

经皮冠状动脉介入治疗术后复发预警模型研究^{*}

王颖晶

倪连超

(上海交通大学医学院附属仁济医院信息中心 上海 200127)

(万达信息股份有限公司 上海 201112)

陈珊黎 邵维君 韩 刚 丁粉华 郑 涛

(上海交通大学医学院附属仁济医院信息中心 上海 200127)

〔摘要〕 提出以术后随访的常规无创检验结果数据为依据,机器学习算法为基础,对患者经皮冠状动脉介入治疗(Percutaneous Coronary Intervention, PCI)术后复发情况进行预警建模,详细阐述模型构建方法、实验评估及分析方法,并对结果进行分析。

〔关键词〕 经皮冠状动脉介入治疗术后复发;机器学习;随访数据

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.06.008

Study on a Prediction Model of Percutaneous Coronary Intervention Postoperative Recurrence WANG Yingjing, Information Center, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China; NI Lianchao, Wonders Information Co. Ltd., Shanghai 201112, China; CHEN Shanli, SHAO Weijun, HAN Gang, DING Fenhua, ZHENG Tao, Information Center, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China

〔Abstract〕 Based on conventional follow-up non-invasive test results and machine learning algorithms, an early warning modeling of Percutaneous Coronary Intervention (PCI) postoperative recurrence is proposed. The model construction method, experimental evaluation and analytical method are expounded in detail, and the results are analyzed.

〔Keywords〕 Percutaneous Coronary Intervention (PCI) postoperative recurrence; machine learning; follow-up data

1 引言

1.1 研究背景

经皮冠状动脉介入治疗(Percutaneous Coronary Intervention, PCI)已成为急性冠脉综合征(Acute Coronary Syndrome, ACS)主要且有效的治疗方法。随着PCI技术的广泛应用和疾病复杂多变性的增加,PCI术后冠脉再狭窄成为临床难点,也是影响患者预后的主要因素。目前机器学习已逐渐应用于

〔修回日期〕 2021-09-27

〔作者简介〕 王颖晶,助理工程师;通信作者:郑涛,正高级工程师。

〔基金项目〕 上海市信息化发展专项资金项目“面向仁济医院医联体的专病临床科研智能辅助决策平台建设”(项目编号:201901007)。

医学辅助诊断领域^[1-2]，机器学习与现代医学融合可用于医疗决策和辅助诊断，能够为临床医务人员提供帮助，为患者提供便利。

1.2 相关研究

当前关于 PCI 预后的研究主要集中在相关危险因素评估以及不良事件或并发症事件发生预测方面。程诚和王鑫^[3]基于行 PCI 术治疗冠心病合并 2 型糖尿病患者的血压、血糖、血脂、支架情况、Th1 及 Th2 细胞因子水平等，采用 Logistic 逐步回归模型分析 PCI 术后发生支架内再狭窄的影响因素。吴欢、薛万国和应俊等^[4]利用随机抽样和合成少数类过采样技术（Synthetic Minority Oversampling Technique, SMOTE）进行样本均衡化处理，再分别利用分类回归树算法（Classification And Regression Tree, CART）、C4.5 条件树，AdaBoost, Bagging, 随机森林（Random Forest, RF）和朴素贝叶斯算法（Naive Bayes）6 种机器学习方法构建 PCI 术后 12 个月主要不良心血管事件的风险预测模型。宋玲花、李保和韩慧媛^[5]对 PCI 术后患者病历资料、定期复诊资料和电话随访资料进行分析，建立 Cox 比例风险模型和预后指数方程，实现对具有不同危险因素的心肌梗死患者的 PCI 术后复发风险预测。

1.3 研究内容

上述模型均需要再次行有创经冠脉造影检查才能予以诊断 PCI 是否复发。针对这一问题，可尝试使用术后随访的常规无创性检验结果和患者基本信息等数据，运用机器学习算法对 PCI 术后复发进行预测。本文主要基于机器学习算法的二分类模型，构建一种多模型融合的 Stacking 算法，并通过医院真实样本数据对患者 PCI 术后的复发情况预警开展研究。

2 数据样本

2.1 数据来源

选取仁济医院 2015–2018 年期间收治、接受过冠状动脉支架植入术且具有完整记录的患者作为样

本集，并以该样本集的医疗电子病历信息数据作为数据源开展研究。

2.2 纳入和排除标准

根据专家知识制定 PCI 术后复发患者纳入和排除标准，在心内科住院和门诊数据中存在多条手术记录的患者即为复发患者，而仅有 1 条手术记录且此次手术时间距离 2021 年已超过 3 年的患者被认为没有复发的患者。基于该标准共采集复发患者即正样本 402 条，没有复发患者即负样本 2 509 条，正负样本共计 2 911 条。通过患者唯一标识，关联得到患者的基本信息、门诊和住院信息以及检查、检验项目等作为分析因素。其中针对多次手术的患者取其最后一次手术信息；针对多次检查和检验的情况以手术时间作为参考时间点，手术前 3 个月内是否存在检验检查记录作为纳入标准，取该时间段内最后一次检验检查项目作为采纳分析数据。

