基于 CatBoost 算法的中青年颈动脉粥样硬化预测方法

丁瑶[1,2] 张小玉[1,2] 许杨[1] 高理升[1] 孙怡宁[1] 王世军[3] 马祖长[1]

摘 要 **目的** 探究 CatBoost 算法在中青年颈动脉粥样硬化预测中的应用价值,为中青年颈动脉粥样硬化早期筛查提供一种可行的技术手段。**方法** 以 2016—2018 年期间在北京某医院体检中心进行健康体检的 2 258 位中青年为研究对象,根据颈动脉彩超检查结果诊断是否有颈动脉粥样硬化。使用下采样技术对样本进行平衡处理。分析变量重要性进行特征选择,构建 CatBoost 模型。利用 Logistic 回归和人工神经网络两类机器学习算法构建模型,并与 CatBoost 模型进行比较分析。以灵敏度、特异性、准确率及受试者工作特征(receiver operating characteristic,ROC)曲线下的面积(area under the ROC curve,AUC)作为模型的评价指标。**结果** CatBoost 模型在测试集上的灵敏度、特异性、准确率和 AUC 均最高,分别为 82.8%、96.7%、90.3%、0.92。Logistic 回归模型和神经网络模型的灵敏度、特异性和准确率均介于 62.4%~73.3%之间,AUC 均介于 0.72~0.78 之间。重要性分析表明影响中青年颈动脉粥样硬化最重要的三个因素依次是年龄、腰高比、高密度脂蛋白胆固醇。**结论** CatBoost 算法在中青年颈动脉粥样硬化预测中的应用具有一定的可行性。相比于其他传统算法,具有较高的诊断价值。

关键词 颈动脉粥样硬化; 特征选择; CatBoost; Logistic 回归; 人工神经网络

DOI:10.3969/j.issn.1002-3208.2020.05.004.

中图分类号 R318.04 **文献标志码** A **文章编号** 1002-3208(2020)05-0470-07

本文著录格式 丁瑶,张小玉,许杨,等. 基于 CatBoost 算法的中青年颈动脉粥样硬化预测方法[J]. 北京生物医学工程, 2020, 39(5):470-476,522. DING Yao, ZHANG Xiaoyu, XU Yang, et al. Carotid arteriosclerosis prediction method based on CatBoost algorithm in young and middle ages[J]. Beijing Biomedical Engineering, 2020, 39(5):470-476,522.

Carotid arteriosclerosis prediction method based on CatBoost algorithm in young and middle ages

DING Yao[1,2], ZHANG Xiaoyu[1,2], XU Yang[1], GAO Lisheng[1], SUN Yining[1], WANG Shijun[3], MA Zuchang[1]

1 *Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei* 230031; 2 *Department of Precision Instrument, University of Science and Technology of China, Hefei* 230026; 3 *Dalian Medical University, Dalian, Liaoning Province* 116044

Corresponding author: MA Zuchang (E-mail: zcma121@126.com)

【Abstract】 **Objective** To explore the application value of CatBoost algorithm in the prediction of carotid atherosclerosis in young and middle-aged people, and to provide a feasible technical means for early screening of carotid arteriosclerosis in young and middle-aged people. **Methods** A total of 2258 young and middle-aged people who underwent a health checkup at a medical checkup center in a Beijing hospital from 2016 to 2018 were selected as the research subjects, carotid

基金项目:国家重点研发计划(2017YFB1002204)、中国科学院科技服务网络计划(KFJ-STS-ZDTP-079)、安徽省科技重大专项(18030801133)资助

作者单位:1 中国科学院合肥物质科学研究院智能机械研究所(合肥 230031)
2 中国科学技术大学(合肥 230026)
3 大连医科大学(辽宁大连 116044)

通信作者:马祖长, 研究员。E-mail:zcma121@126.com

arteriosclerosis was diagnosed based on the results of carotid color doppler ultrasound. Samples were balanced using under-sampling techniques. Feature selection is performed by analyzing the importance of variables. The CatBoost prediction model was built. In addition, models were constructed using two types of machine learning algorithms, Logistic regression and artificial neural network, and compared with CatBoost model. Sensitivity, specificity, accuracy, and the ROC curve areas (AUC) were used as the evaluation indicators of the model. **Results** The CatBoost model had the highest sensitivity, specificity, accuracy, and AUC on the test set, which were 82.8%, 96.7%, 90.3%, and 0.92 respectively. The sensitivity, specificity and accuracy of the models constructed by Logistic regression and neural network were between 62.4% and 73.3%, and the AUCs were between 0.72 and 0.78. Importance analysis showed that the three most important factors affecting carotid arteriosclerosis in young and middle-aged people were age, waist-to-height ratio, and high-density lipoprotein cholesterol. **Conclusions** The CatBoost algorithm is feasible in the prediction of carotid sclerosis in young and middle-aged people. Compared with other traditional algorithms, it has higher diagnostic value.

