# 基于改进的 LightGBM 算法的心脏病预测方法

崔春燕[1] 李宏滨[1]
CUI Chunyan LI Hongbin

**摘 要** 为了优化心脏病预测模型，选取比较流行的 UCI 心脏病数据集为研究对象，提出基于随机森林 - 递归特征消除法（RF-RFE）和 LightGBM 的混合算法——RF-RFE-LightGBM 作为心脏病预测方法。其中，利用 RF-RFE 算法提取较重要的特征，去除对预测结果影响较小的特征，针对优化后的特征建立 LightGBM 模型进行预测，采用主流的模型评价指标进行评估。实验结果表明，RF-RFE-LightGBM 算法的准确率、精度、召回率、$F_1$ 值、AUC 值分别为 0.917 1、0.905 6、0.932 0、0.918 6 和 0.920 3，相比于其他算法建立的模型综合性能更优，具有一定的优势。

**关键词** 随机森林；递归特征消除法；UCI 数据集；LightGBM；心脏病预测

doi：10.3969/j.issn.1672-9528.2024.09.009

## 0 引言

心脏病是全球范围内人类健康的头号杀手之一，对人类健康造成了严重威胁。据统计，全球约有三分之一的人口死亡是由心脏病引起的。在中国，心脏病也是导致大量人口死亡的主要原因之一，每年有数十万人因心脏病而丧生。

近年来，随着人工智能技术的发展，机器学习在医疗领域的应用越来越广泛，其中心脏病预测是一个重要的应用领域[1]。机器学习模型通过分析患者的医疗数据，如年龄、性别、血压、胆固醇水平、血糖水平、运动习惯等信息，预测患者是否患有心脏病、心脏病的类型和严重程度，为医生提供重要的辅助决策信息，从而帮助医生更早地发现患者的风险因素，提前采取预防措施，从而减少心脏病的发病率和死亡率。

目前心脏病预测的模型包括逻辑回归、随机森林、支持向量机、XGBoost 等。陈蒙蒙等人[2]使用逻辑回归模型对心脏病进行预测，结果表明 AUC 值为 0.943，说明此模型预测心脏病的效果较好。朱相奇[3]运用多种模型进行预测，通过对比得出使用 K 近邻模型的预测准确率高达 91%。赵金超等人[4]使用随机森林创建的预测模型准确度达到了 83.2%，AUC 值达到 0.965，结果表明该模型分类效果不错。刘云龙等人[5]利用 GBM 筛除无关特征后的预测模型在支持向量机上的表现较好，准确率达到了 84.62%，在减少特征数量的同时提升了预测准确率。刘宇等人[6]提出了通过 K-means 对数据集进行聚类分块，然后用 XGBoost 算法进行预测，准确率达到了 83.0%，说明了其方法的可行性。秦超超[7]选用 Catboost 算法建立预测模型，并且通过特征筛选和参数优化，使模型的准确率升至 90.76%，AUC 的值升至 0.905 9。王成武等人[8]所提出的改进的支持向量机在对心脏病患者和非心脏病患者的分类预测结果上得到了明显的提升，分类准确率提升到 84.04%。

综上所述，针对各个模型普遍预测准确率较低的情况，提出了一种基于 RF-RFE-LightGBM 的心脏病预测方法。

## 1 数据描述

研究采用的数据来自 UCI 机器学习库提供的心脏病数据集。该数据集由 5 个子数据集组合而成，即克利夫兰、匈牙利、瑞士、弗吉尼亚州长滩和 Stalog（Heart），共有 1025 个样本，其中心脏病患者人数为 526 人，非心脏病患者为 499 人，属于均衡样本。样本特征包括 13 个临床特征和 1 个类别特征，具体的特征描述如表 1 所示。

表 1 数据集特征列表

| 特征属性 | 含义 | 取值范围 |
|---|---|---|
| age | 年龄 | [29,77] |
| trestbps | 静息血压（Hg） | [94,200] |
| chols | 人体胆固醇（mg/dl） | [126,564] |
| thalach | 最大心率 | [71,202] |
| oldpeak | 运动相对于休息引起的 ST 段压低 | [0,6] |
| ca | 主要血管数量 | [0,3] |
| sex | 性别 | 1= 男，0= 女 |
| cp | 胸痛经历 | 1= 典型心绞痛，<br>2= 非典型心绞痛，<br>3= 非心绞痛，4= 无症状 |
| fbs | 空腹血糖（>120 mg/dl） | 1= 真，0= 假 |