2.3 分析因素

分析因素主要包括患者基本信息、临床就诊信息以及检验检查资料等，通过医院信息系统（Hospital Information System, HIS）中的患者病历及其门诊和住院检查数据获得，主要包含以下 3 个方面，共计 263 个特征。一是检验检查特征：住院和门诊中常见的常规检验检查项目，如全血细胞分析、尿液分析、血气分析等 258 个特征。二是患者基本特征：年龄、性别、血型 3 个特征。三是手术相关特征：手术完成时间、手术进行总时长两个特征。

3 基于机器学习的 PCI 术后复发预警模型

3.1 机器学习有监督分类方法

3.1.1 常用方法 基于上述分析因素借助机器学习方法构建 PCI 术后复发预警模型。通过纳入和排除标准确定复发患者和未复发患者，得到明确的样本类别，因此可构建有监督分类方法进行 PCI 术后复发预警。当前机器学习有监督分类方法较多，常用方法包括逻辑回归（Logistic Regression, LR），支持向量机（Support Vector Machine, SVM），K 近邻

(K - Nearest Neighbors, KNN), RF, 极限梯度提升 (Extreme Gradient Boosting, XGB), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost) 等。其中 RF、XGB、LightGBM 和 CatBoost 均为以决策树为基模型构造的集成学习方法, 适合结构化数据建模并具有较强的泛化能力, 而 LR、SVM 以及 KNN 与前几种方法具有较大差异, 构造原理也不同。

3.1.2 各类方法对比分析 逻辑回归受制于模型的线性结构, 无法对非线性信息进行准确刻画, 同时也无法抽取多个特征间的组合关系, 导致使用者必须人工设计复杂特征集合来提升模型预测精度。支持向量机在大规模数据集上难以应用, 当样本数量较大时模型训练会占用大量内存和运算时间。K 近邻的缺点也较为明显, 首先其在预测过程中运算量较大, 需要计算到每个训练样本的距离, 其次模型可解释性较差, 再次模型鲁棒性较差, 容易受到错误数据的影响。随机森林有很强的实用性, 但是在处理某些噪声较大的数据时会出现过拟合。XGBoost、LightGBM 和 CatBoost 均为 Boosting 算法梯度下降树的优化算法。但是这些方法使用门槛较高、模型参数多, 在进行参数优化时较为复杂。

3.2 模型构建方法

为提升 PCI 术后复发预测精度, 基于上述 7 种方法构建一种多模型融合的 Stacking 算法。Stacking 算法能够结合多种不同类型的有监督分类方法作为基学习器, 并在此基础上再构建一个分类器, 预测未知样本类别。该方法最显著的优点是能够明显降低模型方差从而有效提升模型预测精度; 此外能够纳入任何有监督分类方法作为基分类器, 不受方法本身构造原理的约束。因此该方法在实际应用中对计算资源要求更高, 针对同等规模的数据集, 模型训练时间也数倍于单个模型。具体的 Stacking 算法主要分为两层, 第 1 层是由 7 种不同的基学习器组成, 包括 LR、SVM、KNN、RF、XGB、LightGBM 和 CatBoost。第 2 层是将第 1 层输出作为新特征输入到元学习器并进行训练, 形成最终的预测模型, 见图 1。

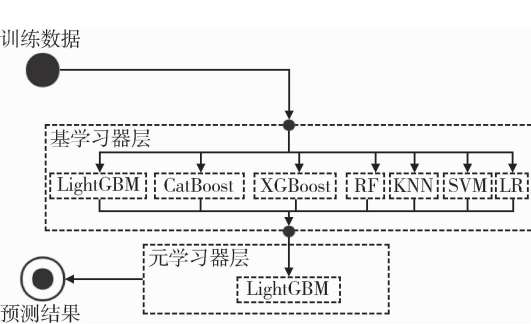


图 1 Stacking 算法架构

4 实验与结果分析

4.1 实验环境和数据预处理

4.1.1 实验环境 为探究基于机器学习的 PCI 术后复发预警模型的有效性, 基于纳入的 2 911 条样本展开实验和分析。借助 Python 及其开源的机器学习框架 scikit - learn 等^[6-8]实现上述 8 种算法。

