生物医学工程研究
Journal of Biomedical Engineering Research 2021,40(2):203~209

机器学习在疾病预测中的应用*

廖华龙[1],曾小茜[2],李华凤[3],于洋[4],赵灿[1],陈宇[1△]
(1.四川大学 生物力学工程省重点实验室,成都 610065;2.四川大学华西医院 华西大数据中心,成都 610041;
3.四川大学华西第二医院麻醉科,成都 610041;4.四川大学华西医院肾脏内科,成都 610041)

摘要:本文通过阐述机器学习的定义及分类,介绍了K近邻、朴素贝叶斯、逻辑回归、支持向量机、决策树、集成的树模型以及人工神经网络等研究者常用于医疗疾病预测的机器学习算法,重点分析了机器学习在预测心血管疾病、糖尿病、肾病、肿瘤、妊娠期疾病等几类常见疾病中的应用情况,主要从特征选择、算法和预测准确性等方面说明机器学习预测疾病的应用特点。

关键词:机器学习;疾病;预测;算法;特征选择
中图分类号:R318　　**文献标识码**:A　　**文章编号**:1672-6278 (2021)02-0203-07

Application of machine learning in disease prediction

LIAO Hualong[1], ZENG Xiaoxi[2], LI Huafeng[3], YU Yang[4], ZHAO Can[1], CHEN Yu[1]
(1. *Laboratory of Biomechanics Engineering, Sichuan University, Chengdu*610065, *China*;
2. *West China Big Data Center, West China Hospital, Sichuan University, Chengdu* 610041;
3. *Department of Anesthesiology, West China Second University Hospital, Sichuan University, Chengdu* 610041;
4. *Department of Nephrology, West China Second University Hospital, Sichuan University, Chengdu* 610041)

Abstract: We expound the concept and classification of machine learning and introduce several common kinds of machine learning methods applied on medical disease prdiction practice, such as K-nearest neighbor, Naive Bayes, logistic regression, support vector machine, decision tree, integrated tree models and artificial neural networks, we focus on the machine learning application in the prediction of some common diseases, such as cardiovascular diseases, diabetes, kidney diseases, tumour, pregnancy diseases and so on. We mainly analyze the characteristics of machine learning in disease prediction from three aspects: feature selection, algorithm and model accuracy.

Key words: Machine learning; Diseases; Prediction; Algorithm; Feature selection

1 引　言

机器学习作为数据挖掘的主要工具之一,被用于医疗领域[1]。机器学习通过对患者现有的医疗检测或调查得到的数据进行学习,建立风险模型,常用于预测疾病,诊断疾病严重程度以及评估疾病预后等[2]。本文阐述了机器学习的定义、分类以及主要的几类算法原理,并以机器学习在预测心血管疾病、糖尿病、肾病、肿瘤、妊娠期疾病的应用为例,从数据特征选择、算法和准确性等方面阐述了机器学习用于疾病预测的特点和效果。本文不涉及医学图像的深度学习算法。

DOI　10.19529/j.cnki.1672-6278.2021.02.17
*国家重点研发计划项目(2018YFC2001805);四川大学华西医院"十三五"高端人才计划项目(ZYGD18027)。
△通信作者　Email:yu_chen@scu.edu.cn
文中彩图请见电子版:www.swyxgcyj.com

2 机器学习的定义与分类

2.1 定义

机器学习是一种能自动构建出数据模型并用来处理数据之间复杂关系的技术[3]。它使用计算机模拟人类学习行为,通过学习现有数据或图像(特征),再根据分类或者回归的任务要求来发现规律,从而获取新经验与新知识,提升性能,实现自我完善。

2.2 分类

机器学习根据是否有人为标记数据分为监督学习和无监督学习。监督学习是用具有分类标签的数据作为现有知识,通过带有标签的数据进行模型训练,并将训练好的模型用来预测新数据的标签结果。无监督学习是用于处理不具有分类标签的数据,通过寻求数据间的内在关联和规律,发现样本数据潜在的结构特征。另外还有针对学习只有少量带有标签的数据而衍生出的半监督学习[4]。

