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 [2]。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%[3]，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，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 [4]，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。

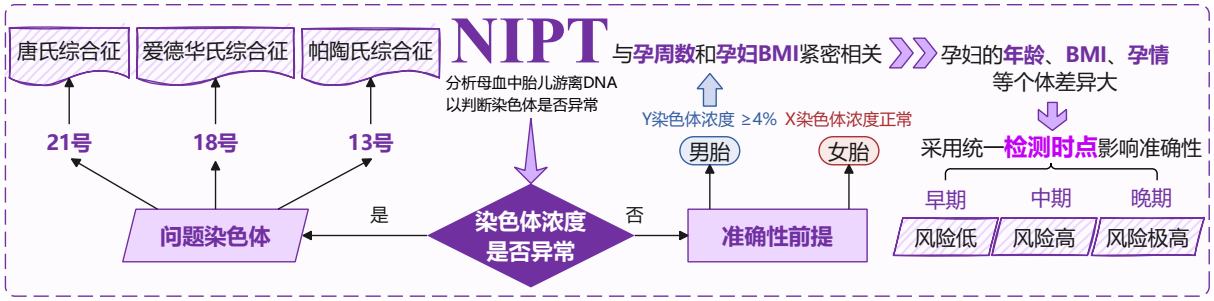

图 **1**　问题背景示意图

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：

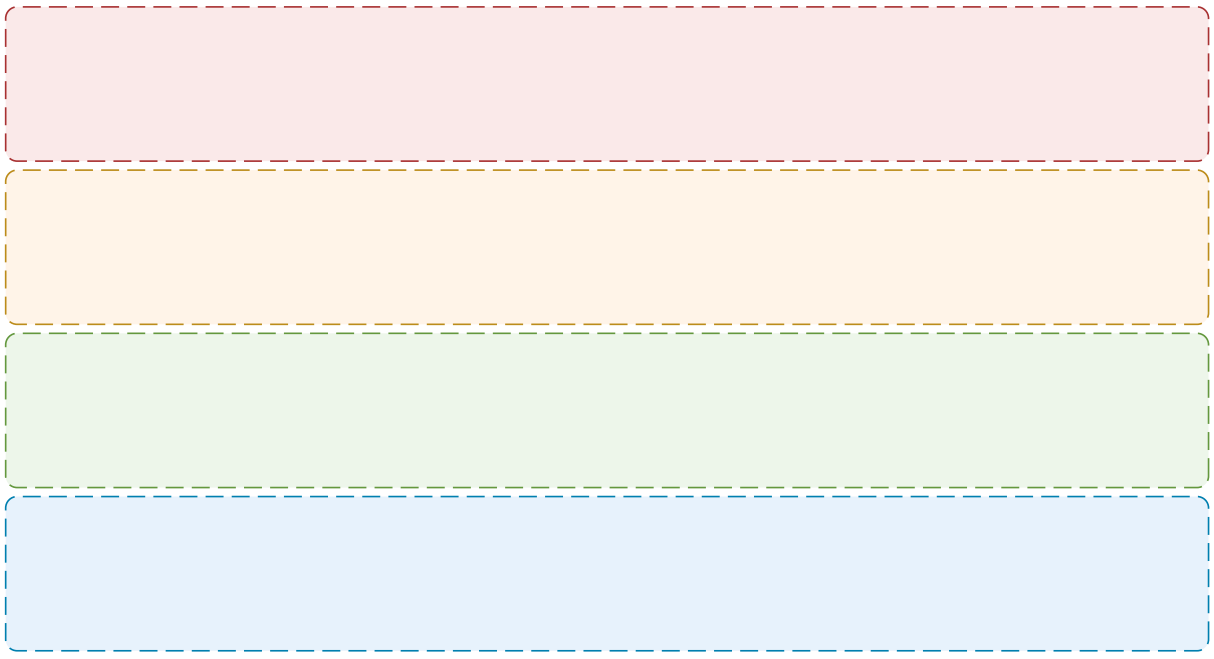

图 **2**　全文框架示意图

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。

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
$\text{clr}(\boldsymbol{x})$	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

