

## 基于改进 Borderline-Smote-GBDT 的冠心病预测

李瑞平<sup>1</sup>, 朱俊杰<sup>2</sup>

1. 河南理工大学电气工程与自动化学院, 河南 焦作 454003; 2. 河南省煤矿装备智能检测与控制重点实验室, 河南 焦作 454003

**【摘要】**针对样本不平衡问题,提出一种基于欧氏距离改进的 Borderline-Smote 过采样算法。首先根据欧式距离判断少数类样本类别;然后根据边界上的少数类样本的k近邻数据找出线性直线,由同侧近邻数据判别是否为噪音;最后重新判别删除噪音的剩余少数类样本的类别,对边界少数类样本和密集的非边界区域的少数类样本过采样合成新样本。等磁场图和二维电流密度图中提取的心磁特征数据集经过改进 Borderline-Smote 过采样处理,结果表明改进 Borderline-Smote-GBDT 冠心病预测模型相比 Borderline-Smote-GBDT 模型准确率提高8.4%,精确率提高2.9%,召回率提高9.1%,AUC提高4.6%。此外,与逻辑回归、随机森林、k近邻、极端随机树模型对比发现,GBDT结果最优,改进 Borderline-Smote-GBDT 准确率、召回率、精确率、AUC分别为91.7%、91.7%、81.8%、87.1%,验证了该模型的可行性。

**【关键词】**冠心病;Borderline-Smote;梯度提升树

**【中图分类号】**R318;TP391

**【文献标志码】**A

**【文章编号】**1005-202X(2023)10-1278-07

## Coronary heart disease prediction based on improved Borderline-Smote-GBDT

LI Ruiping<sup>1</sup>, ZHU Junjie<sup>2</sup>

1. School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454003, China; 2. Henan Key Laboratory of Intelligent Detection and Control of Coal Mine Equipment, Jiaozuo 454003, China

**Abstract:** A Borderline-Smote oversampling algorithm which is improved based on the Euclidean distance is proposed to address the problem of sample imbalance. The category of minority class samples is determined according to the Euclidean distance. Then, the k nearest neighbor data of minority class samples on the boundary is used to find the linear straight-line, and the noise is removed after identifying whether it is the noise misrecognized as boundary samples based on the ipsilateral neighbor data. Finally, the category of the remaining minority class samples is re-determined, and new samples are synthesized through the oversampling for minority class samples on the boundary and those in the dense non-boundary region. The feature datasets extracted from the isomagnetic field map and the two-dimensional current density map are processed with the improved Borderline-Smote oversampling, and the results show that compared with Borderline-Smote-GBDT model, the improved Borderline-Smote-GBDT model for coronary heart disease prediction enhances the accuracy, precision, recall rate and AUC by 8.4%, 2.9%, 9.1%, and 4.6%, respectively. Through the comparison with logistic regression, random forest, k nearest neighbor and extremely randomized tree, it is found that GBDT performs best, and that improved Borderline-Smote-GBDT model has an accuracy, recall rate, precision and AUC of 91.7%, 91.7%, 81.8%, and 87.1%, respectively, which verifies the model feasibility.

**Keywords:** coronary heart disease; Borderline-Smote; gradient boosting decision tree

### 前言

随着科技的发展、社会节奏的变化、老龄化的加剧等,冠心病的患病率和死亡率接连增高。2020年

中国心血管报告指出,2018年中国城市居民冠心病死亡率12.018%,农村居民死亡率12.824%<sup>[1]</sup>。目前,冠心病的诊断仍存在一定困难,多数冠心病患者平时并无任何症状,呈隐性状态<sup>[2]</sup>,即使某些冠心病患者表现出轻微症状,也可能因为病因复杂、诊断复杂等而未能及时治疗。

分类数据中通常会存在样本不平衡问题,如医疗诊断的病例样本、欺诈交易检测、产品质量分析中的不合格样本等,导致分类模型效果较差。常见的处理方式包括数据方面的采样<sup>[3-5]</sup>、数据增强、数据合

