doi: 10.13413/j.cnki.jdxblxb.2024223

基于 Bayes 超参数优化梯度提升树的心脏病预测方法

王海燕, 焦增晨, 赵 剑, 安天博, 鞠 熠
(长春大学 计算机科学技术学院, 残障人士智能康复及无障碍教育部重点实验室, 长春 130022)

摘要: 针对传统机器学习算法在数据集 Cleveland 和 Hungary 上预测准确率低的问题, 提出一种基于 Bayes 超参数优化梯度提升树的心脏病预测方法. 首先, 采用 K-最近邻算法对数据集中的缺失值进行填补, 用 Min-Max 标准化、One-Hot 编码处理数据, 并基于梯度提升树算法进行心脏病预测; 其次, 采用 Bayes 优化和十倍交叉验证的方式搜寻算法的最佳超参数组合. 实验结果表明, 优化后的梯度提升树算法在心脏病数据集 Cleveland 上预测准确率可达 90.2%, 在心脏病数据集 Hungary 上预测准确率可达 81.4%, 优于决策树、支持向量机、K-最近邻等传统机器学习方法, 可辅助医生进行心脏病诊断.

关键词: 心脏病预测; K-最近邻算法; 梯度提升树; Bayes 优化
中图分类号: TP181; TP301.6 **文献标志码**: A **文章编号**: 1671-5489(2025)02-0472-07

Heart Disease Prediction Method Based on Bayesian Hyperparameter Optimization Gradient Boosting Trees

WANG Haiyan, JIAO Zengchen, ZHAO Jian, AN Tianbo, JU Yi
(*Key Laboratory of Intelligent Rehabilitation and Accessibility for People with Disabilities of Ministry of Education, College of Computer Science and Technology, Changchun University, Changchun* 130022, *China*)

Abstract: Aiming at the problem of low prediction accuracy of traditional machine learning algorithms on Cleveland and Hungary dataset, we proposed a heart disease prediction method based on Bayesian hyperparameter optimization gradient boosting trees. Firstly, the K-nearest neighbor algorithm was used to fill in the missing values in the dataset, Min-Max standardization and One-Hot encoding were used to process the data, and the gradient boosting tree algorithm was used to predict the heart disease. Secondly, Bayesian optimization and ten-fold cross validation were used to search for the best combination of hyperparameters of the algorithm. The experimental results show that the prediction accuracy of the optimized gradient boosting tree algorithm can reach 90.2% on the Cleveland heart disease dataset, and the prediction accuracy can reach 81.4% on the Hungarian heart disease dataset, outperforming traditional machine learning methods such as decision tree, support vector machine and the K-nearest neighbor, it can assist doctors in the diagnosis of heart disease.

Keywords: heart disease prediction; K-nearest neighbor algorithm; gradient boosting tree; Bayesian

收稿日期: 2024-06-04.

第一作者简介: 王海燕(1980—), 女, 汉族, 博士, 教授, 从事人工智能、约束求解、约束优化和康复工程的研究, E-mail: wanghy80@ccu.edu.cn.

基金项目: 吉林省教育厅科学技术研究项目(批准号: JJKH20220597KJ)和吉林省科技发展计划项目(批准号: YDZJ202201ZYTS549).

optimization

近年来，心血管病(CVD)已成为危害我国居民健康的重大公共卫生问题[1]. 随着人工智能的高速发展，关于心脏病分类预测的研究也取得了一些进展. Comak 等[2]使用最小二乘支持向量机(SVM)分类器代替人工神经网络设计了一个对心脏病分类的决策支持系统. Parthiban 等[3]对朴素 Bayes 方法诊断糖尿病患者的心脏病工作进行延伸，提出了在 SVM 分类器中使用径向基函数(RBF)核的方法. Weng 等[4]将数据集分为训练队列和验证队列，在训练队列中推导出 CVD 风险算法，在验证队列中应用和测试算法证明，结果表明机器学习显著提高了心血管风险预测的准确率. 王巡等[5]提出了基于 Hadoop 的分布式 C4.5 决策树算法，使用 MapReduce 并行编程模式、HDFS 分布式文件存储系统加快了 C4.5 算法的效率，并应用于心脏病诊断中. 王健等[6]将一维的心脏病数据升维成类似图像格式的三维数据，使数据能适应卷积神经网络，通过对比卷积层数量对准确率和训练时间的影响，得到最优心脏病预测模型. 赵金超等[7]使用 K-最近邻方法(KNN)对数据进行预处理，通过网格搜索的方法对随机森林进行优化，提出了 KNN-RF 模型，该模型对心脏病进行预测的准确率达 83.2%. 王成武等[8]利用网格搜索与交叉验证相结合的方法对模型进行初步的优化，在此基础上使用粒子群优化算法对模型进行进一步优化，优化后模型分类预测的结果得到明显提升. 受上述工作的启发，本文提出一种 Bayes 优化梯度提升树的方法，并对数据集进行归一化、独热编码处理，实验结果表明，模型准确率得到进一步提升.