3 机器学习的主要几类算法

3.1 K近邻

K近邻是一种原理较为简单的机器学习算法。对于给定测试样本,该算法基于距离度量找出训练集中与其最靠近的 k 个训练样本,然后根据这 k 个训练样本(邻居)的信息来进行预测[5]。

3.2 朴素贝叶斯

朴素贝叶斯法是一种基于贝叶斯定理的分类方法。对于给定的训练数据集,其首先基于特征条件独立假设学习输入输出的联合分布概率 $P(x|y)$;然后基于此模型,对给定的输入 x,再利用贝叶斯定理求出其后验概率最大的输出 y。朴素贝叶斯数学表达见式(1)。

$$P(y|x)=\frac{P(x|y)\cdot P(y)}{P(x)},P(\text{label}|\text{features})=\frac{P(\text{features}|\text{label})\cdot P(\text{label})}{P(\text{features})} \tag{1}$$

3.3 逻辑回归

逻辑回归模型可看作是一个被 Sigmoid 函数归一化后的广义线性回归模型,用一条直线区分不同类别的数据,用于分类任务。

3.4 支持向量机

支持向量机把线性不可分数据通过映射函数 ϕ 投射到高维空间,使特征在该空间变得线性可分,再用一个最大边界间隔超平面对特征进行划分[6]。

3.5 决策树

基于训练集的特征,决策树模型通过提出一系列的问题条件来推断样本的分类标签。决策树的建立过程是从根节点(第一个选择点)开始,逐步通过非叶子节点的分支走到叶子节点(最终的决策结果),最终所有的数据都会落到叶子节点。

3.6 集成的树模型

集成学习算法思想是使用弱分类器和多个样本来构建一个强分类器,改善学习效果。集成的树模型包括随机森林、adaptive boosting(AdaBoost)、gradient boosting decision tree(GBDT)、light gradient boosting machine(LightGBM)和XGboost等。

随机森林是以决策树为基础学习器,集成多个决策树的结果,在bagging算法的基础上进行了改动而演化过来的[7]。bagging算法是在原始的数据集上采用有放回的随机取样方式来抽取 m 个子样本,从而利用这些子样本训练 m 个基础学习器,降低模型的方差。在此基础上,随机森林还在训练每个基学习器的时候,随机地选取 k 个特征,从这些特征中选择最优特征来切分节点,从而进一步降低模型的方差。

除了bagging算法之外,boosting也是一种可将弱学习器提升为强学习器的算法,属于集成学习的范畴。其中比较有代表性的是AdaBoost,它会对训练数据集中的每个样本进行训练,并赋予每个样本一个权重。最初这些样本的初始权重相同,然后AdaBoost通过训练数据得出一个弱分类器并计算其错误率,接着在相同的训练数据上,再次训练弱分类器。在弱分类器的第二次训练过程中,每个样本的权重会重新得到调整。AdaBoost对第一次训练分类正确的样本降低其权重,分类错误的样本提高其权重。AdaBoost不断对弱分类器进行训练迭代,最终综合所有弱分类器得到结果。

在使用与Adaboost相同的boosting算法的基础上,GBDT用Gradient Boosting的策略训练出树模型,是一个基于迭代累加的决策树算法。它构造一组弱学习器(决策树),把多棵决策树的结果累加起来,作为最终的预测输出[8]。在GBDT的基础上,LightGBM是一个实现GBDT算法的框架,支持高效率的并行训练,拥有更快的训练速度和更高的准确率等优势[9]。另外,还有学者对GBDT算法进行改进,提出了一种高效灵活,并且可移植性强的最优分布式决策梯度提升库XGBoost[10]。