**【收稿日期】**2023-04-10

**【基金项目】**国家自然科学基金(61601173)

**【作者简介】**李瑞平,硕士,研究方向:交通信息处理与装置, E-mail: 1835507496@qq.com

**【通信作者】**朱俊杰,博士,讲师,研究方向:生物医学信号处理, E-mail: junjiezhu@hpu.edu.cn

成等,算法方面的带权值的损失函数、难例挖掘等<sup>[6-8]</sup>。Wang等<sup>[9]</sup>提出一种基于参差聚类和改进Smote的过采样算法,聚类划分少数子簇,质心法限定的区域内产生新样本。Ning等<sup>[10]</sup>采用Borderline-Smote过采样处理阳性样品和阴性样品之间的不平衡问题,Tomek链接技术滤除噪声数据。Ren等<sup>[11]</sup>提出一种考虑样本数量分布、类收敛趋势和样本收敛趋势的自适应代价敏感学习。机器学习是一门新兴的人工智能学科,已被广泛应用于帮助医生做出客观的预测和判断<sup>[12]</sup>。

本研究通过心磁图获取数据,建立Borderline-Smote-GBDT冠心病预测模型。改进Borderline-Smote过采样处理不均衡数据(对边界少数类样本局部划分边界分割子集,计算同侧近邻数据类别和判断噪音,过采样合成新样本),贝叶斯优化梯度提升树(Gradient Boosting Decision Tree, GBDT)模型重要参数,以构建最优模型进行冠心病二分类预测。

1 数据处理

1.1 数据集获取

实验数据由120人的心磁数据组成,包含74名正常对照者,46名冠心病患者。在人体胸前上方36个位置以1000 Hz的采样频率进行心脏磁场测量,获得心磁数据,根据磁场强度和位置信息,把磁场强度相同的点连接起来,对单一时刻的心磁数据进行3次样条插值处理得到等磁场磁图,正常对照者630 ms时刻的等磁场图见图1,图中不同颜色代表着正负磁极附近的磁场强度。由于冠心病主要与ST波密切相关,对心磁图QRST段进行分割,主要分析ST间隔的数据。T波峰的时间为 $T_t$ 、R波峰的时间为 $T_r$ 、TT波段的起止时间为 $T_t-67$ 到 $T_t+33$ 、ST间隔为 $T_r+49$ 到 $T_t-67$ 。在此基础上求解电流密度图,假设某一点在等磁场图上的坐标为 $(x,y)$ ,该点磁场强度为 $F(x,y)$ ,则该点的电流密度 $J=(\frac{\partial F(x,y)}{\partial y}, -\frac{\partial F(x,y)}{\partial x})$ ,从而得到所有的电流密度,绘制正常对照者二维电流密度图,见图2。图中箭头方向代表着电活动的传播方向,大小代表着电活动在该处的兴奋程度。

1.2 离群值处理

数据中个别偏离预期的数据值会对分类结果造成偏差,使用四分位距法构建箱线图筛选离群值<sup>[13-14]</sup>,超出上下限的数据被定义为离群值。医学统计中,离群值产生的原因未明确前,尤其是数据量很少时,离群值的取舍对分类结果会产生较大的影响。为避免离群值影响分析和统计建模的结果,离群值当作缺失值进行链式方程多重插补<sup>[15-16]</sup>。原始数据

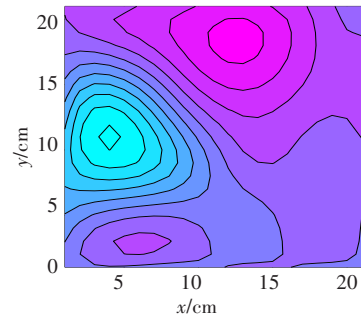


图1 等磁场图  
Figure 1 Isomagnetic field map

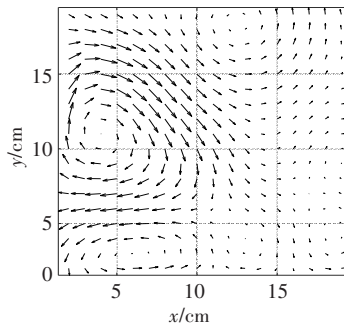


图2 二维电流密度图  
Figure 2 Two-dimensional current density map

与插补数据的分布见图3,红色曲线分别是3个信息熵、TT间期磁场角度最小值、T峰正负磁极磁场角度、TT间期最大电流角度的原始数据,黑色曲线是4次的插补值,根据4次插补值与原始数据的误差对比选择第一次的插补值。

