文章编号：2095-2163(2021)04-0176-04 中图分类号：TP181;R714.252 文献标志码：A

基于随机森林算法的心血管疾病预测研究

石胜源，朱 磊，叶 琳，罗铁清
（湖南中医药大学 信息科学与工程学院，长沙 410208）

摘 要： 血管疾病严重威胁着人类的健康，高发病率、高致残率、高死亡率是心血管疾病的主要特点，因此心血管疾病的预测研究显得尤为重要。本文探讨了随机森林算法在心血管疾病预测中的应用效果。在 Kaggle 网站上下载关于心血管疾病的数据集，用随机森林算法进行训练，实验结果由准确性、精度、召回率、*F*1 - *score* 评价标准来评价其性能的好坏（评价就包括好坏）。本文将其与逻辑回归（Logistic Regression）、K 近邻分类器（K-nearest neighbor classifier）、支持向量机（SVM）进行了比较，实验结果表明，随机森林算法的性能优于其他算法，其准确率为 73.55，精度为 75.51，召回率为 70.11，*F*1 - *Score* 为 72.71。通过基尼重要性评价能从多因素中识别出影响心血管疾病的重要因素，这意味着随机森林算法在心血管疾病预测中具有较大的优势，从而对心血管疾病的预测研究和早期病人的及时有效治疗具有重要意义。

关键词： 随机森林；心血管疾病；疾病预测

Disease Predict of Cardiovascular Based on Random Forest

SHI Shengyuan, ZHU Lei, YE Lin, LUO Tieqing
(School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China)

【Abstract】 Cardiovascular disease is a serious threat to human health, high incidence, high disability rate and high mortality are its main characteristics, thus cardiovascular disease prediction research is particularly important. This paper discusses the effect of stochastic forest algorithm application in cardiovascular disease prediction. Concerning datasets of cardiovascular disease were downloaded from Kaggle and trained using a random forest algorithm, whose performance was evaluated by accuracy, accuracy, recall, and *F*1 - *score*. In this paper, we compare its result with Logistic Regression, K-nearest neighbor classifier and Support Vector Machine. The experimental result shows that the performance of random forest algorithm is better than other algorithms, the accuracy is 73.55, the precision is 75.51, the recall rate is 70.11 and *F*1 - *Score* is 72.71. By Gini importance evaluation, the important factors affecting cardiovascular disease can be identified from multi-factors, which means the stochastic forest algorithm has a great advantage in cardiovascular disease prediction. And this is of great significance for the prediction of cardiovascular disease and the timely and effective treatment of early patients.

【Key words】 Random forest; Cardiovascular disease; Disease prediction

0 引 言

随着中国经济的发展和人民生活水平的提高，人们的饮食结构和生活方式发生了很大的改变，这也给健康带来了很多问题，而健康问题是促进人的全面发展的必然要求，是经济社会发展的基础条件[1]。慢性病是严重威胁中国居民健康的一类疾病，已成为影响国家经济社会发展的重大公共卫生问题[2]。

慢性病是指不构成传染、具有长期积累形成疾病形态损害的疾病的总称。有报告显示，70%国人有过劳死危险，76%的白领处于亚健康状态，20%国人患慢性病[3]。其中，中国心脑血管病现患人数 2.9 亿，心血管疾病仍占城乡居民总死亡原因首位[4]。具体情况见图 1 所示。

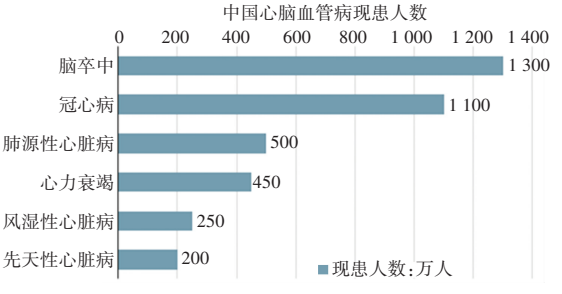

图 1 中国心脑血管病现患人数

Fig. 1 Number of cardiovascular disease patients in China

在心血管病领域，疾病的诊断或是病情的转归原本就十分复杂，更需要依托现代计算机技术来对疾病进行准确评估和预测[5]。

Ambale-Venkatesh 等对比研究了随机生存森林法和传统危险因素对多种族动脉粥样硬化的 6 种心

作者简介： 石胜源（1998-），男，硕士研究生，主要研究方向：机器学习；朱 磊（1998-），男，硕士研究生，主要研究方向：医学图像处理；叶 琳（1995-），女，硕士研究生，主要研究方向：数据治理；罗铁清（1966-），男，硕士，副教授，主要研究方向：软件工程。