3.7　人工神经网络

人工神经网络是模拟人脑神经元结构进行信息处理的一种数学模型。神经网络中的每个神经元接收大量的输入信号,执行输入的加权,通过非线性激活函数产生激活响应,并对随后连接的神经元传递输出信号[11]。还可以设置多个产生激活响应的隐藏层神经元,使其成为多层神经网络。在此基础上进行改进,其又可演变成为深度学习,常用的深度学习算法包括卷积神经网络、深度神经网络与递归神经网络,在医疗领域中主要用来进行疾病诊断和医学影像的分析等。

4　机器学习在疾病预测中的应用

4.1　心血管疾病

心血管疾病是中老年人的常见病。2017 年,国内心血管疾病死亡率高达 40%,已成为我国重大的公共卫生问题[12]。因此,提前预测心血管疾病并积极干预疾病的发生发展,具有重要意义。采用机器学习的方法对心血管疾病进行预测是一种经济、安全、可行的途径。

机器学习在预测心血管疾病方面已得到较多应用,目前相关的预测手段已经较为成熟。Ambale 等[13]收集了 6 814 名来自多民族动脉粥样硬化的数据,用随机生存森林算法预测包括中风、冠心病、心房颤动和心力衰竭等事件的发生,发现随机生存森林算法比已建立的心血管风险评分体系更好,预测准确度更高。因为树模型是通过数据的二元递归分割来生长的,在每次生长分支时,模型会选择一个候选变量,让该变量最大化子节点之间累积危险的差异。数据不能再分割时即停止生长,使得每个终端节点至少有一个唯一的结果,这样层层分支下来能够很好地进行分类。集成的随机生存森林综合了每个树模型的分类预测结果,精确度得到了提高。随机生存森林在其他相关研究中的效果也已得到验证[13]。同时随机生存森林预测模型还发现了可能的患病危险因素,例如空腹血糖水平升高是中风最重要的危险因素,动脉粥样硬化综合指标是冠心病最重要的预测指标之一。另外,左室局部壁厚增加(心肌肥大)、射血分数降低和主动脉横截面积增加等也是冠心病的其他主要预测因素。国内也有针对心血管疾病预测方面的研究。刘宇等[14]收集了来自 300 例患者的 10 000 个健康数据,包括年龄、性别、胸部疼痛、血压、胆固醇等 14 个变量,再使用 K 聚类和 XGBoost 来预测心脏病的发生。结果表明,该预测模型的准确率超过了 0.8,并显示出对患心脏病的影响较为显著的四个变量分别是:年龄、胆固醇、最大心跳和运动后比较心压。相比于支持向量机和随机森林,该预测模型的用时最少,是一种有效预测心脏病的方法。

陈伟伟等[12]预测 2017 年心血管疾病患病人数高达 2.9 亿。随着机器学习使用数据量的增加,相比与其他机器学习算法,适合处理大量数据的神经网络表现出其优势。Weng 等[15]收集了378 256名英国家庭实践患者的常规临床数据(受试者初期没有心血管疾病),数据包括完整的八个核心变量(性别、年龄、吸烟状况、收缩压、血压治疗、总胆固醇、高密度脂蛋白胆固醇和糖尿病),另外还加入了 22 项可能与心血管疾病有关的变量。团队成员采用用随机森林、逻辑回归、梯度增强机(gradient boosting Machine,GBM)和神经网络四种机器学习算法来预测心血管疾病的发生风险。机器学习确定了以前的风险预测工具未发现的其他潜在风险因素,包括慢性阻塞性肺病、严重精神疾病以及甘油三酯水平等。这些潜在发病因素在之后可以纳入预测心血管疾病的模型中,进一步提高预测准确率。同时,与已建立的风险预测模型(美国心脏协会/美国心脏病学院基线模型)相比,四种机器学习算法的表现更好,其中神经网络的效果最好。

