

NIPT 时点选择与胎儿异常判定的优化研究

杨娟¹ 何佳林¹ 徐清¹何家文^{2*}1. 南宁学院信息工程学院, 南宁,
5302002. 南宁学院人工智能学院, 南宁,
530200

摘要 本研究围绕无创产前检测 (NIPT) 在临床实践中的两大核心问题——最佳检测时点选择与胎儿染色体异常判定展开。针对男胎孕妇, 系统分析了胎儿 Y 染色体浓度与孕妇体质指数 (BMI)、孕周等因素的数量关系, 建立了多元回归模型。基于此, 运用生存分析与聚类算法, 将孕妇按 BMI 进行科学分组, 以临床风险最小化为目标, 确定了不同 BMI 区间 (低、中、高) 的最佳 NIPT 检测时点 (分别为 11.5 周、11.0 周、13.0 周)。进一步考虑身高、体重、年龄等多因素影响, 构建了综合评估与多目标优化模型, 验证了模型的鲁棒性。针对缺乏 Y 染色体直接证据的女胎, 创新性地结合 21、18、13 号染色体的 Z 值、GC 含量、读段数等多维指标, 建立了逻辑回归判定模型, 并设计了分层临床决策规则, 在验证集上获得了超过 80% 的准确率。本研究为 NIPT 的个性化、精准化临床应用提供了系统的建模方法与决策支持。

关键字 无创产前检测, 最佳检测时点, 胎儿异常判定, 生存分析, 聚类算法, 逻辑回归

Optimal Timing Selection and Fetal Abnormality Determination in Non-Invasive Prenatal Testing

YangJuan¹ HeJiaLin¹ XuQing¹HeJiawen^{2*}1.College of Information Engineering
Nanning University
Nanning, 530200, China;2.College of Artificial Intelligence
Nanning University
Nanning, 530200, China

Abstract—This research focuses on two core issues in the clinical practice of Non-Invasive Prenatal Testing (NIPT): optimal testing timing selection and fetal chromosomal abnormality determination. For male-bearing pregnancies, the quantitative relationship between fetal Y-chromosome concentration and maternal factors such as Body mass index (BMI) and gestational weeks was systematically analyzed, establishing a multiple regression model. Based on this, survival analysis and clustering algorithms were employed to scientifically group pregnant women according to BMI. Aiming to minimize clinical risk, the optimal NIPT timing for different BMI ranges (low, medium, high) was determined (11.5 weeks, 11.0 weeks, and 13.0 weeks, respectively). Further considering the influence of multiple factors such as height, weight, and age, a comprehensive evaluation and multi-objective optimization model was constructed, and its robustness was verified. For female fetuses lacking direct Y-chromosome evidence, an innovative logistic regression classification model was established by integrating multidimensional indicators including Z-scores of chromosomes 21, 18, and 13, GC content, and read counts. A hierarchical clinical decision rule was also designed, achieving an accuracy rate of over 80% on the validation set. This study provides systematic modeling methods and decision support for the personalized and precise clinical application of NIPT.

Keywords—Non-invasive prenatal testing, Optimal testing timing, Fetal abnormality determination, Survival analysis, Clustering algorithm, Logistic regression

1 引言

近年来, 随着我国产前诊断技术的快速发展与临床应用需求的持续提升, 无创产前检测 (Non-

*通讯作者: 何家文 229043733@qq.com。

Invasive Prenatal Testing, NIPT) 已成为出生缺陷防控体系中的核心技术手段。在实际生产与临床应用中, NIPT 检测面临一系列亟待解决的关键问题: 一是胎儿游离 脱氧核糖核酸 (Deoxyribonucleic Acid, DNA) 浓度受孕周、身体质量指数 (Body Mass Index, BMI) 等多因素影响, 传统统一化检测方案难以适配不同人群特征, 易导致样本检测失败、结果不稳定等问题; 二是染色体异常判定多依赖单一阈值, 在低胎儿游离 DNA 浓度样本中稳健性不足, 漏检风险较高; 三是个性化检测时机选择与多因素综合判定模型构建缺乏系统性研究, 筛查资源配置效率低下; 四是女胎样本分类存在样本不平衡、特征关系复杂等问题, 模型可解释性不足, 难以有效支撑临床决策。

