计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(3):310−317 [doi: 10.15888/j.cnki.csa.008416]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

GRU 和 GBDT 混合模型在早产风险预测中的应用①

吴忆娜[1,2], 张艺超[2,3], 袁贞明[1,2], 胡文胜[4], 卢　莎[4], 孙晓燕[1,2], 吴英飞[1,2]

[1](杭州师范大学 信息科学与技术学院, 杭州 311121)
[2](移动健康管理教育部工程研究中心, 杭州 311121)
[3](杭州师范大学 医学院, 杭州 311121)
[4](杭州市妇产科医院, 杭州 310008)
通信作者: 吴英飞, E-mail: wyf@hznu.edu.cn

摘　要: 早产是新生儿死亡及病残的首要原因, 且影响新生儿的远期健康. 然而早产的准确预测一直是医学上的一个难题. 目前医学上早产的早期筛查多基于特殊检查, 但因成本核算等问题难以大规模临床应用, 而电子病历的普及和人工智能技术的发展, 为产科疾病的早期风险评估提供支持. 本文利用产科电子病历的诊疗信息, 构建 GRU 和 GBDT 的混合模型预测早产. 混合模型利用 GRU 在孕妇多次产检信息中探究早产发生的概率, 并将结果融入孕前和 28 周前末次产检数据, 最后利用 GBDT 对孕妇进行更加精确的早产风险预测. 实验结果表明, 基于 GRU 和 GBDT 的早产预测模型在 AUC 和 ROC 等评估指标上优于其他单一模型, 本研究方法可有效帮助产科医护人员在妊娠早中期判断孕妇是否有早产风险.

关键词: 电子病历; 早产预测; GRU; GBDT; 混合模型

引用格式: 吴忆娜,张艺超,袁贞明,胡文胜,卢莎,孙晓燕,吴英飞.GRU 和 GBDT 混合模型在早产风险预测中的应用.计算机系统应用,2022,31(3):310–317. http://www.c-s-a.org.cn/1003-3254/8416.html

Application of GRU and GBDT Hybrid Model in Risk Prediction of Premature Birth

WU Yi-Na[1,2], ZHANG Yi-Chao[2,3], YUAN Zhen-Ming[1,2], HU Wen-Sheng[4], LU Sha[4], SUN Xiao-Yan[1,2], WU Ying-Fei[1,2]

[1](School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China)
[2](Engineering Research Center of Mobile Health Management System (Ministry of Education), Hangzhou 311121, China)
[3](Division of Health Sciences, Hangzhou Normal University, Hangzhou 311121, China)
[4](Hangzhou Women's Hospital, Hangzhou 310008, China)

Abstract: Premature birth is the primary cause of neonatal death and disability, which can affect the long-term health of newborns. However, the accurate prediction of premature birth is a difficult problem in the medical field. The early screening of premature birth in medicine is mostly based on special examinations, but it is difficult to be applied to large-scale clinical practice due to cost accounting and other problems. The popularization of electronic medical records and the development of artificial intelligence technology provide support for early risk assessment of obstetric diseases. This study uses the diagnosis and treatment information of obstetric electronic medical records and proposes a hybrid model of gate recurrent unit (GRU) and gradient boosting decision tree (GBDT) to predict the risk of premature birth. The hybrid model uses GRU to explore the probability of premature birth in multiple antenatal examination information of pregnant women and integrates the results into the pregnancy data and the last antenatal data before 28 weeks. Finally, GBDT is used to predict the risk of premature birth for higher accuracy. The experimental results show that evaluation indexes such as AUC and ROC of the prediction method based on GRU and GBDT are better than those of other single machine

① 基金项目: 浙江省省级重点研发计划 (2020C03107); 浙江省自然科学基金 (LGF20F020009); 杭州市属高校优秀创新团队; 国家卫生健康委科学研究基金——浙江省医药卫生重大科技计划 (WKJ-ZJ-1911); 杭州市卫生科技计划 (ZD20200035&OO2019054)
收稿时间: 2021-05-28; 修改时间: 2021-07-01; 采用时间: 2021-07-27; csa 在线出版时间: 2022-01-24

learning models. The proposed method can provide a reference for the obstetric medical staff to judge the risk of premature birth in the early and middle stages of pregnancy.

Key words: electronic medical records; premature birth prediction; gate recurrent unit (GRU); gradient boosting decision tree (GBDT); hybrid model

早产是指妊娠达到 28 周但不足 37 周而终止妊娠, 按病因分为自发性早产和医源性早产, 是新生儿死亡及病残的首要原因, 且影响新生儿的远期健康[1].据 WHO 发布的《全球早产儿报告》指出, 全球每年约有 1 500 万早产儿出生, 发生率为 5%–18%, 其中 100 万早产儿发生死亡[2]. 早产是新生儿死亡的主要原因, 随着我国二胎生育政策的实施, 高龄产妇有所增多, 早产儿的发生率呈上升趋势, 出生 1 岁以内死亡的婴儿约 2/3 为早产儿[3]. 此外, 早产儿相较于足月儿, 各器官发育不成熟, 先天畸形、神经系统发育不良如视网膜病变和脑瘫等发病率增高, 远期病死率亦高于足月儿[4]. 因此, 在孕早期预测早产并采取预防性措施对降低早产儿病死率、提高早产儿生存率具有重要意义[5].

