

基于随机森林算法的糖尿病预测研究^{*}

杨美洁 唐建军

(重庆医科大学医学信息学院 重庆 400016)

〔摘要〕 采用随机森林算法建立糖尿病预测模型, 介绍数据建模与评估步骤和方法、选择标准。以受试者工作特征曲线的曲线下面积、灵敏度、特异性、正确率等指标评价模型的预测效果, 结果表明随机森林预测模型对糖尿病具有较强的预测能力。

〔关键词〕 随机森林; 糖尿病; 灵敏度; 特异性; AUC

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2019.09.011

Study on Predictions of Diabetes Mellitus Based on Random Forest Algorithm YANG Meijie, TANG Jianjun, *Medical Informatics College, Chongqing Medical University, Chongqing 400016, China*

〔Abstract〕 The prediction models of diabetes mellitus has been built by taking random forest algorithm. The paper introduces data modeling, steps and methods of evaluation, selection criteria, and evaluates the predictive effect of the models using indicators of the area under curve of receiver operating characteristic curve, sensitivity, specificity, accuracy, etc. The result indicates that the random forest prediction model possesses a relative strong prediction ability as for diabetes mellitus.

〔Keywords〕 random forest; diabetes mellitus; sensitivity; specificity; AUC

1 引言

糖尿病是由于机体不正确产生或使用胰岛素而引起蛋白质、脂肪、糖等物质在血液、肝脏等器官中代谢紊乱的疾病^[1]。据国际糖尿病联盟统计, 2017 年大约 4.25 亿成年人 (20 ~ 79 岁) 患有糖尿病, 到 2045 年糖尿病成年人患者将增加到 6.29 亿。作为 3 大健康杀手之首的糖尿病, 全球的发病率在逐年上升。目前国内未有系统的糖尿病风险预测模型及统一的致病因素, 其发病机制和发病机理尚不

清楚^[2]。近年来国内的学者根据糖尿病的致病因素开始构建疾病风险预测模型。陈松景等利用反向传播神经网络的患病因素对血糖影响的量化进行分析, 筛选出性别、体重等 10 个因素, 该方法可推广到其他疾病危险因素的量化分析^[3]。郭奕瑞等利用神经网络和 Logistic 回归算法建立 2 型糖尿病的预测模型, 使用 ROC 曲线来评估模型, 得出神经网络模型较 Logisti 回归模型具有更好的预测效果的结论^[4]。王勋等采用不同核函数的支持向量机对 2 型糖尿病的患病风险进行预测, 认为径向基核函数支持向量机模型的预测效果最佳^[5]。余丽玲等根据最小方差将支持向量机和自回归积分滑动平均组合得到新的预测模型, 采用多组临床实验数据对比 ARI-MA 模型、SVM 模型、神经网络模型的结果^[6]。Mani S 等利用支持向量机、K 近邻、决策树、随机森林、贝叶斯等方法对电子病历中的 2 型糖尿病的

〔修回日期〕 2019-06-19

〔作者简介〕 杨美洁, 硕士, 讲师, 发表论文多篇。

〔基金项目〕 重庆市社会事业与民生保障科技创新专项 (项目编号: cstc2015shms-ztx10003)。

风险进行预测^[7]。

目前机器学习广泛地应用在图像处理、语言识别、智能控制、文本挖掘、医药等领域。随机森林是机器学习非常重要的一种算法。本研究利用 UCI 数据集之中的 Pima 印第安人糖尿病数据集进行分析研究,对数据集进行特征选择和特征工程确定分析的自变量,再利用随机森林方法建立糖尿病的预测模型,最后获取模型的正确率、灵敏度、特异性、AUC 等指标,得出影响糖尿病的重要因素。

2 糖尿病数据的建模与评估

2.1 概述

本研究采用 UCI 数据集的 Pima 印第安人糖尿病数据集 2 000 例。自变量属性为:怀孕次数、2 小时口服葡萄糖耐量测试的血糖浓度、舒张期血压、肱三头肌皮肤皱壁厚度、2 小时血清胰岛素、BMI、糖尿病家族史、年龄等。因变量属性为:5 年内是否患过或感染过糖尿病。数据显示 1 316 人患有糖尿病,684 人未患病。患有糖尿病与未患有糖尿病的比例为 1.924:1,数据类别较平衡。