国内外学者在产前筛查优化与建模领域已开展大量研究。庞宇^[4]系统分析了 NIPT 在胎儿染色体疾病筛查中的临床应用价值; 陈韵晶、曾如明、崔健等^[7]围绕聚类算法进行研究与改进, 为基于个体特征的人群分层分析奠定了方法学基础; 梁凯毓、张静仪、林凯乐等^[5,6]在非线性回归、高斯混合模型等方向的探索, 也为本研究的特征构建与模型优化提供了重要参考。尽管现有研究已取得一定成果, 但针对低浓度样本判定、个性化检测时机规划及女胎样本可解释分类等核心问题的系统性解决方案仍较为缺乏, 难以满足临床与产业的实际应用需求。

对于检测的可靠性, 本文根据已有数据, 分析求解以下的问题^[7]:

问题 1: 分析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标的相关特性, 给出相应的关系模型, 并检验其显著性。

问题 2: 临床证明, 男胎孕妇的 BMI 是影响胎儿 Y 染色体浓度的最早达标时间 (即浓度达到或超过 4% 的最早时间) 的主要因素。对男胎孕妇的 BMI 进行合理分组, 给出每组的 BMI 区间和最佳 NIPT 时点, 使得孕妇可能的潜在风险最小, 并分析检测误差对结果的影响。

问题 3: 男胎 Y 染色体浓度达标时间受多种因素 (身高、体重、年龄等) 的影响, 综合考虑这些因素、检测误差和胎儿的 Y 染色体浓度达标比例 (即浓度达到或超过 4% 的比例), 根据男胎孕妇的 BMI, 给出合理分组以及每组的最佳 NIPT 时点, 使得孕妇

潜在风险最小, 并分析检测误差对结果的影响。

问题 4: 由于孕妇和女胎都不携带 Y 染色体, 重要的是如何判定女胎是否异常。以女胎孕妇的 21 号、18 号和 13 号染色体非整倍体 (AB 列) 为判定结果, 综合考虑 X 染色体及上述染色体的 Z 值、鸟嘌呤-胞嘧啶含量 (Guanine-Cytosine, GC 含量)、读段数及相关比例、BMI 等因素, 给出女胎异常的判定方法。

本研究立足于 NIPT 生产与临床应用的实际痛点, 针对上述四大核心问题, 构建多维度优化模型, 综合运用特征工程、机器学习与概率校准等技术手段, 提出基于成本效益分析的个性化检测优化策略, 确保检测决策的科学性与合理性。本研究的开展不仅能有效提升 NIPT 检测的准确性与效率, 为临床决策提供科学依据, 还能为同类产前筛查技术的优化提供参考, 对推动出生缺陷防控工作的规范化、个体化发展具有重要的理论与实践意义。

2 问题的建立与求解

2.1 问题1的模型以及求解

针对问题 1, 使用的原始数据来源于某大型产前诊断中心, 包含数千例 NIPT 检测记录。从“男胎检测数据”工作表中提取了孕周转换为总天数、孕妇 BMI、Y 染色体浓度等关键字段。由于样本数据存在异常值, 数据筛选阶段采用、四分位距 (Interquartile Range, IQR) 法确保后续分析可靠, 筛选目标变量需重点关注 NIPT 检测核心指标并特别留意孕妇 BMI 体重指数的异常值筛查, 临床资料显示 BMI 超过 40 kg/m² 会显著影响检测结果准确性需严格控制极端值; 检测孕周限定在 10-25 周的临床适宜区间, 超出此区间的数据需做标记; Y 染色体浓度作为反映胎儿游离 DNA 比例的指标, 其数值过低 < 0.01 或异常偏高 > 0.5 可能源于检测误差。将各变量按升序排列, 计算出第一四分位数 Q1, 即 25% 分位数和第三四分位数 Q3, 即 75% 分位数其次使用公式 $IQR = Q3 - Q1$ 计算四分位距, 结合公式的计算得数据结果清晰明确异常值范围为 $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$; 根据异常值可判断结论是超出此范围的数值判定为异常值。

