ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 4,Jul. 2023,pp. 860－872
DOI:10. 16337/j. 1004-9037. 2023. 04. 010

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种可用于鉴别肝癌呼气信号的改进AdaBoost算法

郝丽俊[1,2]，黄　钢[3,1]

(1. 上海理工大学健康科学与工程学院，上海 200093；2. 上海健康医学院医疗器械学院，上海 201318；3. 上海健康医学院附属嘉定中心医院上海市分子影像学重点实验室，上海 201318)

摘　要：提出一种改进的AdaBoost强化学习算法，并将其应用于鉴别健康者和肝癌患者的呼气信号。首先采集志愿者(包括健康对照组和肝癌患者)的呼气信号，利用Relief算法提取其主要特征；接着融合Stacking模型，基于传统的机器学习算法训练得到若干基分类器组，构建一个个子分类器。为减少训练样本对分类器性能的影响，利用K折交叉，先后得到k个基分类器，形成一个基分类器组；进一步，由投票法得到该基分类器组，即子分类器对测试集的预测结果；然后根据各子分类器对训练集的预测错误率调整训练样本，并获得各子分类器的权重系数；最后将多个子分类器的预测结果进行加权组合，得到最终预测结果。实验结果表明，相比传统的AdaBoost算法，改进的AdaBoost算法在鉴别肝癌呼气和健康对照组呼气时，错误率明显下降，鲁棒性有所提升。该算法在鉴别肝癌呼气时，准确率可以达到90%左右，特异性和精确度也均超过95%。因此，改进的AdaBoost算法可有效提升肝癌呼气鉴别精度，对通过呼气鉴别肝癌、实现早期诊断的研究具有重要意义。

关键词：呼气检测；肝癌鉴别；AdaBoost算法；Stacking模型；基分类器组；Relief算法

中图分类号：TP391　　**文献标志码**：A

An Improved AdaBoost Algorithm for Identifying Breath Signals of Liver Cancer

HAO Lijun[1,2], HUANG Gang[3,1]

(1. School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. Medical Instrumentation College, Shanghai University of Medicine & Health Sciences, Shanghai 201318, China; 3. Shanghai Key Laboratory of Molecular Imaging, Jiading District Central Hospital Affiliated Shanghai University of Medicine and Health Sciences, Shanghai 201318, China)

Abstract: An improved AdaBoost reinforcement learning algorithm is proposed for distinguishing the breath signals of healthy patients and liver cancer patients. First, the breath signals of volunteers, including healthy controls and liver cancer patients, are collected and their main features are extracted by Relief algorithm. Then, based on Stacking model, several groups of base classifiers are trained by traditional machine learning algorithms and some sub-classifiers are then constructed. To reduce the influence of training samples on the classifier performance, a K-fold crossover is applied, and k base classifiers could be successively obtained to form a base classifier group. Further, the prediction results of this base classifier group, i.e., sub-classifiers on the test set, are obtained by the voting method. Then, according to the

基金项目：国家自然科学基金(82127807)；国家重点研发计划(2020YFA0909000)；上海市分子影像学重点实验室建设项目(18DZ2260400)。
收稿日期：2022-07-11；**修订日期**：2022-09-10

prediction error rate of each sub-classifier on the training set, the training set is updated and the weight coefficients of each sub-classifier are obtained according to the prediction error rate of each sub-classifier on the training set. Finally, the prediction results of multiple sub-classifiers are weighted and combined to obtain the final prediction results. Experimental results show that the improved AdaBoost algorithm can achieve an accuracy of about 90% and the specificity and precision are more than 95% in discriminating the breath of liver cancer from the breath of healthy controls. Compared with the traditional AdaBoost algorithm, the proposed algorithm has significantly lower error rate and improved robustness when used for liver cancer breath detection. Therefore, the improved AdaBoost algorithm can effectively improve the accuracy of liver cancer breath identification, which is important for the research of identifying liver cancer by breath for early diagnosis.

