·综述·

基于机器学习的决策树算法在血流感染预后预测中应用现状及展望

范帅华[1] 郭伟[1] 郭军[2]

【摘要】血流感染作为一种严重的全身感染，近年来其患病率逐步升高，是造成患者不良预后的主要成因之一，因此早期识别不良预后的高危患者并及时进行干预尤为重要。但对血流感染预后预测的传统统计学分析在信度及效度上无法满足临床需求，鉴于机器学习算法已在一些临床疾病的预测模型构建中取得了良好的应用效果，展示了其提升临床诊疗精准性的应用前景，本文主要针对其中的决策树算法在血流感染预后预测方面的应用现状进行综述，通过比较其与传统方法的优缺点，对其在血流感染预后预测方面的应用前景进行展望，旨在探索更好的预测方式，以便于临床早期识别高危患者，最大程度降低血流感染的病死率。

【关键词】人工智能；机器学习；决策树；血流感染；预后预测

Current situation and prospect of application of decision tree algorithm based on machine learning in prognosis prediction of bloodstream infection

Fan Shuaihua[1], Guo Wei[1], Guo Jun[2]. [1]School of Clinical Medicine, Tsinghua University, Beijing 100084, China; Department of Respiratory and Critical Care Medicine, Beijing Tsinghua Changgung Hospital Affiliated with Tsinghua University, Beijing 102218, China; [2]Department of Department of Geriatric Medicine, Beijing Tsinghua Changgung Hospital Affiliated with Tsinghua University, Beijing 102218, China; School of Clinical Medicine, Tsinghua University, Beijing 100084, China

Corresponding author: Guo Jun, Email: junguo_med@tsinghua.edu.cn

【**Abstract**】As a serious systemic infection, the prevalence of bloodstream infection has gradually increased in recent years, which is one of the main causes of poor prognosis of patients, so it is particularly important to identify high-risk patients with poor prognosis early and timely. However, the traditional statistical analysis of bloodstream infection prognosis prediction can not meet the clinical needs in terms of reliability and validity, and since machine learning algorithms have achieved good application results in the construction of prediction models for some clinical problems, showing their application prospects to improve the accuracy of clinical diagnosis and treatment, this paper mainly reviews the application status of decision tree algorithm in the prognosis prediction of bloodstream infection, and prospects its application in the prediction of bloodstream infection prognosis by comparing its advantages and disadvantages with traditional methods. This review aims to explore better predictive methods for early clinical identification of high-risk patients and minimize the mortality rate of bloodstream infections.

【**Key words**】Artificial intelligence; Machine learning; Decision tree; Bloodstream infection; Prognosis evaluation

血流感染（bloodstream infection，BSI）定义为全身感染的患者血液培养呈阳性，可能是继发于原发部位明确的感染，或者来源未定；其中败血症是指致病细菌侵入血液循环中生长繁殖引起的急性全身性感染，属于血流感染中较为严重的一类，病死率可达30%～50%。近年来，由于静脉导管留置、机械通气和静脉营养等侵入性设备及治疗的广泛应用，以及免疫抑制剂、广谱抗菌药物的广泛使用，血流感染的发病率呈逐年上升趋势[1]。研究表明，血流感染约占社区获得性败血症和脓毒性休克病例的30%（396/1 341），约占ICU获得性败血症和脓毒性休克病例的40%（653/1 641），此外，血流感染发生与不良预后相关[2]，2008年有近36 000例患者死于血流感染[3]。因此，早期识别

DOI：10.3877/cma.j.issn.1674-1358. 2023. 05. 001

基金项目：北京市卫生健康委员会高层次公共卫生技术人才建设项目培养计划（No. 学科带头人-02-06）

作者单位：100089 北京，清华大学临床医学院，清华大学附属北京清华长庚医院呼吸与危重症医学科[1]；102218 北京，清华大学附属北京清华长庚医院老年医学科，清华大学临床医学院[2]

