

基于多因素回归分析和机器学习算法的 冠心病预测模型构建及比较

陆浩轩¹, 徐瑾妍², 程可爱¹, 谢燕青¹, 王丽¹, 计礼丽¹,
周忠¹, 杨卓¹, 景胜¹, 何文明^{1*}

(1. 宁波大学医学院附属医院 心血管内科, 浙江 宁波 315020;

2. 宁波大学医学院附属医院 神经内科, 浙江 宁波 315020)

摘要: 本文旨在构建基于临床电子病历数据的冠心病预测模型。回顾性收集了2015年至2020年在宁波大学医学院附属医院住院期间, 接受选择性冠状动脉造影的患者的临床数据, 分别应用决策树、朴素贝叶斯和逻辑回归算法构建冠心病预测模型, 比较3种模型的预测性能。共收集354例患者数据, 其中冠心病患者140例, 非冠心病患者214例, 根据逻辑回归、朴素贝叶斯、决策树算法构建的3种预测模型的准确性分别为70.6%、89.5%、90.7%; 曲线下面积分别为0.676、0.869、0.921。所构建的3种预测模型均具备较好的冠心病预测能力, 具有用于冠心病筛查的潜在价值。

关键词: 冠心病; 机器学习; 决策树; 朴素贝叶斯

中图分类号: R541.4

文献标志码: A

文章编号: 1001-5132 (2022) 03-0057-06

冠状动脉粥样硬化性心脏病 (Coronary Atherosclerotic Heart Disease, CAD) 是一种常见的 cardiovascular 疾病, 简称冠心病。一直以来, 选择性冠状动脉造影是诊断冠心病的“金标准”, 但其是一种有创操作, 且价格相对昂贵, 难以作为冠心病的筛查手段。现存的一些冠心病风险评估模型, 包括弗拉明翰风险评分模型^[1-2]等已被证明对诊断冠心病的准确率有限^[3], 也不适用于冠心病早期筛查。

鉴于现有方法的局限性, 本研究尝试构建无创冠心病预测模型, 基于相关的临床数据, 对冠心病进行预测分类。人工智能技术在医学研究领域中被广泛应用于构建疾病预测模型, 其对数据处理具有传统统计学方法不具备的独特优势^[4-7]。目前国内外也报道了基于机器学习算法构建冠心病预测模型的研究, Babaoglu 等^[8]基于支持向量机算法挖掘了23个临床数据特征, 构建了一种冠心病预测模型, 该模型的诊断准确率达到79.71%。

本文以包含354例患者的临床电子病历数据为数据源, 使用逻辑回归分析和两种常见的机器学习算法构建预测模型, 以准确性、曲线下面积 (Area Under the Curve, AUC)、敏感性和特异性作为评价指标, 评估基于冠心病预测模型进行冠心病筛查的可行性。

1 资料收集

本研究回顾性收集了2015—2020年在宁波大学医学院附属医院住院期间行选择性冠状动脉造影的患者资料, 收集的临床数据包括年龄、性别、身高、体重、收缩压、舒张压、高血压病史、高脂血症史、糖尿病史、血管疾病史、肾病病史、吸烟史、饮酒史、实验室检查结果、心超左室射血分数。收集的患者资料符合如下要求: ①既往无冠心病及相关疾病; ②数据资料较完整; ③年龄 ≥ 20 岁。所有患者均签署冠状动脉造影知情同意书。

收稿日期: 2021-11-02.

宁波大学学报(理工版) 网址: <http://journalg.nbu.edu.cn/>

基金项目: 宁波市自然科学基金 (2021J240, 202003N4231); 宁波大学人体生物力学研究院开放基金 (CJ-HBIO202104)。

第一作者: 陆浩轩 (1995-), 男, 浙江杭州人, 医师, 主要研究方向: 心血管医学. E-mail: 271621481@qq.com

*通信作者: 何文明 (1981-), 男, 浙江绍兴人, 副主任医师, 主要研究方向: 心血管医学. E-mail: fyhewenming@nbu.edu.cn

2 数据方法

使用单因素分析筛选与冠心病存在相关性的特征因素,然后使用多因素逻辑回归分析构建冠心病预测模型.同时,使用两种常用的机器学习算法(决策树和朴素贝叶斯)构建冠心病预测模型.比较3种算法构建的模型的预测性能.模型的评价标准为准确性、AUC值、敏感性和特异性.