1　算法描述

1.1　梯度提升树

梯度提升树(GBDT)是 Boosting 策略的一种实现方式，它通过一组分类器的串行迭代得到一个强学习器，以进行更高精度的分类[9]. 无论是处理回归问题还是二分类以及多分类问题，梯度提升树使用的决策树(DT)都是分类回归树(CART)，计算流程如下.

1) 初始化学习器：

$$f(x)=\arg\min_{c}\sum_{i=1}^{N}L(y_i,c), \tag{1}$$

其中 $f(x)$为初始化过程的目标函数，$\sum_{i=1}^{N}L(y_i,c)$ 为损失函数，c 为最小损失常数.

2) 计算残差：

$$r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]f(x)=f_{m-1}(x), \tag{2}$$

将 r_{mi}作为预测值，当 $i=1,2,\cdots,i$ 时拟合得到第 m 棵回归树.

3) 遍历节点，计算回归树每个叶子节点 R_{mj}的输出值 c_{mj}：

$$c_{mj}=\arg\min_{c}\sum_{x\in R_{mj}}L(y_i,f_{m-1}(x_i)+c). \tag{3}$$

4) 更新学习器：

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}c_{mj}I(x), \tag{4}$$

其中 I 表示学习率，J 表示叶子节点个数.

5) 得到强分类器：

$$f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}I(x). \tag{5}$$

1.2　Bayes 优化

Bayes 优化常应用于机器学习的超参数寻优中，其原理是：先通过 Bayes 定理估计目标函数的后验分布，再根据分布选择下一个需要采样的超参数组合，然后利用之前的采样结果完善目标函数，最

终找到全局最优超参数[10].

使用梯度提升树进行心脏病预测时，超参数的选择会影响预测的准确性. 用 Bayes 优化方法优化梯度提升树，对训练过程中梯度提升树的迭代次数(n_estimators)、学习率参数(learning_rate)、弱分类器被允许的最大深度(max_depth)等超参数进行调参，从而获得最佳的超参数组合.

设 $X=X_1,X_2,\cdots,X_n$ 为一组超参数组合，用 $f(x)$表示关于超参数 x 的目标函数，Bayes 优化通过找到 $x\in X$，使得

$$x^* = \underset{x\in X}{\operatorname{argmax}} f(x), \tag{6}$$

其中 x^* 为最优的参数集.

2 数据处理

本文选取 UCI 上公开的心脏病数据集 Cleveland 和 Hungary，这两个数据集均由心脏病检查患者的体质数据组成. 数据集 Cleveland 共有 303 个样本，其中阳性样本 164 个，阴性样本 139 个，每个样本都有 14 个特征，特征属性有连续型和离散型两种，在所有样本中，ca 列有 4 个缺失值，thal 列有 2 个缺失值. 数据集 Hungary 中 slope,ca,thal 列的缺失值过多，因此采用删除法将 3 列数据删除，处理后的数据集共有 294 个样本，其中阳性样本 106 个，阴性样本 188 个，每个样本有 11 个特征. 心脏病数据集的特征信息列于表 1.