1. 太原师范学院计算机科学与技术学院 山西晋中 030600

表 1（续）

| 特征属性 | 含义 | 取值范围 |
|---|---|---|
| restecg | 静息心电图测量 | 0= 正常，1= 有 ST-T 波异常，2= 左心室肥厚 |
| exang | 运动诱发心绞痛 | 1= 是，0= 否 |
| slope | 峰值运动 ST 段的斜率 | 1= 上升，2= 平坦，3= 下降 |
| thal | 地中海贫血的血液疾病 | 1= 正常，2= 固定缺陷，3= 可逆缺陷 |
| target | 是否患有心脏病 | 1= 是，0= 否 |

## 2 实验方法

基于 RF-RFE-LightGBM 建立的心脏病预测模型流程图如图 1 所示，首先对数据进行预处理，包括缺失值处理、异常值检测和数据归一化；其次使用 RF-RFE 算法对数据进行特征提取，去除不重要的特征；然后分别建立 LightGBM、RF、GBM 和 XGBoost 模型进行预测，在训练集上进行五折交叉验证，在测试集中对训练好的模型进行性能评估，其中训练集和测试集的比例为 8:2；最后采用准确率、精度、召回率、$F_1$ 值和 AUC 值 5 项常用指标来评估各模型的性能。实验流程图如图 1 所示。

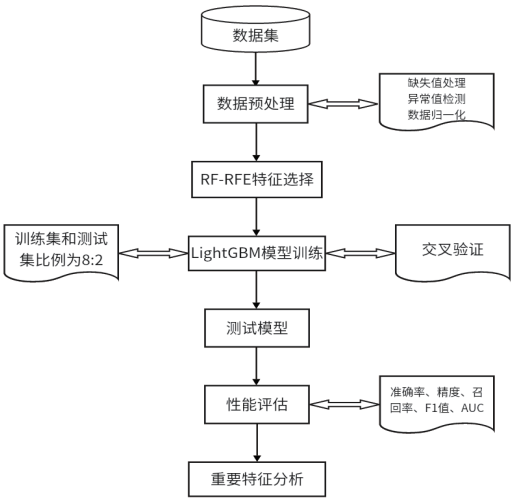


图 1 实验流程图

### 2.1 基于 RF-RFE 的特征选择

RF-RFE（random forest-recursive feature elimination） 是一种贪心搜索算法，通过构建随机森林模型，并反复训练模型来评估每个特征的重要性，按重要性对特征进行排序，从而递归地消除不重要的特征，这个过程不断重复，直到特定数量的特征被保留下来。具体步骤如下。

（1）随机森林模型训练：数据集 $D$ 有一个原始特征集 $X$，包含 $n$ 个特征；使用 RF 对数据集 $D$ 进行训练，其中 $F_1$ 值为：

$$F_1 = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

包含 $m$ 个样本，每个样本 $x_i$ 是一个 $n$ 维特征向量。

（2）计算特征重要性：计算每个特征 $f_j$ 在 RF 中的基尼指数（Gini index），并计算该特征在随机森林中的贡献。特征重要性可以通过如下公式计算：

$$I(f_j) = \sum_{i=1}^{m} Gini(D_i) - \sum_{i=1}^{m} Gini(D_i \mid f_j) \tag{2}$$

式中：$Gini(D_i)$ 是数据集 $D$ 的基尼指数，$Gini(D_i \mid f_j)$ 是数据集 $D_i$ 在特征 $f_j$ 上的基尼指数。

（3）递归特征消除：按照特征重要性从低到高排序，从原始特征集中移除重要性最低的特征 $f_{\min}$，得到一个新的特征集 $X'$，使用新的特征集 $X'$ 对数据集 $D$ 进行训练。重复步骤 2 和 3，直到达到预定的特征数量。

（4）特征选择：选择剩下的特征作为最重要的特征。

RF-RFE 算法通过结合随机森林的特征选择能力和递归特征消除的自动特征选择能力，能够有效地从原始特征集中选择最重要的特征，从而提高模型的性能和泛化能力 [9]。

