机器学习在非酒精性脂肪肝预测中的应用*

蔡偌欣[1]　马亚楠[2]　闻德亮[1]

【提　要】 非酒精性脂肪肝(nonalcoholic fatty liver disease,NAFLD)是全球最常见的慢性肝病,普通成人 NAFLD 患病率在 6.3%~45%,我国大陆一般人群中患病率为 29.81%,在肥胖和 2 型糖尿病人群中发病率更高,会高达 90%。通过运用典型机器学习算法来构建非酒精性脂肪肝的风险预测模型,在肝病研究领域中是比较先进的。本文所归纳的 7 种典型机器学习算法在数据挖掘领域中是比较成熟且稳定的,在各项数据研究当中,基于预测结果的准确率,验证了各个模型的有效性和可行性,为脂肪肝疾病预测提供了基于数据科学的研究方法。

【关键词】 非酒精性脂肪肝　机器学习　预测模型

【中图分类号】 R575.5　　【文献标识码】 A　　DOI 10.11783/j.issn.1002-3674.2024.02.037

非酒精性脂肪肝的流行病学特征和诊断标准

非酒精性脂肪性肝(nonalcoholic fatty liver disease,NAFLD)是指除了酒精和其他明确的肝脏损伤因素所致的肝细胞内脂肪过度沉积为主要特征的获得性代谢应激性肝损伤[1]。成人非酒精性脂肪肝患病率在 6.3%~45%,在我国大陆地区成人的非酒精性脂肪肝病患病率为 29.8%,在肥胖人群和 2 型糖尿病人群中非酒精性脂肪肝的患病率更高,竟会高达 90%[2]。

目前肝穿刺活检仍然是诊断 NAFLD 的“金标准”。然而作为一项有创性检查,容易造成多种并发症或不良影响[3]。目前 NAFLD 的无创定性诊断主要包括:CT、B 超、磁共振和血清生物学标记物等多种方法[4],但是这些方法的缺点是对轻度非酒精脂肪肝的敏感性比较差,对于数据类型要求也比较高。为了解决以上有创性和无创性诊断的问题和实际需求,近年来医学领域引入了机器学习算法模型,用来预测非酒精性脂肪肝的发生,具有节省样本量、节省时间,同时可以兼顾定性定量的优点。

基于机器学习算法的非酒精性脂肪肝研究现状

1.机器学习概况

机器学习作为数据挖掘的一种重要方式,就是从既有的数据库中获取知识,通过测试和训练,对未来的走向和演变趋势进行预测[5]。近年来,机器学习被广泛应用到医学领域中,目前常用的方法主要有决策树(decision tree classifier,DT),线性判别分析(linear discriminant analysis,LDA),随机森林(random forest,RF),朴素贝叶斯(naive Bayes,NB),支持向量机(support vector machine,SVM),K-近邻(K neighbors classifier,KN),logistic 回归法(logistic regression,LR),K 均值聚类(K-means clustering)和主成分分析(principal components analysis,PCA)[6]。

2.机器学习的常用算法

(1)决策树模型

决策树是一种图解方法,在应用直观概率分析的基础上,来求取净现值的期望值大于等于零的概率,可以对多种类型变量进行筛选、检验、分析。它的分析过程是通过画图成像,这个图形类似大树的树干和树枝,所以被形象的称为决策树[7]。与其他分类方法(如 logistic 回归)相比,决策树的显著性优势体现为能够对不同类型的亚组进行分类和判断,以直观的树形结构去表现层次的递进关系,可以十分清晰地演示出整个分析的过程[8]。

(2)随机森林模型

随机森林是一种包括多个决策树的集成式机器学习方法,它的基本单元是决策树,从原始样本集中重复随机抽取 n 个样本生成新的训练样本集合构成一棵决策树,然后重复以上步骤生成 m 棵决策树,成百上千棵决策树就组成了随机森林,新数据的分类结果按分类树系统投票多少继而生成比例分数,再进行系统分类[9]。随机森林法具有泛化能力强,分类速度快、适应非线性数据源并且能够评估各种分类特征的重要性[10]。