表 1 心脏病数据集特征信息

Table 1 Feature information of heart disease dataset

特征名称	特征描述	特征类型
age	年龄	连续型
sex	性别(1=男性，0=女性)	离散值
cp	胸痛类型(1=典型心绞痛，2=非典型性心绞痛，3=非心绞痛，4=无症状)	离散值
trestbps	静息血压	连续型
chol	胆固醇测量值	连续型
fbs	空腹血糖(1=真，0=假)	离散值
restecg	静息心电图(0=正常，1=患有 ST-T 波异常，2=左心室肥大)	离散值
thalach	最大心率	连续型
exang	运动性心绞痛(1=有过，0=没有)	离散值
oldpeak	运动导致的相对于休息的 ST 抑制	连续型
target	是否患心脏病(0=否，1=是)	离散值
slope	最高运动 ST 段的斜率(1=上坡，2=平坦，3=下坡)	离散值
ca	萤光显色的主要血管数目(0～3)	连续型
thal	地中海贫血血液疾病(1=固定缺陷，2=正常，3=可逆缺陷)	离散值

2.1 KNN 缺失值处理

由于得到的数据集含有缺失值，因此需对数据集进行 KNN 缺失值处理. 先将数据初始化，再对缺失值的数据点做 K-最近邻填充，计算含缺失值的数据点与其他不含缺失值数据点的距离矩阵，选出欧氏距离最近的 K 个数据点，欧氏距离公式为

$$D=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}, \tag{7}$$

其中 D 为邻近点的距离. 用选中的 K 个近邻数据点对应的字段均值填充数据中的空缺值[11].

2.2 Min-Max 标准化

数据标准化处理是数据挖掘的一项基础工作，不同数据通常有不同的量纲和量纲单位，这种情况会影响数据分析的结果，为消除数据之间的量纲影响，需进行数据标准化处理，以解决数据之间的可比性.

通过对 UCI 心脏病数据集分析，数据集中各特征属性的尺度不同，因此本文采用 Min-Max 标准化，将数据压缩为 0～1 内，以提高模型分类的准确度. 标准化公式为

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{8}$$

其中 $x_{\min}$ 为该维度数据最小值，$x_{\max}$ 为该维度数据最大值.

2.3　One-Hot 编码

One-Hot 编码又称为独热编码，是一种将离散分类标签转换为二进制向量的方法. 其基本思想是使用 N 位状态寄存器对 N 个状态进行编码，每个状态都由其独立的寄存器位表示，并且在任意时刻，其中只有一位有效. 在 UCI 心脏病数据集中，cp，restecg，slope，thal 等都属于离散值，为适应模型训练，将这些特征进行 One-Hot 编码处理. 通过对各离散特征进行 One-Hot 编码处理，数据集由原来的 14 个特征变为 27 个特征，处理后的数据集在预测准确度上得到提高. 例如，对胸痛类型(cp)(0：典型心绞痛，1：非典型心绞痛，2：非心绞痛，3：无症状)进行 One-Hot 编码处理，处理过程如图 1 所示.

图 1　One-Hot 编码处理胸痛(cp)类型特征

Fig. 1　One-Hot encoding deals with chest pain (cp) type features

3　实验及结果分析

3.1　评价标准

在进行心脏病预测时，通常使用分类的准确率(Accuracy)、精准率(Precision)和召回率(Recall)等指标评价分类器的性能. 其中：准确率表示对给定的测试数据集，分类器预测正确的样本数与总样本数之比；精准率是预测正确的正例样本占预测为正例样本的比例；召回率是预测为正例的样本占实际为正例样本的比例.

设 P 为正样本数量，N 为负样本数量，则准确率、精确率和召回率计算公式分别为

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

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

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

将上述评价指标综合分析，可评价算法的优劣. 这些评价指标都与混淆矩阵相关，混淆矩阵列于表 2.

表 2　混淆矩阵

Table 2　Confusion matrix

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

为能更直观地表征算法的预测准确率，通常绘制 ROC(receiver operating characteristic)曲线. ROC 曲线基于混淆矩阵得出，用于评价模型的预测能力，其由真阳性率作为纵坐标、假阳性率作为横坐标绘制一条曲线. ROC 曲线下的面积为 AUC(area under curve)，是判断二分类预测模型优劣的标准，AUC 值越大，表明模型越好.

3.2 实验环境

实验使用的硬件配置：CPU为i5-8300H，内存为16 GB，显卡为NVIDIA GeForce GTX 1050 Ti. 在Windows10操作系统上，基于Pycharm开发平台，使用Python语言，安装Sklearn，Pandas，Numpy等依赖库作为实验的软件环境.