【**Keywords**】 carotid arteriosclerosis; feature selection; categorical boosting; Logistic regression; artificial neural network

0　引言

心血管疾病是全球死亡和致残的主要原因,也是我国居民死因之首[1-2]。近年来,我国心血管疾病患病率持续上升,且呈现年轻化趋势[3]。部分地区监测报告表明,急性冠心病事件在中青年男性中增幅较大[4],2012—2013 年全国卒中流行病学调查显示,20 岁以上人群卒中年龄标化年发病率为 246.8/10 万[5]。随着心血管疾病防治关口的前移,对于中青年人群及早识别颈动脉粥样硬化,对其危险因素进行临床治疗和综合干预,延缓动脉硬化进程,是心血管疾病防控的关键。研究表明,动脉硬化是心血管疾病的主要病理学原因,颈动脉粥样硬化作为动脉硬化的临床替代终点,与心血管疾病风险的增加直接相关[6-8]。

多年来,科研工作者致力于发展动脉硬化早期筛查技术,当前颈动脉粥样硬化预测模型的建立主要基于 Logistic 回归方法。一项基于中老年人群建立的颈动脉粥样硬化评分模型,C 值检验结果为 0.8,准确性低于 80%[9]。另一项面向工人研究中,在不同性别中建立了两个回归预测模型,但准确性也均未达到 80%[10]。已有的研究受限于 Logistic 回归方法固有的局限性,所建立模型属于广义线性回归模型,在非线性问题的处理上性能较弱。模型输入变量之间的交互作用难以表达,且对模型中自变量多重共线性较为敏感。回归方程的构建过分依赖于训练样本,且易出现过拟合现象,导致模型的泛化能力很差[11],因此建立的模型预测性能不足。

人们进一步尝试利用人工神经网络、决策树、Cox 比例风险回归等方法建立预测模型。一项以 SEER 数据库中的样本为研究对象的研究中,利用决策树建立了乳腺癌患者的预后模型,为医师判断患者预后情况及治疗效果提供了辅助[12]。在另一项研究中,研究人员通过 Cox 比例风险回归建立了 2 型糖尿病发病风险预测模型,并在健康管理人群中获得较好的预测能力[13]。尽管这些方法在某些疾病预测方面具有一定的价值,但它们仍存在许多的局限性。人工神经网络虽然具有较强的非线性映射能力以及较高的泛化能力,但对数据量有一定的要求,并且在模型的训练中收敛速度慢,容易陷入局部极小值而造成网络训练失败[14]。Cox 比例风险回归模型和决策树相比于 Logistic 和人工神经网络,前者在应用时必须满足时间变量已知的前提条件,后者在决策树的生成过程中容易偏向于样本比例过大的特征,而且可能会因样本发生一点点改动而导致树结构的剧烈改变[15-16]。现有的分类算法大都采用单个全局优化模型,单分类器模型性能有限,泛化能力普遍较弱,容错性也较差。

Boosting 技术作为一种集成算法,在训练样本量有限、所需训练时间较短、缺乏调参知识等场景中具有不可或缺的优势。它主要是通过一组分类器的串行迭代训练多个模型,每个分类器采用的样本分布都和上一轮的学习结果有关;然后对模型进行筛选,将一组预测正确率较低的模型组合得到一个整

体预测正确率最强的模型,以此来进行更高精度的分类[17]。

Boosting 技术的主要代表算法有 lightGBM、XgBoost 等。CatBoost 是 Yandex 于 2017 年提出的一种基于决策树梯度提升的新型集成算法,也是 Boosting 策略的一种实现方式。由于 CatBoost 引入了排序提升(ordered boosting)的方式替换传统的梯度估计算法,并且在处理分类特征时使用了更有效的策略,因而它有效克服了过拟合的发生和预测偏移的问题,从而提高了模型的泛化能力。CatBoost 模型减少了对广泛的超参数调优的需求,具有较高的鲁棒性。目前已有研究表明 CatBoost 算法在糖尿病和手足口病的预测中取得了显著的效果[18-19]。

现有的颈动脉粥样硬化预测模型建模的样本年龄区间较大,分布从 30 岁至 80 岁[9,21]。但是随着中青年人群心血管疾病发病情况日益突出,针对中青年人群建立颈动脉粥样硬化预测模型对心血管疾病的早期防控具有重要意义。

本文通过使用 CatBoost 算法实现对中青年颈动脉粥样硬化的预测,并与传统的 Logistic 回归和人工神经网络进行比较分析,探究该算法在中青年颈动脉粥样硬化预测中的应用价值,为中青年颈动脉粥样硬化早期筛查提供一种可行的技术手段。