4.1.2 数据预处理 由于部分方法对特征分布非常敏感, 为更准确地评估各算法的性能表现, 需要对数据进行预处理。首先, 对数据集中每个特征均进行标准化处理, 以此消除特征的量纲。其次, 考虑到除 XGB、LightGBM、CatBoost 3 个较新算法外, 其他传统算法均无法对缺失值进行合适处理, 因此对数据中的缺失值以 -1 数值进行填充。最后, 为了模型训练过程中更加平稳, 防止正负样本严重失衡对模型性能损伤, 对原始数据集中的负样本进行下采样处理, 最终正负样本比例为 1:2。

4.2 实验评估方法

为更好地评估算法表现及其泛化性能, 使用五折交叉验证方式进行模型验证。保证正负样本比例在训练与测试过程中保持一致, 并采用随机划分方式, 以避免人为因素干扰。针对非均衡数据集, 模型评估衡量主要采用受试者操作特征曲线下面积 (Area Under the Curve, AUC), 真正率 (True Positive Rate, TPR), 假正率 (False Positive Rate, FPR) 和 F1 值, 同时为了避免随机划分可能引入的偏差, 以独立的随机划分 5 次后的评估均值为最终结果评价衡准。各项评估指标的具体计算公式为:

$$TPR = \frac{TP}{FN + TP} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \tag{5}$$

其中 TP 、 TN 、 FP 、 FN 分别表示正样本被正确预测、负样本被正确预测、正样本被错误预测、负样本被错误预测的样本数。

4.3 实验结果分析

模型参数通过多次实验效果比较进行选取，见表 1。对于其他未列举的模型参数，其取值均采用计算框架中的默认值。对比本文的 Stacking 算法以及其他 7 种有监督分类方法开展实验情况可见，本文构造的 Stacking 算法性能最佳，AUC 达到 0.901，TPR 达到 78.2%，FPR 为 14.1%， $F1$ 值为 0.744。

表 1 算法与其参数的取值范围

算法	参数	取值
K 最近邻 (KNN)	n_ neighbors	5
	probability	True
	n_ estimators	200
随机森林 (RF)	max_ depth	6
	learning_ rate	0.1
	subsample	0.8
	colsample_ bytree	0.8
LightGBM (LGB)	max_ depth	-1
	learning_ rate	0.1
	feature_ fraction	0.8
	bagging_ fraction	0.8
	min_ data_ in_ leaf	10
CatBoost (CaB)	max_ depth	6
	l2_ leaf_ reg	3
	border_ count	254

4.4 指标因素分析

4.4.1 分析方法 鉴于模型输入变量较多，针对这些指标因素进行重点评估和筛选具有重要研究意义。通过筛选对模型具有重要影响的变量，挖掘对疾病具有提示性的因素，能够为临床研究者提供更多依据以及新的研究思路。为了对建模指标因素进行解读，选取预测精度较高的 Xgboost、LightGBM

和 CatBoost 3 种模型，分别计算指标因素的特征重要性权重。对于这 3 种基于决策树的模型，由于树模型在每一折交叉验证中均对训练集学习得到一个预测模型，同时也得到不同特征的重要性权重，因此将每一折的结果进行平均即可得到其针对整体数据特征重要性的评估。

4.4.2 分析结果 通过计算特征重要性得到 3 种模型各自排名前 20 的特征重要性权重。结果显示载脂蛋白 B 对于 PCI 术后再狭窄的预测评估至关重要，电导率、标准碳酸氢根浓度、促甲状腺激素、糖化血红蛋白 A1C 等指标也有显著影响。因此后期随访中相关检验检查如血脂全套、尿液分析、血气分析、甲状腺功能 5 项等需要重点关注。

5 结语

通过机器学习有监督分类模型，结合检验检查、患者基本信息和手术信息，对 PCI 术后复发进行预测，构建一种基于 Stacking 算法的 PCI 术后复发预测方法，实验结果显示该方法的 AUC 达到 0.901、TPR 达到 78.2%、FPR 低至 14.1%， $F1$ 值达到 0.744。此外特征重要性排序结果显示载脂蛋白 B、电导率、标准碳酸氢根浓度等指标是对 PCI 术后再狭窄预测评估的关键因素，对 PCI 术后再狭窄有重要的提示意义，从而为 PCI 术后复发预测的进一步研究和相关医疗资源的合理使用提供参考。尽管相较于单个基模型，Stacking 算法能够降低模型方差，较大地提升模型预测精度，但该方法存在几点不足有待改进。第一，当建模数据较多时，该方法相比单个模型需要更长的训练和预测时间；第二，由于集成多个模型的预测结果作为特征数据，该方法缺乏对原始指标因素的解释能力，仅凭借该方法难以对临床指标进行分析和筛选；第三，该方法的模型构建较为复杂，考虑到基模型数量较多且每种模型都有多个超参数，模型参数的设定和优化是需要多次迭代和调试进而达到最优，该过程较为繁琐、耗时。针对上述问题，在未来工作中将进一步优化方法流程，在保证模型泛化能力的基础上提升模型训练效率，增强模型的可解释性及易用性。

