化学成分对其进行亚类划分，得出相关化学成分的分类[2-4]。

3 模型假设

（1）数据预处理后，题目所给的数据均是合理的，正确的。（2）题目所提供结果均符合一般规律。（3）题目中所给的各项指标的测定时带来的误差忽略不计。（4）题目不考虑其他因素对玻璃文物风化的影响。（5）题目不考虑随时间影响使其风化产物产生影响。

4 模型建立与求解

4.1 问题一的模型建立与求解

4.1.1 建立 Spearman 和 Logistic 回归模型

1.Spearman 相关系数的具体计算方法：

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

其中，n 是样本的数量，d 代表数据 x 和 y 之间的等级差。

$$r_s\sqrt{n-1}\sim N(0,1) \quad (2)$$

在得到的 p 值中，如果 p 值大于 0.05，则没有显著性差异，也就是说没有理由认为显著性差异存在，即没有相关性，如果 p 值小于 0.05，我们可以认为存在显著性差异。

2.Logistic 回归的原理是用逻辑函数把线性回归的结果 $(-\infty,+\infty)$ 映射到 (0,1)。故先建立线性回归表达式和逻辑函数表达式。

线性回归函数的数学表达式：$y=\theta_0+\theta_1*x_1+\theta_2*x_2+\cdots+\theta_nx_n=\theta_Tx$，其中 x_i 是自变量，y 是因变量，y 的值域为 $(-\infty,+\infty)$，θ_0 是常数项，θ_i 是待求系数，不同的权重 θ_i 反映了自变量对因变量不同的贡献程度。

逻辑函数表达式：

$$g(z)=\frac{1}{1+e^{-z}} \quad (3)$$

逻辑回归函数表达式：

$$g(y)=\frac{1}{1+e^{-y}}=\frac{1}{1+e^{-(\theta_0+\theta_1x_1+\cdots\theta_nx_n)}}=\frac{1}{1+e^{-\theta Tx}} \quad (4)$$

在逻辑回归函数中用逻辑函数把线性回归的结果 $(-\infty,+\infty)$ 映射到 (0,1)，得到的结果类似一个概率值。上式中 x_i 表示给的表中的 14 个特征，y 代表玻璃表面的“风化”“未风化”，当 y 为 1 表示为“风化”，当 y 为 0 表示“未风化”时，这样我们可以进一步地把逻辑函数的值定义为风化的概率：

$$P(1|x)=g_\theta(x)=\frac{1}{1+e^{-y}}=\frac{1}{1+e^{-(\theta_0+\theta_1x_1+\cdots\theta_nx_n)}}=\frac{1}{1+e^{-\theta Tx}} \quad (5)$$

其表示未风化的概率为：

$$P(0|x)=1-g_\theta(x) \quad (6)$$

我们用极大似然数求解逻辑回归中的参数。其中 $y\in\{0,1,\}$。θ_i 为待求参数 [5]。

4.1.2 Spearman 和 Logistic 回归模型的求解

通过 MATLAB 求解，我们计算出变量之间的斯皮尔曼系数，得出表面风化与类型的相关系数为 0.3444 较大，颜色与类型的相关系数为 0.3733 较大。

通过 MATLAB 求解，得出表面风化与类型的显著性较高，类型和颜色的显著性较高，它们的相关性成立。

在样本数量大于 30 的情况下，我们可以通过构建统计量的方式进行假设检验，以下的统计量是符合正态分布的。结合玻璃的类型，我们分别对高钾类和铅钡类的风化前后变化差异进行描述性统计分析，假设检验。最终结果符合原假设，且高钾类和铅钡类风化前后主要化学成分二氧化硅差异比较大，高钾类风化后，二氧化硅含量变多了，铅钡类风化后二氧化硅含量变少了。

最后，我们通过给定的数据，判断出二氧化硅为主要成分，所以针对风化的玻璃文物数据做 Logistic 回归模型，由 MATLAB 求解出解析式为：

$$y=-41.449-0.90346x_1+14.029x_2+15.055x_3-0.57063x_4-10.601x_5+8.1087x_6+3.3293x_7+0.69565x_8+2.4075x_9+3.5724x_{10}+4.1014x_{11}-41.778x_{12}+14.209x_{13}-4.0347x_{14}$$

由模型分析结果得出 p-value 为 0.00575<0.05，则该模型符合，预测出其类别，将实际的成分与前面求得的各类的风化产物成分进行比较，从而得出风化前的化学成分含量。

4.2 问题二的模型建立与求解

4.2.1 建立 K-means 聚类模型

设有 N 个样品，每个样品测得 n 项指标 (变量)，原始资料阵为：

$$X=\begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_N \end{array}\begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix} \end{array}$$

其中 $x_{ij}(i=1,\cdots,N,j=1,\cdots,n)$ 为第 i 个样品的第 j 个指标的观测数据。第 i 个样品，X_i 为矩阵 X 的第 i 行所描述，所以任何两个样品 x_k 与 x_l 之间的相似性，可以通过第 k 行和第 l 行的相似程度来刻画；任何两个指标 x_k 与 x_l 之间的相似性，可以以通过第 k 列和第 l 列的相似程度来刻画。

对 N 个样品进行分类的方法，称为 Q 型聚类法，常用的统计量是用“距离”来表达。对应该题 N 为各类的监测数据，化学成分为指标。

1. 聚类模型的欧式距离。如果把 N 个样品（X 中的 N 个行）看成 p 维空间中的 N 个点，则两个样品间相似程度可用 p 维空间中地两点距离来度量。令 d_{ij} 表示样品 x_i 与 x_j 之间的距离。