2.1 特征筛选与多因素逻辑回归分析

特征选择通过单因素分析进行,即将所有纳入的变量进行分析,检验变量对冠心病分类结果有无显著性影响.单因素分析结果见表1.

表1 单因素分析结果

特征	<i>F</i>	<i>P</i>
年龄	12.358	<0.001
性别	21.063	<0.001
吸烟史	5.521	0.019
饮酒史	0.608	0.436
体重	0.071	0.790
身高	0.708	0.401
收缩压	0.186	0.666
舒张压	1.339	0.248
高血压病史	4.750	0.030
高脂血症史	9.734	0.002
糖尿病史	18.076	<0.001
肾病病史	0.664	0.416
血管疾病史	6.925	0.009
血红蛋白	0.467	0.495
白细胞计数	0.367	0.545
血小板	2.404	0.122
血钠	14.454	<0.001
血钾	1.402	0.237
血钙	4.853	0.028
血糖	15.135	<0.001
血肌酐	16.359	<0.001
尿素氮	8.067	0.005
甘油三酯	0.253	0.615
总胆固醇	2.606	0.107
高密度脂蛋白胆固醇	1.228	0.269
低密度脂蛋白胆固醇	2.031	0.155
促甲状腺素	0.180	0.671
左室射血分数	5.098	0.025

以 $P<0.1$ 作为特征筛选条件,筛选出13个变量:年龄、性别、吸烟史、高血压病史、高脂血症史、糖尿病史、血管疾病史、血钠、血钙、血糖、血肌酐、尿素氮、左室射血分数.通过前向迭代法进行多因素逻辑回归分析并构建预测分类模型.

2.2 决策树算法

决策树算法可基于数据样本的分布概率,通过类似树枝结构形态的方式,实现多层判断,从而判定目标所属类别 n .决策树有3种常用的求解方法:ID3、C4.5、CART.本文选用ID3法求解,其核心原理在于信息熵原理,即通过信息熵增益选择数据集内对于信息增益最大的特征作为分类属性,逐步确定树的分支,最终实现决策树的构建.

信息熵用于反映随机特征的不确定性,信息熵越大,样本的不确定性越大.

假设数据集为 D , $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_i, y_i), \dots, (\mathbf{x}_m, y_m)\}$, 其中: $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)})^T$ 为输入数据,表示特征个数; $y_i \in \{1, 2, 3, \dots, K\}$ 表示类标记; $i = 1, 2, \dots, m$; m 为样本容量.数据集 D 中第 K 类样本所占的比例为 P_K ,则 D 的信息熵为:

$$\text{Ent}(D) = - \sum_{K=1}^{|D|} P_K \log_2 P_K, \quad (1)$$

式中, $\text{Ent}(D)$ 的值越小,反映样本分布的不确定性越小,当 P_K 值为0或1时, $\text{Ent}(D) = 0$.

计算以 a 进行数据划分时的信息增益:

$$\text{Gain}(D) = \text{Ent}(D) - \sum_{v=1}^V \frac{D^v}{D} \text{Ent}(D^v), \quad (2)$$

式中: V 为根据属性 a 划分出的类别数量; D^v 为类别 v 的样本数; $\text{Ent}(D)$ 为数据划分前的信息熵; $\sum_{v=1}^V \frac{D^v}{D} \text{Ent}(D^v)$ 为数据集划分后的信息熵.

2.3 朴素贝叶斯算法

朴素贝叶斯算法是贝叶斯分类算法中最常用的算法之一.