3.3 实验过程

本文采用KNN-GBDT模型作为心脏病预测的方法，实验数据来自心脏病数据集Cleveland和Hungary，由于两个数据集的特征内容相同，通过数据预处理后，两者对数据的表现形式一致，因此，为实验能表现更好的泛化能力，分别对两个数据集的数据进行训练预测，并按7∶3分为训练集和测试集进行建模. 为提高模型的预测准确率，本文对数据集进行缺失值处理、One-Hot编码以及Min-Max标准化等一系列数据预处理操作，得到无噪声、可用性高的数据. 对处理后的数据先采用十倍交叉验证方法划分训练集和测试集，再分别使用KNN-GBDT模型和4种传统的机器学习算法对训练集进行训练，对测试集进行测试，以Accuracy，Precision，Recall和AUC作为评价标准对模型进行评估. 验证本文模型在心脏病预测中的性能. 模型对比流程如图2所示.

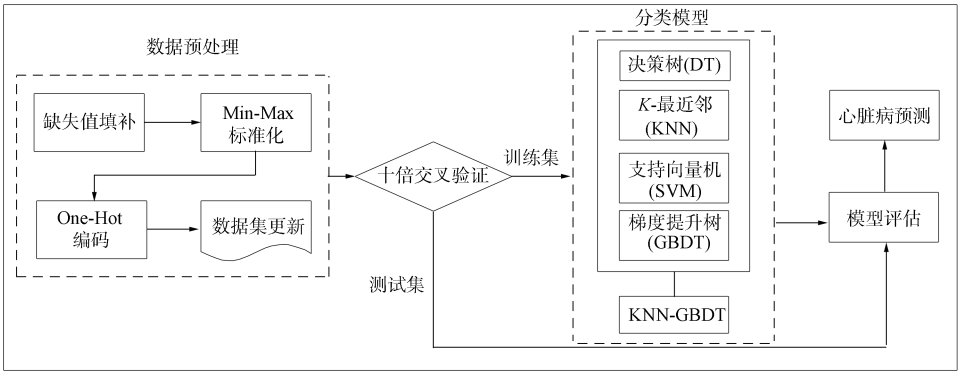

图2 基于优化的梯度提升树预测算法模型对比流程

Fig. 2 Flow chart of model comparison of gradient boosting tree prediction algorithms based on optimization

由于获取公开的心脏病数据集Cleveland和Hungary中含有缺失值，因此需先对数据集进行KNN缺失值填补，填补后的数据集存在离散值特征，而这些离散值特征可能会在梯度提升树上形成稀疏节点，影响整个模型的性能和效率，因此需进一步选用One-Hot编码处理填补后的离散值特征. 数据归一化处理可以将连续型的特征值压缩到0～1内，以提高模型分类的准确率，因此将处理后的数据集使用Min-Max标准化处理. 数据处理后，不同的参数组合对梯度提升树的预测结果有很大影响，因此采用Bayes算法对梯度提升树的迭代次数、学习率、决策树的最大深度等参数进行优化，在设置的参数范围内经过多轮寻优求得一组最佳参数组合，各参数设置范围列于表3.

表3 各参数设置范围

Table 3 Range of each parameter setting

参数名称	参数描述	设置范围
n_estimators	迭代次数，即决策树的个数	(50,250)
min_samples_split	一个节点在被分割前所需的最小样本数	(2,25)
max_features	每个节点在进行分割时所考虑的特征数量	(0.1,0.99)
max_depth	每个决策树的最大深度	(5,12)
min_samples_leaf	一个叶子节点所需的最小样本数	(1,10)
learning_rate	学习率	(0.01,0.1)
random_state	随机种子	(0,50)

获得一组最佳参数后，将最佳参数组合代入梯度提升树算法中，分别在心脏病数据集Cleveland和Hungary上训练预测，用式(9)计算模型的准确率，将KNN-GBDT模型和传统的梯度提升树算法做对比，优化后的梯度提升树模型相对于传统的梯度提升树算法在心脏病数据集Cleveland上准确率提高了9.9%，在心脏病数据集Hungary上准确率提高了7.1%，分类效果得到有效提高. 此外，

KNN-GBDT 模型相对于其他传统机器学习算法在 Recall 和 AUC 评价指标上也有提高.

