DOI:10.3969/j.issn.1005-202X.2023.04.018

医学生物信息

基于不完整数据的IHB-LightGBM心脏病预测模型

赵小强[1,2,3],乔慧[1]
1.兰州理工大学电气工程与信息工程学院,甘肃 兰州 730050;2.兰州理工大学甘肃省工业过程先进控制重点实验室,甘肃 兰州 730050;3.兰州理工大学国家级电气与控制工程实验教学中心,甘肃 兰州 730050

【摘要】提出基于不完整数据的IHB-LightGBM(Improved Hyperband-Light Gradient Boosting Machine)心脏病预测模型。首先,在Hyperband算法超参数采样的基础上引入了权重值,并通过蓄水池法按特征权重对其进行排序,从而筛选出最优参数以提高算法的参数寻优能力;其次,针对心脏病数据样本小且属性缺失的问题,使用K近邻算法对不完整数据进行缺失值插补,再将处理得到的完整数据进行归一化,使数据映射至0~1范围内;最后,对LightGBM采用改进后的IHB优化算法进行全局参数寻优,建立IHB-LightGBM心脏病预测模型。使用UCI心脏病数据集进行实验,结果表明IHB算法的参数寻优效果优于贝叶斯、随机搜索等优化算法,IHB-LightGBM模型在各项评价指标也上明显高于随机森林、极端随机树等算法,可以获得更快的预测速度和更高的预测精度。

【关键词】数据挖掘;心脏病预测;超参数优化算法;LightGBM算法
【中图分类号】R318;R541 【文献标志码】A 【文章编号】1005-202X(2023)04-0512-09

Heart disease prediction using IHB-LightGBM model based on incomplete data

ZHAO Xiaoqiang[1,2,3], QIAO Hui[1]
1. School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China; 2. Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou University of Technology, Lanzhou 730050, China; 3. National Experimental Teaching Center of Electrical and Control Engineering, Lanzhou University of Technology, Lanzhou 730050, China

Abstract: An improved Hyperband-light gradient boosting machine (IHB-LightGBM) model for heart disease prediction based on incomplete data is presented in the study. Based on the sampling of Hyperband algorithm for hyperparameter estimation, the weight values are introduced, and the reservoir method is used to sort the parameters according to the feature weights, so as to screen out the optimal parameters to improve the parameter optimization ability of the algorithm. Subsequently, to overcome the problems of small sample size and missing attributes of heart disease data, K-nearest neighbor algorithm is used to perform the interpolation of missing values for incomplete data, and the obtained complete data are normalized and mapped to the range from 0 to 1. The IHB optimization algorithm is adopted for global parameter optimization, and the IHB-LightGBM model for heart disease prediction is established. Experiments are carried out using the UCI heart disease data sets, and the results show that IHB algorithm is superior to Bayesian, random search and other optimization algorithms in parameter optimization, and that the various evaluation indicators of IHB-LightGBM model are significantly higher than those of random forest, extreme random tree and other algorithms. The proposed model improves both efficiency and accuracy of prediction.

Keywords: data mining; heart disease prediction; hyperparameter optimization algorithm; LightGBM algorithm

前 言

在全球范围内,死于心血管疾病的人数约为癌症死亡人数的两倍[1],而在心血管疾病中,心脏病被认为是世界上最复杂与最致命的疾病之一[2]。随着科技的进步,国内外的研究者们开始将数据挖掘以及机器学习的相关知识进行结合,并且应用到了心脏病预测领域,而目前的心脏病预测算法在预测精

【收稿日期】2022-12-15
【基金项目】国家重点研发计划(2020YFB1713600);国家自然科学基金(61763029);甘肃省教育厅产业支撑计划项目(2021CYZC-02)
【作者简介】赵小强,博士,教授,博士生导师,主要研究方向为数据挖掘、图像处理、故障诊断,E-mail: xqzhao@lut.edu.cn