(3)logistic 回归模型

logistic 回归是研究一个二分类或多分类反应变量与多个影响因素之间关系的多因素分析方法,包括二分类结果的多重 logistic 回归、配对资料的条件 logistic 回归、多分类结果的 logistic 回归、有序结果的累积优势 logistic 回归和有序结果的相邻优势 logistic 回归。logistic 回归也存在变量间共线性不易解决等缺

* 基金项目:国家重点研发计划“重大慢性非传染性疾病防控研究”重点专项(2018YFC1311600)

1.中国医科大学健康科学研究院(110122)
2.中国医科大学公共卫生学院

点[11]。

(4)马尔科夫(Markov)模型

马尔科夫模型是一种关于某种时间发生的概率的预测方法,根据一种事件目前状态来推测一段时间或者未来某一时间点发生结局事件的概率,最后预测的是实际值出现的可能范围。“状态”在马尔科夫模型中是一个非常重要的概念,它是指某一事件在某个时刻(或时期)出现的某种结果。预测概率有多大,只取决于事物现在所处的状态如何,而与以前的状态无关,马尔科夫预测主要研究状态转移过程,推算出状态转移概率。

(5)人工神经网络模型

人工神经网络,它是以人类大脑结构作为模仿对象,基于神经元运行的模式,从信息处理的角度应用数据挖掘的方法对数据进行类似神经元的抽象过程,并建立某种简化模型。它具有非线性、非局限性、自适应性、稳定的非凸性特征[12]。

(6)支持向量机模型

支持向量机方法是建立在统计学习理论的高维特征空间理论和结构风险最小基础上的,通过对学习能力和特定训练样本的学习精度之间寻求最佳的办法,以此获得最好的推广能力。它是一种非线性的分类器,应用二分类模式[13]。

(7)人工贝叶斯模型

贝叶斯分类算法是一类利用概率统计知识进行分类的算法。它的原理是在统计资料的基本特征基础上,依据某些特征,计算每一种类别的概率,从而达到精准分类。这种算法的优点是方法简单、分类准确率高、速度快[14]。

表1 常用机器算法模型的适用条件及特点分析

机器算法模型	适用条件	结局变量	数据特点
决策树模型	评估各个影响因素在不同水平发生的风险	二分类变量	常应用于离散型数据
随机森林模型	评估各个影响因素在不同水平发生的风险	二分类变量	连续变量或者分类变量都可
logistic 回归模型	定量分析多个影响因素和结局的线性关系	二分类变量	连续变量或者分类变量都可,纳入的影响因素比较有限
马尔科夫模型	评估疾病状态转化的影响因素及程度	概率	纵向随访资料,不需要有精确的连续时间的随访资料
人工神经网络模型	发现多个影响因素的未知关系,适用于复杂病因的研究	二分类变量	任何类型资料,线性非线性资料均可
支持向量机模型	解决小样本、非线性及高维模式识别	二分类变量	任何类型资料,尤其是不了解的数据类型
朴素贝叶斯模型	概率推理不确定性因素	概率	适合不完备的资料

3.机器学习在非酒精性脂肪肝病预测中的应用

(1)决策树模型和随机森林模型预测 NAFLD

吕航等[15]应用决策树模型来构建风险模型,探讨糖尿病伴发非酒精性脂肪肝病的危险因素并提出预防措施;Perveen[16]使用决策树模型来对电子病历的相关危险因素进行分类,并对其病情进展进行分析,构建脂肪肝预测模型。Pasolli 等[17]利用随机森林法构建 2 型糖尿病、肝硬化等疾病的预测模型。白江梁等[18]探讨了随机森林在成年体检人群中糖尿病、脂肪肝、肝硬化等风险预测模型的构建。