4.2　糖尿病

糖尿病对人群健康产生巨大危害,消耗大量医疗资源。利用适合大范围使用的糖尿病监测系统,寻找有效的早期检测手段,做到糖尿病的早发现、早诊断和早治疗,以延缓或防止糖尿病及其并发症的发生与发展,能减少患者患病痛苦和医疗负担,提高社会人群健康水平[16]。机器学习可通过学习患者的临床检测资料数据,来对患者是否会发展成为糖尿病进行预测,从而给医生和患者提供参考和建议,有望建立相应的疾病监测系统。Lee 等[17]收集了 11 937名受试者的个体人体测量和甘油三酯等数据,用朴素贝叶斯和逻辑回归来预测 2 型糖尿病的发生,找到了理想的预测因子:针对男性的是腰臀比加甘油三酯的组合,针对女性的是肋臀比加甘油三酯的组合。研究结果显示出了这些预测因子可以组合预测 2 型糖尿病的趋势。

在采用基本的体征测量数据以及甘油三酯数据进行预测的基础上,选取更多的特征数据后,机器学习预测的效果也许会有所提升。Alghamdi 等[18]用三棵不同的决策树(朴素贝叶斯树、随机森林和逻

辑模型树)来学习具有 13 个特征属性的 32 555 名无任何已知冠状动脉疾病或心力衰竭的患者的数据,从而预测糖尿病的发生。模型有着高预测准确度,显示了利用心肺健康数据配合机器学习算法预测糖尿病发病率的优势。特征数据中年龄和恢复心率有最大的信息增益值,说明它们对预测疾病有着最重要的作用。如果在机器学习模型的特征中加入包含血糖类的数据,预测的可靠性还能进一步提高。Ijaz 等[19]使用基于随机森林的混合模型来预测 403 位患者发生 2 型糖尿病的可能性,特征筛选保留了 9 个特征,其中稳定血糖有最大的信息增益值,即对预测 2 型糖尿病的贡献最大。混合模型使用这 9 个特征数据进行预测的结果精度比较高,与其他二分类算法预测糖尿病的模型(支持向量机、多层神经网络、逻辑回归和朴素贝叶斯)相比,具有更理想的预测效果。Zou 等[20]从体检数据中抽取 68 994 名健康人和糖尿病患者资料作为训练集,数据集包括年龄、脉搏、收缩压舒张压以及空腹血糖等 14 项体检指标,再用决策树、随机森林和神经网络来预测糖尿病。结果表明,随机森林预测效果更好,而且仅使用空腹血糖指标即可达到较高的准确度,说明空腹血糖是预测模型最重要的一个特征指标。以往的研究说明,树类模型(决策树,集成的树模型)对糖尿病预测较为有效,预测结果也比常规方法更好。在医院临床中实际预测糖尿病时,考虑以树类模型作为基础算法也许能达到理想的效果。