针对 BMI > 47 kg/m² 与孕周 < 10 周或 > 25 周的极端数据采用删除法剔除, 异常数据不符合临床检测标准; 其次 Y 染色体浓度的异常值用中位数替换,

减轻样本量减少对统计效能的影响。数据处理完成后通过箱线图进行可视化验证，数据离散程度有所降低 Y 染色体浓度标准差从 0.12 降至 0.08，降低了不合格率，增强了达标率。

为了寻找胎儿 Y 染色体浓度与孕周和 BMI 的内在逻辑关系，本研究构建多类回归模型展开分析。首先以多元线性回归作为基准模型 $Y=\beta_0+\beta_1\times\text{GestWeek}+\beta_2\times\text{BMI}+\epsilon$ ，用于初步探索变量间的线性关系；在此基础上进一步构建多项式回归模型，通过加入孕周和 BMI 的二次项 $Y=\beta_0+\beta_1\text{GestWeek}$

$+\beta_2\text{BMI}+\beta_3\text{GestWeek}^2+\beta_4\text{BMI}^2+\epsilon$ ，探索变量间可能存在的非线性关联；同时采用、广义可加模型

（Generalized Additive Model, GAM），借助光滑样条函数拟合变量关系，不预设具体函数形式 $Y=\beta_0+s_1(\text{GestWeek})+s_2(\text{BMI})+\epsilon$ 。最后使用稳健回归模型，通过 Huber 损失函数或 M 估计的方式降低对异常值的敏感度，提供更稳定的系数估计。

基于该模型完成问题 1 求解与显著性检验，结果表明孕周为显著正向影响因素（ $\beta_1=0.0415$ ， $p<0.001$ ），10-25 周内孕周每增加 1 周，Y 浓度平均上升 0.0415%；BMI 为显著负向影响因素（ $\beta_2=-0.0108$ ， $p<0.001$ ），BMI 每增加 1kg/m²，Y 浓度平均下降 0.0108%；模型 R²=0.710、调整 R²=0.705、F=150.3（ $p<0.001$ ），拟合优度高、解释力强，有效量化了 Y 染色体浓度与孕周、BMI 的关联关系，为后续时点优化提供了可靠基础。

2.2 问题2的模型以及求解

针对问题 2，在关系明确的基础上，要解决“何时达标（FF≥4%）”这一时间维度问题。这本质上是

时间与事件数据分析，采用生存分析（Survival Analysis）框架处理。将每位孕妇视为一个观测个体，“达标”视为“事件”，孕周为“时间”。BMI 作为核心协变量，通过聚类形成同质亚组。目标是为每个亚组找到一个最优时点，使得在该时点进行检测，既能获得高达标概率，又能避免过晚检测带来的高临床风险。

采用 K-means 聚类算法对孕妇 BMI 进行聚类，为了从数据本身出发找到与“达标时间模式”最吻合的分组，首先将“Y 染色体浓度首次达到或超过 4%”定义为一个事件（Event）。从研究起始点（孕 10 周）到事件发生的时间（孕周）称为达标时间。其次对于在研究截止（孕 25 周）前仍未达标的样本，其数据为删失数据（Censored Data）。得出数据组是生存分析的最佳理论基础。

