基于特征自动识别的心肌梗死关键因素挖掘研究*

王颖晶 郑 涛 陈珊黎 邵维君 韩 刚 丁粉华

（上海交通大学医学院附属仁济医院信息中心 上海 200127）

〔摘要〕 利用人工智能技术，基于患者既往就诊数据进行机器学习相关算法分析，建立心肌梗死疾病特征自动识别模型，通过特征挖掘找出关键和主要致病因素，为医生提供定性或定量辅助诊断意见。

〔关键词〕 心肌梗死；机器学习；特征重要性

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.01.010

Study on the Mining of Key Factors of Myocardial Infarction Based on the Automatic Feature Recognition *WANG Yingjing, ZHENG Tao, CHEN Shanli, SHAO Weijun, HAN Gang, DING Fenhua, Information Center, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China*

〔**Abstract**〕 Artificial Intelligence (AI) technology is used to analyze the Machine Learning (ML) algorithm based on the patients' previous medical data, and an automatic recognition model for disease features of myocardial infarction is built. The key and main pathogenic factors are found through feature mining to provide qualitative or quantitative auxiliary diagnosis advices for doctors.

〔**Keywords**〕 myocardial infarction; Machine Learning (ML); feature importance

1 引言

1.1 研究背景

目前我国约有2.9 亿心血管病患者，其中心肌梗死患者数量约在250 万人左右。心肌梗死作为一种临床常见的危急重症，近年来死亡率不断攀升[1－2]，加强心血管病防控刻不容缓。心肌梗死疾病发生发展过程缓慢、影响因素较复杂，疾病早期预防及干预需要对发病危险因素进行定位评估，进而针对不同个体风险等级制定相应综合治疗或管理方案，以实现早期预防心肌梗死事件的目标。目前国内有少数研究关注急性心肌梗死（Acute Myocardial Infarction，AMI）患者的危险因素分布情况，但缺乏大规模、前瞻性的临床研究以提供详实数据[2]。随着人工智能、机器学习、数据挖掘等信息技术发展，医学数据挖掘和临床科研工作加速开展。临床大量结构化及半结构化数据为人工智能机器学习相关算法研究提供支撑。机器学习是人工智能领域的重要分支，通过将不同领域数据进行特征提取并自动学习，使模型不断适应数据特征从而提高性能[3]。机器学习逐渐与医学领域相结合，例如医学图像处理的诊断识别技术。利用医疗大数据及

〔修回日期〕 2021－06－11

〔作者简介〕 王颖晶，助理工程师；通讯作者：郑涛，正高级工程师。

〔基金项目〕 上海市信息化发展专项资金项目“面向仁济医院医联体的专病临床科研智能辅助决策平台建设”（项目编号：201901007）。

机器学习技术进行疾病研究，深入医疗决策各环节从而为临床医护人员提供协助将是现代医学探索的方向之一[4]。

1.2 研究目的

利用机器学习建立心肌梗死疾病特征自动识别模型，使用真实临床数据对多种算法模型进行准确性和性能验证，为心肌梗死疾病危险因素挖掘提供有效机器学习方法，通过特征挖掘找出关键和主要致病因素，为医生提供定性或定量辅助诊断意见。

2 数据和方法

2.1 概述

利用心内科患者数据建立机器学习模型，判别和比较各种常见机器学习模型性能并得出性能最佳模型用于特征对应，进而筛选出与心肌梗死相关的变量因素。

2.2 研究对象和数据源

搜集 2015 年 12 月 1 日 – 2020 年 12 月 31 日期间上海仁济医院心内科患者数据作为研究对象，从病案出院诊断中筛选出诊断为心肌梗死的患者，将患者分为心肌梗死组和对照组，利用患者历史诊疗数据进行回顾性研究。预测变量包括患者基本信息、生命体征、肝功能、肾功能、血常规、尿常规以及凝血功能等生化检查检验诊疗数据。初步清洗剔除缺失比例较高的指标后筛选得到 123 个可纳入模型的特征变量。