1 数据采集与预处理

1.1 数据采集

本研究收集了 2016—2018 年在北京某医院体检中心进行健康体检的 2 258 人的脱敏数据进行分析,其中男性有 1 766 人,女性有 492 人。受试者的年龄在 30—39 岁之间,平均年龄 36 岁。数据维度有 118,在咨询相关医师后,排除了与本研究无关的字段;对于反映某个健康状况若出现冗余指标,研究选取最具有诊断意义的一个进行分析。最终,共选出 22 个字段作为模型的初始输入变量。其中分类变量包括性别、吸烟、饮酒、高血压、糖尿病、家族史、家务活等,连续变量包括年龄、身体指数(body mass index, BMI)、血压值、空腹血糖值、血脂四项、腰围身高比等。颈动脉粥样硬化由颈动脉内-中膜厚度(carotid intima media thickness, CIMT)>0.9 mm 或彩超图像显示有硬化斑块存在而进行确诊[20]。颈动脉粥样硬化的有无作为模型的输出变量。

1.2 数据预处理

由于颈动脉粥样硬化人数只有 307 人,占总样本的 13.59%,数据出现了不均衡现象。数据的不平衡性易导致模型输出倾向于无动脉硬化,模型的预测性能会降低。因此本试验对原始数据进行了下采样处理,达到两类样本基本平衡。

2 试验方法

2.1 CatBoost 算法

Boosting 是一种框架算法,由 Schapire 在 1990 年提出。它的基本思想是不断使用基础分类模型进行建模,通过对一系列的基础分类模型进行线性加权组合得到强分类器。Boosting 算法通过联合多个弱识别率的分类器构建高识别率的分类器,提高分类模型的性能。

Boosting 算法在机器学习和数据挖掘中方面有良好的应用效果,但容易出现各种数据偏移问题。2017 年由 Yandex 首次提出了 CatBoost 算法,克服了原有 Boosting 算法的不足,在类别型特征的处理和预测偏移的处理上进行了改善。研究表明,CatBoost 算法的分类性能得到了提升,在慢病预防上得到了良好的应用[18]。

CatBoost 算法的主要步骤为下:(1) 对于训练集 X 中的每一个样本 X_i,CatBoost 都会利用 X_i 之外的全部样本训练并得到模型 M_i;(2) 采用排序提升(ordered boosting)利用 M_i 计算样本 x_i 的梯度估计;(3) 利用新模型重新对样本 x_i 打分,形成一个基学习器;(4) 对基学习器加权处理,获得最终的分类器。

2.2 输入指标集的优化

通过对变量进行筛选,精简变量,利用更为重要的特征构建模型,实现对模型的优化。具体采用重要性分析和单因素分析对变量进行精简。CatBoost 模型通过变量重要性分析进行特征选择。在以树作为基础模型的集成学习算法中,通过分析变量纳入分析和不纳入分析的两种情况下,比较测试集上模型性能的差异,来判断变量的重要程度。如果性能变化很大,说明该变量很重要,对因变量的影响更大。这一部分的数据分析集成在算法里面,可以通过在训练好的模型中,提取 model. feature importance 属性,输出不同输入特征的重要程度结果,分析各输入变量的重要性,从而确定对因变量具有重要影响

的因子。

试验中分析所有输入变量的重要性,剔除对颈动脉粥样硬化没有预测能力的变量,筛选出与颈动脉粥样硬化具有高相关性的变量,构建基于高相关性变量 CatBoost 模型。优化后的模型中,输入指标集减少,对变量进行重要性分析,显示影响最大的变量重要程度进一步提高,影响较小的变量重要程度则有所下降。Logistic 回归算法在构建模型时,可以按照变量的逐步筛选自动选择建模的指标。在已构建好模型中,没有贡献的变量已经被剔除在模型之外,所以模型输入指标的优化已经在模型建立的过程中同步完成。

对于神经网络模型,首先纳入全部变量,构建基于全变量的预测模型。为了进一步简化变量,首先通过单因素分析有颈动脉粥样硬化人群和无颈动脉粥样硬化人群中具有显著性差异的指标,确定与颈动脉硬化具有相关性的变量。利用单因素分析的结果,通过纳入单因素分析筛选后的指标构建基于高相关性变量神经网络,完成输入指标集的优化。

2.3　模型的建立

试验主要基于 CatBoost 算法建立中青年颈动脉粥样硬化预测模型,并与 Logistic 和人工神经网络建立的模型进行比较。将进行下采样后的数据随机划分为 70%的训练样本和 30%的测试样本,其中训练样本用于模型的构建,测试样本用于模型的评估与验证。CatBoost 模型在 Python 上通过 CatBoost 算法包实现。Logistic 回归模型由 SPSS 统计学软件实现,通过组织单因素分析筛选出来的变量作为模型的输入量,构建二元 Logistic 回归模型。人工神经网络通过 SPSS 上集成的 MLP 算法包进行实现。试验分析流程如图 1 所示。

