

基于 Local Cascade Ensemble 方法的胎儿健康自动分类

黄梅佳¹ 李宗辉¹ 郑博伟¹
HUANG Meijia LI Zonghui ZHENG Bowei

摘要

为更好地自动评估胎儿宫内状态, 提出一种基于 local cascade ensemble (LCE) 方法的胎儿健康状态分类模型。选用 UCI 数据集, 使用 ADASYN 方法对不平衡数据集进行填充平衡, 接着结合随机森林算法对数据特征进行选择, 最后使用 LCE 方法对胎儿状态进行自动分类。实验结果表明, 所提出模型使用的方法平均准确率、精确率、召回率和 F1 分数分别达到了 0.955 4、0.905 4、0.955 7 和 0.929 0, 对比传统的机器学习算法能得到更好的分类效果, 有效降低了误判率。

关键词

机器学习; 胎儿监护; 自动分类; Local Cascade Ensemble

doi: 10.3969/j.issn.1672-9528.2024.04.027

0 引言

电子胎心宫缩监护 (CTG) 也被称为电子胎儿监测 (EFM)。EFM 于 1960 年左右首次推出, 并成为第一个使用心音图记录胎儿心率 (FHR) 的工具, 之后被多普勒信号取代, 信号质量显著提高^[1]。CTG 包括对 FHR 变化的无创记录 (使用多普勒超声技术), 并分析胎儿运动与母体子宫收缩之间的关系^[2-3]。由于视觉检查的主观性和 CTG 记录的复杂性, FHR 的解释以及 CTG 的分类受医生主观性和经验的差异影响表现出不一致性, 容易导致误诊率的上升^[4-5]。

因此, CTG 的计算机自动化分析显得尤为重要。近年来, 很多学者在 CTG 的智能分类方法上做了很多尝试。传统的 CTG 分析和解释方法, 如 FHR 记录的时域和频域分析以及非线性方法, 已被广泛使用多年, 许多研究人员在此基础上进行了比较和创新。例如, 国内外学者对 CTG 参数特征定量分析, 通过特征工程从原始特征集中挑选最优特征集, 再运用机器学习算法对胎儿健康状态进行智能分类。G. G. Georgoulas 等人利用从 FHR 信号中提取的一组参数和两个隐马尔可夫模型进行自动分类, 取得了较好的分类结果^[6]。Antoniya Georgieva 等人在使用主成分分析将特征空间降低后, 将 ANN 神经网络模型应用到 FHR 监测中, 得到了较好的分类效果^[7]。Zeng 等人提出了一种新的基于时频 (TF) 特征和代价敏感支持向量机 (ECSVM) 分类器进行胎儿健康分类, 提高了 CTG 异常分类的真阳性率, 降低了假阳性率, 有助于发现和治疗产程异常胎儿^[8]。同年, Chen 等人^[5]采用主成分

分析和可视化技术挖掘 CTG 数据的分布特征, 将随机森林 (RF)、加权随机森林 (WRF)、完全随机森林 (CRF) 和梯度提升决策树 (GBDT) 作为基本分类器, 在级联森林阶段进行深度迭代, 最终得到的模型在公共数据集的准确率值、平均 F1 值和曲线下面积 (area under the curve, AUC) 值分别为 92.64%、92.01% 和 0.990, 与传统机器学习模型、深度神经网络和最先进的 CTG 分类模型相比达到最优。

为了提高胎儿健康分类方法的准确率, 提出了一种基于 Local Cascade Ensemble (LCE) 方法的胎儿健康分类模型。首先, 对胎心数据集进行数据平衡和特征选择; 接着, 使用 LCE 方法用于胎儿健康分类识别, 与其它机器学习算法相比, 取得了较好的分类效果。

1 数据与方法

1.1 实验数据

本文研究选用的数据是 UCI 数据库中的 CTG 胎心数据集 (<http://archive.ics.uci.edu/dataset/193/cardiotocography>)^[9], 该数据集是胎心监护智能分类研究者常用的数据库之一。数据集中包括 2126 条从心电图检查中提取的 CTG 特征记录, 每条记录包含 21 个 CTG 特征 (如表 1 所示), 然后由三位产科专家进行分类和标签, 分为“正常”“可疑”“异常”三类, 其中正常 1655 例, 可疑 295 例, 异常 176 例。