2.2 特征工程和特征选择

删除数据集中的缺失值和异常值后共有 1 153 条数据,其中患有糖尿病的数据 455 条,未患糖尿病的数据 698 条。利用 Shapiro - Wilk 检验显示自变量均不服从正态分布后,通过 Spearman 检验得出自变量间的相关性,结果发现年龄 - 怀孕次数的相关系数为 0.63、血清胰岛素 - 血糖的相关系数为 0.45、血脂厚度指数 - BMI 的相关系数为 0.66。其 p 值分别为 $2.8e - 194$ 、 $7.5e - 25$ 、 $1e - 109$,均小于 0.05,变量间可能存在一定的相关关系。其他相关系数均小于 0.4,不存在共线性。因此需要对年龄、怀孕次数、血清胰岛素、血糖、血脂厚度指数、BMI 等进行共线性分析,结果显示其方差膨胀因子 (Variance Inflation Factor, VIF) 均小于 4。通常情况下当 $VIF \geq 5$ 或 $VIF \geq 10$ 时,可认为自变量间存在严重共线性^[8]。因此自变量间不存在共线性和相关性,全部纳入模型进行分析。

2.3 糖尿病预测模型建立

随机森林是以决策树作为基分类器的集成算法,通过组合多颗决策树投票得出最佳结果,由 Leo Breiman 于 2001 年提出^[9]。随机森林算法具有较高的正确率、不易过拟合、处理高维特征、对噪声和异常数据具有较高的容忍度等优点^[10]。本研究将数据集分为 80% 的训练集和 20% 的测试集,采用 5 折交叉验证获取模型的最优参数。

2.4 糖尿病预测模型评估

本研究采用正确率、灵敏度、特异性、AUC (ROC 曲线下面积) 对预测模型进行指标评估。通过分析得出随机森林预测糖尿病的相关指标:正确率为 0.9,灵敏度为 0.827,特异性为 0.936,ROC 曲线下的面积 AUC 为 0.962。各项指标均较高,具有较强的预测能力。ROC 曲线图,见图 1。

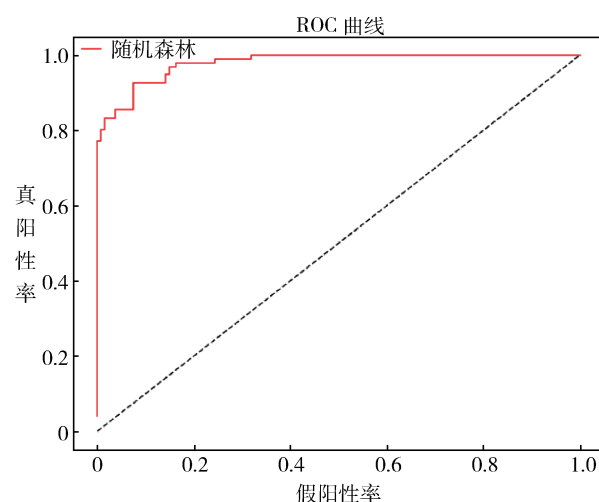


图 1 随机森林的 ROC 曲线

利用随机森林对影响糖尿病患病因素的重要性进行分析,见图 2。可以看出特征的重要程度从高到低依次为血糖值、血清胰岛素、年龄、BMI、糖尿病家族史、肱三头肌皮肤皱壁厚度、怀孕次数、血压。因此,糖尿病患者应多关注血糖值、血清胰岛素,将他们控制在稳定的范围内才会降低发病率;由于患病风险随着年龄的增长而增加,建议年长者日常生活中应控制 BMI 在正常范围内。

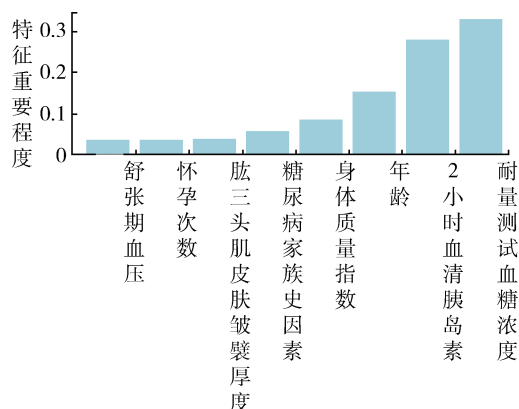


图 2 特征重要程度排序

3 结语

本研究利用随机森林算法对 UCI 数据集中 Pima 印第安人是否患有糖尿病进行预测。首先对 2 000 例数据进行空值和异常值的删除后剩余 1 153 条数据，对怀孕次数、2 小时口服葡萄糖耐量测试的血糖浓度、舒张期血压、肱三头肌皮肤皱壁厚度、2 小时血清胰岛素、BMI、糖尿病家族史、年龄等 8 个自变量进行相关性分析和共线性分析，纳入随机森林模型进行预测研究。将数据集分为 80% 训练和 20% 预测后利用五折交叉验证获取模型的最优参数，采用正确率、灵敏度、特异性、AUC 等指标对预测模型进行评估，结果显示随机森林对此数据集具有较