首先给出贝叶斯概率公式, $P(B)$ 为先验概率, $P(B|A)$ 表示后验概率, $P(A|B)$ 表示条件概率:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \quad (3)$$

对于数据集 (X, Y) ,参照贝叶斯概率公式,算法依照输入-输出的概率分布,求解后验概率最大的类别作为预测输出,假定该类别为 i ,后验概率

为 $P(Y_i|X)$:

$$P(Y_i|X) = \frac{P(X|Y_i)P(Y_i)}{P(X)}.$$

(4)

朴素贝叶斯公式:

$$P(X|Y = y_i) = \prod_{j=1}^m P(x_j|Y = y_i),$$

(5)

$$P(y_i|x_1, x_2, \dots, x_m) = \frac{P(y_i) \prod_{j=1}^m P(x_j|y_i)}{P(x_1, x_2, \dots, x_m)} = \frac{P(y_i) \prod_{j=1}^m P(x_j|y_i)}{\prod_{j=1}^m P(x_j)}.$$

(6)

2.4 统计学处理

使用 SPSS 26.0 统计学软件分析, 计量资料用 $\bar{x} \pm s$ 表示, 计数资料用例数和百分比表示, 单因素和多因素分析采用逻辑回归分析, 以 $P < 0.05$ 为差异有统计学意义.

3 结果

3.1 描述性统计

表 2 列出了符合条件的 354 名患者的人口统计学特征. 根据选择性冠状动脉造影结果显示冠心病 140 例(39.5%), 非冠心病 214 例(60.5%).

表 2 人口统计学特征

变量	(计量)单位或 (计数)分类	统计值	
		冠心病	非冠心病
年龄	岁	65.00±12.502	60.55±11.044
性别	男	104 (74.3%)	108 (50.5%)
	女	36 (25.7%)	106 (49.5%)
吸烟史	不吸烟	77 (55.0%)	147 (68.7%)
	既往吸烟, 已戒	21 (15.0%)	23 (10.7%)
	目前吸烟	42 (30.0%)	44 (20.6%)
饮酒史	有	44 (31.4%)	59 (27.6%)
	无	96 (68.6%)	155 (72.4%)
高血压病史 (HBP)	有	86 (61.4%)	155 (72.4%)
	无	54 (38.6%)	59 (27.6%)
高脂血症史 (HPA)	有	30 (21.4%)	79 (36.9%)
	无	110 (78.6%)	135 (63.1%)
糖尿病史 (DM)	有	41 (29.3%)	25 (11.7%)
	无	99 (70.7%)	189 (88.3%)

续表

变量	(计量)单位或 (计数)分类	统计值	
		冠心病	非冠心病
肾病病史(RF)	有	7 (5.0%)	7 (3.3%)
	无	133 (95.0%)	207 (96.7%)
血管疾病史 (HVD)	有	30 (21.4%)	24 (11.2%)
	无	110 (78.6%)	190 (88.8%)
体重	kg	68.01±9.56	67.68±12.76
身高	cm	165.21±6.64	164.54±7.80
收缩压	mmHg	138.58±24.01	139.57±19.20
舒张压	mmHg	78.98±15.07	80.68±12.39
血红蛋白(HB)	g·L ⁻¹	137.66±17.28	138.87±15.77
白细胞计数 (WBC)	L ⁻¹	7.02±2.36	9.46±47.59
	L ⁻¹	200.81±54.95	209.96±53.79
血钠(Na)	mmol·L ⁻¹	139.70±2.37	140.58±1.99
血钾(K)	mmol·L ⁻¹	4.00±0.35	3.95±0.38
血钙(Ca)	mmol·L ⁻¹	2.27±0.12	2.30±0.14
血糖(Glu)	mmol·L ⁻¹	6.00±1.96	5.36±1.09
血肌酐(Cr)	μmol·L ⁻¹	83.06±26.13	73.08±20.15
尿素氮(BUN)	mmol·L ⁻¹	6.10±2.09	5.54±1.62
甘油三酯(TG)	mmol·L ⁻¹	1.65±1.15	1.58±1.16
总胆固醇(TC)	mmol·L ⁻¹	4.57±1.13	4.76±1.10
高密度脂蛋白胆固醇(HDL)	mmol·L ⁻¹	1.23±0.33	1.26±0.32
	mmol·L ⁻¹	2.71±0.90	2.58±0.80
低密度脂蛋白胆固醇(LDL)	mmol·L ⁻¹	2.71±0.90	2.58±0.80
	mmol·L ⁻¹	2.78±5.93	2.57±3.43
促甲状腺素 (TSH)	mmol·L ⁻¹	2.78±5.93	2.57±3.43
左室射血分数 (LVEF)	%	64.56±6.98	66.1±5.85
	%	64.56±6.98	66.1±5.85

