基于机器学习与关联网络的古代玻璃成分分析

摘要

古代玻璃的化学成分是鉴别其类别与产地的重要依据，然而风化作用会改变原始成分，为定量分析带来挑战。本文基于一批古代玻璃制品的化学成分数据，建立了一套多层次的数学模型，以解决风化影响下的玻璃分类、成分预测与内在关联分析问题。研究综合运用了统计检验、机器学习、智能优化算法与复杂网络理论，为通过化学成分数据理解古代玻璃制品提供了系统性的分析方案。

对于问题一，我们采用**卡方检验**分析了表面风化与玻璃类型、纹饰及颜色的关联性，确认了它们之间存在统计学关联。通过对比风化前后样本的化学成分分布，发现风化主要导致高钾玻璃中的氧化钾与铅钡玻璃中的氧化铅和氧化钡流失。为预测风化前的成分，本文引入**地球化学领域**的**质量平衡分析理论**，构建了基于风化系数的预测模型，该模型在缺乏成对样本的情况下，能够有效恢复文物的原始化学成分。

对于问题二，为研究两大类玻璃的分类规律，我们分别构建了**线性与非线性支持向量机模型**。其中，非线性模型获得了 97.41% 的交叉验证准确率，我们引入**博弈论中的 SHAP 值**对其进行解释，而线性模型的结果则直接验证了氧化铅与氧化钾是分类的核心指标。在亚类划分中，我们依据**变异系数**筛选特征，并结合**层次聚类**与**轮廓系数**，确定铅钡玻璃存在两个亚类，高钾玻璃存在五个亚类。划分结果显示铅钡玻璃可分为高铅钡助熔剂型与高硅基质型，高钾玻璃的亚类则在多种成分上表现出不同特征。

对于问题三，我们构建了基于**改进遗传算法**优化的**支持向量机分类器 IGA-SVM**，以鉴别未知类别文物的所属类型。该模型通过智能化全局寻优确定最优超参数组合，避免了传统模型选择的局限性。应用此模型，我们完成了对全部未知样本的分类。为确保结果的可靠性，我们设计了基于**蒙特卡洛模拟**的数据扰动与基于**参数网格搜索**的模型扰动双重灵敏度分析，验证了分类结果对于数据测量误差和模型参数选择的高度稳健性。

对于问题四，为探寻不同类别玻璃配方的内在结构差异，我们构建了**化学成分关联网络**。为消除成分数据总和恒定带来的虚假相关，我们首先采用化学计量学中的**中心化对数比变换**对数据进行处理。随后利用**图套索算法**计算偏相关系数以构建网络，并应用**鲁汶算法**进行社群发现。网络分析结果表明，铅钡玻璃与高钾玻璃的化学成分关联结构存在显著不同，前者以二氧化硅和氧化铅为核心形成紧密社群，后者则结构较为分散，这些差异反映了两者在原料与烧制工艺上的区别。网络分析结果表明，铅钡玻璃与高钾玻璃的化学成分关联结构存在显著不同，前者以二氧化硅和氧化铅为核心形成紧密社群，后者则结构较为分散，这些差异反映了两者在原料与烧制工艺上的区别。

关键字： 卡方检验　地球化学　质量平衡分析理论　SHAP 值　支持向量机　中心化对数比　图套索算法　鲁汶算法

一、 问题背景

NIPT（无创产前检测）是一种现代产前筛查技术，它通过分析孕妇外周血中来自胎儿的游离 DNA 片段，评估胎儿患有染色体非整倍体疾病的风险 [1]。临床上重点关注三类由染色体数目异常引起的病症，分别为唐氏综合征、爱德华氏综合征与帕陶氏综合征，它们分别对应胎儿 21 号、18 号与 13 号染色体的异常。该检测的有效性依赖于胎儿性染色体浓度：男胎 Y 染色体浓度需达 4%，女胎 X 染色体浓度则需无异常。

检测时点的选择是影响检测有效性的一个重要变量。若检测时间过早，胎儿游离 DNA 浓度可能不足，特别是男胎的 Y 染色体浓度未能达到 4% 的标准，这将导致检测结果无法保证准确性。反之，若检测时间过晚，则可能会延误后续的临床决策。根据临床经验，孕 12 周以内发现异常属于低风险，孕 13 至 27 周发现属于高风险，而孕 28 周以后发现则为极高风险。