强的预测能力。最后对此数据集的糖尿病患病的重要影响因素进行分析，旨在对临床决策提供支持。

参考文献

- 1 林婉媚, 彭成丰, 曾彩云. 糖尿病诊断中采用生化检验与常规检验效果差异 [J]. 中国医药科学, 2016 (13): 139 - 141.
- 2 朱亚楠. 基于数据挖掘技术的 2 型糖尿病患病风险预测研究 [D]. 秦皇岛: 燕山大学, 2017.
- 3 陈松景, 罗森林, 潘丽敏, 等. 2 型糖尿病患病因素对血糖影响的定量分析 [J]. 北京理工大学学报, 2014, 34 (2): 201 - 206.
- 4 郭奕瑞, 李玉倩, 王高帅, 等. 人工神经网络模型在 2 型糖尿病患病风险预测中应用 [J]. 郑州大学学报: 医学版, 2014 (2): 180 - 183.
- 5 王勋, 陈大方. 支持向量机在建立 2 型糖尿病预测模型中的应用 [J]. 中国慢性预防与控制, 2010 (6): 560 - 562.
- 6 余丽玲, 陈婷, 金浩宇, 等. 基于支持向量机和自回归积分滑动平均模型组合的血糖值预测 [J]. 中国医学物理学杂志, 2016 (4): 381 - 384.
- 7 Mani S, Chen Y, Elasy TA, et al. Type 2 Diabetes Risk Forecasting from EMR Data Using Machine Learning [J]. Amia Annu Symp Proc, 2012 (2012): 606 - 615.
- 8 丁儿林, 孔丹莉, 毛宗福. 多重线性回归分析中的常用共线性诊断方法 [J]. 数理医药学杂志, 2004, 17 (4): 299 - 300.
- 9 Breiman, L. Random Forests [J]. Machine Learning, 2001, 45 (1): 5 - 32.
- 10 陆家发, 张国明, 陈安琪. 基于深度学习的疾病诊断 [J]. 医学信息学杂志, 2017, 38 (4): 39 - 43.

(上接第 29 页)

- 6 吴琼. IPv6 简介和远程教育展望 [J]. 开放教育研究, 2003 (1): 51 - 53.
- 7 李瑞俊. 基于 CIDR 的网络 IP 地址规划及应用 [J]. 长春师范学院学报 (自然科学版), 2014, 33 (10): 33 - 36.
- 8 王璐, 代涛, 郑英, 等. 平凉市构建新农合分级诊疗体系的 PEST - SWOT 分析 [J]. 中国数字医学, 2017, 12 (1): 15 - 18.
- 9 张静, 水黎明, 江志琴, 等. 宁波云医院的 SWOT 分析和对策研究 [J]. 中华全科医学, 2017, 15 (2): 338 - 342.
- 10 冯昌琪, 陈文, 汪月. 分级诊疗管理信息系统基本功能的设计研究 [J]. 卫生软科学, 2018, 270 (7): 37 - 42.
- 11 田小庆, 单清, 李扬. 医疗机构之间双向转诊存在的问
- 题及对策思考 [J]. 江苏卫生事业管理, 2012, 23 (4): 31 - 32.
- 12 中华人民共和国国家卫生和计划生育委员会. 国务院办公厅关于推进医疗联合体建设和发展的指导意见 [J]. 中华人民共和国国务院公报, 2017, 24 (13): 14 - 18.
- 13 周新科, 王辉, 崔鹏. 基于 IPv6 的学习资源共享平台设计 [J]. 自动化与仪器仪表, 2017 (10): 63 - 64.
- 14 卿卉. 基于 IPv6 的组播技术及应用系统研究 [D]. 武汉: 武汉理工大学, 2008.
- 15 王虎峰, 王鸿蕴. 关于构建分级诊疗制度相关问题的思考 [J]. 中国医疗管理科学, 2014, 4 (1): 28 - 30.