度上仍然难以满足科研要求。因此，心脏病预测模型及其优化问题也逐渐成为研究热点。

2017年Yekkala等[3]分析了各种集成学习方法以及特征子集选择方法如粒子群优化算法，用来预测特定患者的心脏病发生率，实验结果表明，套袋法和粒子群优化算法达到了更高的精度。Salem等[4]于2018年提出一种基于遗传算法的优化预测模型，该研究构建出心脏病的多种预测模型，并使用遗传算法选择重要的心脏病特征，最终结论得出遗传算法与预测模型的结合提高了心脏病预测的准确率。Dulhare等[5]于2018年提出了基于朴素贝叶斯和粒子群优化的心脏病预测系统，实验结果表明，以粒子群算法为特征选择的模型提高了朴素贝叶斯对心脏病分类的预测精度，准确率达到了87.91%。Gokulnath等[6]于2019年提出一种基于支持向量机（SVM）的优化函数，将其用于遗传算法中，提取出心脏病数据集中更重要的特征来提升心脏病预测的效果，表明所提算法的预测效果更佳。Khourdifi等[7]于2019年提出基于粒子群优化和蚁群优化机器学习算法的心脏病预测和分类研究，通过过滤冗余特征，利用粒子群优化算法和蚁群优化算法对人工神经网络进行优化，并且与其他分类算法进行对比，结果证明了所提出的混合方法在处理心脏病分类数据中的有效性和鲁棒性。Gultepe等[8]于2019年使用朴素贝叶斯与J48分类算法分别对加利福尼亚大学（University of California, Irvine，UCI）数据库中的心脏病数据进行测试，之后又采用集成学习中的Bagging算法对以上两种算法进行袋装优化，结果表明，J48分类算法的预测准确率由初始时的76.92%提升到了81.31%。Sarkar等[9]于2020年将预测模型优化分为两个阶段，第一阶段主要确定并行机上每个数据集的训练集和测试集的并行最佳比例（P_{opt}），之后通过P_{opt}的最佳训练集（T_{best}）再次并行搜索；第二阶段则采用并行遗传算法，通过决策规划顺序法对完美规则归纳生成的规则集进行细化，与基于序贯遗传算法混合模型相比，所提出的分级模型在预测精度上提高了6%。Cherian等[10]于2020年提出基于混合狮子算法和粒子群算法的心脏病预测权值优化神经网络，该方法的特异值均优于其它优化算法与神经网络的结合，能够更快的完成心脏病的预测。赵金超等[11]在2021年使用K近邻算法，对随机森林算法进行优化，提出KNN-RF模型，建立了心脏病预测模型，实验对比验证表明所提模型预测准确率较逻辑回归模型提高了0.3%、较梯度提升树模型提高了1%、较决策树模型提高了10.8%。Valarmathi等[12]于2021年使用3种不同的超参数优化算法对随机森林和极端梯度提升算法进行参数调整与测试，研究发现随机森林通过随机搜索方式进行参数调优的心脏病预测结果更好。

然而上述研究中均使用了数据样本小且较为完整的数据集，而在实际情况下，心脏病数据并非完好无损，而是存在不完整及缺失情况，并且优化算法在进行参数寻优时，依旧存在寻优能力较差以及耗时较长的问题，从而导致预测模型的准确率不令人满意。因此，本文提出基于不完整数据的IHB-LightGBM（Improved Hyperband-Light Gradient Boosting Machine）心脏病预测模型，使用加权随机采样对HB（Hyperband）算法进行改进，计算出采样所得超参数设置的特征值权重，通过排序并筛选，最终得到一组最优超参数。采用K近邻算法对不完整数据进行缺失值插补，避免因数据样本过少而出现预测过拟合情况。利用改进后的IHB算法对Light GBM算法进行超参数优化从而构建出IHB-Light GBM心脏病预测模型。由实验结果可知，IHB-Light GBM模型的预测效果明显优于未优化以及HB算法优化的Light GBM模型，在准确率、召回率、精确率、F_1值、Matthews相关系数、ROC曲线图上均优于随机森林、极端随机树等对比算法。

1 相关算法

1.1 HB超参数优化算法

HB算法是Li等[13]于2017年提出的，该算法在Jamieson等[14]提出的Successive Halving算法的基础之上做了扩展。Successive Halving算法假设有n组超参数组合，对这n组超参数需要给它们均匀地分配预算，然后进行验证评估，根据验证结果淘汰掉一半表现差的超参数组，通过不断地迭代上述过程，直至找到最终一个最优超参数组合。HB正是基于上述的“对半淘汰”算法思路，将超参数寻优的过程视为一个在给定预算（其中的预算主要包括时间、迭代次数、数据量、特征量等）的情况下，在无限臂老虎机上进行探索的过程。HB算法首先是设定好超参数组合的总数与总预算，并且预设尽可能多的超参数组合数量，为每组超参数分配尽可能多的预算，目的是确保尽可能获得最优超参数；其次，将每组超参数组合最多分配的预算与控制每轮进行“对半淘汰”时被保留的超参数组合比例作为输入，通过计算及采样得到超参数组合用于淘汰循环；最后，利用Successive Halving算法进行多次淘汰循环，直至仅剩一组超参数作为最优超参数输出。

1.2 LightGBM

LightGBM是微软在2017年开源的梯度提升框架[15]，与经典的梯度提升决策树（Gradient Boosting

Decision Tree, GBDT)相比，GBDT使用的是传统策略生长树并且不能以小批次的形式训练模型。而LightGBM采用的是Leaf-wise的决策树生长策略，在单个机器不影响速度的情况下，尽可能提高训练效率，同时在多机并行时，尽可能降低通信代价[16]。此外，Light GBM还使用了两种新算法，即梯度的单边采样(Gradient-based One-Side Sampling, GOSS)和互斥特征捆绑(Exclusive Feature Bunding, EFB)，正因为这两种算法的引入，使得LightGBM在许多领域的实际应用中，计算速度超越了大多数算法的同时将误差保持在较小的范围[17]。基于梯度的GOSS算法流程图与EFB算法具体流程如图1、图2所示。