4.3 肾病

肾病中比较常见的是急性肾损伤。急性肾损伤的流行病学数据差异大,发病率和死亡率比较高[21]。急性肾损伤早期症状隐匿,可能被原发疾病所掩盖。近年有报道表明,某些生物标志物与急性肾损伤相关,其中以血清肌酐为主[22]。通过采集和学习患者的相关生物指标数据,机器学习能对急性肾损伤进行精准的早期预测。Kate 等[23]收集了某大型医疗系统 25 521 位 60 岁及以上患者的资料,包括患者的人口统计学信息 12 项、共病 14 项、药物使用 12 项和实验室测定值 9 项,以此作为变量特征,用逻辑回归、支持向量机、决策树和朴素贝叶斯预测急性肾损伤。结果表明,逻辑回归的效果好于其他方法,但所有模型结果都较差,只显示了可以预测的趋势,还未直接用于临床预测。之后 Koyner 等[24]采用梯度增强机来学习 121 158 名患者的数据,希望预测急性肾损伤的发生风险。模型中使用的预测因素特征变量包括人口统计学、生命体征、常规实验室检查结果和生命体征等。梯度增强机发现其中血清肌酐的变化对预测结果的影响最大,该结果也与文献[22]得到的血清肌酐是临床诊断急性肾损伤金标准的结果相符合。该模型最终达到了比较好的预测效果,说明其可以对急性肾损伤高危人群进行早期预测。另外,张渊等[25]从公开的 ICU 医学信息数据库中提取了 1 166 例患者数据作为机器学习的数据集(其中有 884 例患者发展为急性肾损伤),用 LightGBM 预测 ICU 患者发生急性肾损伤,纳入患者 33 项生理生化指标进行预测模型的构建,预测效果非常好。研究也得出了 LightGBM 特征重要性排名前 10 位的特征,分别是液体入量、红细胞比容、患者进行了机械通气、动脉氧分压、乳酸、体温、动脉血 pH、心肌肌钙蛋白、血小板计数、凝血酶原时间,其中对预测结果帮助最大的是液体入量。将模型结果与逻辑回归和随机森林的结果对比发现,LightGBM 模型对急性肾损伤的预测效果最好。LightGBM 与算法原理简单的逻辑回归相比,更适合于处理体量大、维度高的数据。并且 LightGBM 在树模型的基础上增加了一些提升算法,包括采用 leaf-wise 的分裂方式对树模型进一步优化等[9],从理论和实践上提升了性能。基于重症患者的多项生理生化指标数据,机器学习模型预测急性肾损伤可以达到很好的效果,能帮助识别有风险的受试者并实施预防策略,或者帮助医生根据预测结果来管理患者,为临床决策提供辅助支持。

急性肾损伤如果未得到及时控制,可能会发展为慢性肾脏疾病。近年来慢性肾脏病的患病率逐年上升。有研究显示,18 岁以上人群慢性肾脏病的患病率已经超过 10%[26]。若通过机器学习从早期体检资料和实验室检查报告中找到慢性肾病患者的发病规律,即可对疾病进行早期检查,做到及时干预。用于数据挖掘的机器学习可能成为一个有效的肾病预测工具。2018 年 Abdelaziz 等[27]结合线性回归和神经网络两种方法构建了混合预测模型,对慢性肾脏疾病进行诊断与预测,经过特征筛选后找出了 13 项与慢性肾脏疾病的发生相关的关键因素,包括年龄,血压,随机血糖,白细胞计数等。用这 13 项特征数据构建模型,模型预测慢性肾脏病的准确率达 95% 以上,预测效果优于其他研究者构建的预测模型。

4.4 肿瘤

据估计,我国 2013 年新发恶性肿瘤病例约 368.2 万例,死亡病例 222.9 万例[28]。目前早检测、

早预防、早治疗是防治肿瘤、改善预后的重要途径,利用机器学习对早期预测肿瘤具有重要意义。Huang 等[29]采用多种支持向量机及其组合来预测乳腺癌的发生,并且针对不同规模的数据集(一个小规模数据集含有 699 条 11 维数据,另一个大规模数据集含有 102 294 条 117 维数据)和是否进行特征选择的不同情况,找出了最适合的支持向量机算法,预测准确率都非常高,可见使用机器学习预测乳腺癌的技术已经较为成熟。

肿瘤诊断的直接证据是病理检查发现细胞发生了异常增生,在患者的临床生理指标数据基础上,学习细胞的特征数据可提升机器学习预测肿瘤的精度。2019 年苗立志等[30]通过随机森林学习 683 条威斯康星临床科学中心原始数据(数据包含了细胞核特征的 10 个属性),使用这 10 个属性进行模型训练后,发现所建立的预测模型精度非常高。他们同时还通过计算各个属性与致病性(患病)的相关度,发现细胞核周长、灰度值的标准差、轮廓凹面部分的数目与乳腺癌的发展密切相关,可将其作为乳腺癌预后评估的重要指标。细胞核周长、纹理组织和凹点对于乳腺癌的致病性具有较好的特征表述,可用于乳腺癌的诊断与发病规律研究。通过分析乳腺细胞核的特征变量的方法,可在很大程度上降低医患双方的医疗成本,提高医院的工作效率。