六、 模型准备

为了后续模型的建立与求解，本文进行如下模型准备，即数据清洗、数据量化等。具体机制如**??**所示。

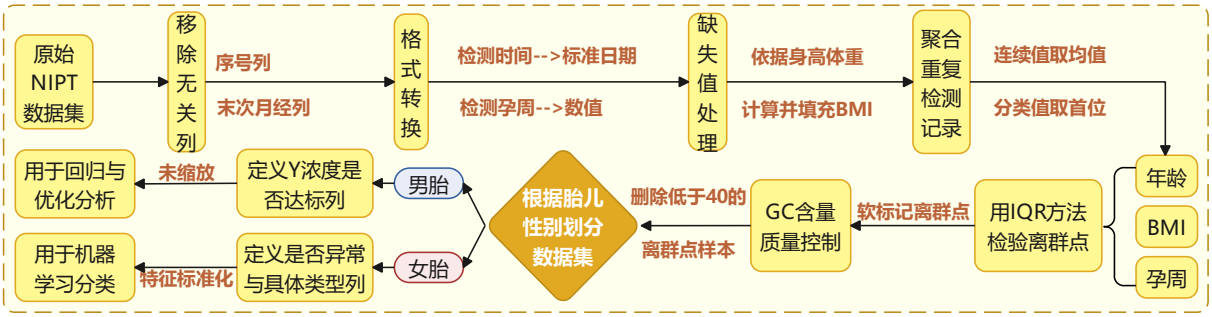

图 **3**　模型准备机制

6.1 数据清洗

6.1.1 缺失值处理与数据整合

为简化数据集，我们移除了与建模任务无直接关联的序号列。随后，我们对数据集中各变量进行缺失值统计，发现末次月经时间列存在部分数据缺失。末次月经时间的主要功能为推算孕周，而数据集中已包含更为直接的检测孕周列，故将其移除。

为使原始特征能够用于定量分析，我们需要将其转化为标准的数值格式。对于纯数字格式的检测时间列，我们将其转换为标准的日期时间格式；对于文本格式的检测孕周列，我们转换为可计算的浮点数值，例如十二周加三天转化为 12.43 周。对于孕妇 BMI 指标列中的缺失值，我们利用同一条记录的身高与体重数据，依据身体质量指数的官方计算公式进行填充，该公式如下所示

$$BMI = \frac{Weight(kg)}{Height(m)^2} \tag{1}$$

与使用均值或中位数等统计量填充相比，利用已有数据进行计算能最大限度地保持数据的真实性。

完成基础清洗与计算后，数据集内核心特征的整体分布如图 4所示。该图展示了孕妇年龄，孕妇 BMI 与检测孕周三个变量的分布状况。从图中可以看出样本的年龄主要集中在 25 至 35 岁之间，孕妇 BMI 分布的峰值位于 30 附近，证实了样本多为高 BMI 的地区特征。检测孕周则在 12 至 25 周之间呈现多个峰值，表明了检测时间点的集中性。

题目中提到存在对同一孕妇进行重复检测的情况，这导致了数据冗余。为消除此冗余，我们以孕妇代码与检测抽血次数为唯一标识，对同一抽血样本的多次检测记录进行数据整合。整合时，对于连续性测量指标，取其算术平均值以减小单次测量的随机误差影响；对于分类性状态指标，则保留其首次出现的值。

图 5展示了数据整合前后的样本数量变化。如左图所示，处理前单个孕妇的样本记录数最高可达 9 条，多数孕妇具有 1 至 6 条记录。经过整合处理后，转变为每次抽血对应唯一记录的形式。如右图所示，单个孕妇的有效检测次数减少至 1 至 5 次，大部分孕妇只有 4 次有效检测，从而消除了冗余信息。