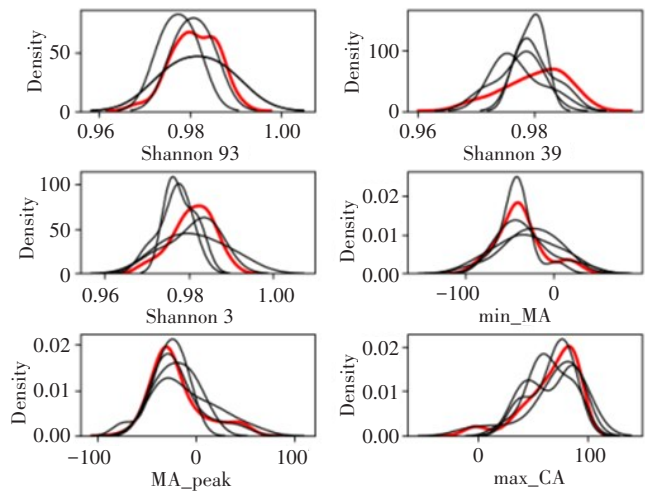


图3 插补值与原始训练集分布图  
Figure 3 Imputation values and the original training set distribution plots

2 改进Borderline-Smote过采样算法

2.1 Borderline-Smote算法

数据集中患者与对照组数据存在不平衡问题,会造成二分类模型严重偏向性。常用欠采样<sup>[17]</sup>和过采样<sup>[18]</sup>处理不平衡数据,欠采样会丢失大量数据,损失一些重要信息。Han等<sup>[19]</sup>提出的 Borderline-Smote 在 Smote 的基础上进行改进,充分考虑近邻不同类别样本的分布特点,使合成样本分布更合理,解决样本重叠问题,避免过拟合。采样过程根据 k 近邻样本把少数类样本分为 3 类: Safe 类、Danger 类、Noise 类,最后对判定为边界上的少数类样本按 Smote 方式随机线性插值合成新样本。

Borderline-Smote 算法流程如下:假设多数类样本集  $M = \{m_1, m_2, \dots, m_i\}$ , 少数类样本集  $N = \{n_1, n_2, \dots, n_i\}$ , k 最近邻多数类样本个数  $m$ , 首先根据不同数据之间的欧氏距离确定每一个  $n_i$  最近邻样本集; 然后判断  $n_i$  近邻样本集  $m$  的大小, 找出  $k/2 < m < k$  的样本集  $N_{\text{danger}}$ ; 最后设置采样倍率,  $N_{\text{danger}}$  中每一个样本与其近邻合成新样本。

## 2.2 改进 Borderline-Smote 算法

传统的 Borderline-Smote 过采样方法没有考虑噪音对少数类样本类别划分的影响。划分类别时, 噪音不仅会导致少数类样本类别划分错误, 而且会致使噪音样本进行过采样, 引入过多的噪音。Borderline-Smote 过采样方法也没有考虑非边界区域的少数类样本信息, 会损失一些重要的信息。针对这些的缺点, 本研究提出一种改进的 Borderline-Smote 过采样方法, 通过对噪音和非边集中的少数类样本的处理, 减少噪音的生成和增加非边界样本信息, 提高模型的泛化能力。

边界上的少数类样本的处理步骤:

设少数类样本集  $S_{\min} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ , 近邻多数类样本个数  $k_1$ , 近邻少数类样本  $k_2$ 。

输入: 少数类样本集  $S_{\min}$ ,

输出: 边界上的少数样本。

首先根据式(1)从  $S_{\min}$  中找出边 k 近邻样本符合  $k_1 > k_2$  且  $k_2 = 1$  的少数类样本集  $S_{\text{danger1}}$ 。

$$d(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2} \quad (1)$$

然后假设每一个样本  $d_i \in S_{\text{danger1}}$ , 其最近邻样本集为  $S_{dd}$ , 从  $S_{dd}$  中找出唯一一个少数类样本  $x_j (x_j \in S_{\min})$ , 从  $S_{dd}$  找出离  $x_j$  最近的两个数据  $x_{j1}, x_{j2}$ 。计算线性方程:

$$(x - x_{j1})(y_{j2} - y_{j1}) = (y - y_{j1})(x_{j2} - x_{j1}) \quad (2)$$

最后根据式(2)计算出  $d_i$  同侧的样本集, 当同侧近邻数据中有且只有一个少数类样本时, 则  $d_i$  视为噪音。

Danger 类样本处理如图 4 所示, 通过线性方程判别是否是噪音。假设  $k_{\text{neighbors}}=5$ , 圆圈代表多数类样本, 五角星代表少数类样本。A 近邻只有一个 B, B 离 A 近邻数据集中 C1、C2 最近, 以 C1、C2 两点绘制 L 直线, 以此为边界, 可以看出以 L 划分的 A 点那侧最近邻的(半圆弧内)数据类别。当半圆弧内无少数类样本(A、B 两点不同侧和 A 点近邻内无少数类样本)或只有一个少数类样本(A、B 两点同侧或 A 点半圆弧内存在少数类样本)时, A 点判定为噪音, 判定为噪音的样本不参与后期的过采样处理。