Key words: breath detection; liver cancer identification; AdaBoost algorithm; Stacking model; base classifier group; Relief algorithm

引　　言

肝癌是目前全球范围内发病率和致死率最高的癌症之一。根据世界卫生组织国际癌症研究机构(IARC)研究数据显示:2020年,肝癌位于世界上发病人数最多的癌症之一,排名第三;在中国,癌症死亡率中肝癌高居第二。无论是全球还是中国,死于肝癌的人数正在越来越接近新诊断的人数[1]。

肝癌常用的主要诊断方法有血清检验、活组织检验和医学影像诊断,其中,影像诊断是辅助肝癌诊断的重要手段之一。该方法能避免对患者造成伤害,但缺点是不够准确,容易受医生主观情绪影响,而且微小的病灶不易被发现[2]。活组织检验是一种监测肝脏组织中可疑病变处以协助诊断的方法,需要通过穿刺和开刀获取组织。实际临床中,肝穿刺活检的应用非常谨慎,因为它有导致癌细胞转移的风险[3]。血清检验简单、易操作,利用肝癌分子标志物(如甲胎蛋白AFP)进行肝癌检测。但由于约30%的肝癌患者AFP水平正常,因此对于那些有临床症状或者高危因素的患者,他们AFP的检测结果不能作为唯一的参考项,诊断效率较低[4]。目前肝癌分子标志物也正在研究中[5]。

电子鼻是近年来一种新型的仪器,可通过采集呼气中的挥发性有机化合物(Volatile organic compounds,VOCs)监测和诊断人体疾病。该方法具有无创、操作简单、检查费用低廉等优点,已成为近年来研究的热点。Mazzone等[6]通过气体化学传感器检测肺癌呼气信号,灵敏度和特异度比较高,结果显示肺癌的正确分类接近100%,健康对照的正确分类接近94%,该研究对肺癌患者呼气中VOCs的分析使人们看到呼气检测有望成为一种新型无创的临床诊断工具;Oakley-Girvan等[7]则作出一个系统评价,确定了与肺、结肠直肠和乳房相关的呼出气VOCs,进一步表明呼气分析在癌症筛查和早期检测方面显示出大好前景;Germanese等[9]研究检测呼出气中的氨区分肝脏损伤严重程度的可能性,证明了基于金属氧化物半导体(Metal-oxide-semiconductor, MOS)气体传感器在检测呼出气氨方面可取得良好效果,发现了一些显著的相关性参数,确定了基于呼出气检测肝脏疾病的可能性;Kitiyakara等[10]通过研究动物的嗅觉来预测肝细胞癌(Hepatocellular carcinoma, HCC)的可能,结果表明具有一定的可行性,准确度为78%,但这只是概念证明,在临床应用之前,需进一步完善检测过程;秦涛[11]通过建立呼气检测方法学,探索了呼气中有机物浓度与肝癌的其他标志物和分期的关系,并尝试建立肝癌的呼气诊断模型,结果显示部分物质诊断肝癌的灵敏性和特异性可分别达到83.3%与91.7%。但该研究是基于固相微萃取/气相色谱/质谱联用技术的,硬件平台昂贵且不易操作,不适宜肝癌的普及型筛查推广。基于此,本文将进一步探索如何基于电子鼻采集的呼气信号,构建高性能的鉴别诊断模型。