2.3 纳入和排除标准

本研究从病案出院诊断中筛选诊断为心肌梗死的患者作为实验组，随机抽取其余数据作为对照组。首先对数据进行均衡性处理，从对照组样本中采用下采样方式抽取样本，得到 2：3 比例的心肌梗死组和对照组数据集。将数据集划分为训练集和测试集分别训练模型和评估模型性能。训练集和测试集以 7：3 比例从数据集中随机划分。最终得到训练集数据量为 557，包括 223 位心肌梗死患者和 334 位对照组患者，得到测试集数据量为 240，包括 96 位心肌梗死患者和 144 位对照组患者。

2.4 数据处理

标准化处理是数据挖掘的基础工作，不同特征具有不同量纲和量纲单位，这一情况会影响数据分析结果。为了消除特征之间的量纲影响，需要进行数据标准化处理以解决数据特征之间的可比性问题。原始数据经过标准化处理后，各指标处于同一数量级从而适合进行综合评价。采用离差标准化方式将连续性特征变量归一化到 0 ~ 1 之间，对训练集和测试集同时进行以下处理：借助正则表达式抽取某一指标信息，转化为数值变量。对部分数值变量进行分层处理，转化为分类变量。例如将患者检查结果转化为正常检查结果与不正常检查结果。将字符类型变量转换为数值变量，方便后续处理。统计建模所需特征和样本缺失情况，剔除缺失比例达到 60% 及以上的特征，以及缺失比例超过 40% 的样本。类别性和连续型特征分别采用众数和均值填充方法进行填补。利用 Pearson 相关系数判断多重共线性。计算两两特征之间的相关系数，相关系数绝对值大于 0.75 的变量之间存在较强共线性，剔除其中 1 个特征。最后进行数据归一化处理。

3 模型训练

3.1 概述

采用逻辑回归（Logistic Regression，LR），决策树（Decision Tree，DT），随机森林（Random Forest，RF），自适应增强（Adaptive Boosting，AdaBoost），梯度提升决策树（Gradient Boosting Decision Tree，GBDT），极限梯度提升（Extreme Gradient Boosting，XGBoost）6 种机器学习算法训练模型[5]。

3.2 决策树

基本分类方法，由节点和有向边组成。决策树包含 1 个根节点、1 个内部节点和 1 个叶节点。其中内部节点代表要素，叶节点代表类。首先根据特

征信息增益对特征进行过滤；然后根据特征值将每个节点划分为子节点。根节点包含样本集。从根节点到每个叶节点的路径对应 1 个决策序列。

3.3 逻辑回归

分类方法，主要用于两分类问题（即输出只有两种，分别代表两个类别）回归模型中。y 是一个定性变量，例如 $y=0$ 或 1，Logistic 方法主要应用于研究某些事件发生概率。逻辑回归过程如下：面对回归或者分类问题建立代价函数，通过优化方法迭代求解出最优模型参数，然后测试验证该求解模型。

3.4 随机森林

基于非线性树的集成学习模型。获取随机森林模型后，当新样本进入时判断随机森林中每个决策树。对于分类问题使用投票方法，最大投票数是最终模型输出。

3.5 自适应增强

集成学习算法。在迭代过程中该算法在训练集上生成新的学习器，以预测所有样本并评估每个样本重要性。区分样本越困难迭代过程中给定的权重越高。当错误率足够小或达到一定数量迭代时将终止整个迭代过程。

3.6 梯度提升决策树

基于决策树的综合学习模型，采用可加模型方法。在迭代训练过程中模型基于最后一次迭代的残差生成弱分类器，通过不断减少训练过程中产生的残差达到数据分类目的。

3.7 极限梯度提升

在 GBDT 基础上对 boosting 算法进行改进而得到的学习模型。内部决策树使用回归树。回归树的分裂结点对于平方损失函数，拟合的就是残差；对于一般损失函数（梯度下降），拟合的就是残差近似值，分裂结点划分时枚举所有特征的值，选取划分点。最后预测结果是每棵树的预测结果相加。

3.8 模型训练方法

对同一组训练和测试数据分别使用 6 种方法进行建模和评估，计算模型在训练集和测试集上的精确率（Precision）、召回率（Recall）、综合评价指标（$F1$ score）、观测者操作特性曲线（Receiver Operating Characteristic，ROC）及曲线下面积（Area Under the Curve，AUC）等指标。计算公式如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$F1$ score = （2 * Precison * Recall）/（Precision + Recall）其中 TP 为真阳性率（True Positive），FP 为假阳性率（False Positive），FN 为假阴性率（False Negative）。