表 1 UCI 数据集特征属性说明

序号	特征	定义
1	LB	FHR 基线 (每分钟心跳节拍)
2	AC	每秒加速次数
3	FM	每秒胎动次数
4	UC	每秒宫缩次数

1. 揭阳职业技术学院信息工程系 广东揭阳 522000

[基金项目] 2021 年揭阳职业技术学院科学研究项目: 胎心监护智能分析算法研究与实现 (项目编号: 2021JYCKY21)

表 1(续)

序号	特征	定义
5	DL	每秒轻微减速次数
6	DS	每秒严重减速次数
7	DP	每秒持续减速次数
8	ASTV	出现异常短变异的时间百分比
9	MSTV	短变异平均值
10	ALTV	出现异常长变异的时间百分比
11	MLTV	长变异的平均值
12	Width	FHR 直方图宽度
13	Min	FHR 直方图最小值
14	Max	FHR 直方图最大值
15	Nmax	直方图峰值数
16	Nzeros	直方图零值数
17	Mode	直方图众数
18	Mean	直方图平均值
19	Median	直方图中位数
20	Variance	直方图方差
21	Tendency	直方图趋势

1.2 数据平衡

使用的 UCI 数据集中，被评估为正常的记录占比为 77.84%，被评估为可疑和异常所占比例分别为 13.88% 和 8.28%，数据集分类极不平衡，不能直接用于分类模型训练。类不平衡是模式识别方法在处理现实生活问题时的一个常见问题，一些方法也相继被提出来解决这种情况，例如使用 SMOTE 与 ADASYN 过采样方法进行处理^[10-11]。

本文采用了 ADASYN 方法处理样本的类不平衡问题。原始数据集中的正常类、可疑类、异常类分别有 1655、295 和 176 个。采用 ADASYN 方法处理数据后，数据集由 2126 例变为 4940 例，其中正常类、可疑类、异常类分别有 1655、1639 和 1646 个，样本数趋于平衡。

1.3 特征选择

在生物医学领域中，特征选择是一种用于选择最相关和最有信息量的特征的方法。生物医学数据通常具有高维度和复杂性，因此特征选择在处理生物医学数据时非常重要。在生物医学数据具有大量的特征情况下，特征选择不仅可以减少维度，还能提高计算效率，提高模型的预测性能和泛化能力^[12]。特征选择在生物医学数据分析中具有重要的意义，可以帮助挖掘数据中的有用信息并构建更准确的预测模型。

本文选用的 UCI 数据集中一共包含 21 维的特征属性，采用随机森林算法^[13]进行特征选择。随机森林是一种集成学习方法，可以给出每个特征的重要性评估，衡量其对目标变量的贡献程度，从而选择出最重要的特征。算法通过组合多个决策树进行特征选择，减少了单个决策树的偏差和方差，提高了特征选择的可靠性。本文采用随机森林算法计算得到的各特征属性重要性分值如图 1 所示。

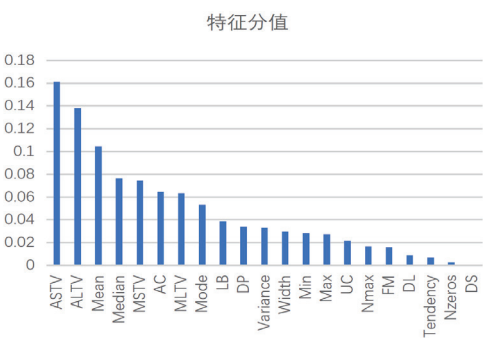


图 1 基于随机森林算法的特征重要性分值

随后对 21 个特征进行相关性分析并生成 Heatmap。通过生成特征之间的热力图，可以直观地展示不同特征之间的相关性。一些特征间存在较高的相关性意味着可能提供了相似或冗余的信息，可以考虑从中选择一个代表性的特征以减少计算复杂度，以降低模型过拟合的风险。生成的 HeatMap 如图 2 所示。从图 3 中可看出，特征 Median 与 LB、Median 与 Mode、Median 与 Mean、Mode 与 Mean 之间呈现强相关，结合上文随机森林算法对特征重要性的排名，去除 Media 特征，最终保留的特征子集共有 20 维特征。

