

基于机器学习的NIPT检测时点优化与异常检测方法

陈思宇

西安科技大学 人工智能与计算机学院, 中国·陕西 西安 710600

摘要: 针对无创产前基因检测 (NIPT) 统一采血方案适配性差、人工筛查主观性强的临床痛点, 本文融合统计分析、聚类与机器学习构建一体化分析模型。先完成数据清洗与特征衍生, 结合皮尔逊相关系数、互信息、随机森林联合筛选核心特征, 采用梯度提升回归实现胎儿 Y 染色体浓度精准预测。依托 K-Means 完成单 BMI、多因素两层人群分层, 通过 L-BFGS-B 约束优化求解各群体最优采血孕周, 并设置 0%、5%、10% 梯度误差完成鲁棒性验证。针对无 Y 染色体参照的女胎样本, 对比随机森林与逻辑回归分类性能, 以 0.5 为概率阈值建立标准化染色体异常判别流程。仿真数据测试表明, 该方法可量化平衡检测精度与妊娠风险, 统一筛查判定标准, 操作简便易落地, 能够有效降低漏诊、误诊概率, 为医疗机构开展个体化产前筛查提供数理支撑。

关键词: 无创产前检测; 机器学习; 聚类分析; 风险优化; 异常判别

Research on NIPT Detection Time Point Optimization and Abnormal Detection Method Based on Machine Learning

Chen Siyu

School of Artificial Intelligence and Computer Science, Xi'an University of Science and Technology, China Shaanxi Xi'an 710600

Abstract: Aiming at the clinical problems of poor adaptability of unified blood collection schemes and strong subjectivity of manual screening in non-invasive prenatal testing (NIPT), this paper constructs an integrated analysis model combining statistical analysis, clustering and machine learning. Data cleaning and feature derivation are carried out firstly. Pearson correlation coefficient, mutual information and random forest are integrated to screen core features, and gradient boosting regression is adopted to predict fetal Y chromosome concentration accurately. On the basis of K-Means algorithm, two rounds of population grouping based on single BMI and multiple factors are implemented. The L-BFGS-B algorithm is used to solve the optimal blood collection gestational weeks, and 0%, 5% and 10% gradient errors are set to verify the model robustness. For female fetuses without Y chromosome reference, random forest and logistic regression are compared, and a standardized chromosome abnormality discrimination system is established with the probability threshold of 0.5. Simulation results show that this method can balance detection accuracy and pregnancy risks quantitatively, unify screening criteria, and reduce missed diagnosis and misdiagnosis. It is simple to operate and easy to implement, providing mathematical support for individualized prenatal screening in medical institutions.

Keywords: Non-invasive prenatal testing; Machine learning; Cluster analysis; Risk optimization; Anomaly detection

0 引言

无创产前基因检测 (Non-invasive Prenatal Testing, NIPT) 是现代产前诊断领域的重要技术手段, 依托母体外周血中游离的胎儿 DNA 完成染色体非整倍体筛查, 凭借无创、安全性高、采样便捷等优势, 目前已在各级医疗机构得到广泛普及。相较于传统唐氏筛查与侵入性羊水穿刺技术, NIPT 能够在保障母体安全的前提下, 更早筛查出 21-三体、18-三体、13-三体等常见染色体疾病, 为临床干预争取宝贵时间, 是优生优育体系中不可或缺的一环。

在男胎筛查过程中, 胎儿游离 Y 染色体 DNA 浓度是

判定检测有效性的核心指标, 临床行业普遍设定 4% 作为浓度临界值, 只有 Y 染色体浓度达到该标准, 才能保证后续基因测序结果真实可靠。大量临床实践与医学研究表明, 孕妇孕周、体质质量指数 (BMI)、年龄、身高、体重等多项生理指标, 都会直接影响母体内胎儿游离 DNA 的富集程度, 进而改变 Y 染色体浓度的变化趋势。现阶段国内多数医疗机构仍采用统一的 10~25 周作为通用检测区间, 并未根据孕妇个体生理差异制定分层、分时的采血方案。对于高 BMI、高龄等特殊群体, 若按照统一时间采血, 极易出现采血过早导致 Y 染色体浓度不足、测序失效, 或是采