3.9 特征重要性

不同特征对模型预测的影响不同，影响大小被称为特征重要性。不同模型计算方法略有不同，均遵循以下原则：谁对模型预测结果准确度贡献越大谁的重要性越高。特征重要性主要通过模型返回的特征系数，或者在建模过程中特征被使用次数、带来的信息增益计算得到。

4 结果

4.1 基线特征

用于反映患者数据在训练集和测试集上各特征集间的组间水平，比较组间统计学显著性差异。将训练集和测试集划分为心肌梗死组和对照组，分别统计心肌梗死组和对照组的均值和标准差，对两组特征变量进行统计学显著性检验得到 P 值，见表 1。

表 1 基线特征

特征名称	训练集心梗组	训练集对照组	P 值训练集	测试集心梗组	测试集对照组	P 值测试集
年龄	66.48 ± 11.33	65.37 ± 11.69	0.28	66.0 ± 9.95	61.89 ± 12.06	0.08
尿素氮	6.27 ± 2.39	6.47 ± 3.19	0.44	6.55 ± 3.28	5.39 ± 1.35	0.07

续表 1

总胆固醇	3. 79 ± 1. 05	4. 12 ± 1. 03	0. 00	3. 90 ± 1. 20	3. 93 ± 0. 82	0. 19
肌酸激酶	172. 24 ± 468. 00	93. 64 ± 49. 15	0. 16	178. 49 ± 424. 41	104. 1 ± 75. 45	0. 21
肌酐	86. 30 ± 28. 32	97. 23 ± 118. 78	0. 00	87. 61 ± 38. 73	72. 16 ± 18. 72	0. 00
高密度脂蛋白胆固醇	1. 01 ± 0. 21	1. 12 ± 0. 31	0. 00	1. 04 ± 0. 29	1. 12 ± 0. 30	0. 05
低密度脂蛋白胆固醇	2. 15 ± 0. 87	2. 34 ± 0. 82	0. 01	2. 22 ± 1. 01	2. 17 ± 0. 68	0. 39
甘油三脂	1. 57 ± 1. 08	1. 49 ± 0. 86	0. 20	1. 55 ± 0. 76	1. 59 ± 0. 95	0. 36
尿酸	385. 28 ± 100. 62	339. 43 ± 104. 16	0. 00	377. 61 ± 101. 67	343. 83 ± 94. 15	0. 05
餐后 2 小时血糖	10. 24 ± 4. 13	9. 22 ± 3. 98	0. 02	10. 01 ± 3. 91	8. 53 ± 3. 78	0. 01
空腹血糖	6. 26 ± 2. 40	5. 79 ± 1. 59	0. 12	6. 14 ± 2. 27	5. 64 ± 1. 81	0. 03
促甲状腺激素（TSH）	2. 67 ± 7. 22	2. 36 ± 1. 45	0. 01	2. 14 ± 1. 31	2. 86 ± 2. 27	0. 05
血小板计数	209. 84 ± 73. 35	207. 60 ± 59. 98	0. 41	211. 91 ± 71. 52	214. 5 ± 50. 99	0. 18
凝血酶原时间	11. 53 ± 2. 20	11. 35 ± 1. 76	0. 49	11. 25 ± 1. 73	11. 72 ± 1. 88	0. 07
凝血酶时间	17. 42 ± 1. 89	17. 87 ± 1. 85	0. 24	17. 64 ± 2. 28	18. 05 ± 1. 69	0. 19
纤维蛋白原	3. 26 ± 1. 07	2. 93 ± 0. 67	0. 02	3. 25 ± 1. 10	3. 02 ± 0. 76	0. 40
纤维蛋白（原）降解物	3. 31 ± 5. 17	2. 08 ± 2. 47	0. 01	3. 34 ± 6. 43	1. 80 ± 1. 43	0. 31
D – D 二聚体	0. 49 ± 0. 88	0. 29 ± 0. 43	0. 00	0. 53 ± 1. 22	0. 25 ± 0. 34	0. 03
部分凝血活酶时间	30. 04 ± 8. 85	28. 81 ± 3. 28	0. 12	28. 21 ± 2. 94	28. 89 ± 2. 72	0. 17
国际标准化比率	1. 01 ± 0. 19	1. 00 ± 0. 16	0. 31	0. 99 ± 0. 15	1. 03 ± 0. 17	0. 15
肌红蛋白	131. 84 ± 322. 47	71. 29 ± 61. 81	0. 00	228. 44 ± 632. 39	52. 96 ± 25. 7	0. 00
肌酸激酶同工酶	11. 44 ± 39. 45	1. 73 ± 3. 84	0. 00	19. 16 ± 64. 33	1. 48 ± 1. 77	0. 00
B 型钠尿肽	283. 86 ± 531. 18	101. 37 ± 158. 31	0. 00	218. 82 ± 303. 51	83. 30 ± 119. 25	0. 00
血清游离脂肪酸	0. 53 ± 0. 29	0. 47 ± 0. 25	0. 05	0. 48 ± 0. 22	0. 47 ± 0. 19	0. 49
非高密度脂蛋白胆固醇	2. 78 ± 1. 05	2. 99 ± 1. 00	0. 02	2. 85 ± 1. 16	2. 81 ± 0. 80	0. 31
肌钙蛋白 I	2. 64 ± 7. 61	0. 07 ± 0. 26	0. 00	2. 68 ± 7. 10	0. 09 ± 0. 32	0. 00
eGFR – MDRD	80. 59 ± 23. 71	86. 18 ± 30. 14	0. 05	82. 67 ± 24. 47	93. 69 ± 16. 73	0. 01
eGFR – EPI Cr	79. 73 ± 21. 09	83. 63 ± 23. 94	0. 04	81. 12 ± 20. 87	92. 32 ± 11. 46	0. 01
游离 T3	4. 31 ± 0. 80	4. 39 ± 0. 59	0. 36	4. 20 ± 0. 66	4. 477 ± 0. 52	0. 03
游离 T4	16. 26 ± 3. 35	16. 24 ± 2. 36	0. 30	16. 34 ± 2. 11	16. 032 ± 2. 47	0. 14
NT – proBNP	1 246. 01 ± 1 494. 66	924. 58 ± 1 336. 60	0. 01	1 326. 71 ± 1 630. 85	310. 57 ± 318. 28	0. 00