目前早产的发生机制尚未明确, 临床对早产的早期风险预警仍缺乏有效的评估手段, 如何设计高效的早产筛查模型是一个具有全球意义的重大产科难题[6]. 超声检查具有操作简便且适用性广等特点, 是临床上评估早产风险的重要手段. 其中, 经会阴超声测量孕妇的宫颈长度是一种全球公认的早产筛查方法[7,8], 但阴道超声受限于多种因素, 例如超声设备的质量、超声医生的技术水平等, 且不适用前置胎盘和阴道出血等孕妇. 此外, 某些研究认为基因检测、某些生物标志物也可用于预测早产[9,10]. 然而以上方法所用的检测指标多为特殊检查项, 检查成本高昂, 难以进行大规模临床验证, 且方法结构简单只考虑了单一因素, 未分析各危险因素间的非线性相互作用[7]. 早产预测模型的建立须考虑疾病的整体性、复杂性和动态性, 而机器学习技术以其独特的整体性、系统性、自学习性和极强的容错性等特点, 为复杂的疾病诊断提供支持, 成为近年来计算机与医学领域研究的热点[11].

流行病学调查显示临床上早产高危因素主要包括社会因素、个人因素、孕妇病史因素以及本次妊娠情况等方面[12], 结合人工智能技术可对电子病历中的体量巨大、类型异构、内部关联复杂的临床大数据进行全面客观地分析[13,14]. 针对早产问题, Koivu 等人[15] 在纽约公开数据集上利用人工神经网络和梯度提升决策树 (gradient boosting decision tree, GBDT) 等算法构建早产预测模型. Luo 等人[16] 利用弹性网络正则化逻辑回归模型, 预测一般性早产 (分娩日期大于 32 周且小于 37 周) 的风险. 但这些机器学习研究均未涉及到时序研究, 且模型的效果相对较差. 门控循环单元 (gate recurrent unit, GRU) 是可用于处理时间序列数据的神经网络, 它是循环神经网络 (recurrent neural network, RNN) 的一种变体, 其结构简单、性能稳定性高, 并解决了 RNN 梯度消失或爆炸的问题[17]. Ljubic 等人[18] 利用 RNN、长短期记忆神经网络 (long short-term memory, LSTM) 和 GRU 预测 2 型糖尿病患者的并发症风险, 其中 GRU 时序模型整体性能最佳. 但 GRU 作为深度学习模型仍难以解释预测结果, 而建立疾病风险评估模型的核心是开发整体和有意义的可解释架构. 由于单一模型的局限性, 众多研究提出了将时序模型与决策树模型结合的混合模型来进行预测分析, 并广泛应用于医学领域[19–21]. 如赖晓莹等人[22] 将加权组合模型应用于预测肺结核发病趋势, 有效提升了模型的预测效果.

针对以上分析, 本文拟将 GRU 与 GBDT 的优势结合, 将 GRU 和 GBDT 混合模型应用于早产风险预测. GRU 模型挖掘产检数据中与早产相关的时间序列隐含信息, 并在 28 周前实现对早产的预测, 同时利用 GBDT 模型探究决策形成的原因, 将预测模型和医学可解释性相结合, 为提早干预和救治、降低早产发生率、改善早产人群的母婴结局提供参考依据.

1 相关工作

1.1 数据获取平台

本项目研究团队前期建立了产科多源异构数据互通共融的产科数据科研平台, 用于孕妇产检数据的获取. 孕妇数据来源包括产检门诊、社区档案、超声检查、实验室检查等, 平台对多源异构的孕妇数据进行清洗、转换、集成, 如文本类型的超声报告结构化, 胎心信号数据的解析等, 最终形成可统一处理的结构化

数据. 产科数据科研平台如图 1 所示.

本文的早产风险预测依托前期集成的产科医疗数据, 基于本研究可形成独立的分析模块嵌入产科数据科研平台的健康分析云模块中, 进行智能的早产风险评估, 为临床决策提供支持, 做到早产的早发现、早诊断、早干预.