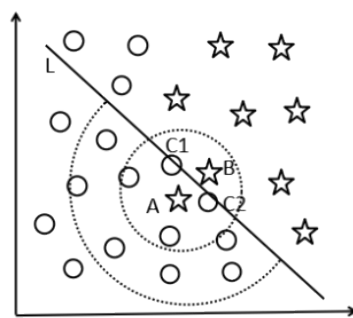


图4 误判为边界样本的噪音识别示意图

Figure 4 Schematic diagram of identifying the noise misrecognized as boundary samples

非边界区域少数类样本合成步骤如下:

设总样本集  $S$ , 近邻多数类样本个数  $k_1$ , 近邻少数类样本  $k_2$ 。

输入: 总样本集  $S$ ,

输出: 合成的非边界少数类样本。

首先假设  $x_i \in S$ , 根据式(1)确定近邻样本集  $S_m$ , 且  $S_m \in S$ 。

然后对每一个  $x_i$  找出符合  $k_1 < k_2$  且  $k_1 = 0$  的少数类样本, 合成新的非边界样本:

$$x_{\text{new}} = x_i + d_{ij} \times \text{rand}(0, R_{ij}) \quad (3)$$

其中,  $R_{ij}$  取值 0.5 或 1.0,  $d_{ij}$  为  $x_i$  与近邻样本对应属性  $j$  中的插值。

非边界区域的少数类样本合成过程如图 5 所示, 圆圈为多数类样本, 五角为少数类样本, 三角为合成样本。当少数类样本比较密集, 周围无多数类样本, 进行插值合成新样本。

改进 Borderline-Smote 算法流程如下:

设训练样本集  $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ 。

输入: 多数类样本集  $S_{\max j}$ , 少数类样本集  $S_{\min}$ , k 近邻多数类样本个数  $k_1$ , 近邻少数类样本  $k_2$ 。

步骤 1: 根据式(1)计算  $S$  样本间的欧氏距离。

步骤 2: 找出每一少数类样本符合  $k_1 > k_2$  且  $k_2 = 1$



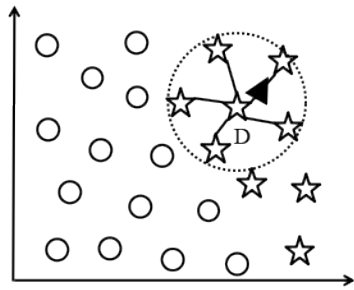


图5 非边界区域少数类样本过采样  
Figure 5 Oversampling of minority class samples in non-boundary region

的少数类样本集  $S_{\text{danger}}$ 。假设  $d_i (d_i \in S_{\text{danger}})$ , 最近邻样本集  $S_{dd}$ , 确定  $S_{dd}$  中少数类样本  $x_i (x_i \in S_{\text{min}})$ ,  $S_{dd}$  中离  $x_j$  最近的两个数据  $x_{j1}, x_{j2}$ 。

步骤3: 近邻值为  $k'$  时, 根据式(3)计算每一个样本  $d_i$  的  $y_{d_i}$  的大小, 当  $y_{d_i} > 0$  或  $y_{d_i} < 0$ ,  $k'$  近邻数据带入式(4)找出同侧数据, 同侧周围无多数类样本时, 判定为噪音并删除。

$$y = \frac{y_{j2} - y_{j1}}{x_{j2} - x_{j1}} x + y_{j1} - \frac{y_{j2} - y_{j1}}{x_{j2} - x_{j1}} x_{j1} \quad (4)$$