构建数学模型表征呼气信号与病症之间的关系，是电子鼻检测算法的核心。基于呼气鉴别肝癌患者和健康者本质上是一个二分类问题。目前应用于电子鼻系统的模式识别算法主要有主成分分析(Principal component analysis, PCA)算法、Fisher判别法、支持向量机、逻辑回归、人工神经网络等[12-14]。但这些算法的性能均与训练样本的数量密切相关。为了降低训练样本数量对检测算法的影响，提升电子鼻呼气检测肝癌的准确度和特异性，本文融合Stacking模型[16]，对AdaBoost算法进行了改进，提出了一种新的强化机器学习算法。首先选择一种传统机器算法，通过K折交叉分组训练，依次得到k个基分类器及对测试集的k个预测值；进一步，基于Stacking模型，得到该组基分类器对训练集的预测值；接着基于投票法[17]，由该基分类器组得到一个子分类器对测试集的预测结果；然后分别选择多个不同的分类算法，并基于前一次对训练样本的预测结果调整样本权重后，依次训练得到更多的基分类器组，并得到多个子分类器对测试集的预测结果；最后将所有子分类器的预测结果进行加权组合，得到最终预测结果。这样做一方面可以减少训练样本的影响，提高分类器的泛化能力；另一方面可保留AdaBoost算法的优点，根据子分类器的训练误差调整其权重[18]，将多个基分类器进行加权组合，提升分类的各项性能指标。

1 鉴别分类器的设计原理

鉴别肝癌患者的呼气信号，本质上是能够设计一种算法将肝癌患者和健康对照组的样本特征进行分类区分，以实现未来利用呼气对肝癌进行早期诊断的目的。

AdaBoost算法是一种自适应增强方法，是集成学习的一种。集成学习是将不同模型通过某些机制或设定标准进行融合，以得到一个更加强大稳健的模型。集成学习分类器的泛化能力更强，且避免了单个模型过拟合等问题。在AdaBoost算法中，对同样的训练集调整样本权重得到不同的训练集，并进一步训练得到多个弱分类器，然后将这些弱分类器加权组合，得到一个最终的分类器。在传统的AdaBoost算法中，多个弱分类器是基于同一个分类算法构建的，一次训练可得到一个弱分类器，本文尝试对此进行改变调整，提出一种改进的AdaBoost算法。

为了获得具有良好泛化性能的高精度分类器，本文尝试将3种常用集成算法的核心思想融合，设计了一种改进的AdaBoost算法。首先借鉴Stacking模型中第一层模型的构建方法，使用K倍交叉划分训练集，得到不到的训练样本，并训练获得多个基学习器[15]；接着融合Bagging模型中最终分类器的形成思想，基于投票方法[16]，由多个基学习器中确定一个子分类器；然后利用AdaBoost理论，根据子分类器的训练误差，调整训练集样本的分布，并得到子分类器的加权系数；之后，进入新一轮的训练，获得新的子分类器。此外，为了融合多个特性的分类器，在新一轮训练中，将加入一种新的机器学习算法来重复上述步骤，基于调整样本分布后的训练集，获得新的子分类器及加权系数。在达到预设训练次数后，停止训练，并对所有子分类器进行加权和组合，实现异质集成，得到最终预测结果。

1.1 基于Stacking模型和投票法的子分类器构建

在AdaBoost算法中，首先从初始训练集训练出一个子分类器，再根据子分类器的表现对训练样本分布进行调整，然后基于调整后的样本分布得到下一个子分类器，如此重复，最终将多个子分类器进行加权组合[17]。因此，子分类器的设计是AdaBoost算法的核心。

对一组训练样本，基于某一机器学习算法，在多次训练中，可获得多个不同的基分类器。假设，将训练集记作TrainSet，测试集记作TestSet，基于Stacking模型构建子分类器的原理如图1所示。图1中，训练集TrainSet按照K折交叉划分为k组，取其中的(k−1)组作为训练样本TrainData，剩余的一组作为测试样本TestData。接着确定一种机器学习算法，基于选择的训练样本得到一个基分类器。然后依次变换测试样本和训练样本，利用同样的分类算法，训练得到更多的基分类器。基于K折交叉验证，对于同一个分类算法，可先后得到k个不同的基分类器。同时，利用各基分类器逐次对相应的k组测试

样本和测试集TestSet分别进行预测。最终可得到k个基分类器、k组测试样本的预测值和k个测试集的预测值。至此,一个由k个基分类器组成的子分类器便构建而成。k组测试样本的预测值集合构成该子分类器对训练集的预测。而基于投票原则,则可得到该子分类器对测试集的一组预测结果。