通信作者：郭军，Email：junguo_med@tsinghua.edu.cn

血流感染人群中预后不佳的患者并针对性治疗尤为重要。

目前，关于血流感染不良预后人群早期识别的研究主要基于回顾性资料、使用统计学方法的危险因素分析[4-6]，使用如Logistics回归、Cox回归和卡方检验等统计学方法，通过计算各危险因素的卡方值、*P*值、*OR*值及95%置信区间等，得到血流感染不良预后相关性较高的危险因素，进行针对性预防。但此类研究有一定局限性：①统计学方法通常计算量较大，当加入新的样本时往往需要重新计算全部数据，工作量较大，因此往往不会重复加入新的样本信息优化危险因素判定；②传统的统计学方法可用于构建统计模型，总结数据规律，但其重点在于刻画自变量与因变量之间的关系，而不是对未来情况进行预测。机器学习的运算目的是构建一个可重复预测的模型，侧重点在于应用而非机制解释，有更强的实践性。

由于传统的统计学方法具有以上局限性，医学研究者们迫切需要寻找一种新的分析方法来预测血流感染不良预后。这时，基于人工智能的机器学习算法进入了研究者们的视线，现针对其研究进展综述如下。

一、人工智能

人工智能（artificial intelligence，AI）是一门集心理认知、机器学习、情感识别、人机交互、数据保存及决策等功能于一身的学科[7]。目前人工智能的发展领域主要包括机器学习、计算机视觉、专家系统和语音识别等方面（图1），随着人工智能在安全防御系统（如人脸识别和指纹识别）、社交软件（语音识别和文字转换）等方面的应用，其强大的分类、识别和预测能力开始被关注，人们开始探索人工智能在医学上的应用。

Jiao等[8]进行了一项研究，研究对象为2020年3月9日至7月20日于宾夕法尼亚大学医院急诊部就诊的新型冠状病毒肺炎（corona virus disease 2019，COVID-19）患者共1 834例，收集患者胸部影像学数据，使用深度神经网络提取重要的影像学特征。对比使用深度神经网络提取出的影响因素和不使用这些影响因素时对疾病预测的准确性，并与放射科医生根据经验分析出的疾病严重程度进行了比较，结果显示在COVID-19患者中，人工智能对疾病严重程度的评估相较于单纯使用临床数据或仅依靠放射科医生的临床经验的判断更加准确。

二、机器学习

机器学习（machine learning，ML）是人工智能的核心子领域[9]，被誉为“机器学习之父”的Arthur Samuel将其定义为“一个赋予计算机学习能力的领域，这种学习能力不是通过显著式编程获得的”。其使算法或分类器能够学习大型复杂数据集中的模式并生成有用的预测输出[10-11]。机器学习分为有监督学习、无监督学习、半监督学习及强化学习（图2）。在已知数据输出（经过标注的）的情况下对模型进行训练，根据输出进行调整、优化的学习方式称为有监督学习；在没有已知输出的条件下，仅根据输入信息的相关性，进行类别的划分的学习方式称为无监督学习。目前常用的机器学习算法包括K近邻算法（K-nearest neighbor，KNN）、逻辑回归（Logistic regression，LR）、决策树（decision tree，DT）、朴素贝叶斯（naive Bayes，NB）、支持向量机（support vector machine，SVM）、卷积神经网络（convolutional neural network，CNN）等。

机器学习的分类算法已经被用于临床问题的分析及预测，并取得了较好的预测结果。Abdulaal等[12]回顾性分析2020年2月2日至2020年4月22日在西伦敦教学医院住院COVID-19患者共398例，使用人工神经网络（artificial neural network，ANN）对收集到的数据进行分析，提取与COVID-19患者出现死亡结局相关度较高的因素，构建入院死亡风险评估系统预测此类患者住院期间死亡风险。结果显示预测模型的受试者工作特征（receiver operating characteristic，ROC）曲线下面积（area under curve，AUC）可达到0.9012，为拟合度优等的预后预测模型，揭示了机器学习算法在病毒感染性疾病预后预测中的应用价值。