当 q=2 时为欧氏距离：

$$d_{ij}(2)=(\Sigma_{a=1}^{P}(x_{ia}-x_{ja})^2)^{\frac{1}{2}} \quad (7)$$

计算任何两个样品 X_i 与 X_j，之间的距离 d_{ij}，其值越小表示两个样品接近程度越大，值越大表示两样品接近程度越小。如果把任何两个样品的距离都算出来后，可排成距离阵 D：

$$D=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix}$$

其中 $d_{11}=d_{22}=\cdots=d_{NN}=0$.$D$ 是一个实对称阵，所以只需计算上三角形部分或下三角形部分即可。根据 D 可对 N 个点进行分类，距离近的点归为一类，距离远的点归为不同的类。

2. 聚类模型的相关系数。通常所说的相关系数，

一般是指变量间的相关系数，作为刻画样品间的相关关系也可类似给出定义，即第 i 个样品与第 j 个样品之间的相关系数定义为：

$$r_{ij}=\frac{\sum_{a=1}^{P}(x_{ia}-\bar{x}_i)(x_{ja}-\bar{x}_j)}{\sqrt{\sum_{a=1}^{P}(x_{ia}-\bar{x}_i)^2\cdot\sum_{a=1}^{P}(x_{ja}-\bar{x}_j)^2}},\ -1\le r_{ij}\le 1 \qquad (8)$$

于是 $R=(r_{ij})$，其中 $r_{11}=r_{22}=\cdots=r_{NN}=1$，可根据 R 对 N 个样品进行分类 [6]。

3.K-means 算法。K-means 算法是最常用的聚类算法，主要思想是：在给定 K 值和 K 个初始类簇中心点的情况下，把每个点（亦即数据记录）分到离其最近的类簇中心点所代表的类簇中，所有点分配完毕之后，根据一个类簇内的所有点重新计算该类簇的中心点（取平均值），然后再迭代地进行分配点和更新类簇中心点的步骤，直至类簇中心点的变化很小，或者达到指定的迭代次数。

4.2.2 K-means 聚类模型求解

具体的划分方法及划分结果：

我们采用聚类分析模型中的 K-means 算法，对于每个类别所有的化学成分对其进行亚类划分。通过 MATLAB 求解得：

高钾类分成亚类划分大致可分为三类：

第一类：二氧化硅 (SiO_2)。

第二类：氧化钾 (K_2O)。

第三类：氧化钠 (Na_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)。

铅钡类成分亚类划分大致主要分为三类：

第一类：二氧化硅 (SiO_2)。

第二类：氧化铅 (PbO)。

第三类：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)。

对分类结果的合理性和敏感性进行分析，通过 MATLAB 求解得出：高钾类和铅钡类化学成分的分类情况与实际情况相吻合，主要成分分类中，二氧化硅都占为一类，该结果比较合理；铅钡类比高钾类的分类多一类，若变动某化学成分含量，则结果可能不准确。

5 模型评价

5.1 模型的优点

1. 建立的 Logistic 回归模型以及判别分析模型可以更好地判别出未知数据的所属特征，简单易懂，有较强的数学基础，且易于应用于现实生活中。

2. 建立的主成分分析模型，可以更好地看出变量之间的相关性，把复杂的数据综合化，使其尽可能地反映原来的信息，降低了复杂性。

3. 建立的聚类分析可以把一堆成分进行进一步的分类，使其更好地展现成分之间的关系。

5.2 模型的缺点

1. 在问题一中， 斯皮尔曼系数对于变量之间的相关性分析不够严谨。

2. 在问题二中，我们用的是聚类分析中的最短距离法，还可以用更好的方法来对主要成分进行亚分类，对结果的敏感性分析不够完善。

6 模型的应用及推广

6.1 模型的改进

在问题一中， 在变量分析中，其实我们可以用卡方检验来分析不同变量之间的相关性。可以用更多的数据来对建立的逻辑回归的模型进行验证，看它是否更贴合实际。

6.2 模型的推广

1. 判别分析与回归分析相似，可用于确定哪些预测变量与因变量相关，并在给定预测变量的某些值的情况下预测因变量的值。在实际生活中，判别分析也被广泛用于预测事物的类别归属。企业营销中，营销人员可通过已有的客户特征数，预测当前的消费者属于哪种类型的顾客，并根据其特点有针对性地采取有效的营销手段，或是根据各成分含量指标，判断特征等。判别分析还可与聚类分析结合使用，如同本文一样。比如，银行确认一些用户的资格之前，可通过此方法判断申请人是否具有良好的信用风险。

2. 聚类用于基于模式识别过程将数据划分为不相交的组；在生物学中聚类是遗传学和分类学的重要工具，有助于理解生物和灭绝生物的进化。还有建立推荐系统、社交媒体网络分析、土地利用分类中的空间分析等。

参考文献：

[1] 马瑞民，姚立飞．回归分析在数学建模中的应用——基于上海世博会参观人数的预测分析模型 [J]. 教育探索，2011,12(04):48-54.

[2] 司守奎．数学建模算法与应用 [M]. 北京：国防工业出版社，2011:595-601.

[3] 林明海．对主成分分析法运用中的十个问题的解析 [J]. 统计与决策，2007(16):16-18.

[4] 楼建华．数学建模与数学实验 [M]. 黑龙江高教研究，2003(03):126-127.

[5] 邢航．回归分析中建立数学模型的方法及其应用 [J]. 职大学报，2008,16(04):68-70.

[6] 阎慈琳．关于用主成分分析做综合评价的若干问题 [J]. 数理统计与管理，1998,17(02):22-25.