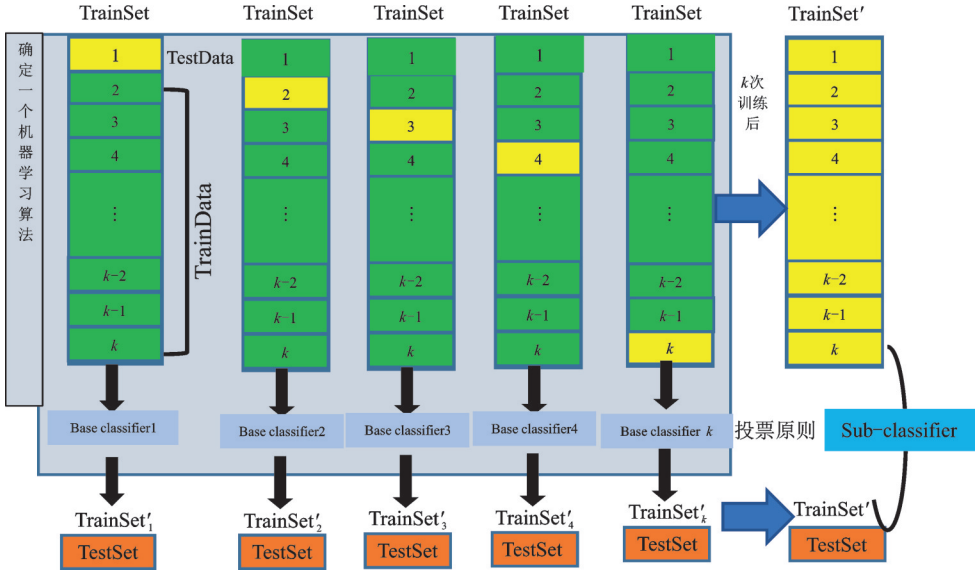

图1 基于Stacking模型和投票原则的子分类器设计

Fig.1 Design of sub-classifier based on Stacking model and voting principle

1.2 改进型AdaBoost分类器的设计

AdaBoost算法的核心是加权组合多个子分类器。在本文算法中,子分类器的设计如1.1小节所述。对同一个训练集,依次选择不同的分类算法,随机进行K折交叉,训练得到多个由k个基分类器投票形成的子分类器,为下一步组合成强化分类器提供分类器组件,如图2所示。