临床实践常依据身体质量指数 BMI 对孕妇分组并推荐统一检测时点，但该方法因忽略年龄、孕情等个体差异，并非最优。因此，需建立数学模型对孕妇进行合理划分，确定各群体的最佳检测时点，以平衡检测可靠性并降低延迟发现的风险。对于女胎，其异常判定则需综合多项生物信息学数据，建立有效的判别规则。

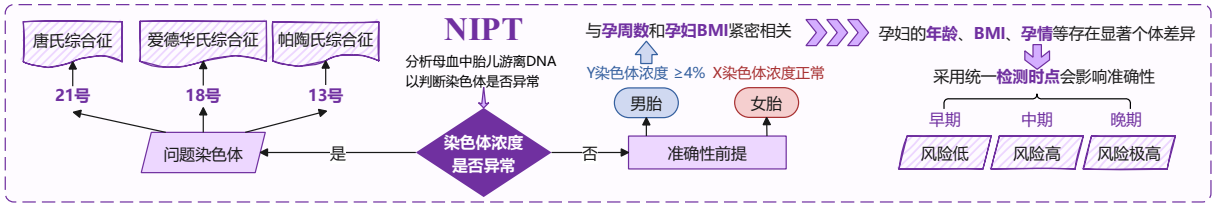

图 **1**　问题背景示意图

二、 问题重述

问题一：分析胎儿染色体浓度与孕妇孕周数和身体质量指数等指标的相互关系，并建立相应的关系模型。

问题二：针对男胎孕妇，依据其身体质量指数进行分组，确定各组的最佳无创产前检测时点以使潜在风险最小，并分析检测误差造成的影响。

问题三：考虑身高，体重，年龄等多重因素与检测误差，根据男胎孕妇的身体质量指数进行分组，并给出每组的最佳无创产前检测时点，以最小化孕妇潜在风险并顾及 Y 染色体浓度达标比例。

问题四：建立女胎染色体异常的判定方法，运用 Z 值，GC 含量，读段数及身体质量指数等多种因素，以识别 21 号，18 号和 13 号染色体的非整倍体异常。

三、 问题分析（需修改）

针对问题一，该问题涉及两个层面，其一是定性关系的判断，其二是定量规律的分析与预测。表面风化与玻璃类型、纹饰、颜色均为分类变量，它们之间的关联性分析适合采用非参数的卡方检验。风化对化学成分含量的影响则需要分组进行统计比较，通过观察含量分布的变化来识别规律。由于缺乏同一文物风化前后的成对数据，直接建立回归预测模型存在困难，因此分析的重点在于依据风化机理，构建一个基于平均变化率的风化系数模型，用以反推风化前的化学成分。

针对问题二，该问题需要从监督学习与无监督学习两个角度展开。高钾与铅钡玻璃的分类规律研究是一个典型的二分类问题，可以基于探索性数据分析发现的关键化学成分，构建支持向量机等分类模型，并对模型进行解释。

针对问题三，该问题是分类模型的实际应用与可靠性验证。分析的核心在于构建一个泛化能力强且稳健的分类器。这需要比较多种模型架构，并论证选择支持向量机的合理性。为使模型性能最优化，需要采用改进遗传算法等智能优化算法进行超参数寻优。最终的鉴别结果需要通过数据扰动和模型参数敏感性分析的双重检验，以证明其结论并非偶然，而是具有高度的可靠性。

针对问题四，该问题要求探寻不同类别玻璃内部化学成分的关联结构。由于玻璃成分数据具有总和恒定的特性，直接计算相关系数会产生误导，因此必须首先采用中心化对数比变换来消除此统计约束。直接计算相关系数会产生误导，因此必须首先采用中心化对数比变换来消除此统计约束。