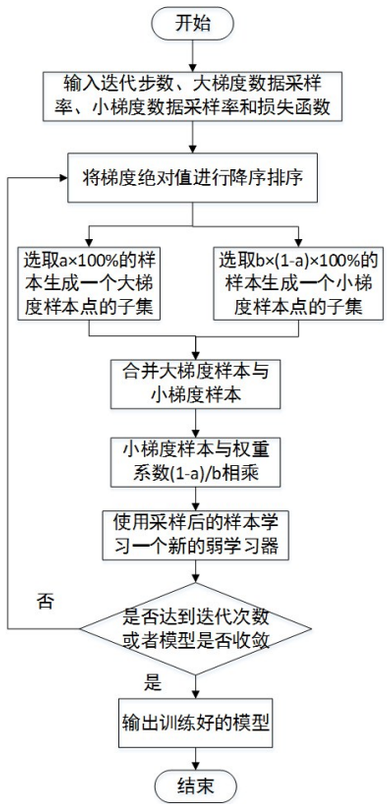

图1 基于梯度的单边采样算法流程图

Figure 1 Flowchart of gradient-based one-side sampling algorithm

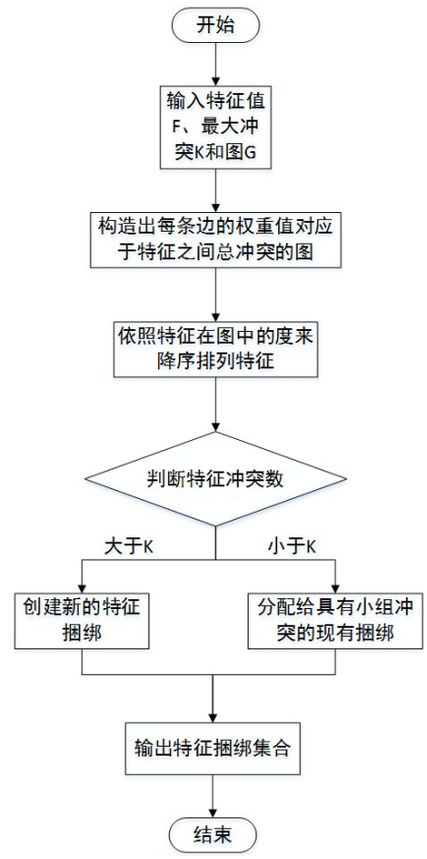

图2 互斥特征捆绑算法流程图

Figure 2 Flowchart of exclusive feature bunding algorithm

2 IHB-LightGBM心脏病预测模型

2.1 改进HB算法

HB优化算法虽然加快了在探索过程中对每种超参数组合的评价速度，但该算法对于超参数设置的采样形式简单，使得各超参数设置差异较大，在进行下一步连续减半过程时会增加循环次数，从而影响寻优效率。为此，笔者在此算法基础上引入了权重值，即计算出采样所得各超参数设置的权重特征值，通过蓄水池法按照权重进行排序[18]，从而选取出更优质的超参数设置，使得接下来的连续减半过程更加快速高效。改进HB算法步骤如下：

输入：单个超参数组合能够分配到的最大预算R以及用来控制每次迭代后淘汰超参数设置的比例η。

输出：最优超参数组合T。

Step 1：初始化。对控制总预算的大小$s_{\max}=\left[\log_{\eta}(R)\right]$与总预算$B=\left(s_{\max}+1\right)R$进行初始化，并将任意可取的$s_{\max}$放入集合$S$中，即$S\in\left\{s_{\max},s_{\max}-1,\cdots,0\right\}$。

Step 2：分配。通过式(1)和式(2)计算出预分配的超参数设置个数n以及每个超参数设置实际所能分配的预算r：

$$n=\frac{B}{R}\frac{\eta}{(s+1)} \tag{1}$$

$$r=R\eta^{-s} \tag{2}$$

Step 3：采样超参数。将均匀随机采样得到的超参数设置集合V的前$n(n\neq 0)$个元素放入结果集合T中，即$T\in\left\{V_1,V_2,\cdots,V_n\right\}$。

Step 4：引入特征权重值。对于集合T中的元素$V_i(i=0,1,\cdots,n)$，选取随机数$u_i=\mathrm{rand}(0,1)$，并通过式(3)计算出各元素的特征值：

$$f_i = u_i^{(1/w_i)} \tag{3}$$

其中，w_i为u_i的权重。

Step 5：选取评判阈值。将集合T中最小的特征值$f_{\min}$作为评判阈值F，对于结果集合T之外的元素$V_j(j = n, n+1, \cdots, m)$同样按式(3)计算出各元素特征值的$f_j$。

Step 6：蓄水池法排序。将计算出的f_j与评判阈值F进行比较，若$f_j > F$，则用此特征值所对应的元素V_j替换集合T中拥有最小特征值的元素。

Step 7：获取最终集合。通过排序，直至将元素全部替换完成，得到最终的集合T，否则返回Step 5继续循环。

Step 8：连续减半循环。根据超参数所对应的验证误差，对其连续减半，将仅剩的最后一组超参数设置作为最优超参数输出。

2.2 K近邻算法插补缺失值

心脏病数据往往存在不完整的情况，这些不完整数据最常以缺失数据与删失数据的形式出现，对于此类数据的常用处理方式是删除缺失部分或通过算法进行插补。而“删除缺失部分”的方式会使数据集样本缩减，从而影响后续实验。本文则采用K近邻算法对不完整数据进行缺失值插补，其原理是寻找出距离缺失值最近的数据记录，将分类型数据采用众数插补，数值型数据则采用距离加权平均插补，具体步骤如下所示。