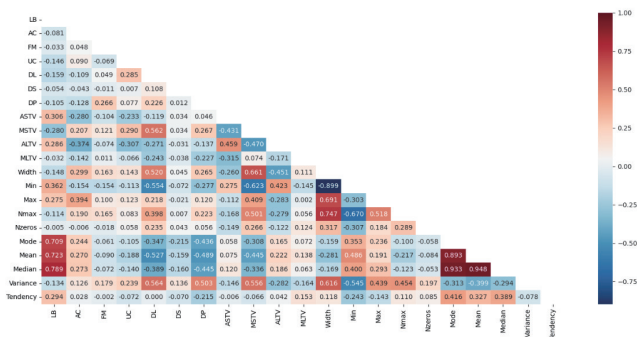


图 2 特征 Heatmap 图

1.4 Local Cascade Ensemble 分类模型构建

1.4.1 Local Cascade Ensemble 算法

Local Cascade Ensemble (LCE) 算法^[14]是一种集成学习算法，它通过构建级联结构来提高分类性能。LCE 算法包括多个层次的分类器，并通过级联方式进行决策。LCE 在局部应用级联泛化，采用了分治策略（决策树）方法，并通过将基于 boosting 的预测器作为基础学习器的方法减少决策树的偏差。LCE 算法的优点在于可以在每个层级上专注于解决特定的子问题，提高了分类器的泛化能力。通过级联结构的特性，可以减少错误传播，提高整体分类性能。算法相对简单，易于实现和理解。

具体的 LCE 算法模型构建过程如图 3 所示。LCE 算法的工作过程如下。

(1) 初始化：将训练数据集分为若干个子集，并在每个子集上训练一个基础分类器（使用当前性能最好的 boosting 算法） C_b 。从第一层开始，依次对每个层级应用以下步骤：

(a) 对于每个样本,使用当前层级的分类器进行预测;(b) 根据预测结果,将样本分配给其相应的子集,向输入数据集中添加新的属性,新的属性从分类器 C_b 中得到;(c) 在下一层级的子集上训练新的分类器,被添加到数据集的输出将被基础分类器视为加权方案,从而更多地关注到先前预测错误的样本。

(2) 集成决策:通过使用 Bagging 方法来减轻由梯度提升决策树产生的过拟合。Bagging 通过对数据集进行替换和随机抽样创建多个预测器,最后使用简单投票法聚合树。

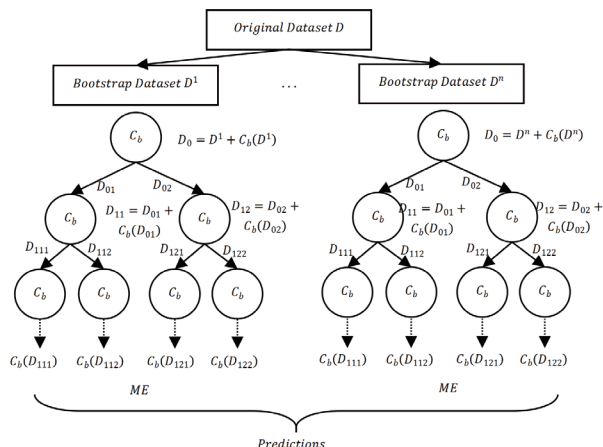


图 3 LCE 模型构建流程图

2 实验结果与分析

2.1 概述

基于 UCI 的 CTG 数据集,本文采用五折交叉验证法计算评估指标,对模型进行性能评估。为了评估 LCE 集成学习算法对 CTG 胎儿健康的分类能力,选用了其它机器学习算法,包括朴素贝叶斯 (NB)^[15]、K 最邻近分类 (KNN)^[16]、支持向量机 (SVM)^[17]、决策树 (DT)^[18]、随机森林 (RF)^[19] 和集成学习 Adaboost^[20] 对相同数据集进行预测,将得到的预测结果与 LCE 模型的预测结果进行对比,以此验证 LCE 模型的有效性。