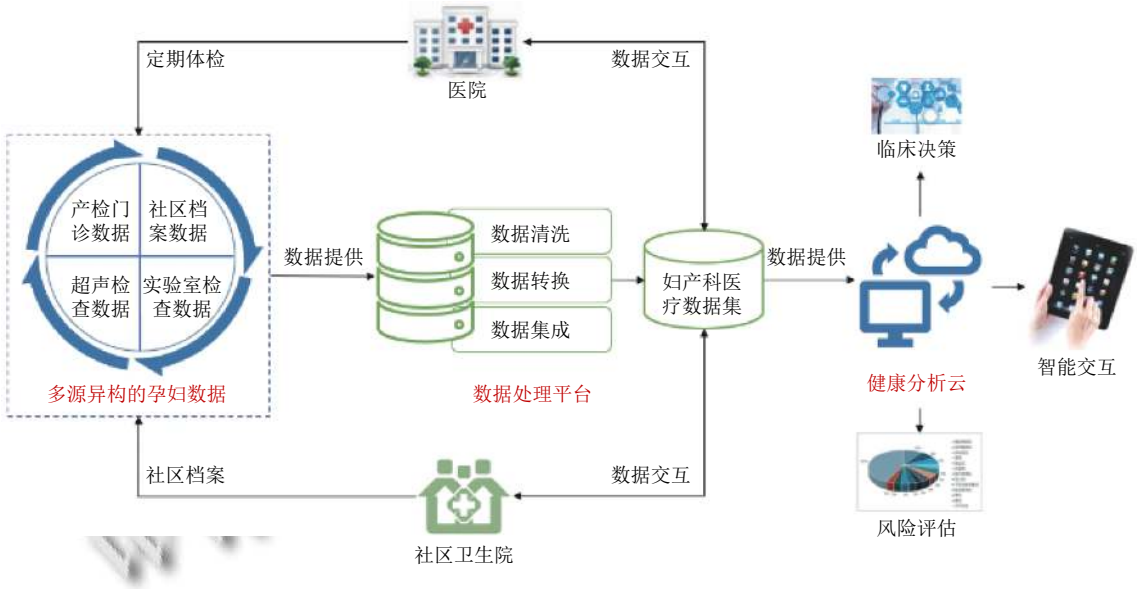

图 1 产科数据科研平台

1.2 GRU 模型

门控循环单元 (GRU) 是 RNN 的变体, 是为了解决长期记忆和反向传播中的梯度等问题而提出的[23,24], 能良好地拟合时序数据. 相比于 LSTM, GRU 只包括更新门和重置门两个门, 简化的结构使得 GRU 在保证预测精度的前提下能有效减少运行时间[25]. GRU 单元内部结构如图 2 所示.

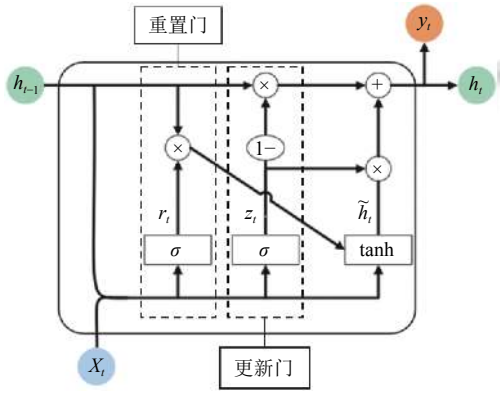

图 2 GRU 单元内部结构

GRU 中更新门、重置门的公式如式 (1)、式 (2) 所示:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{1}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{2}$$

其中, x_t为当前的输入, h_{t-1}代表上一个节点传递下来的隐状态. z_t和r_t分别表示更新门和重置门. W_r和W_z分别代表重置门和更新门的权重矩阵. σ是 Sigmoid 激活函数.

上一时刻隐藏数据经过重置门控得到的重置数据与当前的输入x_t相结合并通过激活函数可以得到当前时刻的候选隐藏状态$\tilde{h}_t$, 公式如式 (3) 所示. 然后结合式 (4) 得到t时刻的隐藏状态h_t, 最后得到 GRU 网络模型在t时刻的输出:

$$\tilde{h}_t = \tanh(W_{\tilde{h}} \cdot [r_t * h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{4}$$

$$y_t = \sigma(W_o \cdot h_t) \tag{5}$$

其中, tanh为双曲正切函数. $W_{\tilde{h}}$代表隐藏状态的权重矩阵. W_o为隐藏层到输出层的代表权重矩阵.

1.3 GBDT 模型

GBDT 是于 2001 年被提出的以 CART 回归树为基学习器的 Boosting 算法, 具有预测精度高, 鲁棒性强, 灵活性高等特点[26]. 其核心思想是通过损失函数的负梯度拟合前一轮基学习器的残差, 具体原理如下:

首先设训练样本为i ($i = 1,2,3,\cdots,n$), 迭代次数j ($j = 1,2,3,\cdots,m$), 损失函数为$L(y_i, F(x_i))$, 设置初始常

数模型来最小化损失函数, 公式如式 (6) 所示.

$$F_0(x) = \arg\min_r \sum_{i=1}^{n} L(y_i, c) \tag{6}$$

负梯度r_{ij}的计算公式如式 (7):

$$r_{ij} = \left[\frac{\partial L(y, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{j-1}(x)} \tag{7}$$