### 2.2 LightGBM 建模

LightGBM（light gradient boosting machine）使用直方图（histogram）算法加速分割点的寻找过程，降低内存消耗，并使用带深度限制的 Leaf-wise 的叶子生长策略提高基学习器精度，更加高效地生成决策树 [10]。在这个基础上分别从减少样本和减少特征的角度进行优化：基于梯度的单侧采样（gradient-based one-side sampling，GOSS）和互斥特征捆绑（exclusive feature bundling，EFB）。LightGBM 的结构图如图 2 所示。

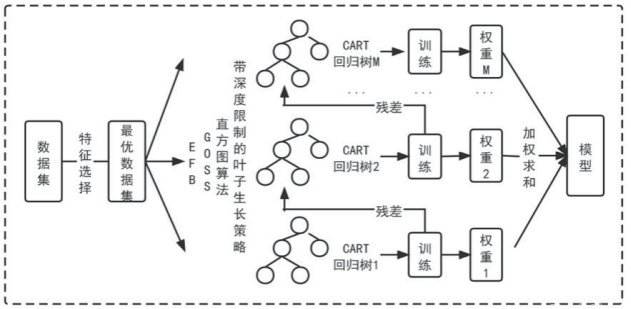


图 2 LightGBM 结构图

GOSS 的原理是保留梯度较大的样本，并对梯度较小的样本进行随机采样或直接丢弃 [11]。这样可以减少梯度较小的样本在模型训练中的作用，从而提高模型的训练效率和预测性能。其采样过程如下。

（1）计算样本梯度：初始化样本集 $D$ 和样本梯度矩阵 $\boldsymbol{G}$。$\boldsymbol{G}$ 是一个 $m \times n$ 的矩阵，$m$ 是样本的数量，$n$ 是特征的数量。计算每个样本 $x_i$ 在每个特征 $f_j$ 上的梯度值 $gradient()$。其公式计算为：

$$gradient(f_j, x_i) = \frac{\partial L(\hat{y}, y)}{\partial \hat{y}} \tag{3}$$

式中：$L(\hat{y},y)$ 是损失函数，其中 $\hat{y}$ 是特征 $f_j$ 在样本 $x_i$ 上的预测值。

（2）计算样本梯度总和：对于每个样本 $x_i$，计算其在所有特征上的梯度大小的总和 $total\_gradient(x_i)$，其计算公式为：

$$total\_gradient(x_i)=\sum_{i=1}^{m}|gradient(f_j,x_i)| \tag{4}$$

（3）样本采样：保留梯度较大的样本，并对梯度较小的样本进行随机采样或直接丢弃。保留的样本集合 $D_{selected}$ 公式计算为：

$$D_{selected}=\{x_i \mid total\_gradient(x_i)\geq \text{threshold}\} \tag{5}$$

式中：threshold 是预设的梯度大小阈值。

EFB 的核心思想是将互斥的特征捆绑在一起，视为一个整体进行处理，而不是独立地考虑每个特征。互斥特征是指在同一数据点上，特征的值相反或完全不同的特征。在某些情况下，这些互斥特征可能对模型的预测没有额外的好处。其原理如下。

（1）计算特征互斥性：初始化特征集 $F$ 和特征捆绑矩阵 $\boldsymbol{B}$。$\boldsymbol{B}$ 是一个 $n\times k$ 的矩阵，其中 $n$ 是特征的数量，$k$ 是捆绑的特征组数。对于每个特征 $f_i$，计算与其他特征的互斥性 $mutex(f_i)$。其计算公式为：

$$mutex(f_i)=\frac{1}{|D|}\sum_{x_j\in D}|gradient(f_i,x_j)\times gradient(f_j,x_j)| \tag{6}$$

式中：$gradient(f_i,x_j)$ 是特征 $f_i$ 在样本 $x_j$ 上的梯度值，$|D|$ 是数据集 $D$ 的样本数量。

（2）特征捆绑：根据互斥性 $mutex(f_i)$ 对特征进行排序，将互斥性最高的特征捆绑在一起形成一个新的特征组，更新特征捆绑矩阵 $\boldsymbol{B}$ 和特征集 $F$。重复步骤 1 和 2，直到达到预定的特征捆绑组数 $k$。

（3）特征捆绑优化：对于每个特征组，计算捆绑特征的加权平均梯度值 $bundled_{gradient(f_i)}$。其计算公式为：

$$bundled_{gradient(f_i)}=\frac{1}{|group(f_i)|}\sum_{f_i\in group(f_i)}gradient(f_i,x_j) \tag{7}$$