Step 1：初始化数据集并构建数据矩阵，其中需要将连续型属性进行离散化处理。

Step 1.1：给定训练集T以及连续的属性s，选取区间并将其初始化，若属性在训练集中出现了多种取值，则需要先将它们进行从小到大的排序。

Step 1.2：通过将连续属性离散化，可基于d划分点将T分为T_d^-与T_d^+。

Step 1.3：将$\left[s^i \cdot s^{i+1}\right)$的中心点作为划分点即$H_s = \left\{\dfrac{s^i + s^{i+1}}{2} \middle| 1 \leqslant i \leqslant n-1\right\}$，对属性$s$进行划分处理。

Step 1.4：通过式(4)计算出二分后的信息熵，并且取最大值作为最佳划分点，该点则会划分概率属性值。

$$\mathrm{Gain}(T,s) = \max_{h \in Hs} \mathrm{Gain}(T,s,h) \tag{4}$$

Step 2：通过式(5)计算矩阵中数据的欧氏距离D，从算出的D中选出最小的K个数据。

$$D = \left(\left|x_1 - x_{i1}\right|^2 + \left|x_2 - x_{i2}\right|^2 + \cdots + \left|x_n - x_{in}\right|\right)^{\frac{1}{2}} \tag{5}$$

Step 3：根据式(6)分别计算出K个数据的权值w_k，从而对缺失数据进行插补。

$$w_k = \frac{1/D}{\sum_{i=1}^{k} 1/D} \tag{6}$$

2.3 预测模型的构建

随着模型复杂度的提升，传统的超参数优化算法在进行参数调优时，会极大地增加计算资源的消耗，从而导致寻优过程相对耗时。LightGBM模型拥有内存占用较少、计算复杂度较低、回归准确度较高等优点，但其众多的参数对预测结果有着重要的影响，因此对于其参数的优化就显得格外重要，使用改进后的IHB算法优化LightGBM模型能够达到更精准的预测效果，IHB-LightGBM模型具体流程如下所示。

Step 1：数据预处理。将所获取的心脏病数据初始化，通过K近邻算法将数据缺失值进行插补，得到实验所需的完整数据集。将完整数据按照70%与30%的比例分为训练集与测试集，并按式(7)对数据进行归一化处理，使处理后的完整数据映射至0~1范围内，其中x_t为数据实际值，$x_{\min}$为数据集中最小值，$x_{\max}$为数据集中最大值。

$$x_t' = \frac{x_t - x_{\min}}{x_{\max} - x_{\min}} \tag{7}$$

Step 2：使用IHB算法优化LightGBM模型。

Step 2.1：初始化LightGBM模型参数，结合式(1)、(2)、(3)定义超参数搜索空间并采样得到超参数设置集合T，调用LightGBM模型，当实际的分配预算为$rd = r\eta^d$时，将各超参数设置的验证误差放入集合$L = \{(t, rd), t \in T\}$中，其中，$d \in \{0, 1, \cdots, s\}$。

Step 2.2：根据所输入集合T与集合L，通过式(8)与式(9)选取出前k个超参数设置代替原本的集合T，判断集合T中的超参数设置是否淘汰到只剩最后一组，若否，则返回上一步继续循环减半。

$$nd = \lfloor n\eta^{-d} \rfloor \tag{8}$$

$$k = \frac{nd}{\eta} \tag{9}$$

Step 3：建立IHB-LightGBM预测模型。基于训练集，将循环得到的超参数设置作为最优超参数输出给LightGBM模型，得到最终的IHB-LightGBM预测模型。

Step 4：使用测试集测试IHB-LightGBM预测模型的性能，得到心脏病预测结果。

IHB-LightGBM预测模型的流程图如图3所示。

该模型需要调优的参数主要包含：学习率、叶子数量、叶片最小数据量、迭代次数、每次迭代时的数据比例、每次迭代时随机使用的参数比例等，6个参数的默认值以及参数含义如表1所示。