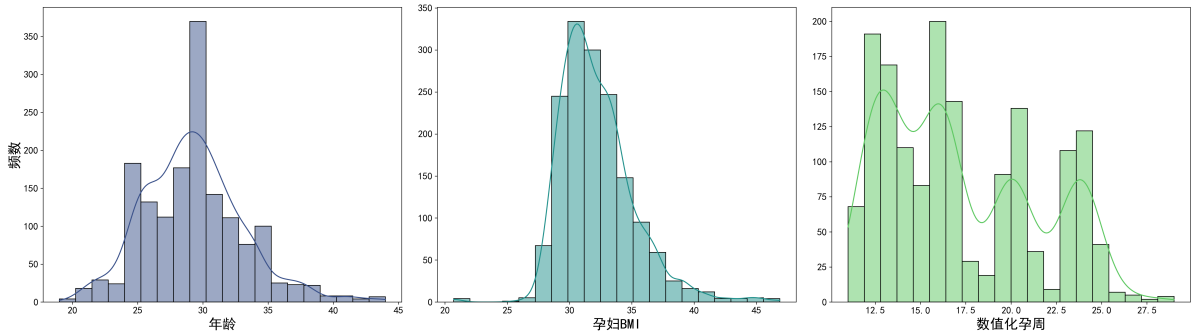

图 **4**　核心特征分布

根据题目说明，原始数据中存在同一孕妇多次采血多次检测或一次采血多次检测的情况，这种情况造成了数据冗余。为消除此冗余并为每个样本建立唯一的检测记录，本文以孕妇代码与检测抽血次数的组合为标识，对重复的检测条目进行聚合处理。在聚合过程中，对于连续型测量指标，采用算术平均值以平滑单次测量的随机波动。对于分类性质的状态指标，则保留其初次记录的值。

该聚合过程的效果通过图 5中的前后对比得以展示。左图显示在处理前，单个孕妇对应的所有样本的所有检测记录数从 1 条至 9 条不等，其中多数孕妇具有 3 至 6 条记录。经过聚合处理后，如右图所示，每位孕妇的有效检测次数范围缩小至 1 至 5 次，其中具有 4 次有效检测的孕妇数量最多。

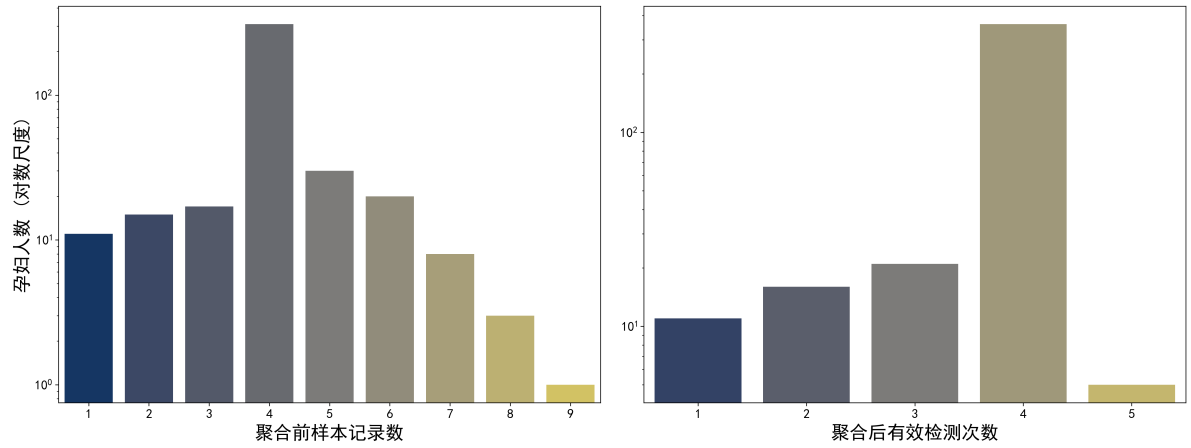

图 **5**　数据整合效果对比

6.1.2 异常值处理

我们首先针对年龄，孕妇 BMI 与孕周这三个核心特征执行四分位距法检验。对于识别出的离群点，本文采用软标记方法，即在数据集中新增一列用以标记样本是否为离群点，而不直接删除样本记录。离群点可能代表了真实存在的极端生理状况或数据录入错误，采用软标记策略为后续建模提供了灵活性。图 6展示了离群点分析的过程，图中

每个子图的小提琴形状表示了数据的分布，红色的圆点则标记出被标准四分位距准则判定为离群点的样本。图中可见，年龄的离群点主要分布在 40 岁以上的高龄区间，孕妇 BMI 的离群点则同时出现在低于 20 的低值区域与高于 40 的高值区域，而孕周的离群点相对较少。