全文框架如图 2所示：

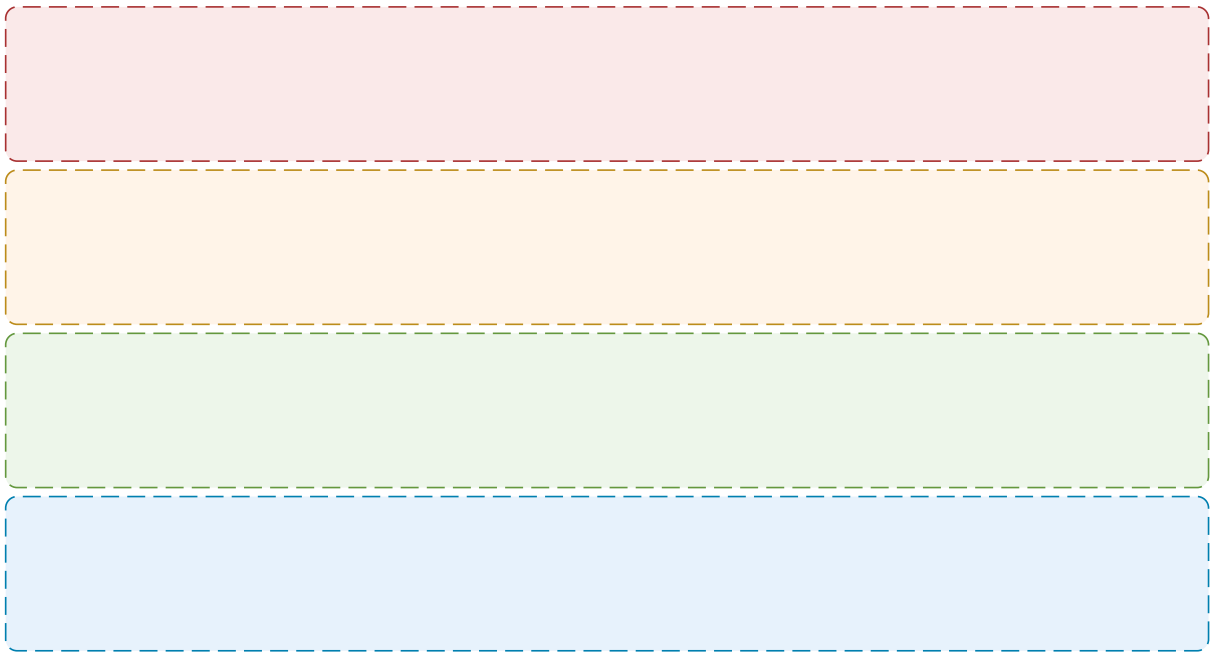

图 **2**　全文框架示意图

四、 模型假设（需修改）

为保证模型构建的合理性与分析过程的严谨性，我们提出以下基本假设。

1. 所有用于分析的数据均为成分总和在百分之八十五至一百零五区间的有效数据。
2. 数据中的空白项表示该化学成分未被检测到其含量可视为零。
3. 同一类型玻璃的风化过程与化学成分变化规律具有统计上的一致性。
4. 文物的化学成分组合能够有效地区分高钾玻璃与铅钡玻璃两大类别。
5. 同一文物多个采样点的平均化学成分可以代表该文物的整体特征用于亚类划分。
6. 化学成分间的偏相关关系能够反映其在玻璃配方或制作工艺中的直接关联。
7. 化学成分间的偏相关关系能够反映其在玻璃配方或制作工艺中的直接关联。

五、 符号说明（需修改）

本文后续章节中所使用的主要数学符号及其说明如表所示。

表 **1**　符号说明表

符号	说明
$k_{t,j}$	t 类玻璃中 j 成分的风化系数
$\bar{C}_{t,j,\text{weathered}}$	t 类玻璃风化样本中 j 成分的平均含量
$\bar{C}_{t,j,\text{unweathered}}$	t 类玻璃未风化样本中 j 成分的平均含量
$C'_{\text{unweathered},j}$	某一样本风化前 j 成分的预测含量
$\boldsymbol{w}, b$	支持向量机分类超平面的法向量与位移项
C	支持向量机的正则化系数
γ	径向基核函数的参数
ξ_i	支持向量机的松弛变量
CV	变异系数，用于衡量数据的相对离散程度
$\mathbf{clr}(\boldsymbol{x})$	对样本向量 $\boldsymbol{x}$ 进行中心化对数比变换
S	样本协方差矩阵
Θ	精度矩阵，即逆协方差矩阵
λ	图套索算法的正则化参数
$\rho_{ij\cdot\text{rest}}$	变量 i 和 j 之间的偏相关系数
S	样本协方差矩阵
Θ	精度矩阵，即逆协方差矩阵
λ	图套索算法的正则化参数
$\rho_{ij\cdot\text{rest}}$	变量 i 和 j 之间的偏相关系数

参考文献

[1] 张亮亮, 卓召振, 黄盛文, 任凌雁, 牟静, and 匡颖. 贵州省多中心 16798 例 nipt-plus 结果回顾性分析. 贵州医药, 49(08):1296–1299, 2025.