步骤4: 除噪音的剩余少数类样本集  $S_{\text{min}1}$  重新划

分类别, 选择每一个样本  $d_j \in S_{\text{min}1}$  的近邻样本集  $S_{nn}$  符合  $k/2 < |S_{nn} \cap S_{\text{max}j}| < k$  的少数类样本。从  $S_{nn}$  中随机选择一个样本  $d_{i(nn)}$ , 从而合成新样本  $S_{\text{new}1}$ :

$$S_{\text{new}1} = d_j + \text{rand}(0,1) (d_{i(nn)} - d_j) \quad (5)$$

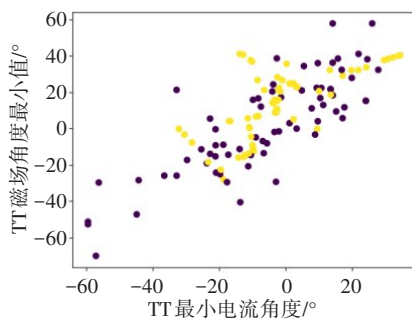
得到 Danger 类合成的新样本  $S_{\text{new}1}$ 。

步骤5: 对 Safe 类中  $k_1 = 0$  的少数类样本过采样合成新样本  $S_{\text{new}2}$ 。

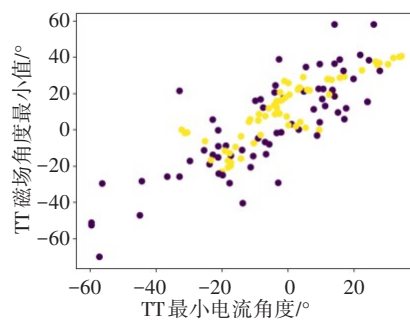
输出: 少数类样本过采样合成的新样本:

$$S_{\text{new}} = S_{\text{new}1} \cup S_{\text{new}2} \quad (6)$$

传统的 Borderline-Smote 算法和改进的 Borderline-Smote 算法合成少数类新样本的分布情况, 见图6, 图中紫色圆圈代表正常对照者, 黄色圆圈代表冠心病患者。改进 Borderline-Smote 算法根据局部分类边界优化了判定为边界的少数类样本, 删除了被判定为边界样本的噪音, 同时避免了噪音对类别划分的影响。改进过采样算法对密集的非边界样本的处理也保留了部分非边界样本信息。从图6中可以看出, Borderline-Smote 合成样本的决策边界比较复杂, 患者数据和正常对照者数据难以区分。改进 Borderline-Smote 合成的样本更加聚集, 非线性决策边界会更容易区分。



a: Borderline-Smote 算法合成新样本



b: 改进 Borderline-Smote 算法合成新样本

图6 过采样合成新样本分布图

Figure 6 Distributions of new samples synthesized through oversampling

### 3 基于改进 Borderline-Smote-GBDT 的冠心病模型构建

#### 3.1 GBDT 模型原理及构建

本研究选择 GBDT<sup>[20]</sup> 构建冠心病预测模型。单独使用决策树时, 容易发生拟合。GBDT 是一种迭代的决策树算法, 通过构造一组弱的学习器, 并把多颗决策树的结果累加起来作为最终的预测输出, 将决策树与集成思想进行有效的结合, 解决拟合的问题。

基于改进 Borderline-Smote 算法和 GBDT 的冠心

病预测模型流程见图7。提取相关特征: T 峰正负磁极的磁场角度、T 峰最大电流角度、TT 最小电流角度、TT 磁场角度最小值、TT 最大电流角度变化值、TT 最大电流角度、ST 磁场角度最大值、3 个 TT 期间信息熵。原始数据集预处理: 重复值、缺失值、离群值处理及特征选择。测试过程使用原始数据进行测试, 过采样之前对原始数据集进行划分; 改进 Borderline-Smote 算法处理不平衡数据; 处理后的训练集和测试集标准化; 贝叶斯优化 GBDT 重要参数, 建立最优模型; 测试集对训练好的模型进行预测评估。