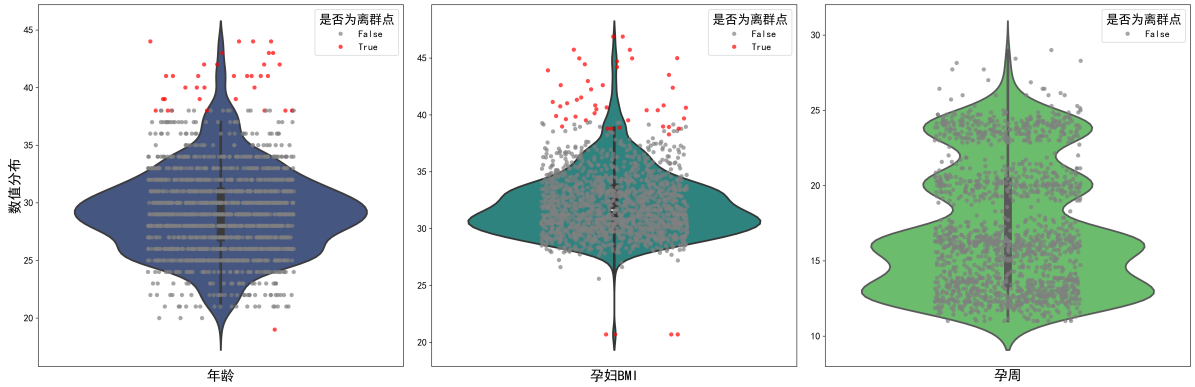

图 **6**　**IQR** 离群点分析

其次我们对 GC 含量进行质量控制。GC 含量是衡量测序质量的重要指标。处理过程结合了统计学检验与业务规则，先对 GC 含量列执行标准的 1.5 倍四分位距检验以识别出所有统计学上的离群点。然后，仅将这些离群点中数值低于 40% 的样本从数据集中直接删除。此方法能够剔除那些不仅在统计上异常，且低于业务经验常规下限的低质量测序样本。图 7展示了处理后的数据，其中直方图展示了 GC 含量的整体分布形态近似正态分布，频数在 0.400 处达到峰值，绝大多数样本的 GC 含量分布在 0.395 至 0.405 之间，分布集中，数据质量较高。