2.4　模型的评价指标

临床上用于评价预测模型的指标通常包括灵敏度(sensitivity)、特异性(specificity)、准确率(accuracy)和受试者工作特征(receiver operating characteristic,ROC)曲线下的面积(area under the ROC curve, AUC)。灵敏度是指实际为正的样本中被预测为正的比例,特异性是指实际为负的样本中被预测为负的比例,准确率是指分类正确的样本占总样本的比例,AUC 是综合灵敏度和特异性对模型进行的评价指标, AUC 大于 0.7 说明诊断价值较高。其中灵敏度、特异性和准确率可由分类结果的混淆矩阵计算,数值越大表明模型的性能越好。

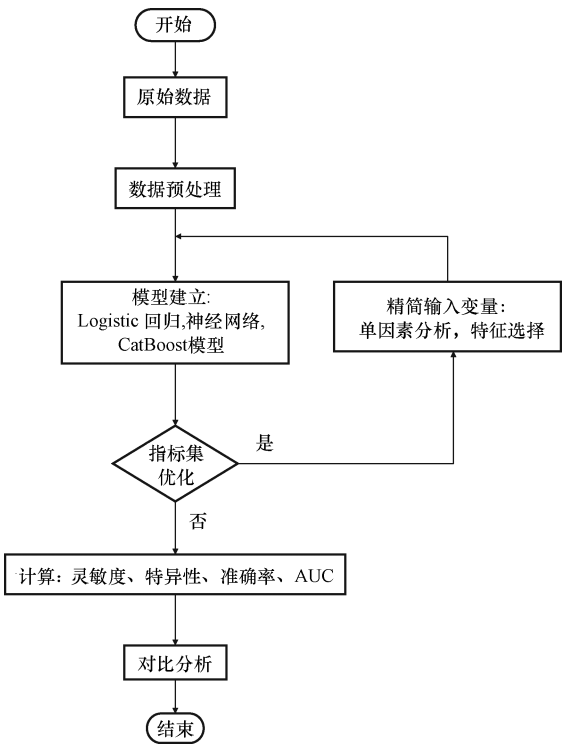

图 1　试验流程

Figure 1　The experimental procedure

表 1　分类结果的混淆矩阵

Table 1　Confusion matrix for classification results

实际	预测		合计
	1	0	
1	TP	FN	TP+FN
0	FP	TN	FP+TN
合计	TP+FN	FN+TN	TP+FN+FP+TN

注:TP(true positive)为真正,表示实际样本为正预测也为正;FP(false positive)为假正,表示实际样本为负预测为正;FN(false negative)为假负,表示实际样本为正预测为负;TN(true negative)为真负,表示实际样本为负预测也为负。

$$灵敏度 = \frac{TP}{TP + FN} \tag{1}$$

$$特异性 = \frac{TN}{FP + TN} \tag{2}$$

$$准确率 = \frac{TP + TN}{TP + FN + FP + TP} \tag{3}$$

3 试验结果

3.1 模型的构建与优化

对全部 22 个输入变量进行分类特征处理后，采用 Ordered bosting 梯度提升算法训练模型，得到基于全变量的 CatBoost 模型。进一步通过分析输入变量的重要性进行特征选择，剔除变量文化程度、是否被动吸烟、交通方式、收缩压、舒张压、是否糖尿病、是否高血压、糖尿病家族史、高血压家族史后，选出了 13 个对颈动脉粥样硬化具有预测能力的变量，分别是年龄、性别、饮酒、是否吸烟、工作类型、家务活、身体指数(body mass index，BMI)、腰高比、空腹血糖和血脂 4 项，构建了基于高相关性变量的 CatBoost 模型。

3.2 模型的验证分析

在开始对样本进行切分时，有 30%的样本构成测试集，用来对建立的模型进行验证分析。每一个样本由输入的变量大小计算预测概率，再利用阈值进行是否有颈动脉粥样硬化的判断。通过对比实际结果，建立测试样本的混淆矩阵，计算模型评价指标的大小，确定模型的预测性能。结果显示基于全变量的 CatBoost 模型在测试集上的 ROC 曲线下面积为 0.89，准确率为 86.5%；基于高相关性变量的 CatBoost 模型在测试集上的 ROC 曲线下面积为 0.92，准确率为 90.3%。

3.3 模型预测性能比较

利用测试样本对 Logistic 回归和神经网络方法建立的模型进行验证，通过输出值分析所有模型的预测性能，结果如表 2 所示。CatBoost 模型的准确率和 AUC 明显高于 Logistic 回归模型和神经网络模型。基于全变量的 CatBoost 模型准确率和 AUC 分别为 86.5% 和 0.89；而基于高相关性变量的 CatBoost 模型准确率和 AUC 最高，分别为 90.3%和 0.92。单独分析 Logistic 回归和神经网络模型，发现 Logistic 回归模型具有最低的准确率和 AUC，分别为 64.1%和 0.72。神经网络模型性能有所提高，其中基于高相关性变量神经网络模型的准确率在 71.3%，AUC 为 0.78。