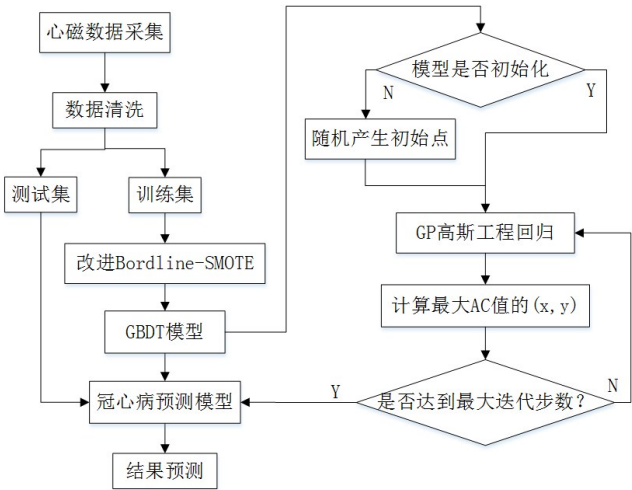


图7 冠心病预测模型流程图  
Figure 7 Flowchart of coronary heart disease prediction model

3.2 贝叶斯优化

模型参数优化能让模型输出和实际观测数据之间达到最佳的拟合程度。贝叶斯优化调参采用高斯过程,拟合优化目标函数,充分利用被测试点忽略的前一个点的信息<sup>[21-22]</sup>,迭代次数少,针对非凸问题依然稳健,可以在有限计算次数内更快地找到近似最优解,提高调参效率。贝叶斯调参的核心过程是:先验函数(PF)<sup>[23]</sup>和采集函数(AC)<sup>[24]</sup>,通过从目标函数获取信息,找到下一个评估位置,从而找到最优超参数组合。

4 结果与分析

4.1 分类指标

对于样本分类来说,混淆矩阵是衡量分类模型最直观、最基本的方法,可以更好地分析每个类别的误分情况。混淆矩阵中将少数类认为负例,多数类认为是正例。真阳性(TP)为检测患病,实际患病;假阳性(FP)为检测患病,实际无病;真阴性(TN)为检测无病,实际无病;假阴性(FN)为检测无病,实际患病。使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、接受者操作特征(Receiver Operating Characteristic, ROC)曲线、ROC曲线下面积(AUC)作为二分类模型的评价指标。如式(7)~(9)所示:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

4.2 模型优化

GBDT模型建立之后,主要对n\_estimators(弱学

习器的最大迭代次数)、max\_depth(决策树最大深度)、loss(损失函数)、learning\_rate(学习率)等主要参数进行优化,得到最优的参数组合。根据max\_depth、min\_samples\_leaf等控制树的复杂性和大小,避免过度拟合。模型调优结果见表1。

表1 GBDT模型参数调优结果  
Table 1 GBDT model parameter tuning results

参数名称	含义	取值
n_estimators	弱学习器的最大迭代次数	286
max_depth	决策树最大深度	4
loss	损失函数	deviance
learning_rate	学习率	0.05

4.3 GBDT模型结果分析

采用不同方法处理不平衡样本,GBDT构建集成学习模型,贝叶斯优化参数寻找最优解。分类指标见表2。对比发现不平衡数据过采样后建立的GBDT模型的分类指标都有一定的提高,显然,处理不平衡样本后构建的模型分类效果较好。对比Borderline-Smote算法和改进Borderline-Smote算法,数据经过改进Borderline-Smote过采样算法后建立的GBDT模型的训练效果比较好,各个评价指标都有所提高;相较于Borderline-Smote-GBDT模型,改进Borderline-Smote-GBDT模型的准确率、召回率、精确率以及AUC分别提高了8.4%、9.1%、2.9%、4.6%。对比GBDT集成模型的ROC曲线,结果发现改进Borderline-Smote-GBDT模型的分类效果更好(图8)。

表2 GBDT模型分类效果  
Table 2 Prediction performances of GBDT model

模型	准确率	精确率	召回率	AUC
GBDT	0.750	0.875	0.636	0.800
Borderline-Smote-GBDT	0.833	0.888	0.727	0.825
改进 Borderline-Smote-GBDT	0.917	0.917	0.818	0.871

4.4 机器学习模型对比分析

为进一步验证改进Borderline-Smote算法的有效性和Borderline-Smote-GBDT模型的分类效果,分别对采用Borderline-Smote算法及改进Borderline-Smote算法处理后的数据进行RF<sup>[25]</sup>、GBDT、KNN<sup>[26]</sup>、ET<sup>[27]</sup>、LR<sup>[28]</sup>模型训练,贝叶斯对各个模型主