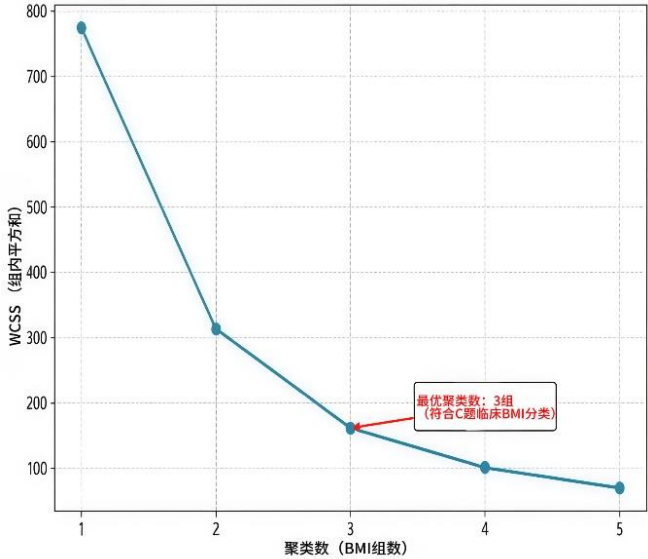


图 1 BMI 聚类肘部法则图

表 1 基于K-means的BMI 分组结果及达标特征

分组名称	BMI范围	样本量 (占比)	临床近似分类	达标特征描述
低BMI组	[18.5, 29.9]	286 (36.6%)	正常/超重	Y浓度上升最快，孕11周时累积达标率已超过90%。
中BMI组	[30.0, 34.9]	423 (54.1%)	I度肥胖	Y浓度上升速度中等，孕11周时累积达标率约为85%。
高BMI组	[35.0, 47.0]	73 (9.3%)	II-III度肥胖	Y浓度上升速度较为缓慢，孕13周时累积达标率才达到80%以上。

采用肘部法则 (Elbow Method) 确定 BMI 最佳聚类数 K ，通过计算不同 K 值下的簇内平方和 (Within-Cluster Sum of Squares, WCSS) 并绘制曲线 (见图 1)，从图 1 可知当 $K=3$ 时 WCSS 下降的边际效益显著降低，形成明显“肘部”，且该结果与临床三类划分逻辑一致，最终确定 $K=3$ ；基于 K -means 算法完成孕妇 BMI 分组后，其描述性统计及达标特征见表 1 最终划分为三组，使用 Log-rank 检验比较三组的生存曲线，得到 $\chi^2=62.3$ ， $p<0.001$ ，表明三组间的达标时间分布存在统计学上的显著差异，分组具有合理性。

综合 K -means 聚类分组与最佳检测时点优化模型可知，聚类方法实现孕妇 BMI 精准分层，优化模型则在分层基础上量化平衡检测成功率与临床风险，输出各组最优 NIPT 检测时点。因此，样本量较大时，先聚类分层再优化建模可实现个性化检测策略；样本量较小时，可直接依据分组特征参考对应最优检测时点，兼顾效率与效益。

2.3 问题3的模型汇总与求解

在问题 2 的基础上，考虑到现实情况基于多因素影响中，身高、体重、年龄、孕产史等都可能影响 FF 动态。采用特征工程和多元聚类，在众多特征中筛选关键变量，将孕妇划分为特征更相似、预后更一致的群组。随后在组内进行时点优化。由于聚类数 K 与模型的复杂度和实用性存权衡，引入多目标优化的需求。

为了更全面地捕捉影响胎儿游离 DNA 比例的潜在因素，本文进一步扩充了特征空间，构建了多维度的孕妇特征集。具体而言，除了纳入年龄、身高、体重等用于派生体型指标的基础特征外，还引入了产次与孕次等反映孕产史的生理特征。此外，因为考虑到变量间的非线性关联，本文增设了 BMI 与年龄的交互项，并引入体表面积作为扩展特征。为消除不同物理量纲对模型收敛速度及权重分配的影响，我们随后对所有数值型特征进行了 Z-score 标准化处理，确保了特征在同一数量级下参与模型运算。

在模型构建阶段，本文对比了三种不同原理的聚类算法，目的是通过多维度验证寻求稳健的分组共识。首先，采用 K -means 聚类作为基础模型，并综合利用肘部法则与轮廓系数协同确定最优聚类数

K 。其次，引入基于概率模型的高斯混合模型进行软聚类分析，以处理可能存在的复杂簇形状并提供概率解释。最后，通过层次聚类生成的树状图直观审视样本间的亲疏关系，辅助确定切割水平。

本研究的实验结果得到三种方法在样本分类趋势上表现出高度的一致性，均倾向于将样本划分为四个主要簇，验证了本研究 BMI 分组结果的可靠性。最终，本文以 K -means 聚类的 4 簇结果为基础框架，并参考 GMM 与层次聚类的聚类边界进行了微调与互证，形成了由 G1 至 G4 组成的最终四分组方案。