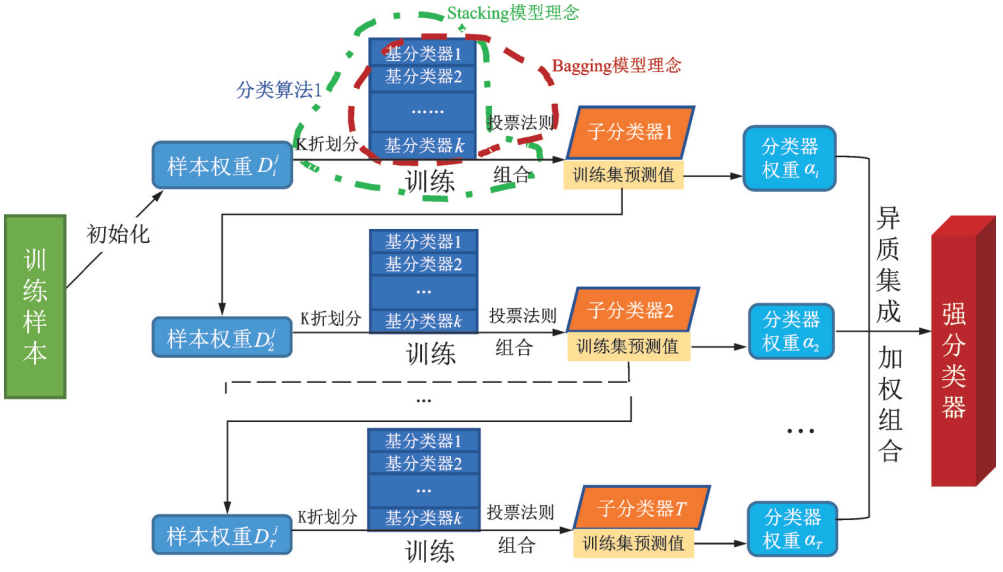

图2 改进型AdaBoost分类器的设计原理图

Fig.2 Design principle diagram of improved AdaBoost classifier

算法的主要过程如下[18]：

(1) 确定T个机器学习算法和训练样本空间，并初始化训练样本数据的权重，即

$$D_i^j=\frac{1}{m}\quad i=1,2,\cdots,T \tag{1}$$

式中：i代表第i个子分类器，i的取值为$1-T$；j代表第j个训练样本，如果训练集样本数为m，则j的最大值为m。

(2) 按照1.1所述，将训练集K折交叉分组，依次选择一折数据作为测试样本，剩余的($k-1$)折数据作为训练样本，基于一个机器分类算法，进行k次训练，逐次得到基于该机器学习算法的k个不同的基分类器，形成一个基分类器组，记作第i个子分类器。

同时，利用该基分类器组对训练集的预测值$\boldsymbol{g}_i(j)$和真实值y_j，由式(2)和(3)计算对应该子分类器的错误率e_i。

$$g_i(j)=[g_{i.}^1,g_i^2\cdots,g_i^k] \tag{2}$$

式中：$g_{i.}^1,g_i^2,\cdots,g_i^k$分别为$k$个基分类器对训练集中一折数据的预测，将其合并，构成一个子分类器对训练集全部样本的预测[15]。

$$e_i=\sum\nolimits_k D_i^j(k)\quad k=1,2,\cdots,m(g_i(j)\neq y_j) \tag{3}$$

式中：k遍历训练集所有样本中，预测值与真实值不相符的样本。对于二分类问题，错误率本质上就是这些样本的权重之和。

(3) 进一步，计算第i个子分类器的权重系数α_i。在AdaBoost算法中，采用指数函数作为损失函数[19]，可得到该子分类器的权重系数α_i。

$$\alpha_i=\frac{1}{2}\lg\left(\frac{1-e_i}{e_i}\right) \tag{4}$$

由式(4)可知，分类误差率e_i越大，则对应的子分类器的权重系数α_i越小，反之亦然。

(4) 根据第i个子分类器的预测情况，调整下一轮算法设计中训练集样本的权重D_{i+1}^j。

$$D_{i+1}^j=D_i^j\mathrm{e}^{-\alpha_i y_j g_i(j)}/\mathrm{Dsum} \tag{5}$$

式中：D_i^j为第i个基分类器组对应的训练集样本权重，而D_{i+1}^j则为调整后的第$i+1$个基分类器组对应的训练集样本权重系数；α_i为第i个子分类器的权重系数，而y_j和$g_i(j)$分别为训练集中第j个样本的真实值和预测值。Dsum为归一化因子，可表示为

$$\mathrm{Dsum}=\sum_{j=1}^{m}D_{i+1}^j \tag{6}$$

从式(5)可以看出，如果第j个样本分类错误，则$y_j g(j)<0$，该样本的权重系数在第$i+1$基分类器组中增大，而如果分类正确，则权重在第$i+1$个基分类器组中减少。

(5)依次选择T个机器学习算法中剩余的算法，重复上述步骤(2)至(4)，并按照不断调整的训练集样本权重计算得到各个子分类器的误差率e_i和权重系数α_i。

(6)将各子分类器的预测结果加权组合，便可实现多个子分类器的异质集成，形成最终的集成强分类器的预测值，即

$$\boldsymbol{H}(l)=\mathrm{sign}\left[\sum_{i=1}^{T}\alpha_i\cdot\boldsymbol{h}_i(l)\right]\quad l=1,2,\cdots,n \tag{7}$$

式中：$\boldsymbol{h}_i(l)$为第i组基学习器对测试集的一组预测值；α_i为第i个子分类器的权重系数；$\boldsymbol{H}(l)$为集成强分类器的预测结果。