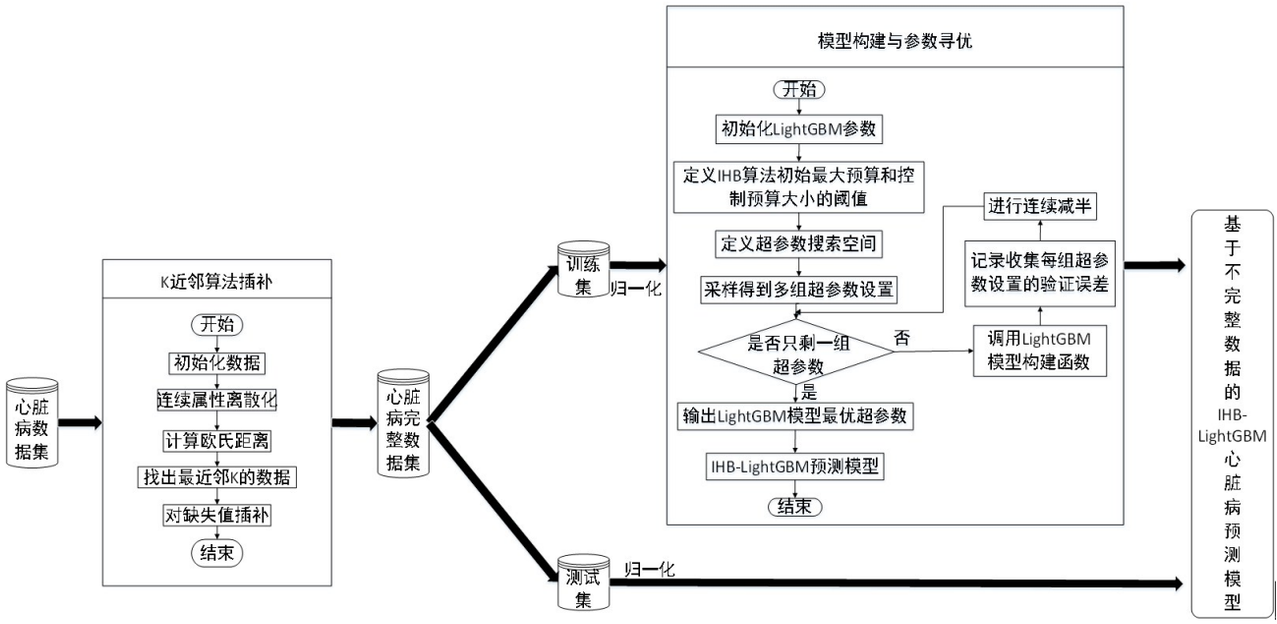

图3 IHB-LightGBM预测模型流程示意图

Figure 3 Flowchart of IHB-LightGBM prediction model

表1 LightGBM的默认参数

Table 1 Default parameters for LightGBM

参数名称	默认值	参数含义
学习率(learning_rate)	0.1	梯度下降的步长参数,目的是提高准确率
叶子数量(num_leaves)	31	用来控制树结构从而防止过拟合
叶片最小数据量(min_data_in_leaf)	20	用于指定叶子节点向下分裂的最小样本数,避免出现过拟合
迭代次数(n_estimators)	100	用来控制决策树的数量,通常与学习率构成最佳组合以提高准确率
每次迭代时的数据比例(bagging_fraction)	1.0	指训练每棵树时要采样的特征百分比,目的是避免过拟合
每次迭代时随机用的参数比例(feature_fraction)	1.0	用于训练每棵树的训练样本百分比,与bagging _ fraction作用相似

3 实验结果与分析

3.1 实验数据集

本文选取5个独立可用的数据集,它们均来源于UCI机器学习存储库[19],分别为:克利夫兰心脏病数据集(303例)、匈牙利心脏病研究所数据集(294例)、瑞士苏黎世大学数据集(123例)、瑞士巴塞尔大学数据集(270例)和弗吉尼亚长滩数据集(200例)。尽管每个数据集都广泛用于测试实验,但也存在缺点。例如,瑞士苏黎世大学数据集样本量较小;匈牙利数据集中的3个属性存在大量删失数据;瑞士巴塞尔大学数据集中存在较多缺失数据,这些问题均会降低研究水平与预测难度。因此,笔者将以上数据集进行合并,共计1 190例数据,其中,有272例数据出现重复,因此需要将它们删减,最终保留918例数据。

3.2 数据预处理

笔者利用上述K近邻算法结合式(4)、(5)、(6)对918例不完整数据进行缺失数据插补,得到一组包含12个共同属性的完整数据集,其中11条用于预测心脏病的特征属性,剩余1条用作标记样本,具体属性描述如表2所示。

3.2.1 特征转换 处理得到的完整数据中,有性别、胸痛类型、静息心电图、运动诱发心绞痛以及运动高峰心电图5个属性为离散特征,笔者需要将它们转换为数值型特征,便于进行后续分类预测,笔者利用时间成本低的One-hot编码将数据集中的文本型特征转化为数值型特征,经过处理后,数据集由原来的11个特征属性变为16个特征属性。

3.2.2 特征分析 通过对数据集中各特征之间进行分析,可以更深层次地了解每个特征之间的关系,此种关系可以称为“相关”,通过图4可以清晰地看出各个特征之间的相关性,横纵坐标形成的色块代表着特征的相关程度,颜色越浅表明相关系数越小、关联度

表2 数据集属性描述

Table 2 Data set attribute description

属性名称	描述	类型
年龄（Age）	只能是整数	数值型
性别（Sex）	M为男，F为女	分类型
胸痛类型（Chest Pain）	TA为典型心绞痛，ATA为非典型心绞痛，NAP为非心绞痛，ASY为无症状	分类型
静息血压（Resting BP）	单位为mm/Hg	数值型
血清胆固醇（Cholesterol）	单位为mg/dL	数值型
空腹血糖（Fasting BS）	0为小于120 mm/dL，1为大于120 mg/dL	数值型
静息心电图（Resting ECG）	N为正常，ST为ST-T波异常，LVH为Estes标准下可能或明确的左室肥厚大	分类型
最大心率（Max HR）	单位为beat/min	数值型
运动诱发心绞痛（Exercise Angina）	N为否，Y为是	分类型
运动诱导ST段（Oldpeak）	相较于静息时，运动诱导的ST段下移值	数值型
运动高峰心电图（ST_Slope）	UP为向上倾斜，Flat为平坦，Down为向下倾斜	分类型
患心脏病风险（Heart Disease）	0为否，1为是	数值型

越低，说明它们之间互不干涉，相互独立。当两种特征之间相互独立时，相关系数趋近为0，否则相关系数趋近为1，会出现特征冗余，影响预测结果。由图4可看出，本文数据集的特征之间相关度较低，不存在特征冗余，因此不必进行特征选择。