优化过程调用 bayes_opt 库的 bayes_opt.BayesianOptimization 函数通过步长调整参数，在设置的参数范围内进行寻优，采用十倍交叉验证方法评估. 通过 Bayes 算法多次寻优，训练数据集 Cleveland 最优超参数组合为[n_estimators：78，min_samples_split：4，max_features：0.14，max_depth：5，min_samples_leaf：7，learning_rate：0.02，random_state：44]时模型准确率最高，训练数据集 Hungary 时，得到最优超参数组合为[n_estimators：63，min_samples_split：3.2，max_features：0.10，max_depth：6，min_samples_leaf：4.6，learning_rate：0.02，random_state：25]时模型准确率最高.

DT,KNN,SVM,GBDT 几种常见的机器学习算法在分类预测中应用的较多，为验证 KNN-GBDT 模型的有效性，在数据集 Cleveland 上将其与 4 种算法进行对比，计算各算法的评价指标，对比结果列于表 4.

表 4　在数据集 Cleveland 上不同预测算法的性能比较

Table 4　Performance comparison of different prediction algorithms on Cleveland dataset

预测算法	准确率	精准率	召回率	AUC
DT	0.711	0.571	0.615	0.69
KNN	0.803	0.778	0.800	0.85
SVM	0.829	0.806	0.829	0.91
GBDT	0.803	0.800	0.821	0.88
KNN-GBDT	0.902	0.933	0.875	0.94

为进一步验证 KNN-GBDT 模型的泛化能力，在数据集 Hungary 上进行训练，训练后的预测准确率有明显提升，对比结果列于表 5. 由于心脏病数据集 Hungary 含有 106 个阳性样本，188 个阴性样本，经过处理后每个样本剩余 11 个特征，数据集中阳性、阴性类别的样本数量差异较大，因此，处理后的特征未能很好地捕捉样本之间的差异，导致数据集的类别不平衡及特征表示不足，进而导致 KNN-GBDT 模型的精准率略低于 KNN 算法. 综合实验结果表明，KNN-GBDT 模型能进行心脏病的预测.

表 5　在数据集 Hungary 上不同预测算法的性能比较

Table 5　Performance comparison of different prediction algorithms on Hungarian dataset

预测算法	准确率	精准率	召回率	AUC
DT	0.689	0.560	0.538	0.65
KNN	0.811	0.850	0.607	0.87
SVM	0.743	0.708	0.586	0.76
GBDT	0.743	0.647	0.458	0.81
KNN-GBDT	0.814	0.750	0.714	0.92

由表 4 和表 5 可见，将 KNN-GBDT 模型与 DT,KNN,SVM,GBDT 几种算法对比，优化后的梯度提升树在两个数据集上都表现出高性能，预测准确率得到提升，在召回率、精准率、AUC 各项评价指标上都优于其他传统的机器学习算法. 图 3 为在数据集 Cleveland 上各算法的 ROC 曲线. 图 4 为在数据集 Hungary 上各算法的 ROC 曲线. 其中 FPR 表示假阳性率，TPR 表示真阳性率. 由图 3 和图 4 可见，KNN-GBDT 模型的 ROC 曲线下的面积最大，AUC 值分别为 0.94,0.92，证明模型的性能最佳，充分说明了优化后的梯度提升树算法可有效预测心脏病.

综上所述，本文采用 UCI 上的心脏病数据集 Cleveland 和 Hungary 进行实验，针对原有模型预测准确率低、不稳定的问题，提出了一种基于 Bayes 优化梯度提升树的预测方法，先用 K-最近邻算法进行缺失值填补、Min-Max 标准化、One-Hot 编码对数据进行处理，再结合 Bayes 优化和交叉验证方法对模型的最优超参数组合进行寻优. 将本文模型在准确率、召回率、精准率和 AUC 上与传统机器学习方法进行对比的实验结果表明，本文提出的 KNN-GBDT 模型在数据集 Cleveland 上准确率可达 90.2%，AUC 可达 0.94；在数据集 Hungary 上准确率可达 81.4%，AUC 可达 0.92. 实验结果验证了