三、决策树算法

决策树是一种常见的机器学习方法，属于有监督学习

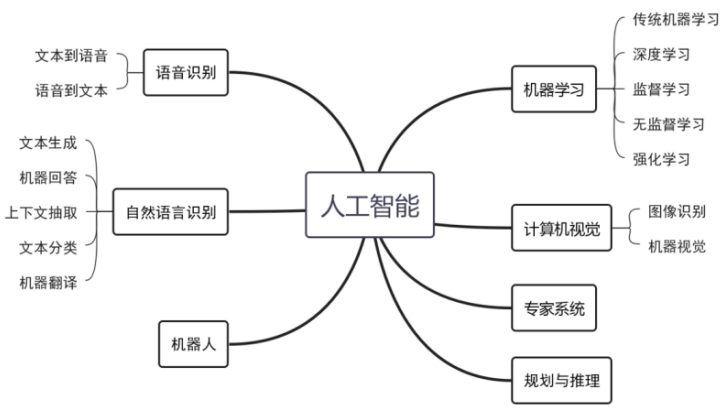

图1　人工智能分支领域概述

的一种[13]，其核心思想是相同（或相似）的输入产生相同（或相似）的输出，通过树状结构来进行决策，其目的是通过对样本不同属性的判断、决策，将具有相同属性的样本划分到同1个叶子节点中，从而实现分类或回归。一般来说，1棵决策树包含1个根节点、若干个内部节点和若干个叶子节点（图3）。叶子节点对应最终的决策结果，其他每个节点则对应与1个属性的测试。最终划分到同一个叶子节点上的样本，具有相同的决策属性，可对这些样本的值求平均值来实现回归，对这些样本进行“投票”（选取样本数量最多的类别）实现分类。

分类决策树是一种在二分类问题中常用且较为高效的算法模型[14]，常使用的方法包括ID3算法、C4.5算法及CART算法[15]，其中ID3算法使用信息增益作为标准来划分属性；C4.5算法使用增益率作为标准来划分属性；而CART分类算法使用基尼系数来选择划分属性（选择属性时，选择划分后基尼值最小的属性作为最优属性）。

在决策树模型的基础上衍生出一些集合模型，通过迭代多棵树来共同决策，从而产生1 + 1 ＞ 2的效果。此类集成模型包括随机森林（random forest，RF）、梯度提升（gradient boosting，GBDT）、极限梯度提升（extreme gradient boosting，XGBoost）和自适应提升（adaptive boosting，Adaboost）等。

四、决策树算法在血流感染预后预测中的应用现状

Zoabi等[16]进行了一项回顾性队列研究，收集自2014年1月至2020年1月就诊于以色列特拉维夫医学中心血培养阳性（仅细菌）的住院患者信息，使用决策树的递增算法Light GBM（light gradient boosting machine）构建血流感染的预后预测模型。获得革兰染色结果后进行包容模型构建，产生的AUC_{ROC}为0.82，准确召回率曲线（precision recall curve，PRC）为0.65。在紧凑模型中，预测仅基于培养时可用的25个特征，不包括革兰染色结果；AUC_{ROC}为0.81，PRC为0.63。研究中使用查尔森合并征指数（Charlson Comorbidity Index，CCI）作为血流感染病死率的标准评分系统，结果显示两种模型的表现均优于CCI指数。

Roimi等[17]基于美国马萨诸塞州波士顿的Beth Israel Deaconess医疗中心（Beth Israel Deaconess Medical Center，