表 2 不同模型的预测性能比较

Table 2 Comparison of prediction performance of different models

模型	灵敏度	特异性	准确率	AUC
Logistic 回归	65.7%	62.4%	64.1%	0.72
基于全变量神经网络	72.2%	66.7%	70.7%	0.74
基于高相关性变量神经网络	73.3%	69.0%	71.3%	0.78
基于全变量 CatBoost 模型	76.3%	96.7%	86.5%	0.89
基于高相关性变量 CatBoost 模型	82.8%	96.7%	90.3%	0.92

各预测模型的 ROC 曲线如图 2 所示。

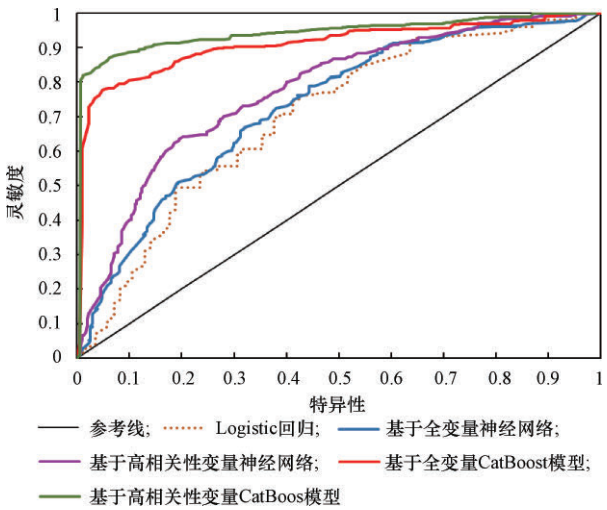

图 2 不同模型的 ROC 曲线

Figure 2 ROC curves of different models

3.4 特征重要性分析

分析各变量的特征重要性，明确各变量对颈动脉粥样硬化的影响程度。特征重要性排名前 10 的变量如图 3 所示。

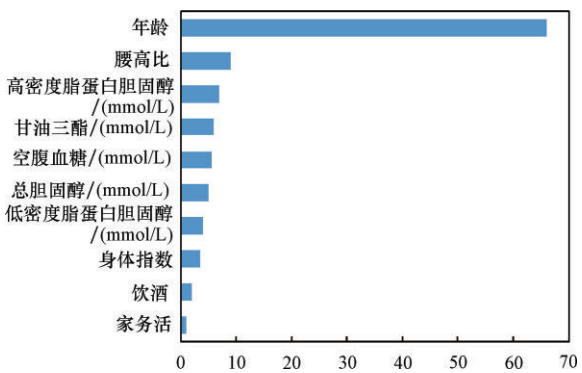

图 3 特征重要性

Figure 3 Feature importance

重要性排名前三的依次是年龄、腰高比和高密度脂蛋白胆固醇,且年龄对颈动脉粥样硬化的重要性远高于其他特征变量,得分近 70,其他特征值均在 10 分以下。空腹血糖水平和其他三项血脂指标排第 4 位至第 7 位。身体指数 BMI 虽然也是一个特征变量,但其得分低于腰高比。饮酒和是否有干家务活的习惯则是影响较大的两种生活方式因素。

4　讨论

本试验在评估多个模型的预测结果后发现,CatBoost 模型对于中青年颈动脉粥样硬化预测的性能最好,准确性最高,诊断效果优于传统的 Logistic 回归模型和人工神经网络模型。同时根据重要性分析对特征变量进行影响程度的判断,发现年龄、腰高比、高密度脂蛋白胆固醇是影响颈动脉粥样硬化最重要的 3 个因素。

CatBoost 作为新的集成性算法,因其特有的梯度估计算法,得到的模型稳定性高,泛化能力强。由于模型中减少了对超参数的调整,因此降低了过拟合发生的情况。CatBoost 模型因其特有的对分类变量的处理方法,可以直接自动处理分类特征而不需要对数据进行预处理,因此对于各类问题的适用性较高。

在对特征变量的重要性进行分析后,通过筛选精简了模型的输入变量。当输入变量减少后,基于高相关性变量的 CatBoost 模型比基于全变量的 CatBoost 模型的性能更好。说明了特征选择对于 CatBoost 模型的建立至关重要,在恰当地剔除噪声从而进行正确的特征选择后会提高模型的准确性。

已有的研究大多基于 Logistic 回归方法建立颈动脉粥样硬化预测模型。一项对杭州社区的 1000 例抽样调查数据建立的颈动脉粥样硬化预测评分系统中, AUC 为 0.8,灵敏度和特异度分别为 71.3% 和 75.7%[9]。另一项针对随访人群的劲动脉硬化预测模型分析了在不同性别中的诊断效果,其中男性和女性的 AUC 分别为 0.835 和 0.809[10]。与本研究相比,其他基于 Logistic 回归方法建立的模型的结果较好,这可能是由于在建模的时候纳入了更多的实验室测量指标以及生活方式因素。同时本研究中的模型是基于横断面研究,相比于基于随访研究的结果性能较差。另外,其他模型未考虑模型的泛化能力,仅评估了模型在训练样本中的预测效果。