除了乳腺癌以外,机器学习还可以通过学习基因表达数据来预测其他肿瘤。肺癌是常见的恶性肿瘤[28],目前也有将机器学习应用于肺癌的研究。冷菲等[31]收集了 474 例肺腺癌样本和 491 例肺鳞癌样本的数据,使用 XGBoost 学习 1 099 个差异表达的 mRNA 数据,希望对肺癌亚型肺鳞状细胞癌(肺鳞癌)和肺腺癌进行预测。XGBoost 表现出了非常高的预测精度和良好的稳定性,优于逻辑回归和支持向量机,为肺鳞癌和肺腺癌的早期诊断和治疗提供了试验依据,同时机器学习也找到了对建立预测模型的贡献率排名前三的 mRNA。

因为肿瘤的发病机理与其他疾病不同,是在致癌因素的作用下细胞的基因发生了改变,导致细胞产生异常增殖。使用机器学习算法来预测癌症时,以细胞或者基因层面的数据为主来进行学习和预测为主,或许能改善预测效果,甚至找出与癌症发生有关的病变细胞或者基因,帮助医生对患者进行诊治。

4.5　妊娠期疾病

目前研究较多的妊娠期疾病是妊娠期高血压类疾病,妊娠期高血压又可划分为子痫、妊娠合并慢性高血压、慢性高血压并发子痫前期。妊娠期高血压严重危害母婴身体健康,是孕产妇和产儿死亡的主要原因之一。如果能用机器学习来确定妊娠期高血压疾病的高危因素,提前对妊娠期疾病进行预测,进而及时进行疾病的早期干预诊断和治疗,可能会帮助改善母婴结局。Poon 等[32]总结之前的研究发现大多数在怀孕 11 到 13 周发展为先兆子痫的孕妇,其平均动脉压、子宫动脉搏动指数、母体血清中胎盘因子妊娠相关血浆蛋白 - a,以及胎盘生长因子浓度发生了明显变化,研究团队收集了 7 797 条包含这四种指标的单胎妊娠数据,希望用逻辑回归学习孕妇的这四种特征数据,来预测先兆子痫(preeclampsia,PE)的早期和晚期以及妊娠高血压。结果早期先兆子痫的的预测精度非常高,然而 PE 晚期对应的预测效果比较差,妊娠高血压的效果最差。该项研究采用的是发生先兆子痫的四种明显变化的母体特征数据,能很好地预测先兆子痫早期。但是对于先兆子痫晚期和妊娠高血压,目前还需要筛查其他的特征数据来进行学习,找出对预测晚期先兆子痫和妊娠高血压最有帮助的一项或者几项数据指标。另外,在预测妊娠期疾病方面,少有研究验证和对比多种机器学习模型的效果和作用,常见的简单分类算法也许并未达到理想的预测效果,因此,还需要进一步研究。

4.6　其他类疾病

除了以上几类疾病之外,机器学习在其他一些疾病的预测中也有涉及,并且取得了较好的效果。例如非综合征性唇裂伴或不伴腭裂是一种多因素、部分遗传的先天性疾病,涉及多个基因和遗传与环境的复杂相互作用。Zhang 等[33]发现该病在国内的发病率超过了世界平均水平,于是收集了 587 名对照组和非综合征性唇裂伴或不伴腭裂婴儿的血液样本数据,希望结合 43 个单核苷酸多组合形式与机器学习来建立预测模型帮助预测疾病。他们对比了支持向量机、逻辑回归、朴素贝叶斯、随机森林、K 近邻、决策树以及人工神经网络的效果,最后发现逻辑回归预测新生儿发病的效果最好,并且找出了可能与发病相关的基因。