使用基学习器$h_j(x)$拟合损失函数的负梯度r, 求出使损失函数最小的最佳拟合值:

$$r_j = \arg\min_r \sum_{i=1}^{n} L(y, F_{j-1}(x) + h_j(x_i)) \tag{8}$$

接着进行模型更新, 本轮的强学习器如下:

$$F_j(x) = F_{j-1}(x) + r_j h_j(x_i) \tag{9}$$

输出最终的结果:

$$F_M(x) = \sum_{j=1}^{m} F_{j-1}(x) \tag{10}$$

GBDT 的特征重要性计算是基于计算决策树分裂节点的增益, 并用累积求和来评估特征的重要性. 其中特征j的全局重要性由特征j重要性平均值衡量, 公式如下:

$$\hat{I}_j^2 = \frac{1}{M}\sum_{m=1}^{M} \hat{I}_j^2(T_m) \tag{11}$$

其中, M表示树的数量, $\{T_m\}_1^M$表示决策树的集合. 特征j在单棵树中的重要度如下:

$$\hat{I}_j^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 1(v_t = j) \tag{12}$$

其中, $\hat{i}_t^2$代表节点分裂后平方损失的减少值, v_t代表与j关联的特征, $J-1$代表非叶子节点的数量.

2 基于 GRU-GBDT 的早产风险预测方法

GRU 模型结构简单, 运行速度快, 在时间序列预测方面具有较高的拟合能力和良好的预测效果. 然而深度学习模型难以对预测结果与输入特征之间的关系做出解释. GBDT 模型能计算每个输入特征对最终预测结果的重要性, 特征重要性级别越高, 表明该特征对预测结果的影响越大, 以此解释 GBDT 模型的预测结果. 鉴于 GRU 和 GBDT 模型的优点, 本文旨在利用 GRU 和 GBDT 的混合模型在孕妇 28 周前预测早产风险, GRU 模型在孕妇时序产检数据中学习并预测早产发生的概率, 结合 GBDT 模型实现更准确的早产风险预测, 而且在提升预测性能的同时分析输入特征在模型中的贡献程度, 实现模型的可解释性. 早产风险预测总体流程如图 3 所示.

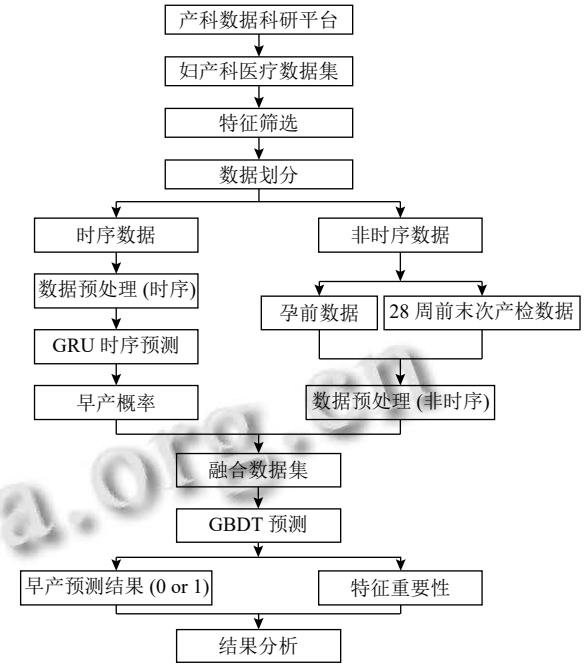

图 3 早产风险预测总体流程

首先通过产科数据科研平台获取妇产科医疗数据集, 并对获取的数据进行特征筛选、数据划分等处理. 然后将数据分为时序数据和非时序数据并分别根据两个数据集的特点进行数据预处理.

接着针对时序数据利用 GRU 模型得到早产发生的概率. GRU 输入层为预处理后的孕妇 28 周前的 5 次产检数据, 输入序列为:

$$X = (x_1, x_2, x_3, x_4, x_5) \tag{13}$$

其中, x_t代表孕妇第t次的产检记录, GRU 神经网络的隐藏层数为 2 层, 隐藏层的第二层连接了上一个隐藏层中保留下来的信息. 最后将输出层中最后一个时刻h_5的结果作为模型的输出, 并经过 Softmax 激活函数得到孕妇在不同分类结果下的概率y. 双层 GRU 网络结构图如图 4 所示.

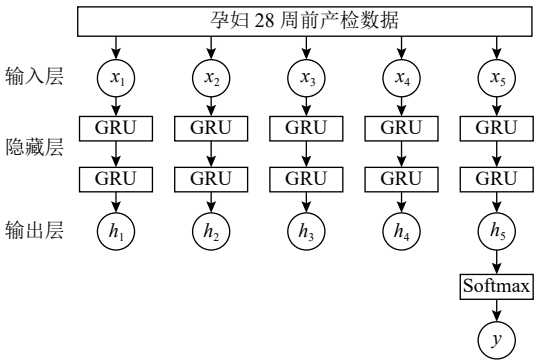

图 4 双层 GRU 网络结构图

GRU 模型的目标是利用时序数据预测出早产发生的概率. 将该早产概率作为新特征与孕前数据和 28 周前末次产检等非时序数据融合得到新数据集. 将新数据集输入 GBDT 模型中, 实现进一步的早产预测. 预测过程中 GBDT 模型计算输入每一个特征在预测时的贡献度. 接着, GBDT 模型在得到分类结果的同时获取输入数据中每个特征的重要性. 最后对实验结果进行对比分析, 验证所用方法的有效性.