4. 2 模型比较

分别计算 6 种方法在验证集上的 Precision、Recall、*F*1 和 AUC 值，见表 2。其中 XGBoost 模型对应的 ROC 曲线，见图 1。XGBoost 模型对应的箱线图，见图 2。

表 2 6 种模型指标

模型	Precision	Recall	*F*1 – score	AUC
Logistic	0. 66	0. 67	0. 66	0. 785 3
Decision Tree	0. 73	0. 72	0. 72	0. 761 9
Random Forest	0. 82	0. 8	0. 79	0. 852 2
AdaBoost	0. 82	0. 8	0. 79	0. 892
GBDT	0. 81	0. 81	0. 8	0. 913 5
XGBoost	0. 82	0. 82	0. 82	0. 914 8

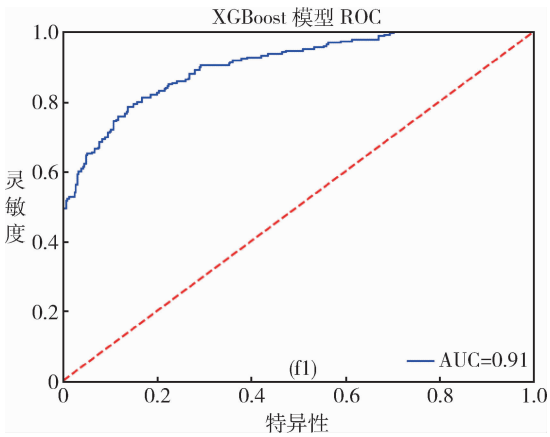

图 1 XGBoost 模型的 ROC 曲线

4. 3 模型分析

4.3.1 XGBoost 性能最佳 利用训练好的 6 个模

型分别对验证集进行特征重要性分析并对结果按降序排列，如 XGBoost 模型的特征重要性排序，见图3。在利用机器学习模型进行心肌梗死特征分析实验中，XGBoost 的 Precision、Recall、*F*1、AUC 值分别达到 0.82、0.82、0.82、0.91，均为各方法中的最高值，可见 XGBoost 在本研究中性能最佳。