(下转第 71 页)

需要进一步研究。

参考文献

- 1 国务院办公厅. 关于推进分级诊疗制度建设的指导意见 [J]. 中华人民共和国国务院公报, 2015 (27): 27-31.
- 2 谢桦, 曹剑锋, 汪森然. 区域医疗信息互联互通跨省应用探索 [J]. 中国卫生信息管理杂志, 2020, 17 (5): 565-569, 688.
- 3 丁乙芳, 李群, 吴银姬. 县域医共体医疗同质化管理实践与思考 [J]. 中国农村卫生事业管理, 2021, 41 (4): 283-286.
- 4 施咏月, 鲍瀛, 张晓平. 基于医联体的远程医学平台建设及应用 [J]. 医学信息学杂志, 2020, 41 (6): 58-62.

- 5 黄发云, 陈以寒. 医联体背景下远程医疗应用现状研究 [J]. 现代医院管理, 2020, 18 (6): 36-39.
- 6 孙煜, 谢丽娟, 李文敏, 等. 同质化视角下我国县域医疗共同体建设的探讨及思考 [J]. 中国卫生事业管理, 2021, 38 (5): 331-333, 338.
- 7 谢子秋, 夏海朋, 李刚, 等. 多院区医院同质化医疗质量管理体系构建与思考 [J]. 医学与社会, 2021, 34 (5): 84-87.
- 8 李为民, 宋张骏, 翁国星, 等. 优质医疗资源扩容要在同质化上下功夫 [J]. 中国卫生, 2021 (4): 89-91.
- 9 沈婉婉, 高欢, 杜杏利, 等. 基于新医改背景下同济医院构建医联体的实践与思考 [J]. 中国医院, 2021, 25 (4): 19-21.

(上接第 39 页)

- 28 邵尤伟. 基于深度学习的智能舌诊方法研究 [D]. 厦门: 厦门大学, 2018: 27-62.
- 29 陆萍. 基于图像处理和模式识别的齿痕舌象研究 [D]. 上海: 华东理工大学, 2014: 11-50.
- 30 宋超, 王斌, 许家伦. 基于深度迁移学习的舌象特征分类方法研究 [J]. 计算机工程与科学, 2021, 43 (8): 1488-1496.
- 31 龚一萍, 陈海燕, 卜佳俊, 等. 常见病理舌形的定量研究及与高血压病和证型相关性的分析 [J]. 中华中医药

杂志, 2005, 20 (12): 730-731.

- 32 邱童. 基于深度学习与多特征融合的舌象诊断算法 [J]. 现代信息科技, 2020, 4 (1): 63-65.
- 33 Shamim M, Syed S, Shiblee M, et al. Automated Detection of Oral Pre-Cancerous Tongue Lesions Using Deep Learning for Early Diagnosis of Oral Cavity Cancer [J]. The Computer Journal, 2022, 65 (1): 91-104.
- 34 刘召婕. 基于感兴趣区域的舌象特征提取与病理分析 [D]. 哈尔滨: 哈尔滨工业大学, 2018: 26-51.

(上接第 43 页)

参考文献

- 1 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述 [J]. 模式识别与人工智能, 2014, 27 (4): 327-336.
- 2 兰欣, 卫荣, 蔡宏伟, 等. 机器学习算法在医疗领域中的应用 [J]. 医疗卫生装备, 2019, 40 (3): 101-105.
- 3 程诚, 王鑫. 冠心病合并 2 型糖尿病患者 PCI 术后支架内再狭窄危险因素分析 [J]. 中国循证心血管医学杂志, 2018, 10 (2): 196-198, 202.
- 4 吴欢, 薛万国, 应俊, 等. 基于机器学习方法的 PCI 术预后主要不良心血管事件预测模型研究 [J]. 中国数字医学, 2018, 13 (8): 2-5.
- 5 宋玲花, 李保, 韩慧媛. 经皮冠状动脉介入治疗术后冠

心病患者复发风险的预测与评价 [J]. 中国药物与临床, 2013, 13 (3): 311-313.

- 6 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python [J]. JMLR, 2011 (12): 2825-2830.
- 7 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System [C]. San Francisco: SIGKDD International Conf. on Knowledge Discovery and Data Mining 2016 (SIGKDD 2016), 2016.
- 8 Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [C]. Long Beach: Advances in Neural Information Processing Systems 2017 (NIPS 2017), 2017.