Logistic 回归模型最大的局限性在于非线性的处理。通过广义线性回归表达了各自变量对因变量的关系,忽略了各自变量之间的交互作用。Logistic 回归对模型输入变量要求是数值型变量,对于字符串型的多分类变量需要转换成哑变量形式,在建模之前对变量的预处理降低了建模的效率。

人工神经网络相对于 Logistic 回归模型在各个性能指标上均得到了一定的提升。这是因为人工神经网络由于它的结构中引入了非线性函数,因此对于非线性问题的处理较好。同时在精简输入变量之后,基于高相关性变量的神经网络模型相对于基于全变量的神经网络模型,性能也有所提升。这说明变量的相关性程度会影响神经网络模型的建立,利用少量有价值的信息能够建立更可靠的模型。但神经网络模型同样会因为对分类变量进行数值型转化的预处理工作而影响建模速度。此外,神经网络在训练过程中容易陷入局部极小值,导致训练失败。

特征重要性分析结果指出年龄是最重要的特征变量。年龄作为公认的心血管疾病危险因素,同样对颈动脉粥样硬化的发生发展有着重要的影响[20]。一项横断面研究中发现仅年龄对颈动脉粥样硬化的 ROC 曲线下面积达 0.773,风险比值比为 1.15[21]。随着年龄的增加,人体血管内膜逐渐发生肌性纤维增厚,血管弹性减弱,硬度增加[22]。腰高比是评价中心型肥胖的指标,有研究中表明,其对颈总动脉和颈内动脉的内中膜厚度的比值比分别是 1.71 和 1.56,高于其他肥胖评价指标[23]。腰高比反映了内脏脂肪的堆积,而内脏脂肪组织具有易于动脉硬化的细胞因子和脂肪因子[24],因此它与动脉硬化之间的相关性更高[25-26]。既往研究表明低的高密度脂蛋白胆固醇水平是冠状动脉疾病的独立危险因素[27]。高密度脂蛋白是一种具有多种生物学活性的异质脂蛋白,主要由蛋白质和脂质构成。高密度脂蛋白胆固醇代表了高密度脂蛋白上的胆固醇含量,由于它的逆转作用,具有重要的抗动脉粥样硬化功能[28]。

5　结论

本研究提出了一种基于 CatBoost 算法的中青年颈动脉粥样硬化预测方法。通过和其他几种模型进行对比发现,基于该算法的模型在中青年颈动脉粥样硬化诊断中取得了良好的效果。由该模型分析出

的颈动脉粥样硬化影响因素与临床经验相一致。这表明 CatBoost 算法在中青年颈动脉粥样硬化预测中具有良好的应用价值,可以作为中青年颈动脉粥样硬化预测的一种可行性技术。

参考文献

[1] Timmis A, Townsend N, Gale C, et al. European society of cardiology: cardiovascular disease statistics 2017[J]. European Heart Journal, 2017, 39(7): 508-579.

[2] 胡盛涛, 高润霖, 刘力生, 等.《中国心血管病报告 2018》概要[J]. 中国循环杂志, 2019, 34(3): 6-17.
Hu SS, Gao RL, Liu LS, et al. Summary of the 2018 report on cardiovascular diseases in China [J]. Chinese Circulation Journal, 2019, 34(3): 6-17.

[3] 中国心血管病预防指南(2017)写作组, 中华心血管病杂志编辑委员会. 中国心血管病预防指南(2017)[J]. 中华心血管病杂志, 2018, 46(1): 10-25.

[4] 邓木兰, 李河, 石美玲, 等. 广州市番禺区农民急性冠心病事件发病率及 20 年变化趋势[J]. 中华心血管病杂志, 2014, 42(3): 236-240.
Deng ML, Li H, Shi ML, et al. Prevalence of acute coronary heart disease among farmers in Panyu, Guangzho: a 20-year population-based study [J]. Chinese Journal of Cardiology, 2014, 42(3): 236-240.

[5] Wang WZ, Jiang B, Sun HX, et al. Prevalence, incidence and mortality of stroke in china: results from a nationwide population-based survey of 480687 adults[J]. Circulation, 2017, 135(8): 759.

[6] 田进伟, 符亚红. 动脉粥样硬化易损斑块快速进展机制与临床治疗进展[J]. 中国动脉硬化杂志, 2019,27(4): 277-280.
Tian JW, Fu YH. The mechanism of progression and clinical intervention of atherosclerotic vulnerable plaque [J]. Chinese Journal of Arteriosclerosis, 2019,27(4): 277-280.

[7] Pang HY, Ye YC, Ding FM, et al. Risk factors for progression of carotid intima-media thickness in patients with systemic lupus erythematosus: protocol for an observational cohort study in China [J]. BMJ Open, 2019, 9(9): e030721 .