3.3 评价指标

为评估模型的性能，笔者挑选了5种常用的评价指标：准确率（Accuracy）、召回率（Recall）、精确率（Precision）、F_1值、Matthews相关系数（MCC）以及接受者操作特性（Receiver Operating Characteristic, ROC）曲线对模型进行评价。各项评价指标计算公式如下：

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{10}$$

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

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

其中，TP为真正类（True Positive），TF为假负类（False Negative），TN为真负类（Ture Negative），FN为假正类（False Positive），TP与TN表示分类结果正确，FP与FN表示分类结果错误。

准确率、召回率、精确率、F1值、MCC 5种指标的数值大小与预测效果成正比关系。ROC曲线图是一种反应敏感度（真阳性率）与特异度（假阳性率）连续变量的综合指标，用曲线图的方式表现出两度之间的关系，横坐标越接近0，准确度越高；纵坐标越接近1，精度越好，即各条曲线越靠近左上角表明预测效果越好。而ROC曲线下面积（AUC）就是ROC曲线与横坐标轴所覆盖的区域面积，它所代表的是分类器的平均性能值，AUC值越大则表明预测性能越好。

3.4 结果分析

3.4.1 改进后的IHB算法性能对比分析 为进一步验证IHB算法参数寻优的效果，笔者将其与贝叶斯优化（Bayesian Optimization, Bayes）算法[20]以及树状结构帕仁估计（Tree-structured Parzen Estimator, TPE）算法[21]以及随机搜索优化（Random Search Optimization, RS）算法[22]使用相同的数据集对LightGBM模型进行参数寻优。

从表3、表4、图5可以看出，IHB-LightGBM预测模型准确率高于Bayes-LightGBM模型11.21%、高于TPE-LightGBM模型10.89%、高于RS-LightGBM模型14.07%，在召回率、精确率、F_1值、MCC值与AUC值上也最优，并且用时最短。这表明，IHB算法在参数寻优过程中表现更佳，能够更快速地找到适合LightGBM模型的参数。

3.4.2 基于IHB算法优化模型的预测结果对比 笔者分别将未优化的LightGBM模型、HB算法优化的LightGBM模型以及改进后的IHB算法优化的LightGBM模型进行实验对比验证，由图6可以看出，IHB算法优化的LightGBM模型在准确率、AUC、召回率、F_1值上都超过了95%，并且精确率与MCC值也都在90%以上，这说