基于平衡分组精度与系统复杂性的核心目标，通过将问题抽象为多目标优化模型，最小化包含组内预测风险、管理成本惩罚及达标时间方差的综合损失与聚类数 K ，最终实现了组内一致性与方案实用性的兼顾。

根据图 2 中帕累托前沿曲线的演化规律来看，发现当聚类数 K 从 3 增至 4 时综合损失显著下降，而继续增至 5 时边际改善趋于平缓；针对临床操作便利性与模型复杂度的平衡需求，可推断出 $K=4$ 为全局最优折衷点；通过选取该折衷点，找到累积达标概率达到 95% 对应的最早孕周，作为该组的推荐检测时点。可在不显著增加管理复杂性的前提下，实现模型效能的最优化，最终临床建议方案总结于表 2。

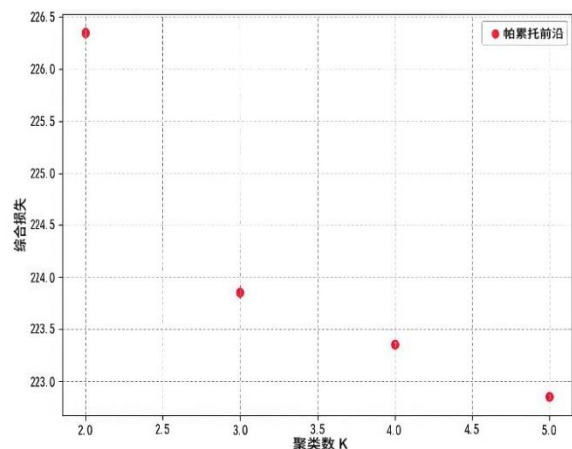


图 2 达标率与孕周的帕累托前沿

敏感性分析表明，核心分组与推荐时点稳定，模型鲁棒性良好。根据 python 运行结果可知，四组中推荐方案随 BMI 与风险梯度递增：G1（11 周，>96%）至 G4（14 周，>93%），匹配不同风险等级。

表 2 多因素综合优化后的NIPT检测推荐方案

分组标签	主要特征描述 (平均BMI等)	推荐检测孕周	预估达标率	风险等级	临床行动建议
G1	年轻, 低BMI	11周	>96%	z低风险	常规检测, 1次即可
G2	中年, 中低BMI	12周	>95%	低风险	常规检测, 1次即可
G3	年龄较大, 中高BMI	13周	>94%	中风险	建议检测, 可考虑2次
G4	高BMI, 或伴其他风险因素	14周	>93%	中风险	重点关注, 建议2次或结合其他筛查

2.4 问题4的模型汇总与求解

针对问题 4，这是一个独立的监督学习分类问题。输入是女胎样本的多维特征向量，输出是“正常”或“异常”的标签。根据样本不平衡、特征间存在复杂关系、需要高可解释性以支持临床决策。因此，需要综合运用特征选择、处理类别不平衡、模型比，与集成、概率校准及决策阈值优化等技术。

针对女胎样本 “正常/异常”二分类问题，本文基于 604 例有效样本构建包含核心判别、测序质量及衍生交互的多维度特征集，采用 L1 正则化逻辑回归、随机森林与 极端梯度提升（eXtreme Gradient Boosting, XGBoost） 模型进行建模，并通过概率校

准优化输出，以缓解样本不平衡与特征复杂问题，最终为临床决策提供高可解释性的分类方案。

针对女胎样本异常率仅 11.1%的严重类别不平衡问题，采用合合成少数类过采样技术合成少数类过 采 样 技 术 （ Synthetic Minority Oversampling Technique, SMOTE）过采样平衡正负样本，构建 L1 正则化逻辑回归、随机森林与 XGBoost 模型并进行概率校准。经分层五折交叉验证，由表 3 可知，L1 正则化逻辑回归曲线下面积（Area Under the Curve, AUC）达 0.884，兼具高可解释性与均衡性能，被选为最终模型。