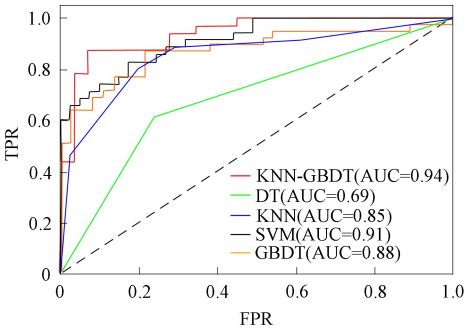

图3 在数据集 Cleveland 上各算法的 ROC 曲线

Fig. 3 ROC curves of each algorithm on Cleveland dataset

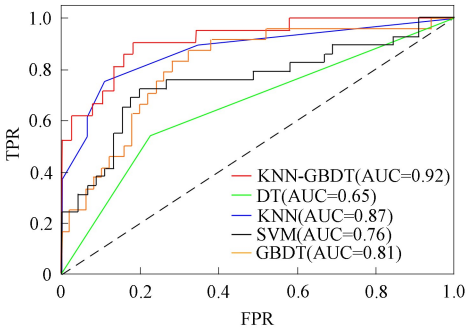

图4 在数据集 Hungary 上各算法的 ROC 曲线

Fig. 4 ROC curves of each algorithm on Hungarian dataset

优化算法可有效提高模型的分类能力，可辅助医生进行心脏病预测的诊断. 但在对数据集 Hungary 进行预测评估时，本文模型在精准率上略低于 KNN 算法，这可能是因为数据类别不平衡以及特征表示不足所致. 在后续的工作中，会收集更多的样本数据，以提高模型的精准率并改善模型的性能.

参 考 文 献

[1] 王增武，胡盛寿.《中国心血管健康与疾病报告2022》要点解读[J]. 中国心血管杂志，2023，28(4)：297-312.(WANG Z W，HU S S. Chinese Journal of Cardiovascular Health and Disease Report 2022：Key Points Interpretation[J]. Chinese Journal of Cardiovascular，2023，28(4)：297-312.)

[2] COMAK E，ARSLAN A，TURKOGLU I. A Decision Support System Based on Support Vector Machines for Diagnosis of the Heart Valve Diseases[J]. Computers in Biology and Medicine，2007，37(1)：21-27.

[3] PARTHIBAN G，RAJESH A，SRIVATSA S K. Diagnosing Vulnerability of Diabetic Patients to Heart Diseases Using Support Vector Machines[J]. International Journal of Computers and Applications，2012，48(2)：45-49.

[4] WENG S F，REPS J，KAI J，et al. Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?[J]. PloS One，2017，12(4)：0174944-1-0174944-14.

[5] 王巡，杜方辉. 基于 Hadoop 的 C4.5 决策树算法在心脏病诊断中的应用[J]. 信息技术与信息化，2017(Z1)：36-40.(WANG X，DU F H. Application of C4.5 Decision Tree Algorithm Based on Hadoop in Heart Disease Diagnosis[J]. Information Technology and Information Technology，2017(Z1)：36-40.)

[6] 王健，李孝虔. 一种基于特征组合和卷积神经网络的心脏病预测新方法[J]. 黑龙江大学自然科学学报，2019，36(1)：115-120.(WANG J，LI X Q. A New Method of Heart Disease Prediction Based on Feature Combination and Convolutional Neural Network[J]. Journal of Natural Science of Heilongjiang University，2019，36(1)：115-120.)

[7] 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 J C，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.)

[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.(WANG C W，GUO Z H，YAN J F. Research on Improved Support Vector Machine in Heart Disease Prediction[J]. Computer Technology and Development，2022，32(3)：175-179.)

[9] YANG T，CHEN W T，CAO G T. Automated Classification of Neonatal Amplitude-Integrated EEG Based on Gradient Boosting Method[J]. Biomedical Signal Processing and Control，2016，28：50-57.

[10] 崔佳旭，杨博. Bayes 优化方法和应用综述[J]. 软件学报，2018，29(10)：3068-3090.(CUI J X，YANG B. Bayesian Optimization Algorithm and Application Review[J]. Journal of Software，2018，29(10)：3068-3090.)

[11] HUANG J L，KEUNG J W，SARRO F，et al. Cross-Validation Based *K* Nearest Neighbor Imputation for Software Quality Datasets：An Empirical Study[J]. Journal of Systems and Software，2017，132：226-252.

(责任编辑：韩 啸)