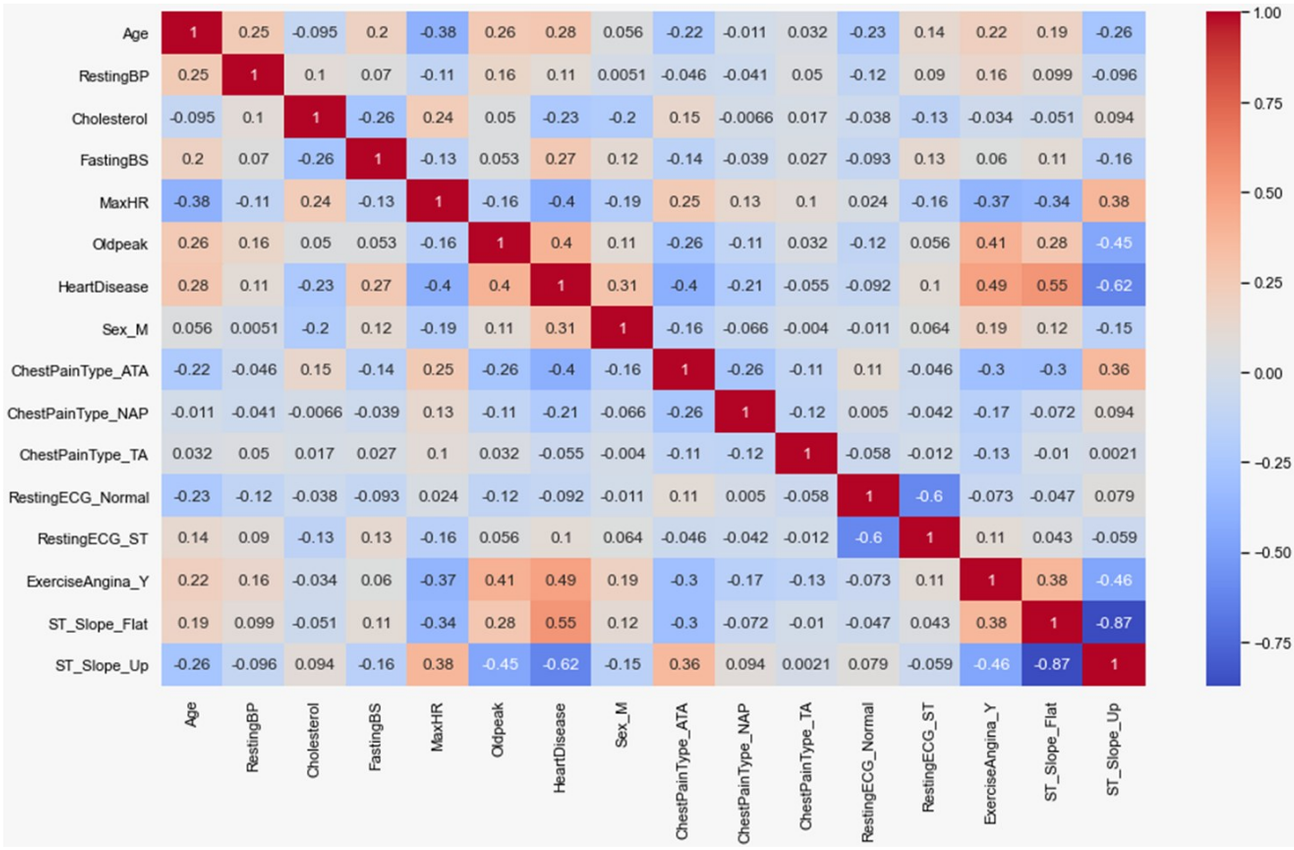

图4 特征相关性热力图

Figure 4 Characteristic correlation thermal map

表3 不同优化算法寻优的最优参数

Table 3 The optimal parameters of different optimization algorithms

参数名称	IHB	Bayes	TPE	RS
学习率	0.016 0	0.019 9	0.349 9	0.400 0
叶子数量	57	199	173	10
最小数据量	14	20	22	26
迭代次数	289	123	32	80
每次迭代时的数据比例	0.861 2	0.834 5	0.794 2	0.915 6
每次迭代时随机用的参数比例	0.763 2	0.454 7	0.575 3	0.437 8

明IHB算法对预测模型有着更好的优化能力。

3.4.3 与常用分类模型的预测结果对比 明确IHB优化算法的能力后，再验证IHB-LightGBM模型的预测效果。将9种常用算法使用插补后的完整数据集进行模型搭建，得到相应预测结果用于对比。由表5可看出，IHB-LightGBM模型的预测结果在各项指标上都远超于其他对比模型，值得注意的是，仅本文模型的F_1值达到97%以上，这表明本文模型对于心脏病预测的能力更加准确且可靠。

表4 不同优化算法的预测结果对比

Table 4 Comparison of prediction results among different optimization algorithms

模型	准确率/%	AUC/%	召回率/%	精确率/%	F_1值/%	MCC/%	时间/s
RS-LightGBM	82.81	90.39	88.24	81.08	84.51	65.58	624.16
TPE-LightGBM	85.99	91.21	88.66	86.98	87.70	71.71	504.66
Bayes-LightGBM	85.67	92.72	91.52	83.41	87.23	71.49	400.68
IHB-LightGBM	96.88	98.91	99.98	94.74	97.29	93.79	268.82

4 结 论

由于检测设备故障、临床医护人员忘记记录、部分数据只在特定情况下进行收集等原因，会导致心脏病数据集不完整。此外，针对Hyperband超参数优

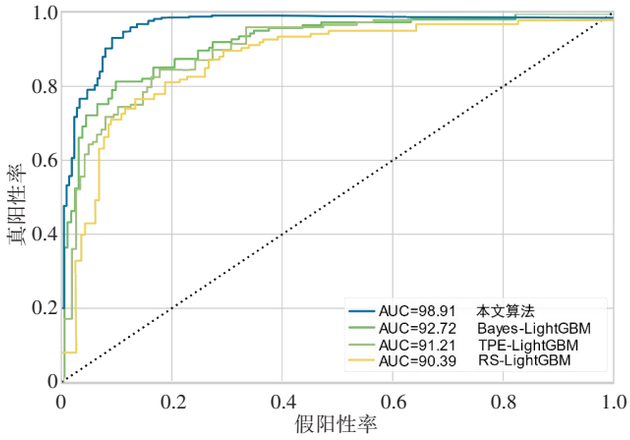

图5 不同优化算法的ROC曲线图

Figure 5 ROC curves of different optimization algorithms

化算法寻优能力不足以及效率较低导致心脏病预测模型效果欠佳的问题，本文提出了基于不完整数据的IHB-LightGBM心脏病预测模型。首先，针对HB算法采样得到的超参数寻优效率较低的问题，使用加权随机采样将其改进，对采样所得的超参数设置加入权重值，并通过蓄水池法按权重排序，实现更高效寻优；其次，采用K近邻算法对不完整数据进行缺失值插补，并采用one-hot编码等方式对得到的完整数据进行特征工程；最后，基于强鲁棒性的LightGBM，使用IHB算法对其进行参数调优，从而构建IHB-LightGBM心脏病预测模型。结果表明，改进后的IHB算法对于参数调优有更好的效果，构建出的

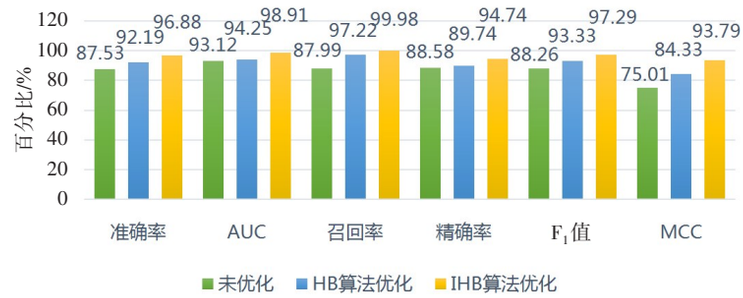

图6 未优化、HB算法优化与IHB算法优化模型的评估指标对比

Figure 6 Comparison of evaluation indexes among unoptimized, HB algorithm optimization and IHB algorithm optimization models