通讯作者： 罗铁清 Email：tieqingluo@163.com

收稿日期： 2020-12-18

血管事件的预测差异[6]；Li 等利用 SVM 来构建预测心血管疾病的模型，研究对象为甲状腺功能正常且非糖尿病的 538 位患者，并挖掘疾病影响变量[7]；郑晓燕基于机器学习的算法，建立了心血管疾病预测模型并开发了相对应的预测 Web 系统[8]；李孝虔利用卷积神经网络构建了心脏病预测模型[9]；王振飞等提出了一种自适应模块化神经网络结构模型，采取聚类的方法预测心血管疾病[10]。

基于心血管疾病数据维数高和数据之间关系复杂的特点，且已有研究利用随机森林在其它疾病的预测上取得了良好的效果，本文提出采取随机森林来预测心血管疾病，从而防止模型过拟合，并进一步提高预测的稳定性和准确率。

1　对象与方法

1.1　研究对象

选取知名机器学习竞赛网站 Kaggle 上关于心血管疾病的数据集，数据信息包括：年龄、身高、体重、性别、收缩压、舒张压、胆固醇、葡萄糖、是否吸烟、是否饮酒、是否经常运动等信息，共计 70 000 例。

1.2　数据预处理

参考中国成人健康体检基本数据集标准（HRC00.04），对以上数据集进行清洗，剔除重复项及数据异常者，如身高和体重异常，舒张压高于收缩压，血压为负值数据等。此次构建 CVD 疾病预测模型，最终纳入分析的样本总量为 60 142。对数据进行数据归一化（Scaler）处理，具体方法为式（1）：

$$x_{scale} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{1}$$

其中，x 代表数据集当前数据的值；x_{min} 代表数据集的最小值；x_{man} 代表数据集的最大值。

1.3　随机森林模型的建立

随机森林是一种机器学习方法，本质上是许多决策树的集合，可以对数据进行综合分类，关联性检验、预测和解释[11]。研究的结局变量为研究对象是否发生 CVD，是研究的根本目的，解释变量为一系列对 CVD 发生率有影响的危险因素，如血压，胆固醇的检测值，用于支持结局变量的准确性。建立随机森林模型的步骤如下：

（1）通过简单交叉验证方法对样本随机划分为训练子集和测试子集，其中训练集样本为 70%，测试集为 30%；

（2）通过 sklearn 库方法对特征进行重要性评估，并根据评分对特征降序排序，运行 SWSFS 过程筛选最优变量数；

（3）利用网格搜索的方式对随机森林在局部范围内找出最优参数；

（4）建立决策树，不对树进行修剪，根据所有决策树的投票结果决定数据的分类。

1.4　特征重要性评估

随机森林在拟合训练的同时也完成了特征的重要性评估，其目的是对解释变量在结局发展中的重要性进行评价，评估基于基尼指数式（2）：

$$Gini(P) = \sum_{k=1}^{K} P_K(1 - P_k). \tag{2}$$

其中，k 代表 k 个类别，P_K 代表类别 k 的样本权重。

在随机森林模型的建立过程中，结局变量的分类主要是依据解释变量的分类程度，变量的重要性评分越高，则表明该变量对模型准确率贡献越多。

2　结　果

2.1　纳入病例的基本信息

共计纳入 60 142 例研究对象，年龄为 30～65（53.33）岁，其中女性为 39 254 例（其中 12 553±1 例为患者），男性为 20 888 例（其中 3578±1 例为患者），随机选取 30%为测试组，70%为训练组，两组一般资料差异无统计学意义（$p > 0.05$），具有可比性。

2.2　随机森林筛选变量

根据重要性评分，对特征进行降序排序见表 1，并运行（SWSFS）过程，即从重要性得分最大的变量开始，逐个引入变量，每加入一个变量即运行一次随机森林并绘制准确率图，如图 2 所示。结果显示，在变量数为 8 时具有最优的分类准确性，所以将重要性评分排在前 8 位的特征纳入随机森林模型进行分析，所选变量为收缩压、胆固醇、性别、葡萄糖、舒张压、年龄、体重、吸烟情况。

表 1　变量重要性得分

Tab. 1　Importance of variables score

序号	变量名	编码	变量重要性得分
1	收缩压	ap_lo	0.423 4
2	胆固醇	cholesterol	0.165 0
3	性别	gender	0.152 7
4	葡萄糖	gluc	0.079 7
5	舒张压	ap_hi	0.060 6
6	年龄	age	0.051 4
7	体重	weight	0.034 6
8	吸烟情况	smoke	0.011 9
9	身高	height	0.007 7
10	饮酒情况	alco	0.005 6
11	运动情况	activie	0.002 8