血过晚导致妊娠风险大幅上升等问题，最终造成假阳性、假阴性结果增多，不仅浪费医疗资源，也会给孕产妇带来不必要的心理负担。

与男胎检测相比，女胎 NIPT 筛查存在更大的技术难点。由于女胎不存在 Y 染色体，临床无法借助该指标判断采血时机与检测有效性，现阶段主要依靠染色体测序 Z 值、GC 含量等数据结合医师临床经验完成异常判定。这种经验主导的筛查模式缺乏统一、量化的判定标准，主观依赖性较强，面对复杂测序数据时，漏诊与误诊的发生概率显著提升，难以满足精准医疗的发展需求。

梳理现有国内外相关研究可以发现，当前针对 NIPT 的数据分析研究大多存在研究维度单一的问题。一部分文献聚焦于胎儿 DNA 浓度影响因素的相关性分析与预测建模，仅实现指标关联规律挖掘；另一部分研究则专注于染色体异常分类算法的设计与优化，缺少对采血时点这一前置环节的考量。同时，多数研究并未结合孕妇多维度生理特征开展人群分层，也未量化分析检测误差对最终结果的干扰，研究成果与临床实际应用场景存在脱节。

针对以上行业现状与研究短板，本文以 NIPT 仿真临床数据集为研究载体，融合数理统计、集成学习、聚类分析与数值优化算法，搭建特征筛选—浓度预测—人群分层—时点优化—异常判别一体化分析框架。首先通过多方法组合策略筛选核心特征，构建回归模型实现 Y 染色体浓度精准预测；其次利用 K-Means 聚类完成单指标与多因素双重人群分组，以检测风险最小化为目标，借助 L-BFGS-B 算法求解最优采血窗口，并完成多梯度误差模拟实验验证模型稳定性；最后针对女胎检测难点，对比不同分类模型性能，建立标准化染色体异常判别体系。本研究旨在用数据驱动的量化方案替代传统经验模式，为临床推行个体化 NIPT 筛查提供技术参考，助力产前筛查技术向精细化、标准化、智能化方向发展。

1 数据来源与预处理

1.1 数据来源

本文实验数据取自全国大学生数学建模竞赛公开 NIPT 仿真数据集，模拟真实临床产前检测场景，共包含 1082 条样本，分为男胎、女胎两类。数据集涵盖孕妇生理指标（孕周、BMI、年龄等）与胎儿测序指标（Y 染色体浓度、染色体 Z 值、GC 含量、测序读段等），维度完整，同时规避临床数据隐私问题，可支撑全流程建模分析。

1.2 数据清洗

原始数据集在采集与整理过程中，存在文本格式不规范、数据缺失、异常极值等问题，若直接用于建模会大幅

降低模型精度与可靠性，因此本文首先开展系统性数据清洗工作。第一步为样本筛选，结合临床 NIPT 检测规范，限定有效孕周区间为 10~25 周，Y 染色体浓度合理区间为 0~16.5%，孕妇 BMI 有效区间为 18~45，剔除超出上述范围的异常样本。第二步处理缺失值，对孕周、BMI、Y 染色体浓度等核心指标存在空缺的样本进行直接删除，避免缺失数据干扰变量相关性分析。经过两轮清洗后，最终得到有效男胎样本 1066 条，样本完整性满足建模要求。

针对原始数据中“11w+6”这类孕周文本格式，统一进行格式转换，将周、天组合格式换算为十进制小数格式，例如将 11 周 6 天转换为 11.857 周，实现全数据集格式标准化，方便后续数值计算与统计分析。对孕周、BMI、Y 染色体浓度三项核心指标进行完整性统计，数据整体质量良好。

1.3 特征工程

基于原始指标构造 40 余项衍生特征（平方项、交互项、比值项），挖掘变量非线性关联。由于各指标量纲差异较大，对全部特征做标准化处理，消除量纲干扰，为后续特征筛选与模型训练奠定基础。