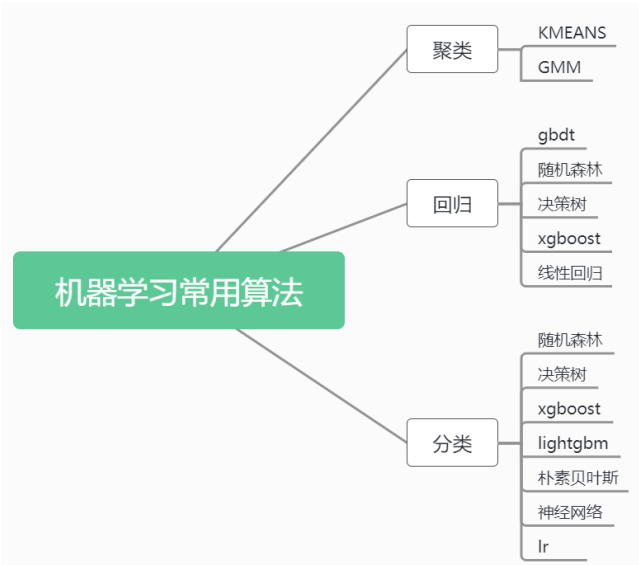

图2　机器学习常用算法概括

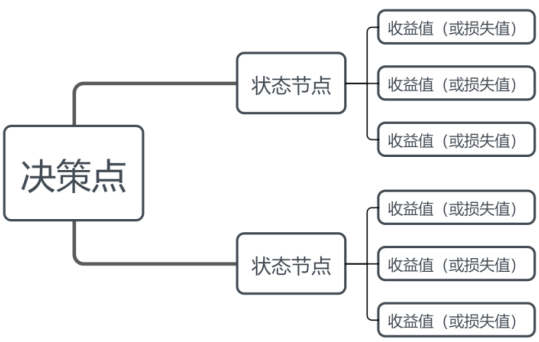

图3　决策树的决策模型概括

BIDMC）和以色列海法Rambam医疗保健园区（Rambam health care campus，RHCC）的患者电子健康记录，使用XGBoost算法构建ICU获得性血流感染（bloodstream infectious，BSI）早期诊断的预测模型。结果显示，交叉验证AUC_{ROC}的平均值在BIDMC中为（0.87 ± 0.02），在RHCC中为（0.93 ± 0.03）。内部验证显示，两个中心的AUC_{ROC}分别为（0.89 ± 0.01）和（0.92 ± 0.02），外部验证结果不甚理想，模型的AUC_{ROC}分别恶化至（0.59 ± 0.07）和（0.60 ± 0.06）。模型在交叉验证及内部验证阶段均取得不错的效果，但外部验证结果不是很理想，提示模型的外延能力较差，有待进一步研究。

Hu等[18]使用MIMIC-Ⅳ数据库，纳入了所有在ICU再入院前18 h或ICU再入院后前24 h内被诊断为败血症的成年人（＞ 24岁），分别构建LR、KNN、DT、RF、AdaBoost、NB、线性判别分析（linear discriminant analysis，LDA）、CNN和XGBoost模型，模型评估显示RF模型实现了最佳性能，验证队列中的AUC_{ROC}为0.81，准确率为85%，精度为62%。研究表明，基于机器学习方法、使用常规收集的临床数据准确预测ICU再入院脓毒症患者的病死率是可能的，证实了应用人工智能预测脓毒症患者病死率的重要作用。

Li等[19]收集2013年1月至2018年1月侵袭性念珠菌感染合并细菌性血流感染的246例住院患者信息，使用RF、LR和SVM算法来构建预测模型。结果显示，RF、LR、SVM模型的准确率分别为78.4%、71.6%和62.2%，RF、LR和SVM模型的AUC_{ROC}分别为0.919、0.753和0.777，因此，RF模型在此研究中获得了最佳预测效能。

Gao等[20]收集自2013年1月至2018年1月就诊于中国医科大学第一附属医院的367例合并念珠菌血症住院患者的临床资料，分别使用RF、LR及SVM算法构建念珠菌血症患者的预后预测模型。结果显示，RF模型具有较好的AUC_{ROC}为0.8505（0.7443，0.9567），且其预测准确性高于LR模型及SVM模型。这些研究表明机器学习的分类算法在临床应用的可行性，可以更直观看到机器学习算法在临床预测方面的应用前景。

五、展望

随着人工智能热潮的不断推进，机器学习、深度学习和神经网络等词汇开始涌入人们的视线。相较于传统的研究方法，使用机器学习算法对血流感染预后预测具有以下优点：①机器学习算法是传统统计学方法的外延，通常仅需要几行代码，就可以提取样本中与血流感染预后相关度较高的影响因素，并对这些因素进行赋值，用于血流感染预后预测，大大节省了计算时间；②可不断加入新的样本对构建好的预测模型进行优化，使其对血流感染预后的预测更加精准。同时解决了传统统计方法对不同样本差异较大的缺点，可以将不同样本全部纳入模型进行分析；③机器学习算法可以得到完整的预测模型，将其封装后可以作为预测工具，输入相关影响因素的数值，即可对此患者是否易发生不良预后进行预测，相较于传统统计学方法更加客观。

目前，国内关于机器学习算法在血流感染方面应用的研究较少，国外已发表的研究证实机器学习算法在血流感染预后预测方面的科学性及可行性。不难发现，基于机器学习的算法模型较于传统统计学方法更加简便、程序化，准确率更高[21-23]，更加适合构建预测模型来解决临床问题，因此，期待更多此类的研究来探索机器学习算法模型在血流感染预后预测方面的应用前景。

参 考 文 献

[1] 翟如波, 李云慧, 孙跃岭, 等. 某院连续三年医院血流感染病原菌分布特征及耐药性分析[J/CD]. 中华实验和临床感染病杂志(电子版),2016,10(1):36-40.