3.2 多因素逻辑回归分析

多因素逻辑回归分析允许同时分析线性变量和分类变量, 将筛选出的 13 个变量纳入算法. 本文使用前向迭代法, 算法迭代 7 次后趋于平稳, 最终纳入了 7 个变量(年龄、性别、高血压病史、糖尿病史、血钠、血糖、血肌酐), 多因素逻辑回归分析结果见表 3. 表 4 显示了模型的迭代情况及最终的准确性. 结果显示, 迭代 7 次之后最终模型的准确性为 70.6%.

表 3 多因素逻辑回归分析结果

变量	回归系数(B)	标准误差	Wald 检验值	P	exp(B)	exp(B)的 95%置信区间	
						下限	上限
年龄	0.042	0.011	13.648	<0.001	1.043	1.020	1.067
性别	0.750	0.291	6.660	0.010	2.118	1.198	3.744
高血压病史	-0.823	0.271	9.248	0.002	0.439	0.258	0.746
糖尿病史	0.823	0.365	5.085	0.024	2.278	1.114	4.659
血钠	-0.125	0.061	4.171	0.041	0.882	0.782	0.995
血糖	0.200	0.100	3.970	0.046	1.222	1.003	1.487
血肌酐	0.015	0.007	5.061	0.024	1.015	1.002	1.028

表 4 回归模型迭代情况

步骤	实测	预测			总体正确率/%
		0	1	正确率/%	
1	0	214	0	100.0	60.5
	1	140	0	0	
2	0	167	47	78.0	63.8
	1	81	59	42.1	
3	0	176	38	82.2	66.1
	1	82	58	41.4	
4	0	179	35	83.6	70.9
	1	68	72	51.4	
5	0	179	35	83.6	70.6
	1	69	71	50.7	
6	0	176	38	82.2	70.6
	1	66	74	52.9	
7	0	176	38	82.2	70.6
	1	66	74	52.9	

3.3 模型预测结果

图 1 为 3 种模型的受试者工作特征(Receiver Operating Characteristic, ROC)曲线.

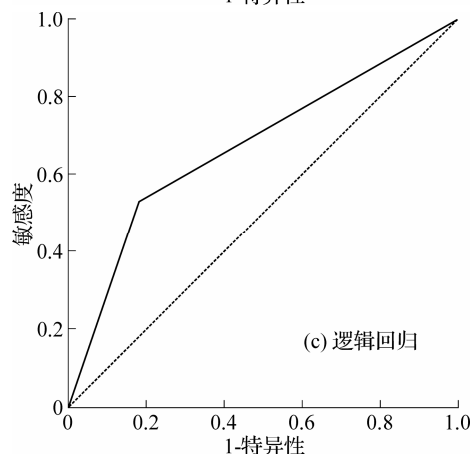
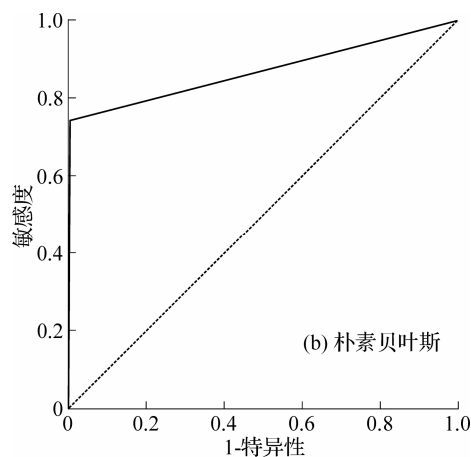
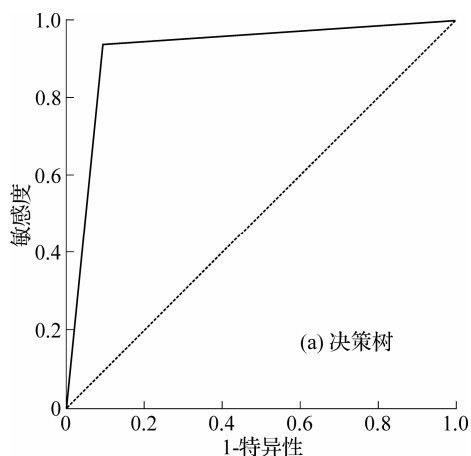