2 研究方法

2.1 多方法融合特征筛选与 Y 染色体浓度预测模型

冗余与无效特征会提升模型复杂度、削弱泛化能力，因此特征筛选是建模的重要环节。本文融合皮尔逊相关系数、互信息、随机森林特征重要度开展联合筛选：依次设置阈值 0.15、0.05、0.005 筛选特征，取结果并集后通过加权评分（相关系数 40%、特征重要度 40%、互信息 20%）排序，最终确定 20 项核心特征。随后对比六种机器学习模型的拟合优度与均方误差，选定 HistGradientBoosting 回归模型。数据集按 8:2 分层抽样划分，并采用五折交叉验证训练，有效规避过拟合问题。

2.2 基于 K-Means 的单指标分组与时点优化

不同 BMI 区间孕妇的 Y 染色体浓度变化特征差异显著，统一采血方案适用性有限。本文以 BMI 为分组指标，采用 K-Means++ 改进聚类算法，遍历 $k=2\sim5$ 并结合轮廓系数确定最优聚类数目。参照相关临床指南，将孕周划分为三级风险区间。以全样本平均检测风险最小为目标，设置分组、时间窗约束，采用 L-BFGS-B 算法求解各组最优采血区间。

2.3 多因素加权聚类与误差敏感性分析

单一 BMI 分组存在局限，因此纳入年龄、身高、体重等指标开展精细化划分。本文构建多元线性回归模型，以回归系数绝对值为权重计算综合得分，再结合 K-Means

算法完成二次聚类，并通过 L-BFGS 算法优化采血区间。为验证模型鲁棒性，设置 0%、5%、10% 三档误差，采用正态分布扰动数据，分析不同误差下检测窗口与风险值的变化规律。

2.4 基于机器学习的女胎染色体异常判别模型

女胎检测缺少 Y 染色体参照，本文选取染色体 Z 值、GC 含量、测序读段比例及孕妇 BMI 等特征构建判别模型，将正常、异常样本分别编码为 0 和 1。数据集存在约 8:1 的类别不平衡问题，因此采用 7:3 分层抽样划分训练集与测试，保留原始数据分布。分别搭建随机森林与逻辑回归模型，借助网格搜索完成参数调优，以精确率、召回率、F1 分数、AUC-ROC 为评价指标对比模型性能。确定最优模型后分析特征重要性，结合 Sigmoid 函数设置 0.5 为判定阈值，建立标准化异常判别流程。

3 实证结果与分析

3.1 Y 染色体浓度预测结果

经过多方法融合筛选后，最终确定 20 项核心特征参与模型训练，有效剔除了大量冗余与干扰特征。HistGradientBoosting 回归模型训练完成后，整体拟合效果表现优异，模型决定系数 $R^2=0.6442$ ，调整后决定系数 $R^2=0.6374$ ，说明模型可以解释 63.74% 以上的 Y 染色体浓度变异规律；模型均方误差 $MSE=0.0004$ ，均方根误差 $RMSE=0.0200$ ，平均绝对误差 $MAE=0.0158$ 。Y 染色体浓度实际取值范围为 0~16.5%（均值约 3.2%）， MSE 为 0.0004 对应 $RMSE$ 约为 0.02（即 0.2% 的绝对浓度误差），相对于浓度均值 3.2% 而言误差约为 6.25%，预测精度较高。作为基线对比，若采用始终预测均值（3.2%）的简单模型， MSE 将达到 0.0012，本文模型相比基线降低了约 67% 的误差。F 检验结果显示模型整体显著性 $p<0.0001$ ，证明模型整体回归效果显著，具备统计学意义。

为进一步验证模型可靠性，本文开展残差分析。残差散点图结果显示，模型残差整体围绕 0 轴随机分布，无明显趋势性变化，不存在系统性预测偏差；残差分布直方图与正态 Q-Q 图表明，残差近似服从正态分布，符合回归模型的统计假设。特征相关性热力图显示，部分衍生交互特征存在多重共线性，但 HistGradientBoosting 集成算法对多重共线性具备较强容忍度，因此并未对最终预测效果造成明显影响。综合各项指标可知，该预测模型稳定、有效，能够精准量化各项指标与 Y 染色体浓度的关联关系。