2 方 法

2.1 肝癌呼气信号采集

电子鼻是一种基于气体传感器和模式识别技术、模拟生物嗅觉系统,实现气体检测和识别等功能的系统。本研究采用德国UST公司研发的电子鼻系统采集志愿者呼出气体。该电子鼻内含3个传感器,可同时采集3组数据[20]。

本研究已经获得上海长征医院生物医学研究伦理委员会的批准。共采集120例志愿者,包括69例肝癌患者和51例健康对照组的呼气数据。呼气采集在空腹状态下进行,经由口腔呼气完成。采集过程中仅使用一次性吹气嘴,无任何介入性装置使用,对人体无任何伤害。志愿者的纳入标准是患者必须为原发性肝癌,无其他转移癌证,近3个月无抽烟酗酒史。表1所示为志愿者的基本信息。

表1 志愿者基本信息

Table 1 Basic information of volunteers

实验对象	采集样本数/个		平均年龄/岁
	男(平均年龄/岁)	女(平均年龄/岁)	
肝癌患者	57(56.25±10.35)	12(57.18±12.49)	56.40±10.63
健康参照组	34(53.09±14.47)	17(52.26±14.88)	51.60±14.57

将电子鼻系统的采样率设为2,对每一个志愿者,连续采集30 s的呼气数据,3组传感器可同步采集到其3个采样点均为60的波形。图3所示为基于电子鼻系统传感器B采集到的所有志愿者的波形信号。图中,横坐标为采样点,纵坐标为传感器阵列对不同呼出气体的响应电阻。

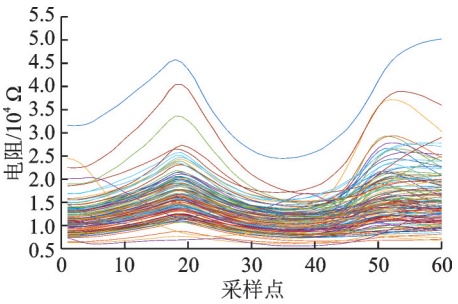

图3 所有志愿者的呼出气体原始波形

Fig.3 Original waveform of exhaled gas of all volunteers

2.2 信号的特征提取及优化

如图3所示,由传感器采集到的输出信号数值和幅度变化较大,且每次采集的数值变化也较大。为便于比较,在此进行归一化处理。在不改变波形状态的情况下,由式(8)将采集到的每一组信号数值,转变成[0,1]范围内的相对值。

$$y_k(i)=\frac{y_k(i)-\min\left(y_k\right)}{\max\left(y_k\right)-\min\left(y_k\right)} \tag{8}$$

式中:k可取A、B、C,分别代表3个传感器;i表示某一传感器采集的第i个样本;$\min\left(y_k\right)$和$\max\left(y_k\right)$分别代表同一传感器采集到的所有样本信号的最小值和最大值。

归一化预处理后产生新的60×120×3数据集,随后对该数据集进行进一步的数据分析,但由于数据的特征不够明显,无法进行高效分类识别,因此需要先进行特征提取,以便提高分类器的准确度。在此,对3个传感器提取的信号分别提取时域、频域和统计等特征[21],具体包括:时域特征14个(最大值及对应位置,最小值及对应位置,平均值、峰峰值、整流平均值、方差、标准差、波形因子、脉冲因子、峰值因子、裕度因子和面积),频域特征14个(重心频率、频率方差、均方根差、频谱和各种方法计算得到的功率谱)和统计特征10个(极差、中位数、分位数、众数、变异系数、偏度、峰度、自相关系数和信息熵),并进一步计算3个传感器信号之间的两两相关性R_{xy},R_{yz}和R_{zx},获得3个特征。将对3组传感器信号提取的所

有特征进行组合,得到对应一个样本的一组高维特征。

为了避免维数灾难问题和提高运行速度,进一步基于Relief算法[22]选择特征,将特征降至不同维度后,再分别利用优化后的特征,逐次构建分类器,并进行分类器性能的计算。