(2)logistic 回归模型预测 NAFLD

logistic 回归是预测并发症研究中应用最多的模型[11]。高福来[19]基于血清 Betatrophin 水平,构建了非酒精性脂肪肝列线图预测模型,模拟效能良好。Sorino 等人[20]采用多因素 logistic 回归分析非酒精性脂肪肝炎进展到肝硬化的危险因素分析,并且构建了预测模型。

(3)马尔科夫模型预测 NAFLD

Wang 等人[21]分别使用决策树和马尔可夫模型,基于社区人群对成人非酒精性脂肪肝患病风险及对经济效应进行评估。陈潇潇等人[22]基于马尔可夫模型对代谢综合征及脂肪肝的转归进行风险预测。

(4)人工神经网络模型预测 NAFLD

Heinemann[23]使用人工神经网络模型对非酒精性脂肪肝病患者的病理切片进行评分并分配离散分数,对非酒精性脂肪肝进行组织病理学特征分类。陈菲等[24]应用人工神经网络对非酒精性脂肪肝进行分类,以此分为轻、中、重度型脂肪肝。

(5)支持向量机模型预测 NAFLD

韩秀芝等[25]通过收集患者的肝脏超声图像特征,采用了支持向量机法对非酒精脂肪肝类型进行了分类;Perakakis[26]采用 SVM 模型分析 29 种脂质或脂质与聚糖和/或激素的组合区分是否存在肝纤维化(准确度为 98%)并判断其健康状况。

(6)人工贝叶斯模型预测 NAFLD

张永媛[27]通过使用人工贝叶斯算法预测模型,对代谢综合症与非酒精性脂肪肝之间的患病风险及双向因果关系进行推断;徐磊[28]用贝叶斯模型尝试建立了非酒精性脂肪性肝病的预测模型,结果提示模型对于脂肪性肝病的判断正确率为 86.7%,实现了对弥漫性脂肪肝的分类。

小结与展望

通过运用典型机器学习来构建非酒精性脂肪肝的风险预测模型,是目前整合医学资源和数学模型一种新的数据挖掘方法。本文根据既往国内外相关研究,基于流行病学研究结果,归纳出 7 种比较成熟和应用广泛的机器学习算法。结合临床实践数据和国内外研究分析验证了不同机器学习算法不同模型的分类原则和纳入标准。

在脂肪肝研究领域运用典型机器学习来构建非酒精性脂肪肝的风险预测模型，是整合医学资源和数学模型一种新的创新方法。本文通过整理归纳，为非酒精脂肪肝预测提供了基于大数据科学的研究方法。为了进一步提高模型效能和准确性，下一步有必要开展大规模、大样本的非酒精性脂肪肝的长时间随访研究，建立起区域性甚至全国性的大型队列。再一点，我们需要获得多时点数据，并结合医学人工智能的体系构建发病风险评估，最终达到提高风险评估模型的机器算法应用性和精确性的目的，从而建立起完善的评价体系，为非酒精性脂肪肝防控策略制定提供理论基础和科学依据。

参 考 文 献

[1] Diehl AM, Day C. Cause, Pathogenesis, and Treatment of Nonalcoholic Steatohepatitis. N Engl J Med,2017,377(21):2063-2072.

[2] Li J, Zou B, Yeo YH, et al. Prevalence, incidence, and outcome of non-alcoholic fatty liver disease in Asia, 1999-2019:a systematic review and meta-analysis. Lancet Gastroenterol Hepatol,2019,4(5):389-398.

[3] 中华医学会肝病学分会脂肪肝和酒精性肝病学组. 中国非酒精性脂肪性肝病诊疗指南(2010 年修订版). 中国医学前沿杂志(电子版),2012,4(7):4-10.

[4] Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. Clin Microbiol Infect,2013,19(11):999-1005.

[5] 张润, 王永滨. 机器学习及其算法和发展研究. 中国传媒大学学报自然科学版,2016,23(2):10-24.