图 1 3 种模型的 ROC 曲线

由图 1 可知决策树算法构建的预测模型的 AUC 值最高, 为 0.921, 其次是朴素贝叶斯算法, AUC 为 0.869, 逻辑回归的 AUC 值最低, 为 0.676. 不同算法的预测分类结果见表 5. 可以看出, 2 种机器学习算法的准确率均在 85% 以上, 预测效果良好. 准确性更高的为决策树, 其准确性达到了 90.7%, 朴素贝叶斯次之, 准确性为 89.5%, 逻辑回归分析的准确性最低, 仅为 70.6%. 3 种模型的混淆

矩阵即冠心病的实际预测结果见表6。

表5 3种模型的评价

模型	准确性/%	AUC	敏感性/%	特异性/%
逻辑回归	70.6	0.676	52.9	82.2
决策树	90.7	0.921	93.6	90.7
朴素贝叶斯	89.5	0.869	74.3	99.5

表6 3种模型的混淆矩阵

结果	逻辑回归		朴素贝叶斯		决策树	
	阳性	阴性	阳性	阴性	阳性	阴性
阳性	74	66	104	36	131	9
阴性	38	176	1	213	194	20

4 结语

冠心病因其高发病率、高住院率、高死亡率,已然成为了倍受重视的国民健康问题^[9]。准确诊断冠心病依赖冠状动脉造影,然而,由于选择性冠状动脉造影操作的有创性和复杂性,其并不适合作为临床筛查冠心病的工具。以弗拉明翰风险评分为代表的心血管风险评分工具筛查冠心病的准确性十分有限,也难以满足临床筛查冠心病的需求。

以人工智能算法为代表的大数据挖掘技术的兴起为解决上述问题提供了新的思路。近年来,基于医疗大数据训练机器学习算法并构建疾病预测模型的研究不断涌现。Gould等^[10]基于常规临床和实验室检查数据,利用机器学习算法建立了肺癌预测筛查模型,且证明了该模型相比于现有的筛查标准更准确(AUC为0.86)。Dinh等^[11]基于机器学习构建了糖尿病和心血管疾病预测模型,所构建的两种预测模型的AUC分别为0.831和0.862,表现出了不俗的预测性能。Krishnamurthy等^[12]从台湾健康保险研究数据库获得的共病和药物数据进行慢性肾脏病的发病预测,使用卷积神经网络算法构建的预测模型在6个月和12个月预测的测试集AUC分别为0.957和0.954。

以上研究为本研究构建冠心病预测模型提供了思路,本研究分别应用机器学习算法和逻辑回归算法构建了冠心病预测分类模型,其中决策树算法显示了最高的准确性和AUC,决策树及朴素贝叶斯算法相比于传统逻辑回归方法均表现出明显的优越性。因此,机器学习算法适用于冠心病分

类预测,对于冠心病的筛查可能具有一定的应用前景。

本研究也存在一些局限性。首先,样本量相对较少,未来仍需进一步扩大样本量以进行外部验证该预测模型的准确性。其次,选取的变量中,关于心血管影像学的变量只纳入了心超显示的左室射血分数,未来将尝试进一步纳入更多的影像学变量并构建模型。最后,本研究使用的冠心病数据集存在一定程度的样本缺失,虽使用了均值填补法对缺失值进行了填补,但仍可能对模型的准确性产生了一定程度的影响。

综上,基于机器学习的冠心病预测分类模型是一种很有应用前景的工具,有利于早期筛查冠心病。本研究初步证明了机器学习算法对预测冠心病具有较好的准确性。未来可在此基础上进一步改良机器学习算法,同时发掘更多与冠心病具有相关性的临床数据,随着大规模心脏中心的建立和冠心病数据库的完善,这项研究的临床应用可推广性将被进一步评估。