式中：$group(f_i)$ 是特征 $f_i$ 所在的特征组，$|group(f_i)|$ 是特征组 $group(f_i)$ 的特征数量。

（4）模型训练：使用捆绑后的特征组 $F_{bundled}$ 和原始目标变量 $y$ 进行模型训练。捆绑后的特征组 $F_{bundled}$ 可以通过以下公式计算：

$$F_{bundled}=\{bundled\_gradient(f_i)\mid f_i\in F\} \tag{8}$$

通过 GOSS 和 EFB 这两种优化方法，可以减少模型中的特征数量，从而简化模型，提高训练效率，并可能减少过拟合的风险。

### 2.3 评价指标

通常使用混淆矩阵展示模型对样本进行分类的结果，帮助理解模型在不同类别上的表现，如表 2 所示。

表 2 分类结果混淆矩阵

| 患病状态 | 预测患病 | 预测未患病 |
|---|---|---|
| 实际患病 | TP | FN |
| 实际未患病 | FP | TN |

其中，TP 表示真正例（true positive），指的是模型将正例预测正确的次数；FN 表示假反例（false negative），指的是模型将正例预测错误的次数；FP 表示假正例（false positive），指的是模型将反例预测错误的次数；TN 表示真反例（true negative），指的是模型将反例预测正确的次数。

评价分类器性能的优劣常用的指标包括分类准确率（Accuracy）、精度（Precision）、召回率（Recall）、$F_1$ 值（$F_1$ Score）和 AUC 值（Area Under the ROC Curve，即 ROC 曲线下的面积）。各指标公式如下：

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$\text{Precision}=\frac{TP}{TP+FP} \tag{10}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{11}$$

$$F_1=\frac{2\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{12}$$

## 3 实验结果

在进行相关性分析时，采用相关系数衡量两个连续变量之间的线性相关性，取值范围为 −1 ～ 1。相关性值被映射为热力图，一般使用渐变色带表示。通常，较高的正相关性值会用较深的颜色表示，较高的负相关性值会用较浅的颜色表示，无相关性则使用中间的颜色表示 [12]。特征相关性热力图如图 3 所示。

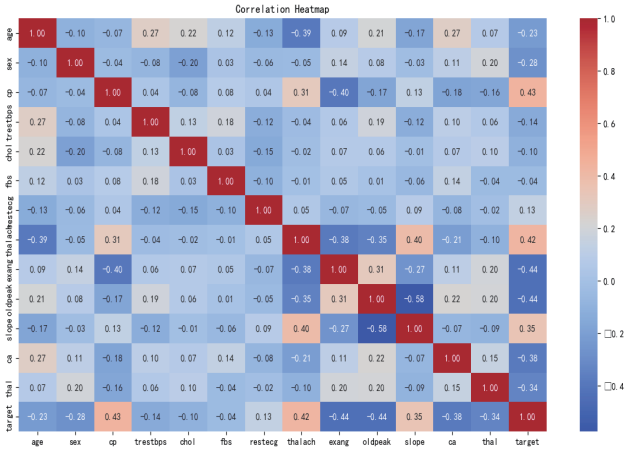


图 3 特征相关性热力图

首先获取特征重要性排名，排名越靠前的特征，对于模型的预测影响越大。如图 4 所示。

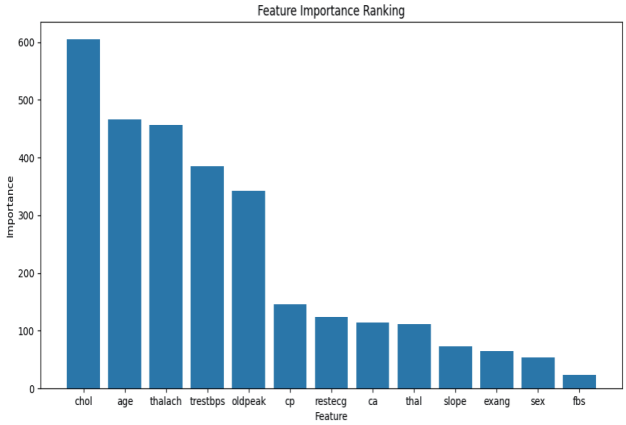


图 4 特征重要性排名

然后通过 RF-RFE 算法去除与预测结果相关性不大的特征变量后，得到的最优特征子集包括 9 个特征，分别是：chol、age、thalach、trestbps、oldpeak、cp、restecg、ca 和 thal。