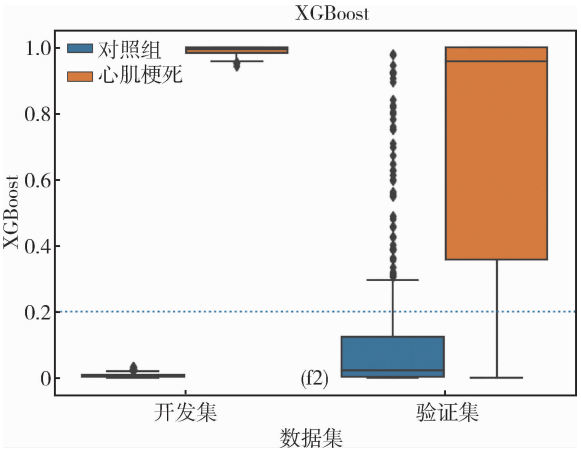

图2　XGBoost 模型的箱线图

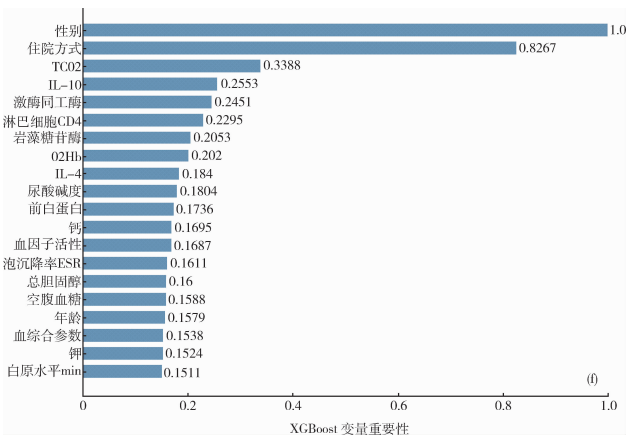

图3　XGBoost 模型特征重要性排序

4.3.2　心肌梗死发病影响因素分析　通过对比观察 6 种方法的特征重要性可以与心肌梗死关键和致病因素关联起来。载脂蛋白 A1、激酶同工酶、空腹血糖、纤维蛋白原、胆固醇等因素均排在前列且占据重要位置。在性能最佳的 XGBoost 方法预测结果中，空腹血糖、总胆固醇、激酶同工酶、尿酸碱度等因素均排在前列，与临床判断血糖、血脂、心肌酶谱和心功能指标等相关指标与心肌梗死具有重要相关性相符合[1-2]。同时研究结果表明性别、年龄等因素对心肌梗死发病具有重要影响。

4.3.3　研究优势与局限　优势在于使用大量真实心肌梗死患者临床数据，其中包含临床评估和生化变量数据。模型具有从机器学习算法和常规回归中得出的易于使用的临床数据。同时研究存在一定局限性。首先，上述数据均来自同一机构，缺乏外部验证，在研究中可能会出现潜在偏见，需要进行进一步的前瞻性研究和其他人群研究。其次，研究未包括社会经济状况、职业、饮食和体育活动方面数据，此类数据同样可能是心肌梗死的危险因素。

5　结语

通过机器学习模型建立逻辑回归、决策树、随机森林、AdaBoost、GBDT、XGBoost 6 种心肌梗死疾病特征自动识别模型，挖掘心肌梗死关键和主要致病因素，为疾病发病可能性提供数据支持。其中 XGBoost 方法的 AUC 值最高，达到 0.914 8。在该方法预测中，空腹血糖、总胆固醇、激酶同工酶、尿酸碱度等因素占据前列，与临床经验相符合。可见利用机器学习算法建立特征自动识别模型对心肌梗死致病因素重要性进行研究具有可行性。

参考文献

1　高晓津，杨进刚，杨跃进，等．中国急性心肌梗死患者心血管危险因素分析［J］．中国循环杂志，2015，30（3）：206－210.

2　中华医学会心血管病学分会，中华心血管病杂志编辑委员会，中国循环杂志编辑委员会．急性心肌梗死诊断和治疗指南［J］．中华心血管病杂志，2001，29（12）：710－725.

3　何清，李宁，罗文娟，等．大数据下的机器学习算法综述［J］．模式识别与人工智能，2014，27（4）：327－336.

4　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）：101－105.

5　周志华．机器学习：Machine Learning［M］．北京：清华大学出版社，2016.