3 实验设计及分析

3.1 数据预处理

通过产科数据科研平台, 实验收集了 2017 年 1 月–2020 年 5 月于某三甲医院产科分娩且临床资料完整的孕妇数据, 孕妇的排除标准如下: (1) 妊娠合并子宫体肿瘤; (2) 有严重的心、脑、血管、肾等内外科合并症及妊娠并发症; (3) 妊娠期间行宫颈环扎术; (4) 妊娠结局为剖宫产、引产的孕妇.

3.1.1 特征筛选

通过文献分析、专家小组会议并结合临床知识, 共纳入 32 个可能影响早产的相关因素, 包括孕前数据、产检数据和超声检查数据, 其中孕前数据包括年龄、孕次、产次、身高、孕前体重、文化程度、孕前收缩压、孕前舒张压、末次月经、初潮、经期、周期、月经量、痛经、是否自然妊娠、血型、流产史、早产史和其他既往史; 产检数据包括孕期体重、BMI、宫高、腹围、孕期血压和孕期血常规; 超声检查包含双顶径、胎儿头围、股骨长、胎儿腹围、羊水指数、脐动脉的血流指数和颈项透明层. 所有纳入的特征均由产科数据科研平台获取.

3.1.2 数据划分

本研究纳入的特征中产检数据和超声检查数据呈现明显的时序分布, 产检指南要求产妇在孕 12 周前进行登记和初检, 并在孕 28 周前每月 1 次通过门诊随访. 因此, 本研究根据产检指南规定将孕检时间初步划分为孕周 12 周前、13–16 周、17–20 周、21–24 周以及 25–28 周, 共计 5 次产检信息. 但实验发现这种选取方法会丢失近 80% 的产妇, 2 万多名孕妇中只有 5 680 名左右孕妇在各个划分的横断面均有孕检记录. 图 5 为孕妇 29 周前真实产检分布. 由图可知, 造成此类误差的原因可能是 13–16 周的部分孕妇是属于 17 周滞后的检查, 因此本研究根据实际产检分布略微调整, 将 13–16 周、17–20 周和 21–24 周分别调整为 13–17 周、18–21 周和 22–24 周. 根据调整后的横断面, 本研究最终共纳入 8 140 名孕妇, 包括 40 700 条产检记录. 此外, 本研究将纳入的 8 140 名孕妇的孕前数据和孕 28 周前末次产检的数据抽取出来作为实验所用的非时序数据.

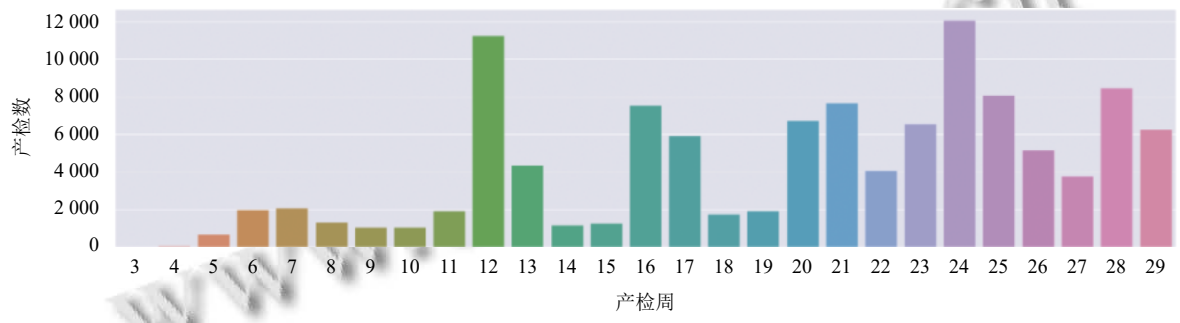

图 5 妊娠早中期产检分布

3.1.3 数据标准化

本研究收集的诊疗数据具有明显的时序特性, 构建孕产妇早期产检的时序特征集合, 表 1 为其中一例产妇的时序特征集合示例, 包括其常规的体格检查、超声检查等, 其中化验数据暂未纳入时序特征集合, 原因是孕产妇早期的化验多在各自社区进行, 其检验手段和标准难以统一.

按照第 3.1.2 节的 5 个横断面筛选定时门诊且产检资料完整的孕产妇数据作为时序数据. 将孕前数据和 28 周前末次产检的数据作为非时序数据. 在数据使用前, 需要对数据中的缺失值和异常值进行处理.时序数据的缺失值采用线性插值法进行填充, 非时序数据采用均值填充. 结合临床中各个指标的范围, 将数据中观测极大值或极小值作为异常值, 处理方法同缺失值. 此外, 由于样本特征数据具有不同的量纲和量纲单位, 数值间的差距会对模型造成影响, 因此需要对数据进

行归一化处理, 避免值域较大的特征影响其他特征, 同时提升模型的收敛速度. 本文采用 min-max 标准化, 使得结果映射到 [0, 1] 之间, 如式 (14) 所示:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{14}$$

$$y = y_{\text{predict}}(x_{\max} - x_{\min}) + x_{\min} \tag{15}$$

其中, x为当前特征值, $x_{\min}$, $x_{\max}$分别为当前特征的最小值和最大值, x^*为标准化后的特征值. 模型得到预测结果后, 通过式 (15) 对结果进行反归一化处理得到真实值, 其中y为真实值, y_{predict}为预测值.