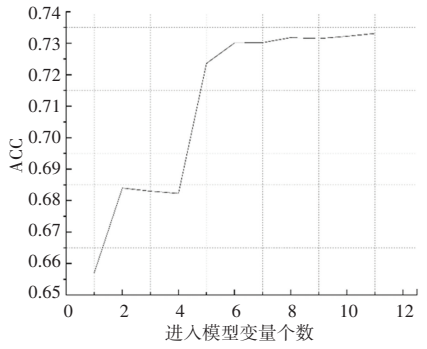

图 2 SWSFS 过程分类准确图

Fig. 2 SWSFS process classification accuracy chart

2.3 网格搜索局部最优参数

运行网格搜索,分类器个数从 10~100,以 10 为步长,最大深度从 10~100,以 5 为步长。结果显示当分类器个数为 60,决策树最大深度为 10 时,模型准确率最高,即为局部最优参数。

2.4 模型预测性能评价

为了衡量该模型的性能,将其与逻辑回归(Logistic Regression)、K 近邻分类器(k-nearest neighbor classifier)、支持向量机(SVC)进行了比较。

实验使用混淆矩阵进行预测结果的分类,共分为 *TP*、*TN*、*FP*、*FN* 4 类,见表 2。

表 2 混淆矩阵

Tab. 2 Confusion Matrix

	正例(患有 CVD)	反例(没有 CVD)
预测正确	真正例(TP)	真反例(TN)
预测错误	伪正例(FP)	伪反例(FN)

实验结果由准确性、精度、召回率、*F*1 - *score* 这几个指标来表示。具体公式如式(3)~(6):

准确性:

$$acc = \frac{TP + TN}{TP + FP + TN + FN}. \tag{3}$$

精度:

$$pre = \frac{TP}{TP + FP}. \tag{4}$$

召回率:

$$rec = \frac{TP}{TP + FN}. \tag{5}$$

*F*1 - *score*:

$$F1 = 2 * \frac{pre * rec}{pre + rec}. \tag{6}$$

本文实验的预测结果见表 3,可以看出随机森林模型在 4 个指标上都优于其它 3 种方法。

表 3 不同预测方法的性能指标

Tab. 3 Performance index of different prediction methods

预测方法	准确性	精度	召回率	*F*1 - *score*
随机森林	73.55	75.51	70.11	72.71
逻辑回归	72.02	73.61	68.25	70.83
K 近邻	62.76	62.88	61.29	62.07
SVC	67.04	67.77	63.64	65.64

为了更清晰地比较各预测方法的性能,将表 3 中的各数据表示成图 3 形式。

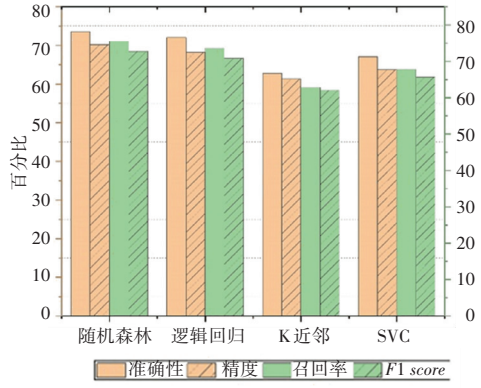

图 3 不同预测方法的性能指标

Fig. 3 Performance index of different prediction methods

显然,在心血管数据集上进行的实验表明本文提出方法有更好的性能。分析原因如下:

(1)随机森林采用了集成算法,本身精度比大多数单个算法要好;

(2)两个随机性的引入,使得随机森林具有一定的抗过拟合能力和抗噪声能力,对比其他算法具有一定的优势。

3 结束语

本文提出将随机森林算法应用在心血管疾病预测上,并与其他主流机器学习分类算法作了比较,证明随机森林相对于其他算法更适用于该心血管疾病数据集,且预测准确率为 73.5(±1),对于识别心血管病人,并对其进行及时、有效的医疗有一定的现实意义。

但本研究还存在一定的局限性,打算在后续的工作中进一步改善以下问题:首先,对于数据的清洗过于笼统,未能咨询相关医学背景的学者达到更深层次的清洗;其次,对特征的选择上,只能简单地利用现有的特征进行重要性评估,可能更为重要的特征未被纳入数据集;最后,本研究采用数据来源为北美地区,所得结果具有局限性,需要进一步扩大样本来源验证结果的适用性。 (下转第 181 页)