为进一步提升模型可解释性，本文采用方法对最终模型进行分析。计算各特征对 Y 染色体浓度预测的 SHAP 值，结果显示 SHAP（Shapley Additive Explanations）孕周、

BMI 及其交互项是影响浓度预测的最主要因素，且均呈现正向贡献；年龄的 SHAP 值分布较为分散，其影响存在个体差异。该结果与临床认知一致，可为医生提供直观的特征影响方向。

3.2 单 BMI 分组与时点优化结果

经轮廓系数检验，确定最优聚类数为 2，BMI 分割点为 31.25。低 BMI 组（20.7~32.1）、高 BMI 组（32.3~45.0）对应的最优采血区间分别为 11.0~13.71 周、12.14~26.43 周。分组结果显示低 BMI 人群占比更高，多数样本达标孕周均落在最优区间内。经 L-BFGS-B 算法优化，整体平均风险值为 1.8442，相较于传统统一采血方案，个性化时间窗更贴合人群特征，应用价值更高。BMI 分组分布见图 1。

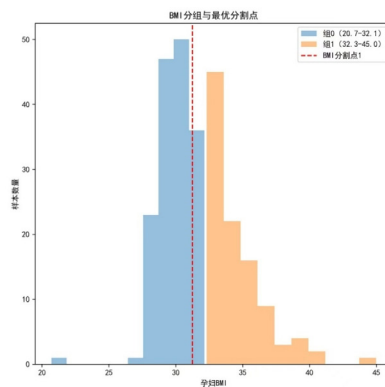


图 1 孕妇 BMI 分组分布及最优分割点

不同 BMI 分组样本的达标孕周与最优检测区间匹配情况见图 2。

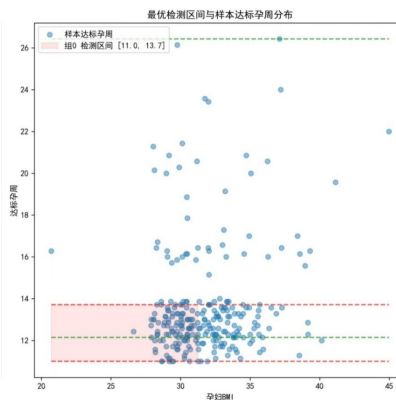


图 2 达标孕周与最优检测区间分布散点图

3.3 多因素分组及误差分析

多元线性回归结果显示，体重、BMI 对 Y 染色体达标孕周影响最为显著，身高次之，年龄影响最弱，该结论与临床认知相符。多因素聚类后样本分组区分度进一步提升。在 0%~10% 梯度误差测试中，采血区间与分组界限基本稳定，仅风险值小幅波动，表明本文模型抗干扰能力强，

临床容错性良好。

3.4 女胎染色体异常判别结果

综合 AUC 值、F1 分数、精确率与召回率等多项指标对比,逻辑回归模型的综合分类性能优于随机森林,因此确定逻辑回归为女胎染色体异常判别核心模型。特征重要性排序结果显示,21 号染色体 GC 含量、13 号染色体 GC 含量、18 号染色体 GC 含量是影响异常判定的三大核心特征,X 染色体浓度、孕妇 BMI、染色体 Z 值也起到重要作用。