除了可以预测单个疾病之外,机器学习还能对病人的整体健康状况进行预测,判断病人是否患病以及最可能患什么病。Miotto 等[34]采用一种新的无监督深度特征学习方法对 76 214 名受试者进行评估,以期预测个体的整体健康状况水平,发现该模型对严重糖尿病、精神分裂症和各种肿瘤的预测效

果较好,为临床决策系统提供了一个很好的框架。

另外,如果患者已经发病,机器学习还能对患者疾病的发展趋势进行预测。Shah 等[35]收集了一年期慢性阻塞性肺病临床试验的110例患者的资料数据,用逻辑回归预测慢性阻塞性肺疾病发生恶化的概率。结果逻辑回归预测该疾病加剧的效果较好,同时还发现了脉搏率、血氧饱和度和呼吸频率这三项生命体征,均能预测慢性阻塞性肺病加剧,将这些生命体征与基于机器学习的鲁棒算法相结合则可以进一步提高预测精度。此外,Finkelstein 收集了成人哮喘患者在家庭远程监测期间提交的7 001份记录数据,包括呼吸症状、哮喘引起的睡眠障碍、体力活动受限、感冒和药物使用等信息,采用朴素贝叶斯、自适应贝叶斯网络和支持向量机预测哮喘恶化,结果自适应贝叶斯网络的预测准确率非常高[28]。

5 总结

机器学习方法较多,可以仅通过学习患者的临床检测数据或监测记录数据来预测疾病的发生风险,帮助医生对患者的疾病进行提前干预和治疗,减少患病风险,降低医疗成本,对疾病防控具有重要意义。机器学习还能发现对患病有巨大影响的潜在特征指标项,为疾病的诊断和治疗提供新的依据。但是,不同的机器学习算法具有各自适宜预测的疾病,在实际应用时,需要找到最适合预测某种疾病的对应算法。

致谢

科技部国家重点研发计划项目(2018YFC2001805),四川大学华西医院"十三五"高端人才计划项目(ZYGD18027)。

参考文献:

[1]叶雷. 机器学习算法在医疗数据分析中的应用[D]. 华中师范大学,2017.

[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):93-97.

[3]Deo R C. Machine learning in medicine[J]. Circulation,2015,132(20):1920-1930.

[4]Baştanlar Y,Özuysal M. Introduction to machine learning[M]. miRNomics: MicroRNA Biology and Computational Analysis. Totowa,NJ: Humana Press,2013:105-128.

[5]Cover T,Hart P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory,1967,13(1):21-27.

[6]Cortes C,Vapnik V. Support-vector networks[J]. Machine Learning,1995,20(3):273-297.

[7]Breiman L. Random forests[J]. Machine Learning,2001,45(1):5-32.

[8]Friedman J H. Machine[J]. The Annals of Statistics,2001,29(5):1189-1232.

[9]Ke GL, Meng Q, Finley T,et al. LightGBM: A highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems,2017:3149-3157.

[10]Chen T Q,Guestrin C. XGBoost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:785-794.

[11] Lo Y C,Rensi S E,Torng W,et al. Machine learning in chemoinformatics and drug discovery[J]. Drug Discovery Today,2018,23(8):1538-1546.

[12]陈伟伟,高润霖,刘力生,等.《中国心血管病报告2017》概要[J]. 中国循环杂志,2018,33(1):1-8.

[13]Ambale-Venkatesh B,Yang X,Wu C O,et al. Cardiovascular event prediction by machine learning: The multi-ethnic study of atherosclerosis[J]. Circulation Research,2017,121(9):1092-1101.

[14]刘宇,乔木. 基于聚类和XGboost算法的心脏病预测[J]. 计算机系统应用,2019,28(1):228-232.

[15]Weng S F,Reps J,Kai J,et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? [J]. PLoS One,2017,12(4):e0174944.