[8] 张萌, 郑慧, 张敏, 等 . 颈动脉不稳定型斑块、血脂、血压与急性脑梗死关系的病例对照研究[J]. 中华疾病控制杂志, 2016, 20(8): 831-834.
Zhang M, Zheng H, Zhang M, et al. Case-control study on association of carotid artery unstable carotid plaque, blood lipid and blood pressure with acute cerebral infarction[J]. Chinese Journal of Disease Control & Prevention, 2016, 20(8): 831-834.

[9] 童璐莎, 姜雯红, 严慎强, 等. 基于社区抽样调查数据的颈动脉疾病预测模型[J]. 中华急诊医学杂志, 2014, 423(7): 801-805.
Tong LS, Jiang WH, Yan SQ, et al. The predictive model of carotid angiopathy set from randomly sampled community data [J]. Chinese Journal of Emergency Medicine, 2014, 23(7): 801-805.

[10] 王琪, 李娟生, 蒲宏全, 等. 某随访人群颈动脉粥样硬化发生影响因素及风险预测能力研究[J]. 中华疾病控制杂志, 2019, 23(4): 382-386.
Wang Q, Li JS, Pu HQ, et al. Influence factors and predictive ability of a risk prediction model for carotid atherosclerosis in a follow-up population[J]. Chinese Journal of Disease Control & Prevention, 2019, 23(4): 382-386.

[11] 牟冬梅, 任珂. 三种数据挖掘算法在电子病历知识发现中的比较[J]. 现代图书情报技术, 2016(6): 102-109.

[12] Zhang MH, Zhang X, Guo X, et al. Prognostic factors of breast cancer with machine learning method based on SEER database[J]. Beijing Biomedical Engineering, 2019, 38(5): 486-491,497.

[13] 苏萍, 杨亚超, 杨洋, 等. 健康管理人群 2 型糖尿病发病风险预测模型[J]. 山东大学学报(医学版), 2017, 55(6): 82-86.
Su P, Yang YC, Yang Y, et al. Prediction models on the onset risks of type 2 diabetes among the health management population [J]. Journal of Shandong University (Health Sciences), 2017, 55(6): 82-86.

[14] 尤晓东, 苏崇宇, 汪毓铎. BP 神经网络算法改进综述[J]. 民营科技, 2018 (4): 152-153.

[15] 严若华, 李卫, 谷鸿秋,等. Cox 比例风险回归模型 C 统计量的计算方法及其 SAS 实现[J]. 中华疾病控制杂志, 2016, 20(9): 953-956,961.
Yan RH, Li W, Gu HQ, et al. Calculation of C statistics for the Cox proportional hazards regression models and its implementation in SAS[J]. Chinese Journal of Disease Control & Prevention, 2016, 20(9): 953-956,961.

[16] 马晓梅, 徐学琴, 闫国立, 等. BP 神经网络和决策树分析在重症手足口病临床早期预警指标中的应用[J]. 中国卫生统计, 2019, 36(3): 381-383.

[17] 徐继伟, 杨云. 集成学习方法:研究综述[J]. 云南大学学报(自然科学版), 2018, 40(6): 36-46.
Xu JW, Yang Y. A survey of ensemble learning approaches[J]. Journal of Yunnan University(Natural Science), 2018, 40(6): 36-46.

[18] 苗丰顺, 李岩, 高岑, 等. 基于 CatBoost 算法的糖尿病预测方法[J]. 计算机系统应用, 2019, 28(9): 215-218.
Miao FS, Li Y, Gao C, et al. Diabetes Prediction Method Based on CatBoost Algorithm[J]. Computer Systems & Applications, 2019, 28(9): 215-218.

[19] 王斌, 冯慧芬, 王芳, 等. 基于机器学习的 Cat Boost 模型在预测重症手足口病中的应用[J]. 中国感染控制杂志, 2019, 18(1): 12-16.

(下转第 522 页)