本文以 0.5 作为异常判定概率阈值,流程简洁高效,可适配各级医疗机构,有效统一筛查标准、降低误诊风险。各特征贡献度排序见图 3。

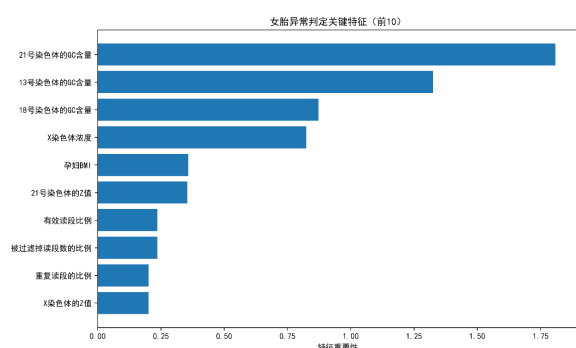


图 3 女胎异常判定特征重要性排序

4 讨论

本文整合统计、机器学习与数值优化算法,构建了覆盖特征挖掘、浓度预测、人群分层、时点优化及异常判别的 NIPT 全流程分析方案,相比传统经验式筛查具备多重优势。研究通过多方法联合筛选锁定核心影响指标,利用聚类与约束优化实现采血窗口量化,并完成误差验证,同时针对女胎检测难点建立标准化判别体系,既贴合临床需求,也可同类医疗数据挖掘提供参考。

实验结果表明,逻辑回归的综合性能优于随机森林。究其原因,数据集特征与标签存在较强线性关联,逻辑回归可充分利用该特征;而随机森林在小样本、高维场景下易出现过拟合,对数据噪声更为敏感。这也说明在临床场景中,兼顾可解释性的简易模型更具实用价值。

本研究仍存在一定局限:实验采用仿真数据集,无法完全复刻真实临床的噪声、平台差异等复杂情况,模型泛化能力有待临床样本进一步验证;尽管采用 k-means++ 优化初始化方式,聚类算法仍存在收敛至局部最优的风险;线性的逻辑回归模型对复杂非线性数据、极端样本的识别能力尚有不足。

针对以上问题,提出后续优化思路:一是合作采集真

实临床数据,重新训练并迭代模型;二是优化聚类策略,引入 SHAP、LIME 等工具提升模型可解释性;三是融合支持向量机、轻量化深度学习等算法,强化非线性数据处理能力。未来还可将本模型拓展至更多染色体疾病筛查场景,进一步拓宽应用范围。

5 结语

本文基于无创产前基因检测仿真数据集,综合数理统计、机器学习与约束优化算法,完成 NIPT 全流程建模分析。研究融合皮尔逊相关系数、互信息与随机森林完成特征筛选,剔除无效冗余变量,并采用 HistGradientBoosting 算法构建回归模型,可精准量化各生理指标与 Y 染色体浓度的关联规律,模型拟合效果与统计学表现良好。

研究先后基于单一 BMI、多维度生理指标完成两阶段人群分组,借助 L-BFGS-B 算法求解最优采血区间。多组误差实验证实,该检测窗口抗干扰能力强,能够适配临床常规检测偏差。针对女胎检测特点,经模型对比选用逻辑回归作为分类器,结合 Sigmoid 函数设置判定阈值,搭建起标准化异常判别流程,改善了传统筛查主观性强、标准不一的问题。

整套方案原理简单、易落地,可为医疗机构开展个体化 NIPT 筛查提供理论与技术支撑。受仿真数据限制,模型仍有优化空间,后续将依托真实临床样本迭代完善算法,推动技术落地,助力产前筛查精细化发展。

参考文献:

- [1] 何选森,郑茜,许凤. K-Means 算法最优聚类数量的确定[J]. 电子科技大学学报, 2022, 51 (06): 904-912.
- [2] 李欣海. 随机森林是特点鲜明的模型,不是万能的模型 [J]. 应用昆虫学报, 2019, 56 (01): 170-179.
- [3] 张婷,刘阳. 约束优化算法在医疗决策中的应用研究 [J]. 数学的实践与认识, 2023, 53 (08): 98-106.
- [4] 侯伟,周红辉,谢潇潇等. 无创产前检测适用人群的探讨及其临床应用评价 [J]. 解放军医学院学报, 2021, 42 (03): 246-250.
- [5] 王健,李军,陈静. 机器学习在医疗数据挖掘中的应用研究进展 [J]. 计算机工程与应用, 2023, 59 (10): 1-10.
- [6] 中华医学会妇产科学分会产科学组. 孕前和孕期保健指南 (2018)[J]. 中华妇产科杂志, 2018, 53(1): 7-13.

作者简介:陈思宇(2006-),女,陕西西安人,西安科技大学人工智能与计算机学院本科在读,研究方向:机器学习、生物医学数据分析、无创产前基因检测算法优化。