表 3 女胎异常分类模型性能对比（5折CV平均）

模型	AUC	准确率	灵敏度	特异度	关键优势
逻辑回归 (L1)	0.884	0.803	0.781	0.809	可解释性极强，性能均衡
随机森林	0.812	0.902	0.692	0.935	非线性拟合好，特异度高
XGBoost	0.868	0.851	0.763	0.865	综合性能优秀
校准后逻辑回归	0.849	0.885	0.775	0.902	预测概率更可靠

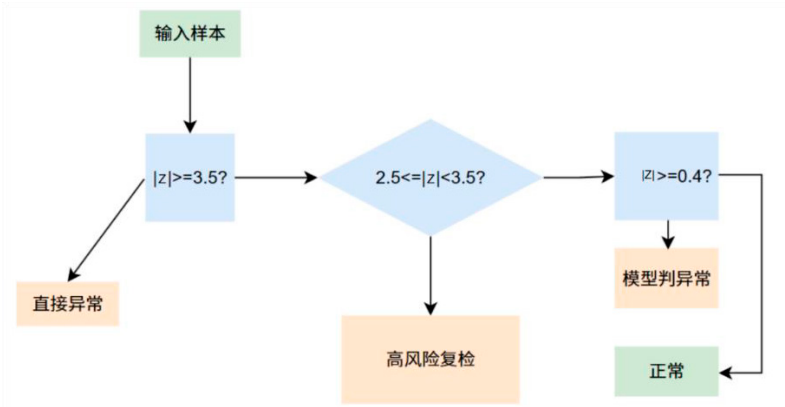


图3女胎NIPT异常判定的临床决策流程

由图 3 结合临床 Z 值阈值: Z13、Z18、Z21 任一 ≥ 3.5 判定高风险, $2.5 \leq Z_{\max} < 3.5$ 判定灰区与经成本效益优化的模型。预测概率阈值 0.4, 依托校准后逻辑回归 AUC 0.884 的高检测性能, 设计分层临床决策规则, 实现女胎 NIPT 从高风险、灰区、模型提示异常到低风险的精准分流, 既兼顾数据驱动的精准性, 又符合医疗实践的安全需求, 实用性强。

3 结论

本文深入探讨无创产前检测 (NIPT) 临床应用中的核心优化问题, 围绕个性化检测时机规划与胎儿染色体异常精准判定两大关键需求, 构建多维度特征集与系列机器学习模型, 为 NIPT 技术的精准化、个体化应用提供理论支撑与实践方案。针对男胎样本, 通过多因素回归与聚类分析明确不同 BMI 人群的最佳检测时点, 有效提升了检测成功率; 针对女胎样本, 通过特征工程与概率校准构建高可解释性的异常判定模型, 弥补传统单一阈值方法的不足。本文主要研究为 NIPT 个性化检测提供理论与实践指导, 但需结合临床实际灵活应用并持续优化。未来研究可进一步扩大样本量、融合多组学临床数

据, 拓展模型应用场景, 考虑更多影响因素, 为临床产前筛查提供更全面的决策支持。

参考文献

- [1] 陈蔚昌.面向特定数据的密度峰值聚类算法优化研究[D].南昌工程学院,2024.
- [2] 梁凯毓.非线性自回归模型误差方差估计的极限性质的若干结果[D].吉林大学,2025.
- [3] 曾如明.K-means聚类算法的改进及其应用研究[D].西华师范大学,2022.
- [4] 庞宇.无创产前检测在胎儿染色体疾病筛查中的临床应用[D].安徽医科大学,2022.
- [5] 张静仪.基于高斯混合模型的多重假设检验研究[D].华南理工大学,2022.
- [6] 林凯乐.基于因素分析的因果森林算法研究[D].辽宁工程技术大学,2024.
- [7] 全国大学生数学建模组委会. 2025 高教社杯全国大学生数学建模竞赛赛题[EB/OL]. [2025-09-04]. https://www.mcm.edu.cn/html_cn/node/03c91a444e62eee81a3740fa97a461a6.html.
- [8] 崔健,徐健.K-means++聚类算法研究与实现[J].电脑知识与技术,2024,20(17):78-81.
- [9] 彭博,吕文英,熊睿.基于TFIDF与XGBoost的新闻分类研究[J].计算机技术与教育学报,2022.