表5 不同算法预测性能对比(%)

Table 5 Comparison of prediction performance among different algorithms (%)

模型	准确率	AUC	召回率	精确率	F_1值	MCC
IHB-LightGBM	96.88	98.91	99.98	94.74	97.29	93.79
Random Forest Classifier	87.53	93.58	90.92	86.66	88.67	75.07
Extra Tress Classifier	86.91	92.71	88.87	86.94	87.84	73.78
Logistic Regression	86.59	92.93	87.97	86.97	87.39	73.22
Linear Discriminant Analysis	86.44	93.35	87.69	87.08	87.29	72.91
Naive Bayes	86.43	92.46	89.16	86.16	87.54	72.91
Gradient Boosting Classifier	86.43	92.66	88.27	86.59	87.39	72.76
AdaBoost Classifier	84.09	88.91	85.34	84.95	85.09	68.14
Decision Tree Classifier	77.71	77.72	77.76	80.32	78.96	55.40
K Neighbors Classifier	71.18	76.30	75.12	72.19	73.37	42.39

IHB-LightGBM预测模型不仅在各项评价指标上最好，并且预测速度更快，能够更高效地预测心脏病患病风险。

【参考文献】

[1] Harding S, Silva MJ, Molaodi OR, et al. Longitudinal study of cardiometabolic risk from early adolescence to early adulthood in an ethnically diverse cohort[J]. BMJ Open, 2016, 6(12): e013221.

[2] Ansarullah SI, Kumar P. A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method[J]. Int J Recent Technol Eng, 2019, 7 (6S): 1009-1015.

[3] Yekkala I, Dixit S, Jabbar MA. Prediction of heart disease using ensemble learning and Particle Swarm Optimization [C]//2017 International Conference on Smart Technologies for Smart Nation (Smart Tech Con). IEEE, 2017: 691-698.

[4] Salem T. Study and analysis of prediction model for heart disease: an optimization approach using genetic algorithm[J]. Int J Pure Appl Math, 2018, 119(16): 5323-5336.

[5] Dulhare UN. Prediction system for heart disease using naive Bayes and particle swarm optimization[J]. Biomed Res, 2018, 29(12): 2646-2649.
[6] Gokulnath CB, Shantharajah SP. An optimized feature selection based on genetic approach and support vector machine for heart disease[J]. Cluster Comput, 2019, 22(6): 14777-14787.
[7] Khourdifi Y, Bahaj M. Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization[J]. Int J Intell Syst, 2019, 12(1): 242-252.
[8] Gultepe Y, Rashed S. The use of data mining techniques in heart disease prediction[J]. International Journal of Computer Science and Mobile Computing, 2019, 8: 136-141.
[9] Sarkar BK. Hybrid model for prediction of heart disease[J]. Soft Comput, 2020, 24(3): 1903-1925.
[10] Cherian RP, Thomas N, Venkitachalam S. Weight optimized neural network for heart disease prediction using hybrid lion plus particle swarm algorithm[J]. J Biomed Inform, 2020, 110: 103543.
[11] 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.
Zhao JC, Li Y, Wang D, et al. Heart disease prediction algorithm based on optimized random forest[J]. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2021, 42(2): 112-118.
[12] Valarmathi R, Sheela T. Heart disease prediction using hyper parameter optimization (HPO) tuning[J]. Biomed Signal Proces, 2021, 70: 103033.
[13] Li L, Jamieson KG, De Salvo G, et al. Hyperband: bandit-based configuration evaluation for hyperparameter optimization[C]//ICLR (Poster), 2017.
[14] Jamieson K, Talwalkar A. Non-stochastic best arm identification and hyperparameter optimization[C]//Artificial Intelligence and Statistics. PMLR, 2016: 240-248.
[15] Ma X, Sha J, Wang D, et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning[J]. Electron Commer Res Appl, 2018, 31: 24-39.
[16] Shaker B, Yu MS, Song JS, et al. LightGBM: computational prediction model of blood-brain-barrier penetration based on LightGBM[J]. Bioinformatics, 2021, 37(8): 1135-1139.
[17] Wang Y, Chen J, Chen X, et al. Short-term load forecasting for industrial customers based on TCN-LightGBM[J]. IEEE Trans Power Syst, 2020, 36(3): 1984-1997.
[18] Efraimidis PS, Spirakis PG. Weighted random sampling with a reservoir [J]. Inform Process Lett, 2006, 97(5): 181-185.
[19] Apache. Index of /ml/machine-learning-databases/heart-disease [EB/OL]. https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/
[20] Pelikan M, Goldberg DE, Cantú -Paz E. BOA: the Bayesian optimization algorithm [C]//Proceedings of the Genetic and Evolutionary Computation Conference, 1999, 1: 525-532.
[21] Bergstra J, Yamins D, Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures[C]//International Conference on Machine Learning. PMLR, 2013: 115-123.
[22] Spall JC. Introduction to stochastic search and optimization: estimation, simulation, and control[M]. John Wiley & Sons, Inc., 2005.

（编辑：薛泽玲）