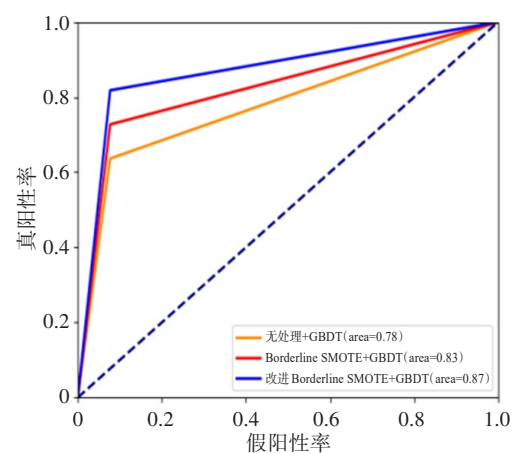


图8 ROC曲线  
Figure 8 ROC curve

要参数调优,得到最优组合参数模型并进行预测。  
各分类器分类指标结果见表3。

表3 不同机器学习模型分类效果

Table 3 Prediction performances of different machine learning models

过采样算法	模型	准确率	精确率	召回率	AUC
Borderline-Smote	GBDT	0.833	0.888	0.727	0.825
	KNN	0.791	0.750	0.818	0.794
	ET	0.833	0.888	0.727	0.825
	RF	0.750	0.727	0.727	0.748
	LR	0.708	0.643	0.818	0.717
改进 Borderline-Smote	GBDT	0.917	0.917	0.818	0.871
	KNN	0.833	0.833	0.727	0.825
	ET	0.875	0.875	0.727	0.725
	RF	0.833	0.833	0.636	0.800
	LR	0.750	0.750	0.818	0.755

GBDT集成方式采用Boosting,通过加法模型,不断减小训练过程中产生的残差进行分类或者回归,模型具有较好的解释性和鲁棒性。在改进Borderline-Smote处理数据的基础上建立的GBDT的指标都优于KNN、RF等分类器,准确率提高4.2%~16.7%,精确率提高1.1%~20.8%,召回率提高0%~18.2%。

过采样算法对数据处理可以避免预测偏向样本较多的分类。改进Borderline-Smote算法处理数据后建立的5种模型的指标都要优于Borderline-Smote算法。改进的过采样算法不仅避免了Noise类的干扰,而且避免了误判的噪音合成新的噪音,同时考虑了非边界区域的少数类样本,为分类器提供更多的样本信息。

5 结 语

本研究依据心磁检测干扰因素少、对切向电流和体表电位差异敏感等特点,从心磁图和二维电流密度图中提取ST波段135个特征,根据3种特征选择方法选择最优的特征组合。基于传统Borderline-Smote算法缺陷,提出改进Borderline-Smote算法处理样本不均衡,使用最优参数组合构造GBDT分类器对冠心病进行预测。实验对比结果验证了改进Borderline-Smote算法可以有效地提高模型的分类能力。对比不同机器学习方法,集成学习思想的GBDT模型可以很好地对冠心病进行预测。

【参考文献】

[1] 国家心血管病中心. 中国心血管健康与疾病报告2020[J]. 心肺血管病杂志, 2021, 40(9): 885-889.  
National Center for Cardiovascular Diseases. China cardiovascular health and disease report 2020 [J]. Journal of Cardiopulmonary Vascular Diseases, 2021, 40(9): 885-889.

[2] 邓丽君. 颈动脉超声在诊断老年冠心病患者中的临床应用研究[J]. 现代诊断与治疗, 2022, 33(1): 99-102.  
Deng LJ. Clinical application of carotid ultrasound in the diagnosis of elderly patients with coronary heart disease [J]. Modern Diagnosis and Treatment, 2022, 33(1): 99-102.

[3] Ren JJ, Wang YP, Mao MQ, et al. Equalization ensemble for large scale highly imbalanced data classification[J]. Knowl-Based Syst, 2022, 242(22): 108295.

[4] Liqaa MS, Jamila HS. Adaptation proposed methods for handling imbalanced datasets based on over-sampling technique [J]. Mustansiriyah J Sci, 2020, 31(2): 25-29.

[5] Priyanka R, Arcot S, Erik M, et al. Data augmentation with improved regularisation and sampling for imbalanced blood cell image classification.[J]. Sci Rep, 2022, 12(1): 18101.

[6] Huang C, Huang X, Fang Y, et al. Sample imbalance disease classification model based on association rule feature selection[J]. Pattern Recogn Lett, 2020, 133: 280-286.

[7] 万建武, 杨明. 代价敏感学习方法综述[J]. 软件学报, 2020, 31(1): 131-136.  
Wan JW, Yang M. Review of cost-sensitive learning methods [J]. Journal of Software, 2020, 31(1): 131-136.

[8] Lin CT, Chen SP, Santoso PS, et al. Real-time single-stage vehicle detector optimized by multi-stage image-based online hard example mining[J]. IEEE Trans Veh, 2020, 69(2): 1505-1518.

[9] Wang X, Yang Y, Chen MS, et al. AGNES-SMOTE: an oversampling algorithm based on hierarchical clustering and improved SMOTE[J]. Sci Program, 2020, 2020(242): 108295.