最后针对特征子集分别建立几种模型进行预测，包括 RF-RFE-RF、RF-RFE-GBM、RF-RFE-XGBoost 和 RF-RFE-LightGBM，各模型性能对比如表 3 所示。

表 3 特征选择后各模型性能对比

| 模型 | Accuracy | Precision | Recall | $F_1$ | AUC |
|---|---|---|---|---|---|
| RF-RFE-RF | 0.873 2 | 0.829 0 | 0.941 7 | 0.881 8 | 0.939 6 |
| RF-RFE-GBM | 0.836 1 | 0.815 8 | 0.911 7 | 0.861 1 | 0.890 5 |
| RF-RFE-XGBoost | 0.819 7 | 0.810 8 | 0.882 3 | 0.845 1 | 0.927 0 |
| RF-RFE-LightGBM | 0.917 1 | 0.905 6 | 0.932 0 | 0.918 6 | 0.920 3 |

可以看出，RF-RFE-LightGBM 的准确率、精度和 $F_1$ 值分别为 0.917 1、0.905 6 和 0.918 6，在所有模型中是最高的，尤其准确率较其他算法有较大提升。而召回率为 0.932 0，略低于 RF-RFE-RF 的 0.939 6；AUC 值为 0.920 3，在四种模型中还算比较优秀。总体上看，RF-RFE-LightGBM 较其他模型来说综合性能更优。

## 4 结语

为了优化心脏病预测模型，提出了一种心脏病预测方法——RF-RFE-LightGBM。经过实验发现，基于 RF-RFE-LightGBM 算法建立的心脏病预测模型综合性能更优，表明其在心脏病预测方面的有效性和可行性。

由于实验所使用的数据量不算太大，无法直接证明该算法在大数据集上的适用性，下一步将收集更多的数据样本，构建一个更大的数据集，对算法模型进行进一步调优，以提高模型在大数据集上的预测准确性和稳定性。

**参考文献：**

[1]SHATENDRA K D, SITESH S, ANURAG J.Heart disease prediction classification using machine learning[J]. International journal of inventive engineering and sciences,2023,10(11):290-293.

[2] 陈蒙蒙，方振红，涂文怡，等．基于 Logistic 回归模型的心脏病预测模型构建及效果分析 [J]. 医学管理论坛，2022,39(2):32-35.

[3] 朱相奇．基于机器学习的心脏病风险预测和风险因素分析 [J]. 信息与电脑 ( 理论版 ),2023,35(4):166-169.

[4] 赵金超，李仪，王冬，等．基于优化的随机森林心脏病预测算法 [J]. 青岛科技大学学报 ( 自然科学版 ),2021,42(2):112-118.

[5] 刘云龙，周怡君，罗晨．基于 GBM 的特征选择在心脏病预测中的研究 [J]. 现代电子技术 ,2023,46(19):101-106.

[6] 刘宇，乔木．基于聚类和 XGboost 算法的心脏病预测 [J]. 计算机系统应用 ,2019,28(1):228-232.

[7] 秦超超．基于 Catboost 模型的心脏病预测研究 [D]. 曲阜：曲阜师范大学 ,2022.

[8] 王成武，郭志恒，晏峻峰．改进的支持向量机在心脏病预测中的研究 [J]. 计算机技术与发展 ,2022,32(3):175-179.

[9] 白莲，刘平．基于 RF-RFE 算法的地铁车站洪涝灾害预测研究 [J]. 铁道标准设计 ,2024,68(3):192-197+207.

[10] 陈维刚，张会林．基于 RF-LightGBM 算法在风机叶片开裂故障预测中的应用 [J]. 电子测量技术 ,2020,43(1):162-168.

[11] 王选，刘祥伟．集成特征选择算法和 LightGBM 融合的分类模型 [J]. 福建电脑 ,2022,38(4):12-15.

[12] 辛瑞昊，董哲原，苗冯博，等．基于机器学习的心脏病预测模型研究 [J]. 吉林化工学院学报 ,2022,39(9):27-32.

**【作者简介】**

崔春燕（1997—），女，山西忻州人，硕士研究生，研究方向：计算机技术、移动互联应用及开发。

李宏滨（1968—），男，山西晋中人，硕士，副教授，研究方向：智能图像处理。

（收稿日期：2024-05-29）