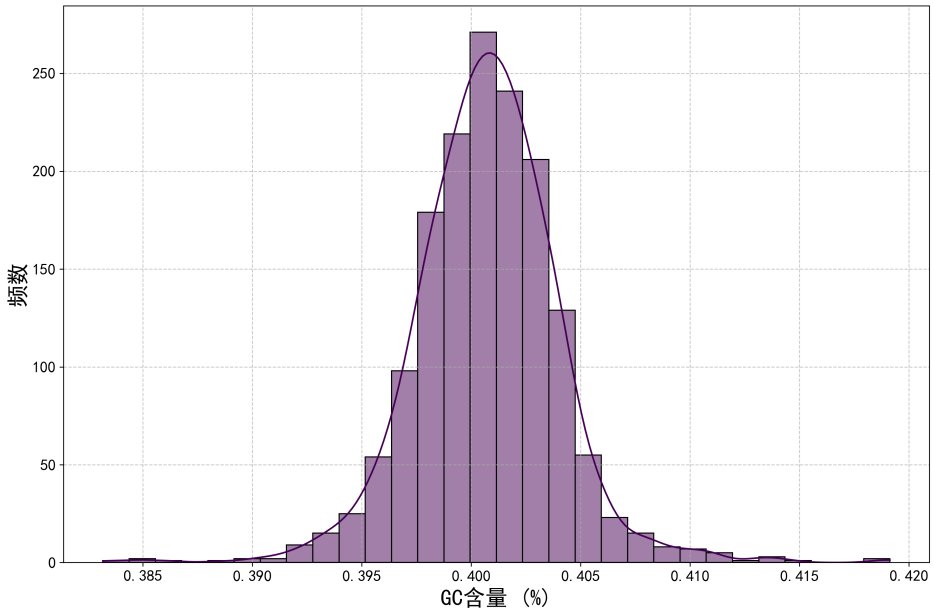

图 **7**　**GC** 含量分布探索

6.2 数据量化

为了便于后续相关数学模型的建立与求解，增强可理解性，我们进一步进行了数据量化。

针对不同问题, 我们构建了相应的目标变量。对于男胎数据，依据题目定义的 4% 阈值，我们创建了“Y 浓度是否达标”的二元分类变量。对于女胎数据，则基于染色体的非整倍体列中空白即为无异常的定义，创建了“是否异常”以及各类异常的分类目标。图 8展示了女胎样本中正常与各类异常样本的数量分布。图中可见正常样本数量为 488 例，而 18 号，13 号与 21 号染色体三体综合征的样本数量分别为 45 例，22 例与 13 例。正常样本远多于异常样本，表明这是一个数据不平衡问题。

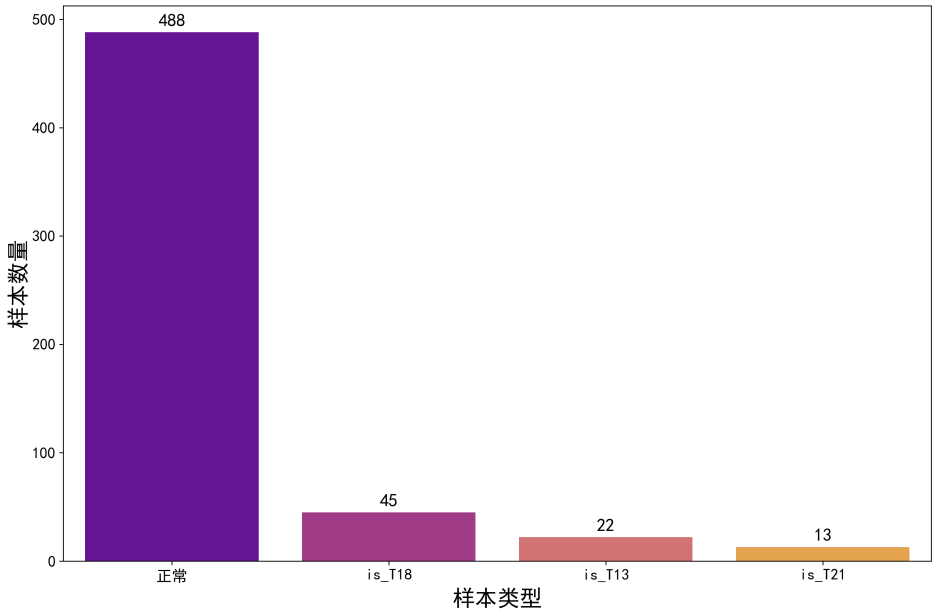

图 **8**　染色体异常类型分布

考虑到问题四需要使用机器学习分类模型，我们仅针对女胎数据的所有数值型预测特征应用了标准化处理。标准化将所有特征转换到同一尺度，其计算方法如下

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

其中 x 为原始值，μ 为特征均值，σ 为特征标准差。此操作避免了模型的训练过程被原始读段数等具有极大数值范围的特征所主导，使模型能够更稳定地学习所有特征的贡献。男胎数据因主要用于回归和优化分析，保留其原始数值的物理意义更为重要，故不进行缩放。图 9展示了女胎数据中所有数值型预测变量在缩放前后的分布对比。每个子图呈现了同一特征在处理前后的分布形态，处理前各变量的分布中心与尺度范围各不相同，而经过标准化处理后，所有变量的分布均被调整至以 0 为中心，具有相似的尺度，这为后续机器学习模型的训练提供了同质化的输入。