[10] Ning Q, Zhao XW, Ma ZQ. A novel method for identification of glutarylation sites combining borderline-SMOTE with Tomek links technique in imbalanced data [J]. IEEE/ACM Trans Comput Bi, 2021, 19(5): 2632-2641.

[11] Ren ZJ, Zhu YS, Kang W, et al. Adaptive cost-sensitive learning: improving the convergence of intelligent diagnosis models under imbalanced data[J]. Knowl Based Syst, 2022, 241: 108296.

[12] 张静, 赵洛沙. 高敏等原发性高血压合并冠心病患者脂蛋白(a)的变化[J]. 中国综合临床, 2011, 5: 471-474.  
Zhang J, Zhao LS. Study on blood lipoprotein (a) in patients with essential hypertension and coronary heart disease [J]. Clinical Medicine of China, 2011, 5: 471-474.

[13] Gaura EI, Brusey J, Allen M, et al. Edge mining the internet of things[J]. IEEE Sens J, 2013, 13(10): 3816-3825.

[14] Choi H, Poythress JC, Park C, et al. Regularized boxplot via convex clustering[J]. J Stat Comput Sim, 2019, 89(7): 1227-1247.

[15] Van BV, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis [J]. Stat Med, 1999, 18(6): 681-694.

- [16] Holmes Finch W, Hernandez Finch ME, Singh M. Data imputation algorithms for mixed variable types in large scale educational assessment: a comparison of random forest, multivariate imputation using chained equations, and MICE with recursive partitioning[J]. *Int J Educ Res*, 2016, 3(3): 129-153.
- [17] Lee YS, Bang CC. Framework for the classification of imbalanced structured data using undersampling and convolutional neural network[J]. *Inform Syst Front*, 2021, 24(6): 1-15.
- [18] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority oversampling technique[J]. *J Artif Intell Res*, 2002, 16: 321-357.
- [19] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [C]// *International Conference on Intelligent Computing*, 2005, 3644(5): 878-887.
- [20] Xu X, Lin MK, Xu TT. Epilepsy seizures prediction based on nonlinear features of EEG signal and gradient boosting decision tree[J]. *Int J Env Res Pub He*, 2022, 19(18): 11326.
- [21] 邓帅. 基于改进贝叶斯优化算法的CNN超参数优化方法[J]. *计算机应用研究*, 2019, 36(7): 1984-1987.
- Deng S. Hyper-parameter optimization of CNN based on improved Bayesian optimization algorithm [J]. *Application Research of Computers*, 2019, 36(7): 1984-1987.
- [22] Salim L. Integrating convolutional neural networks, KNN, and Bayesian optimization for efficient diagnosis of Alzheimer's disease in magnetic resonance images[J]. *Biomed Signal Proces*, 2023, 80(1): 104375.
- [23] Lu Y, Herbei R, Kurtsek S. Bayesian registration of functions with a Gaussian process prior[J]. *J Comput Graph Stat*, 2017, 26(4): 894-904.
- [24] 崔佳旭, 杨博. 贝叶斯优化方法和应用综述[J]. *软件学报*, 2018, 29(10): 3068-3090.
- Cui JX, Yang B. Review of Bayesian optimization methods and applications[J]. *Journal of Software*, 2018, 29(10): 3068-3090.
- [25] Ilhem T, Akila D, Farida Hayet M, et al. An improved random forest based on feature selection and feature weighting for case retrieval in CBR systems: application to medical data [J]. *Int J Software Innovation*, 2022, 10(1): 1-20.
- [26] Arti R, Ankur D, Rajesh S, et al. An efficient machine learning approach for diagnosing Parkinson's disease by utilizing voice features[J]. *Electronics*, 2022, 11(22): 3782-3782.
- [27] 李宏彬, 贺太平. 三种决策树同源算法在肝部B超计算机辅助诊断中的应用比较[J]. *医学信息*, 2021, 34(19): 13-18.
- Li HB, He TP. Comparison of three decision tree homology algorithms in computer aided diagnosis of liver B ultrasound [J]. *Journal of Medical Information*, 2021, 34(19): 13-18.
- [28] Anitha P, Srimathi C. Isogeny Hosmer-Lemeshow logistic regression-based secured information sharing for pharma supply China[J]. *Electronics*, 2022, 11(19): 3170.

(编辑:谭斯允)