参考文献:

- [1] D'Agostino R B Sr, Vasan R S, Pencina M J, et al. General cardiovascular risk profile for use in primary care: The Framingham heart study[J]. *Circulation*, 2008, 117(6): 743-753.
- [2] Pencina M J, D'Agostino R B Sr, Larson M G, et al. Predicting the 30-year risk of cardiovascular disease: The Framingham heart study[J]. *Circulation*, 2009, 119(24): 3078-3084.
- [3] Reriani M K, Flammer A J, Jama A, et al. Novel functional risk factors for the prediction of cardiovascular events in vulnerable patients following acute coronary syndrome[J]. *Circulation Journal: Official Journal of the Japanese Circulation Society*, 2012, 76(4):778-783.
- [4] Johnson K W, Torres Soto J, Glicksberg B S, et al. Artificial intelligence in cardiology[J]. *Journal of the American College of Cardiology*, 2018, 71(23):2668-2679.
- [5] Kann B H, Hosny A, Aerts H J W L. Artificial intelligence for clinical oncology[J]. *Cancer Cell*, 2021, 39(7):916-927.
- [6] He J X, Baxter S L, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine[J]. *Nature Medicine*, 2019, 25(1):30-36.

- [7] Topol E J. High-performance medicine: The convergence of human and artificial intelligence[J]. *Nature Medicine*, 2019, 25(1):44-56.
- [8] Babaoğlu I, Findik O, Bayrak M. Effects of principle component analysis on assessment of coronary artery diseases using support vector machine[J]. *Expert Systems with Applications*, 2010, 37(3):2182-2185.
- [9] 《中国心血管健康与疾病报告 2020》编写组. 《中国心血管健康与疾病报告 2020》要点解读[J]. *中国心血管杂志*, 2021, 26(3):209-218.
- [10] Gould M K, Huang B Z, Tammemagi M C, et al. Machine learning for early lung cancer identification using routine clinical and laboratory data[J]. *American Journal of Respiratory and Critical Care Medicine*, 2021, 204(4): 445-453.
- [11] Dinh A, Miertschin S, Young A, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning[J]. *BMC Medical Informatics and Decision Making*, 2019, 19(1):211.
- [12] Krishnamurthy S, Ks K, Dovgan E, et al. Machine learning prediction models for chronic kidney disease using health insurance claim data in Taiwan[J]. *Healthcare (Basel, Switzerland)*, 2021, 9(5):546.

Construction and comparison of coronary heart disease prediction models based on multivariate regression analysis and machine learning algorithm

LU Haoxuan¹, XU Jinyan², CHENG Ke'ai¹, XIE Yanqing¹, WANG Li¹, JI Lili¹,
ZHOU Zhong¹, YANG Zhuo¹, JING Sheng¹, HE Wenming^{1*}

(1.Department of Cardiology, Affiliated Hospital of Medical School, Ningbo University, Ningbo 315020, China;

2.Department of Neurology, Affiliated Hospital of Medical School, Ningbo University, Ningbo 315020, China)

Abstract: The current study was designed to construct coronary heart disease (CHD) prediction models based on clinical electronic medical record data. The clinical data was collected from the patients who underwent elective coronary angiography during hospitalization at the Affiliated Hospital of Medical School, Ningbo University from 2015 till 2020. Decision tree, naive Bayes, and logistic regression algorithms were applied to construct prediction models for coronary heart disease. The predictive performance of these three models was compared. A total of 354 patients, including 140 CHD patients and 214 non-CHD patients, were recruited. Using logistic regression, naive Bayes, and decision tree algorithms, the accuracies of the three prediction models were 70.6%, 89.5% and 90.7% respectively. The AUCs were 0.676, 0.869, 0.921, respectively. All the three prediction models had good predictive ability for CHD and may have screening potential for CHD.

Key words: coronary heart disease; machine learning; decision tree; naive Bayes

(责任编辑 韩 超)