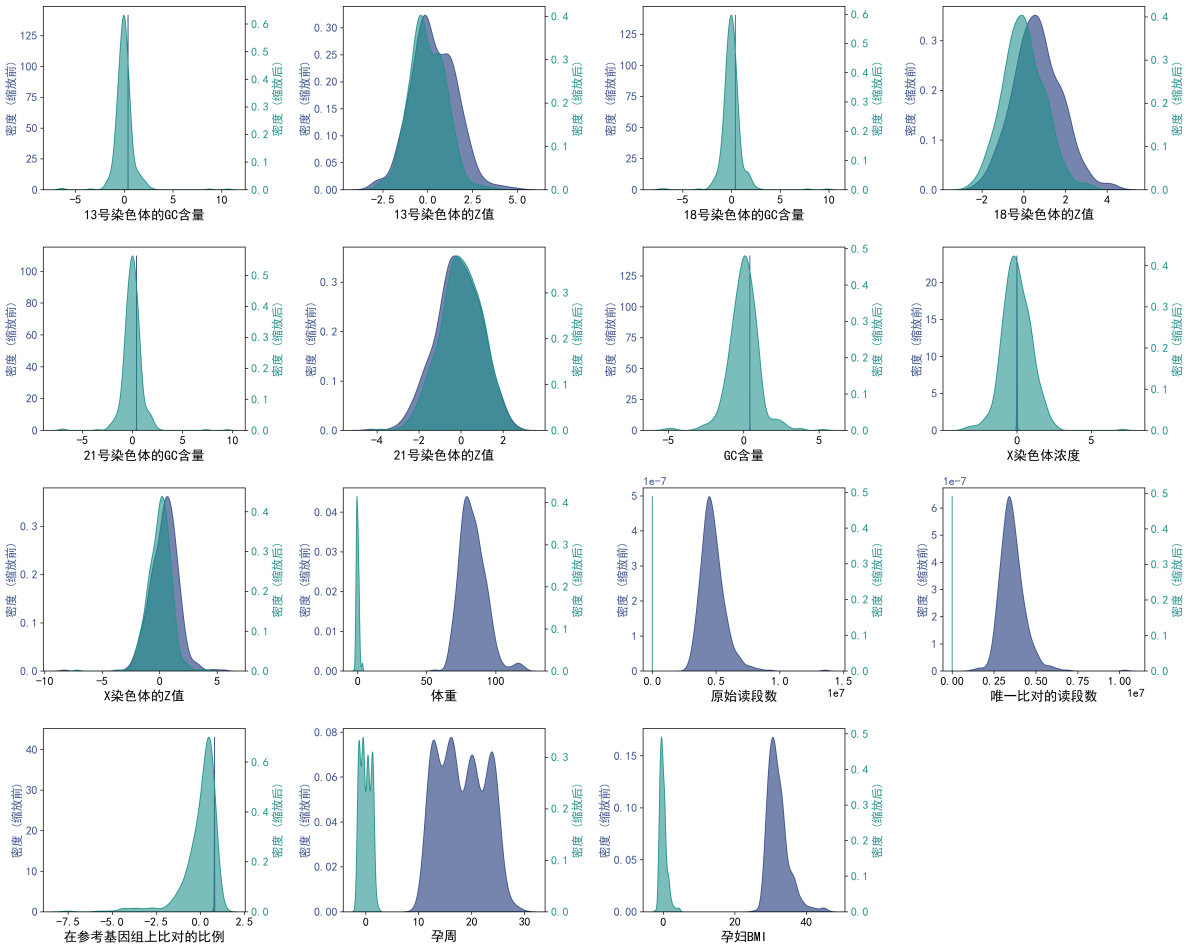

图 **9**　所有数值特征缩放效果对比

此外，我们还构建了部分衍生特征。我们依据题目对风险的定义，创建了与孕周对应的风险等级特征。同时，为进行时序分析，还计算了每次检测距首次检测的天数。

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.

[2] 张鹏, 莫伟英, 蒙明慧, and 张红燕. 无创产前检测对性染色体非整倍体检出情况的影响及相关伦理思考. 中国临床新医学, 18(06):690–695, 2025.

[3] 钟佳通, 胡亮, 温丽娟, 李双武, 陈晓杭, 裴元元, and 刘维强. 双胎妊娠中无创产前检测的胎儿游离 dna 浓度特征及检测失败原因分析. 中国产前诊断杂志 (电子版), 15(03):25–30, 2023.

[4] 蒋丽雅, 卢劭侃, 杜佳恩, 陈阔阔, 张思恬, 伍圣洁, 柳天玉, and 叶佳昊. 无创产前检测技术的发展与应用. 临床医学研究与实践, 10(23):191–194, 2025.