[2] Timsit JF, Ruppé E, Barbier F, et al. Bloodstream infections in critically ill patients: an expert statement[J]. Intensive Care Med,2020,46(2):266-284.

[3] Miniño AM, Murphy SL, Xu J, et al. Deaths: final data for 2008[J]. Natl Vital Stat Rep,2011,59(10):1-126.

[4] 刘小婷, 杨欢, 姚红, 等. 碳青霉烯类耐药肺炎克雷伯菌感染死亡风险预测模型的建立及其对患者预后的预测价值研究[J]. 中国全科医学,2020,23(30):3789-3797.

[5] 张安汝, 王启, 周朝娥, 等. 碳青霉烯类耐药肠杆菌目细菌院内感染危险因素和临床预后分析[J]. 中华医学杂志,2021,101(21):1572-1582.

[6] 游锦燕, 王素萍. 老年住院患者院内感染危险因素分析[J]. 中国卫生统计,2020,37(2):281-283.

[7] 李玉环. 人工智能综述[J]. 科技创新导报,2016,13(16):77-78.

[8] Jiao Z, Choi JW, Halsey K, et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study[J]. Lancet Digit Health,2021,3(5):e286-e294.

[9] Goodswen SJ, Barratt JLN, Kennedy PJ, et al. Machine learning and applications in microbiology[J]. FEMS Microbiol Rev,2021,45(5):1-19.

[10] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects[J]. Science,2015,349(80):255-260.

[11] Noble WS. What is a support vector machine?[J]. Nat Biotechnol, 2006,24(12):1565-1567.

[12] Abdulaal A, Patel A, Charani E, et al. Prognostic modeling of COVID-19 using artificial intelligence in the United Kingdom: Model development and validation[J]. J Med Internet Res,2020,22(8):e20259.

[13] 申明尧, 韩萌, 杜诗语, 等. 数据流决策树集成分类算法综述[J]. 计算机应用与软件,2022,39(9):1-10.

[14] 林文怡, 宛小燕, 刘元元. 常见新近决策树算法及其在卫生领域中的应用[J]. 现代预防医学,2019,46(23):4233-4237, 4242.

[15] 贾涛, 韩萌, 王少峰, 等. 数据流决策树分类方法综述[J]. 南京师大学报(自然科学版),2019,42(4):49-60.

[16] Zoabi Y, Kehat O, Lahav D, et al. Predicting bloodstream infection outcome using machine learning[J]. Sci Rep,2021,11(1):20101.

[17] Roimi M, Neuberger A, Shrot A, et al. Early diagnosis of bloodstream

infections in the intensive care unit using machine-learning algorithms[J]. Intensive Care Med,2020,46(3):454-462.

[18] Hu C, Li L, Li Y, et al. Explainable machine-learning model for prediction of in-hospital mortality in septic patients requiring intensive care unit readmission[J]. Infect Dis Ther,2022,11(4):1695-1713.

[19] Li Y, Wu Y, Gao Y, et al. Machine-learning based prediction of prognostic risk factors in patients with invasive candidiasis infection and bacterial bloodstream infection: a singled centered retrospective study[J]. BMC Infect Dis,2022,22(1):150.

[20] Gao Y, Tang M, Li Y, et al. Machine-learning based prediction and analysis of prognostic risk factors in patients with candidemia and bacteraemia: a 5-year analysis[J]. Peer J,2022,10:e13594.

[21] Peng L, Peng C, Yang F, et al. Machine learning approach for the prediction of 30-day mortality in patients with sepsis-associated encephalopathy[J]. BMC Med Res Methodol,2022,22(1):183.

[22] 刘建模, 罗颢文, 俞鹏飞, 等. 基于机器学习的急性缺血性脑卒中医院感染预测模型建立与评价[J]. 中国感染控制杂志,2023,22(2):129-135.

[23] 冯颖, 时克, 王宪波. 肝硬化合并肠系膜上静脉栓塞患者2年预后预测模型的构建与验证[J]. 首都医科大学学报,2023,44(1):107-114.

（收稿日期：2023-03-19）

（本文编辑：孙荣华）

范帅华，郭伟，郭军．基于机器学习的决策树算法在血流感染预后预测中应用现状及展望[J/CD]. 中华实验和临床感染病杂志(电子版), 2023,17(5):289-293.