表 1 时序特征构建示意图

孕周	宫高	腹围	体重	收缩压	舒张压	BMI	…	胎儿头围	股骨长
12周前	10	68	52	118	72	17.99	…	—	—
13–17周	12	72	53	99	66	18.34	…	—	—
18–21周	20	77	109	109	69	19.38	…	—	—
22–24周	22	82	60	126	66	20.76	…	21.7	4.6
25–28周	26	88	61	131	74	21.11	…	25.9	5.5

3.2 模型设置和评价指标

本研究最终纳入 8 140 名孕妇, 其中早产 342 名, 足月 7 798 名. 因本实验数据集样本分布不平衡, 会使预测的分类结果偏向于多数类样本点集合, 使得少数类样本点的分类正确率低. 针对样本不平衡问题, 在训练过程中将少数类样本进行随机过采样处理. 实验时使用网格搜索法选择模型的最优超参数, 并采用五折交叉验证的方法来验证结果的可靠性.

调整后的 GRU 和 GBDT 的最优参数设置如表 2 所示, GRU 神经网络批次大小为 256, 输入特征维度为 32, 隐藏层层数为 2, 将学习率设置为 0.001, 并使用 Adam 优化器进一步加速训练过程, Adam 优化器可以使用动量和自适应学习率来加速收敛速度. GBDT 模型的损失函数采用对数似然损失函数“deviance”, 设置学习率 0.01, 弱学习器的最大迭代次数为 200, 子采样取值为 0.8, 防止过拟合.

此外本研究采用加权平均敏感性 (weighted average sensitivity, *WA_Sensitivity*)、加权平均特异性 (weighted average specificity, *WA_Specificity*)、AUC 和 ROC 曲线下面积对模型的性能进行评价. 加权平均是将每一个类别样本数量在总样本中的占比作为权重, 在样本不平衡情况下可获得更加客观的总体评价, 加权平均敏感性和加权平均特异性公式如式 (16)、式 (17) 所示, 其中se_0、se_1分别代表 0、1 样本的灵敏度, num_0、num_1分别为 0、1 样本的数量, num_{all}为总样本数, sp_0、sp_1分别代表 0、1 样本的特异性.

$$WA\_Sensitivity = \frac{se_0 \times num_0}{num_{\text{all}}} + \frac{se_1 \times num_1}{num_{\text{all}}} \tag{16}$$

$$WA\_Specificity = \frac{sp_0 \times num_0}{num_{\text{all}}} + \frac{sp_1 \times num_1}{num_{\text{all}}} \tag{17}$$

表 2 GRU 和 GBDT 模型各参数的含义及取值

模型	参数	值	参数含义
GBDT	loss	‘deviance’	损失函数
	learning_rate	0.01	学习率
	n_estimators	200	最大迭代次数
	max_depth	2	树的最大深度
	subsample	0.8	子采样
GRU	Loss function	CrossEntropy	损失函数
	Num_layers	2	GRU层个数
	Optimizer	Adam	优化器
	Hidden_size	55	隐层状态维数
	Input size	32	输入特征维数
	Learning rate	0.001	学习率
	Batch-size	256	批处理大小
	Epochs	20	迭代次数

3.3 实验结果和分析

实验中分别选用逻辑回归 (logistic regression, LR)、随机森林 (random forest, RF)、GBDT、LSTM 、GRU 和 GRU-GBDT 进行对比, 数据集均按照 8:2 划分为训练集和测试集, 并通过五折交叉验证实验结果. 表 3 为各模型的预测结果.

由表可知, 与单一模型相比, GRU-GBDT 模型对早产的预测能力最佳, 其中加权平均敏感性为 0.77, 加权平均特异性为 0.84, AUC 为 0.647, 均优于其他方法.此外, GBDT 相比较于 LR 和 RF, 在牺牲少许运行时间的情况下各项指标均有所提高. 时序模型的 AUC 明显优于非时序模型, 其中 GRU 相对于 LSTM 结构更加简单, 可在保持模型性能的前提下显著提升算法运行速度.

表 3 各模型预测结果表

模型	WA_Sensitivity	WA_Specificity	AUC	运行时间 (s)
LR	0.65	0.76	0.587	2
RF	0.72	0.80	0.585	3
GBDT	0.75	0.82	0.61	9
LSTM	**0.77**	0.80	0.631	193
GRU	0.75	0.83	0.628	86
LSTM-GBDT	0.75	0.82	0.645	209
GRU-GBDT	**0.77**	**0.84**	**0.647**	97

图 6 为 GBDT、 GRU 和 GRU-GBDT 混合模型的 ROC 曲线, 曲线下面积值越高模型的预测性能越佳. 由图可知, 本研究的 GRU-GBDT 混合模型优于对应的单一模型.