[6] Albhaisi S, Sanyal AJ. Applying Non-Invasive Fibrosis Measurements in NAFLD/NASH:Progress to Date. Pharmaceut Med,2019,33(6):451-463.

[7] 吕文娣, 赵广高,付近梅,等. 基于决策树模型的幼儿超重关键因素研究. 成都体育学院学报,2020,46(1):86-93.

[8] 陈辉林, 夏道勋. 基于 CART 决策树数据挖掘算法的应用研究. 煤炭技术,2011,30(10):164-166.

[9] 何清, 李宁, 罗文娟. 大数据下的机器学习算法综述. 模式识别与人工智能,2014,4(7):197-198.

[10] Ai L, Tian H, Chen Z, et al. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. Oncotarget,2017,8(6):9546-9556.

[11] 王琦琦, 于石成, 亓晓. logistic 回归及其应用. 中华预防医学杂志,2019,53(9):955-960.

[12] 陈飞彦, 田宇驰, 胡亮. 物联网中基于 KNN 和 BP 神经网络预测模型的研究. 计算机应用与软件,2015,32(6):127-129.

[13] 宋月婵. 数据降维算法在铀矿堆浸工艺中的应用研究. 东华理工大学硕士论文.

[14] 李玥. 机器学习的分类、聚类研究. 电脑知识与技术,2020,16(4):160-162.

[15] 吕航, 王昊, 刘媛,等. 基于决策树的中医人格体质对 2 型糖尿病患者伴发非酒精性脂肪肝病风险的预测研究. 中国中医基础医学杂志,2017,23(9):1257-1259.

[16] Perveen S, Shahbaz M, Keshavjee K, et al. A Systematic Machine Learning Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression. Sci Rep,2018,8(1):2112.

[17] Pasolli E, Truong DT, Malik F, et al. Machine Learning Meta-analysis of Large Metagenomic Datasets:Tools and Biological Insights. PLoS Comput Biol,2016,12(7):e1004977.

[18] 白江梁, 张超彦, 李伟. 某医院体检人群糖尿病预测模型研究. 实用预防医学,2014,25(1):116-119.

[19] 高福来, 谢长顺, 张利利. 基于血清 Betatrophin 水平的非酒精性脂肪肝列线图预测模型的建立与分析. 中国医药导报,2019,4(16):103-106.

[20] Sorino P, Caruso MG, Misciagna G, et al. Selecting the best machine learning algorithm to support the diagnosis of Non-Alcoholic Fatty Liver Disease:A meta learner study. PLoS One,2020,15(10):e0240867.

[21] Wang J, Xu C, Xun Y, et al. ZJU index:a novel model for predicting nonalcoholic fatty liver disease in a Chinese population. Sci Rep,2015,5:16494.

[22] 陈潇潇. 基于马尔科夫模型的代谢综合征描述和风险预测研究. 济南:山东大学,2015.

[23] Heinemann F, Birk G, Stierstorfer B. Deep learning enables pathologist-like scoring of NASH models. Sci Rep,2019,9(1):18454.

[24] 陈菲. 中医药治疗非酒精性脂肪肝的研究现况. 全国第九次中西医结合传染病学术会议暨深圳市医学会肝病专业委员会 2018 年学术年会,2018.

[25] 韩秀芝, 赵希梅, 于可歆. 一种基于 LBP 特征提取和稀疏表示的肝病识别算法. 中国生物医学工程学报,2017,36(6):647-653.

[26] Perakakis N, Polyzos SA, Yazdani A, et al. Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning:A proof of concept study. Metabolism,2019,101:154005.

[27] 张永媛. 基于队列设计纵向监测与概率图模型的非酒精性脂肪肝与代谢综合征双向因果推断研究. 济南:山东大学,2013.

[28] 徐磊. 脂肪性肝病与代谢综合征的现况调查研究. 现代实用医学,2009,21(6):223-226.

(责任编辑:郭海强)