and serum albumin[J]. Kidney International,2004,65:1449-1460.
[2] 周福德,王梅. 北京市血液透析单位透析用水及透析液质量的持续性质量改进[J]. 中国血液净化,2006,5(4):178-181.
Zhou FD, Wang M. The continuous quality improvement of hemodialysis water and dialysate in Hemodialysis Centers in Beijing[J]. Chinese Journal of Blood Purification,2006,5(4):178-181.
[3] Rao M, Jaber BL, Balakrishnan VS. Inflammatory biomarkers and cardiovascularrisk: association or causeandeffect? [J]. Seminars in Dialysis,2006,19:129-135.
[4] Gordon SM, Oettinger CW, Bland LA, et al. Pyrogenic reactions in patients receiving conventional, high-efficiency, or high-flux hemodialysis treatments with bicarbonate dialysate containing high concentrations of bacteria and endotoxin [J]. Journal of the American Society of Nephrology :JASN, 1992, 2:1436-1444.
[5] 魏媛媛,马迎春. 透析用水及透析液的微生物检测[J]. 中国血液净化,2014,4(4):335-339.
[6] 国家食品药品监督管理总局. 血液透析及相关治疗用水(YY 0572-2015)[S]. 北京:中国标准出版,2015.
[7] 国家食品药品监督管理总局. 血液透析及相关治疗用浓缩物(YY0598-2015)[S]. 北京:中国标准出版社,2015.
[8] 国家药典委员会. 中华人民共和国药典[M]. 北京:中国医药科技出版社,2015:155-156.
[9] Dawids SG,Vejlsgaard R. Bacteriological and clinical evaluation of different dialysate delivery systems [J]. Acta Medica Scandinavica, 1976,199:151-155.
[10] Lonnemann G,Krautzig S, Koch KM. Quality of water and dialysate in hemodialysis [J]. Nephrology Dialysis Transplantation,1996,11(6):946-949.
[11] 梅长林,叶朝阳,赵学智. 实用透析手册[M]. 北京:人民卫生出版社,2003:37.
[12] 朱雪峰,蒋惠云,张德力,等. 血液透析用水及透析液细菌污染情况监测分析[J]. 蚌埠医学院学报,2006,31(6):660-662.
[13] 朱笠, 邹梅, 梁玉红, 等. 医院透析液污染现况调查[J]. 中国消毒学杂志, 2008, 25(5):539-540.
[14] 杨云海,姜佳莹,周艳霞,等. 血液透析液细菌污染的调查[J]. 中华医院感染学杂,2002,12(7):529-532.
[15] 王静, 黄文治, 陈军军,等. 血液透析导管相关性感染影响因素研究及其护理对策[J]. 中国医药导报, 2019, 16(16):174-177.
Wang J,Huang WZ,Chen JJ,et al.Study on influencing factors of catheter-related infection in hemodialysis patients and nursing countermeasures[J]. China Medical Herald, 2019, 16(16):174-177.
[16] 田爱辉,曹立云. 血液透析设备的消毒[J]. 中国血液净化,2009,8(1):5-7.

(2020-04-21 收稿,2020-08-10 修回)

(上接第 476 页)

Wang B, Feng HF, Wang F, et al. Application of CatBoost model based on machine learning in predicting severe hand-foot-mouth disease[J]. Chinese Journal of Infection Control, 2019, 18(1):12-16.
[20] Pan XF, Lai YX, Gu JQ, et al. Factors significantly associated with the increased prevalence of carotid atherosclerosis in a northeast chinese middle-aged and elderly population [J]. Medicine, 2016, 95(14):e3253.
[21] 钟金鹏. 基于实验室指标的颈动脉粥样硬化模型的建立与评价[D]. 重庆:重庆医科大学, 2011.
Zhong JP. Establishment and evaluation of the predictive model for carotid arteriosclerosis based on laboratorial parameters [D]. Chongqing:Chongqing Medical University, 2011.
[22] Sun Z. Aging, arterial stiffness, and hypertension [J]. Hypertension, 2015, 65(2):252-256.
[23] Zhang ZQ, He LP, Xie XY, et al. Association of simple anthropometric indices and body fat with early atherosclerosis and lipid profiles in chinese adults[J]. Plos One, 2014, 9(8):e104361.
[24] Lee HJ, Hwang SY, Hong HC, et al. Waist-to-hip ratio is better at predicting subclinical atherosclerosis than body mass index and waist circumference in postmenopausal women [J]. Maturitas, 2015, 80(3):323-328.
[25] Ge WZ, Faruque P, Fen W, et al. Association between anthropometric measures of obesity and subclinical atherosclerosis in Bangladesh[J]. Atherosclerosis, 2014, 232(1):234-241.
[26] 陈玉香, 吴晓秋, 凌云,等. 心外膜脂肪定量及脂肪因子水平与冠脉粥样硬化斑块易损性的相关性研究[J]. 中国医药导报, 2019, 16(2):45-48.
Chen YX, Wu XQ, Ling Y, et al. Correlation between epicardialadipose quantification and adipose factor levels and vulnerability of coronary atherosclerotic plaques[J]. China Medical Herald, 2019, 16(2):45-48.
[27] Nakajima H, Momose T, Misawa T. Prevalence and risk factors of subclinical coronary artery disease in patients undergoing carotid endarterectomy: a retrospective cohort study [J]. International Angiology, 2019, 38(4):312-319.
[28] 刘蕾, 姜涛. 高密度脂蛋白胆固醇和高密度脂蛋白颗粒与颈动脉粥样硬化发生及严重程度的相关性[J]. 岭南心血管病杂志, 2017, 23(6):673-676.
Liu L, Jiang T. Correlations between HDL-C, HDL-P with the incidence and severity of carotid arterial atherosclerosis[J]. South China Journal of Cardiovascular Diseases, 2017, 23(6):673-676.

(2019-12-11 收稿,2020-03-20 修回)