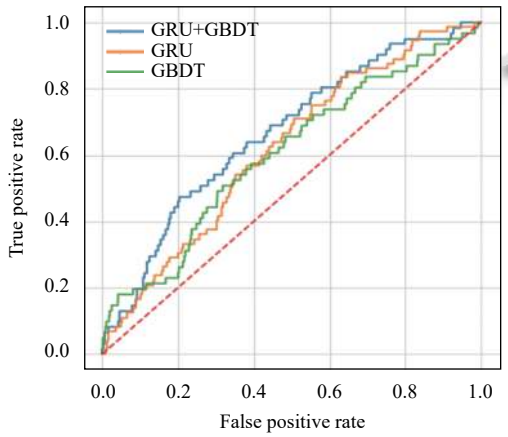

图 6 GBDT、GRU 和 GRU-GBDT 的 ROC 曲线图

3.4 特征重要性分析

根据 GBDT 模型输出的特征重要性排序如图 7 所示, 由于所用特征较多, 本文只列出重要性排序前 15 名的特征.

由图 7 可知, GRU 输出的新特征在早产预测中的重要性最高, 可见本文采用的 GRU-GBDT 混合模型在早产预测中的有效性. 此外, 宫高、BMI、血红蛋白、舒张压、双顶径等都是早产的重要因素. 结合 GBDT 模型输出的特征重要性结果, 可辅助医生临床决策, 便于医生对存在早产风险的孕妇进行及时有效的干预.

4 结论与展望

针对早产风险预测问题, 本文通过分析孕妇历次产检数据特征, 利用 GRU-GBDT 混合模型预测孕妇早产风险. 本实验整合多源异构的产科诊疗数据并根据产检指南以及实际情况合理获取多次产检的信息, 通过 GRU 模型捕捉孕妇历次产检的生理特点并得到早产发生的概率, 然后采用 GBDT 模型在融合数据的基础上预测最终分类结果, 并获取特征重要性. 通过与其他方法对比分析, 验证了该混合模型在早产分类效果上的优越性, 其中 GRU 对于时间序列信息有较强的学习能力, 该混合模型保证了孕妇多次产检数据的合理利用, 同时 GBDT 模型能在预测时获取每一特征对预测结果的贡献度, 特征重要性可为医生判断孕妇早产风险提供辅助决策.

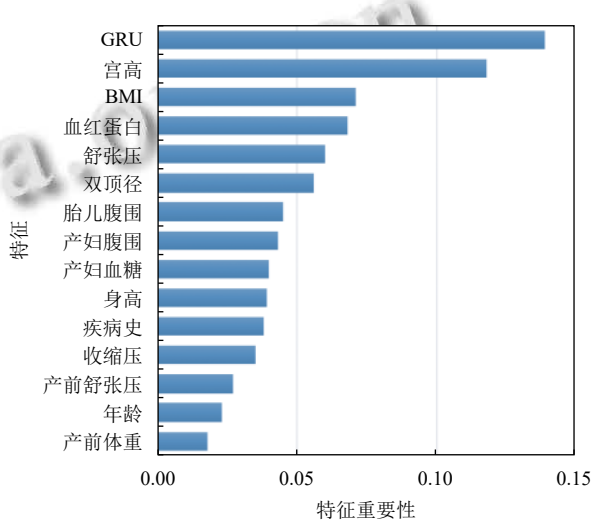

图 7 GBDT 模型重要性排序

本文可实现早期的早产高危人群筛选, 以建议该部分人群进行进一步的早产项目检查. 本研究不涉及特殊化验项和检查项, 在不进行额外检查项的同时基于历史诊疗数据进行早期筛查, 可节省大量资源. 然而, 本研究未加入孕妇常见的化验数据, 未来将对 GRU-GBDT 模型结构进一步优化, 并添加化验项来提高整体预测效果.

参考文献

1 Mboya IB, Mahande MJ, Obure J, *et al*. Predictors of singleton preterm birth using multinomial regression models accounting for missing data: A birth registry-based cohort study in northern Tanzania. PLoS One, 2021, 16(4): e0249411. [doi: 10.1371/journal.pone.0249411]

2 Howson CP, Kinney MV, McDougall L, *et al*. Born too soon: Preterm birth matters. Reprod Health, 2013, 10Suppl1(S1): S1. [doi: 10.1186/1742-4755-10-S1-S1]

3 谢幸, 苟文丽. 妇产科学. 8 版. 北京:人民卫生出版社, 2013: 221.