2.2 评估标准

本文采用 4 个评估指标对 CTG 健康分类算法性能进行评估,分别为:准确率 (Accuracy),精确率 (Precision)、召回率 (Recall)、综合评价指标 (F_1 Score)。各指标计算公式为:

$$Accuracy = \frac{T_p + T_n}{T_p + F_p + F_n + T_n} \quad (1)$$

$$Precision = \frac{T_p}{T_p + F_p} \quad (2)$$

$$Recall = \frac{T_p}{T_p + F_n} \quad (3)$$

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \quad (4)$$

式中: T_p 表示正确样本被分类为正确的样本数, T_n 表示错误样本被归类为错误的样本数, F_p 表示错误样本被归类为正确的样本数, F_n 表示正确样本被归类为错误的样本数。通过以上评价指标的综合评估,可以分析该自动分类模型的整体性能,进一步验证所提出的胎儿健康分类模型的实际应用意义。

2.3 结果分析

使用 UCI 的 CTG 数据集得到的分类结果如表 2 所示。从表 2 中可以看出, LCE 模型相对于所列出的其它算法,正常类、可疑类、异常类和总体平均的 Accuracy、Precision、Recall 和 F_1 均达到最高,与表现较好的 RF 算法相比,平均 Accuracy 提高了 0.032 9,平均 Precision 提高了 0.055 36,平均 Recall 提高了 0.075 1,平均 F_1 提高了 0.046,而异常类的召回率在以上所有算法中同样达到最高,可以更好地避免将异常类误判为正常或可疑类,提高对异常情况的识别能力。

表 2 算法分类评价对比

评价指标		Accuracy	Precision	Recall	F_1
NB	正常		0.989 0	0.807 8	0.889 3
	可疑		0.513 5	0.890 6	0.651 4
	异常		0.488 4	0.724 1	0.583 3
	平均	0.814 6	0.663 6	0.807 5	0.708 0
KNN	正常		0.965 9	0.936 9	0.951 2
	可疑		0.714 3	0.781 3	0.746 3
	异常		0.727 3	0.827 6	0.774 2
	平均	0.906 1	0.802 5	0.848 6	0.823 9
SVM	正常		0.985 6	0.822 8	0.896 9
	可疑		0.469 9	0.609 4	0.530 6
	异常		0.415 4	0.931 0	0.574 5
	平均	0.798 1	0.623 6	0.787 7	0.667 3
DT	正常		0.978 2	0.942 9	0.960 2
	可疑		0.771 4	0.843 8	0.806 0
	异常		0.771 4	0.931 0	0.843 8
	平均	0.927 2	0.840 4	0.905 9	0.870 0
RF	正常		0.990 3	0.921 9	0.954 9
	可疑		0.694 1	0.921 9	0.791 9
	异常		0.871 0	0.931 0	0.900 0
	平均	0.922 5	0.851 8	0.924 9	0.882 3
Adaboost	正常		0.981 2	0.981 2	0.871 5
	可疑		0.467 3	0.781 3	0.584 8
	异常		0.415 1	0.758 6	0.536 6
	平均	0.781 7	0.621 2	0.774 6	0.664 3
LCE	正常		0.984 6	0.961 0	0.972 6
	可疑		0.852 9	0.906 3	0.878 8
	异常		0.878 8	1.0	0.935 5
	平均	0.955 4	0.905 4	0.955 7	0.929 0

3 结语

本文提出了一种基于 Local Cascade Ensemble (LCE) 方法的胎儿健康状态分类模型。针对 UCI 的 CTG 数据集不平衡问题,该模型使用 ADASYN 方法对数据集进行了填充,

接着使用随机森林算法,结合热力图对特征集进行筛选,最后使用LCE方法对胎儿健康状态进行自动分类。实验结果表明,本文使用的方法对比传统的机器学习算法能得到更好的分类效果。本文所构建的模型将机器学习技术与产前胎儿电子监护紧密结合,可为构建新型的产前智能胎儿监护系统提供有力支持,在远程医疗监护中也可得到更广泛的应用,带来更方便准确的远程胎儿监护服务,缓解地域差异带来的医疗资源不均衡问题。

本文在数据集的平衡处理、特征选择和数据集的规模上还存在改进空间。在后续研究中,一方面,需要对数据集不平衡问题探索更多元的方式以改善训练效果,并在特征选择方法上进行改进,筛选出与胎儿健康状态关联更紧密的特征集;另一方面,需要使用更多的CTG数据库优化识别效果,提高模型的健壮性,以期将算法应用到临床中,提高临床胎监的自动识别效果。

参考文献:

- [1] PINAS A, CHANDRAHARAN E. Continuous cardiotocography during labour: analysis, classification and management[J]. Best pract res clin obstet gynaecol, 2016, 30: 33-47.
- [2] 郭晓辉,陈敦金,漆洪波. 产前和产时电子胎心监护临床实践专家共识[J]. 中国实用妇科与产科杂志, 2022, 38(7): 714-725.
- [3] HAWHEEL T, BANGASH J. Volterra neural analysis of fetal cardiotocographic signals. In: 1st international conference on communications, signal processing, and their applications (ICCSPA). Piscataway: IEEE, 2013: 1-5.
- [4] FENG G, QUIRK G J, HEISELMAN C, et al. Estimation of consecutively missed samples in fetal heart rate recordings[C]//Estimation of Consecutively Missed Samples in Fetal Heart Rate Recordings. Piscataway: IEEE, 2021: 1080-1084.
- [5] CHEN Y, GUO A, CHEN Q, et al. Intelligent classification of antepartum cardiotocography model based on deep forest[J]. Biomedical signal processing and control, 2021, 67(2): 102555.
- [6] GEORGOULAS G G, STYLIOU C D, NOKAS G, et al. Classification of fetal heart rate during labour using hidden markov models[C]//2004 IEEE International Joint Conference on Neural Networks, V.3. Piscataway: IEEE, 2004: 2471-2475.
- [7] GEORGIEVA A, PAYNE S J, MOULDEN M, et al. Artificial neural networks applied to fetal monitoring in labour[J]. Neural computing & applications, 2013, 22(1): 85-93.
- [8] ZENG R, LU Y, LONG S, et al. Cardiotocography signal abnormality classification using time-frequency features and ensemble cost-sensitive SVM classifier[J]. Computers in biology and medicine, 2021, 130: 104466.
- [9] Campos D, Bernardes J. Cardiotocography [EB/OL]. UCI Machine Learning Repository, 2010. [2023-10-21]. <https://doi.org/10.24432/C51S4N>.
- [10] CHAWLA N V, JAPKOWICZ N, KOTCZ A. Editorial: special issue on learning from imbalanced data sets[J]. ACM SIGKDD explorations newsletter, 2004, 6: 1-6.
- [11] 徐玲玲,迟冬祥. 面向不平衡数据集的机器学习分类策略[J]. 计算机工程与应用, 2020, 56(24): 12-27.
- [12] 施启军,潘峰,龙福海,等. 特征选择方法研究综述[J]. 微电子学与计算机, 2022, 39(3): 1-8.
- [13] 方匡南,吴见彬,朱建平,等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.
- [14] FAUVEL K, FROMONT E, VÉRONIQUE M, et al. XEM: an explainable-by-design ensemble method for multivariate time series classification[J]. Data mining and knowledge discovery, 2022, 36(3): 917-957.
- [15] CHEN S, WEBB G I, LIU L, et al. A novel selective naïve Bayes algorithm[J]. Knowledge-based systems, 2020, 192: 105361.
- [16] ZHANG S, LI X, ZONG M, et al. Efficient kNN classification with different numbers of nearest neighbors[J]. IEEE transactions on neural networks and learning systems, 2017, 29(5): 1774-1785.
- [17] CERVANTES J, GARCIA L F, RODRÍGUEZ M L, et al. A comprehensive survey on support vector machine classification: applications, challenges and trends[J]. Neurocomputing, 2020, 408: 189-215.
- [18] SONG Y, YING L. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.
- [19] SPEISER J L, MILLER M E, TOOZE J, et al. A comparison of random forest variable selection methods for classification prediction modeling[J]. Expert systems with applications, 2019, 134: 93-101.
- [20] CAO Y, MIAO Q, LIU J, et al. Advance and prospects of AdaBoost algorithm[J]. Acta automatica sinica, 2013, 39(6): 745-758.

【作者简介】

黄梅佳(1993—), 通信作者(email: 1422494039@qq.com), 男, 广东揭阳人, 硕士, 助教, 高级工程师, 研究方向: 软件工程、人工智能。

李宗辉(1982—), 男, 山东郯城人, 硕士, 副教授, 研究方向: 网络工程、机器学习。

郑博伟(1995—), 男, 广东揭阳人, 硕士, 助教, 研究方向: 电子信息工程、图像识别。

(收稿日期: 2024-02-05)