[16]汪会琴,胡如英,武海滨,等. 2型糖尿病报告发病率研究进展[J]. 浙江预防医学,2016,28(1):37-39.

[17]Lee B J,Kim J Y. Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning[J]. IEEE Journal of Biomedical and Health Informatics,2016,20(1):39-46.

[18]Alghamdi M,Al-Mallah M,Keteyian S,et al. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project[J]. PLoS One,2017,12(7):e0179805.

[19]Ijaz M,Alfian G,Syafrudin M,et al. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection,synthetic minority over sampling technique (SMOTE),and random forest[J]. Applied Sciences,2018,8(8):1325.

[20]Zou Q,Qu K Y,Luo Y M,et al. Predicting diabetes mellitus with machine learning techniques[J]. Frontiers in Genetics,2018,9:515.

[21]郎夏冰,杨毅,陈江华. 中国住院患者急性肾损伤流行病学调查现状[J]. 浙江大学学报(医学版),2016,45(2):208-213.

[22]陈沐林,杨陈,韩焕钦,等. 急性肾损伤生物标志物的研究现状及新进展[J]. 医学综述,2019,25(9):1761-1765.

[23]Kate R J,Perez R M,Mazumdar D,et al. Prediction and detection models for acute kidney injury in hospitalized older adults[J]. BMC Medical Informatics and Decision Making,2016,16:39.

[24]Koyner J L,Carey K A,Edelson D P,et al. The development of a

machine learning inpatient acute kidney injury prediction model[J]. Critical Care Medicine,2018,46(7):1070 – 1077.

[25]张渊,冯聪,李开源,等. ICU 患者急性肾损伤发生风险的 Light-GBM 预测模型[J]. 解放军医学院学报,2019,40(4):316 – 320.

[26]上海慢性肾脏病早发现及规范化诊治与示范项目专家组,高翔,梅长林. 慢性肾脏病筛查 诊断及防治指南[J]. 中国实用内科杂志,2017,37(1):28 – 34.

[27]Abdelaziz A, Elhoseny M, Salama A S, et al. A machine learning model for improving healthcare services on cloud computing environment[J]. Measurement,2018,119:117 – 128.

[28]陈万青,郑荣寿,张思维,等. 2013 年中国恶性肿瘤发病和死亡分析[J]. 中国肿瘤,2017,26(1):1 – 7.

[29]Huang M W, Chen C W, Lin W C, et al. SVM and SVM ensembles in breast cancer prediction[J]. PLoS One,2017,12(1):e0161501.

[30]苗立志,刁继尧,娄冲,等. 基于 Spark 和随机森林的乳腺癌风险预测分析[J]. 计算机技术与发展,2019,29(8):142 – 146.

[31]冷菲,李巍. 基于 XGBoost 对肺鳞癌和肺腺癌的分类预测[J]. 首都医科大学学报,2019,40(6):889 – 893.

[32]Poon L C, Kametas N A, Maiz N, et al. First – trimester prediction of hypertensive disorders in pregnancy[J]. Hypertension,2009,53(5):812 – 818.

[33] Zhang S J, Meng P Q, Zhang J N, et al. Machine learning models for genetic risk assessment of infants with non – syndromic orofacial cleft [J]. Genomics, Proteomics & Bioinformatics, 2018, 16(5):354 – 364.

[34] Miotto R, Li L, Kidd B A, et al. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records[J]. Scientific Reports,2016,6:26094.

[35] Shah S A, Velardo C, Farmer A, et al. Exacerbations in chronic obstructive pulmonary disease: Identification and prediction using a digital health system[J]. Journal of Medical Internet Research,2017,19(3):e69.

[36] Finkelstein J, Jeong I C. Machine learning approaches to personalize early prediction of asthma exacerbations[J]. Annals of the New York Academy of Sciences,2017,1387(1):153 – 165.

(收稿日期:2020 – 09 – 24)