4 张小松, 杨慧霞. 早产发生的影响因素及其流行病学研究

进展. 中华妇产科杂志, 2017, 52(5): 344–347. [doi: 10.3760/cma.j.issn.0529-567x.2017.05.013]

5 Kalengo NH, Sanga LA, Philemon RN, *et al*. Recurrence rate of preterm birth and associated factors among women who delivered at Kilimanjaro Christian Medical Centre in Northern Tanzania: A registry based cohort study. PLoS One, 2020, 15(9): e0239037. [doi: 10.1371/journal.pone.0239037]

6 Grantz KL, Hinkle SN, Mendola P, *et al*. Differences in risk factors for recurrent versus incident preterm delivery. American Journal of Epidemiology, 2015, 182(2): 157–167. [doi: 10.1093/aje/kwv032]

7 Zhang J, Pan M, Zhan WQ, *et al*. Two-stage nomogram models in mid-gestation for predicting the risk of spontaneous preterm birth in twin pregnancy. Archives of Gynecology and Obstetrics, 2020, 303(6): 1439–1449. [doi: 10.1007/s00404-020-05872-0]

8 Reicher L, Fouks Y, Yogev Y. Cervical assessment for predicting preterm birth—Cervical length and beyond. Journal of Clinical Medicine, 2021, 10(4): 627. [doi: 10.3390/jcm10040627]

9 Hezelgrave NL, Kuhrt K, Cottam K, *et al*. The effect of blood staining on cervicovaginal quantitative fetal fibronectin concentration and prediction of spontaneous preterm birth. European Journal of Obstetrics & Gynecology and Reproductive Biology, 2017, 208: 103–108. [doi: 10.1016/j.ejogrb.2016.11.027]

10 Glover AV, Manuck TA. Screening for spontaneous preterm birth and resultant therapies to reduce neonatal morbidity and mortality: A review. Seminars in Fetal and Neonatal Medicine, 2018, 23(2): 126–132. [doi: 10.1016/j.siny.2017.11.007]

11 Bhattad PB, Jain V. Artificial intelligence in modern medicine—The evolving necessity of the present and role in transforming the future of medical care. Cureus, 2020, 12(5): e8041. [doi: 10.7759/cureus.8041]

12 Koullali B, Oudijk MA, Nijman TAJ, *et al*. Risk assessment and management to prevent preterm birth. Seminars in Fetal and Neonatal Medicine, 2016, 21(2): 80–88. [doi: 10.1016/j.siny.2016.01.005]

13 Prema NS, Pushpalatha MP. Prediction of preterm birth using data mining—A survey. IIOAB Journal, 2019, 10(2): 13–17.

14 Włodarczyk T, Płotka S, Szczepański T, *et al*. Machine learning methods for preterm birth prediction: A review. Electronics, 2021, 10(5): 586. [doi: 10.33904/electronics10050586]

15 Koivu A, Sairanen M. Predicting risk of stillbirth and preterm pregnancies with machine learning. Health Information Science and Systems, 2020, 8(1): 14. [doi: 10.1007/s13755-020-00105-9]

16 Luo W, Huning EYS, Tran T, *et al*. Screening for post 32-week preterm birth risk: How helpful is routine perinatal data collection? Heliyon, 2016, 2(6): e00119. [doi: 10.1016/j.heliyon.2016.e00119]

17 Zhao K, Shao HD. Intelligent fault diagnosis of rolling bearing using adaptive deep gated recurrent unit. Neural Processing Letters, 2020, 51(2): 1165–1184. [doi: 10.1007/s11063-019-10137-2]

18 Ljubic B, Hai AA, Stanojevic M, *et al*. Predicting complications of diabetes mellitus using advanced machine learning algorithms. Journal of the American Medical Informatics Association, 2020, 27(9): 1343–1351. [doi: 10.1093/jamia/ocaa120]

19 Huan J, Li H, Li MB, *et al*. Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: A study of Chang Zhou fishery demonstration base, China. Computers and Electronics in Agriculture, 2020, 175: 105530. [doi: 10.1016/j.compag.2020.105530]

20 罗计根, 杜建强, 聂斌, 等. 基于双向 LSTM 和 GBDT 的中医文本关系抽取模型. 计算机应用研究, 2019, 36(12): 3744–3747. [doi: 10.3969/j.issn.1001-3695.2018.07.0420]

21 Dairi A, Harrou F, Zeroual A, *et al*. Comparative study of machine learning methods for COVID-19 transmission forecasting. Journal of Biomedical Informatics, 2021, 118: 103791. [doi: 10.1016/j.jbi.2021.103791]

22 赖晓莹, 钱俊. ARIMA-LSTM-XGBoost 加权组合模型在肺结核发病趋势预测的研究. 现代预防医学, 2021, 48(1): 5–9.

23 Chung J, Gulcehre C, Cho KH, *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv: 1412.3555v1, 2014.

24 Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural Computation, 1989, 1(2): 270–280. [doi: 10.1162/neco.1989.1.2.270]

25 蔡慧, 罗佳伟. ICU 病人数据对 LSTM 和 GRU 预测模型效果及收敛速度的对比分析. 绵阳师范学院学报, 2020, 39(11): 1–10. [doi: 10.16276/j.cnki.cn51-1670/g.2020.11.001]

26 Friedman J H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 2001, 29(5): 1189–1232. [doi: 10.2307/2699986]