3 实 验

本研究中,首先将两组样本,包括69名肝癌患者和51名健康对照组的呼气信号进行特征提取,形成120×1 560的高维特征数组,然后利用Relief算法提取主成分,分别得到不同维度优化特征数据集,并依次将其作为样本数据集。为了便于构建二分类算法,将肝癌患者呼气样本和健康参照组的呼气样本标签分别记作1和0。

在设计改进的AdaBoost肝癌呼气鉴别算法时,为了综合应用各分类算法的优点,在子分类器的构建时,利用了KNN近邻法、随机森林(Random forest, RF)、逻辑回归模型(Logic regression, LR)、线性判别法分析(Linear discriminant analysis, LDA)、反向传播神经网(Back propagation neural network, BP)和贝叶斯分类器(Bayes)6种不同的机器学习算法模型[23-25]。在设计每一个子分类器时,融合Stacking模型理论,基于K折交叉构建k个基分类器形成一个基分类器组,并进一步利用投票法得到一个子分类器。为了增加分类评估结果的保真性,常见的做法是将$\frac{2}{3}\sim\frac{4}{5}$的样本用于训练,剩余样本用于测试。因此,在本研究中,k值设为5,即进行五折交叉,每次取$\frac{1}{5}$的样本用于测试,剩余样本用于训练。具体过程为:将训练集的正负样本分别进行五折划分,依次选择其中的一折数据为测试样本,剩余样本为训练样本,依据所选择的机器学习算法,先后训练得到5个基分类器,形成一个基分类器组;将5个基分类器对5组不同的测试样本的预测值集合,得到训练集的预测值。而基于投票原则,根据5个基分类器对测试集的预测结果可确定子分类器对测试集的预测值;在得到子分类器和其对训练集的预测值后,基于AdaBoost算法,计算误差值及损失函数,调整得到每个子分类器的加权系数,并更新训练集的权重系数;依次应用6个机器学习算法,形成6个异质子分类器;进一步,在各子分类器预测值的基础上,利用集成思想加权构建出一个强分类器。

为了定量评价改进AdaBoost算法鉴别肺癌呼气的性能。首先将样本数据集进行了随机划分,取其中的20%作为测试集,剩余的80%作为训练集。接着,分别以传统AdaBoost算法和改进型AdaBoost算法构建分类器,并进行性能对比。为了得到较为客观的结果,利用选择的训练集先后10次设计构建分类器,并计算每个分类器对测试集预测的性能指标。

图4所示为在将特征维度降为40后,改进型AdaBoost集成分类器、基于不同子分类器算法的6个传统AdaBoost集成分类器和集成前各子分类器的10次预测误差对比。该图以预测测试集的错误率为衡量参数,对比了各算法的稳定性。在每一次测试中,训练样本的划分均随机且独立。图中,横坐标对应6个不同的子分类器算法(依次分别为KNN、RF、LR、LDA、BP和Bayes),纵坐标为错误率。曲线中,红色标记为改进型AdaBoost分类器的错误率,黑色为传统AdaBoost分类器的错误率,紫色为各个子分类器的预测误差,蓝色为6个子分类器的平均误差率。由图4可以看出,相比其他算法,改进AdaBoost肝癌呼气鉴别算法有效降低了肝癌呼气的检测误差,且错误率比较稳定,基本在10%左右,算法的鲁棒性较好[26]。

图5则以分类器常用的5个性能指标的平均值对比了改进前后分类器在鉴别肝癌呼气时的表现。这5个指标依次为准确率、敏感性、特异性、精准率和F_1-score指标[27]。其中准确率为检测正确的百分比;敏感性,也称为召回率,为肝癌患者能够能正确检出的百分比,敏感性越高,漏诊的可能性越小;特异性为正确检测为正常人的百分比;精准率为正确判断为肝癌的百分比,精准率越高,误诊的可能性越

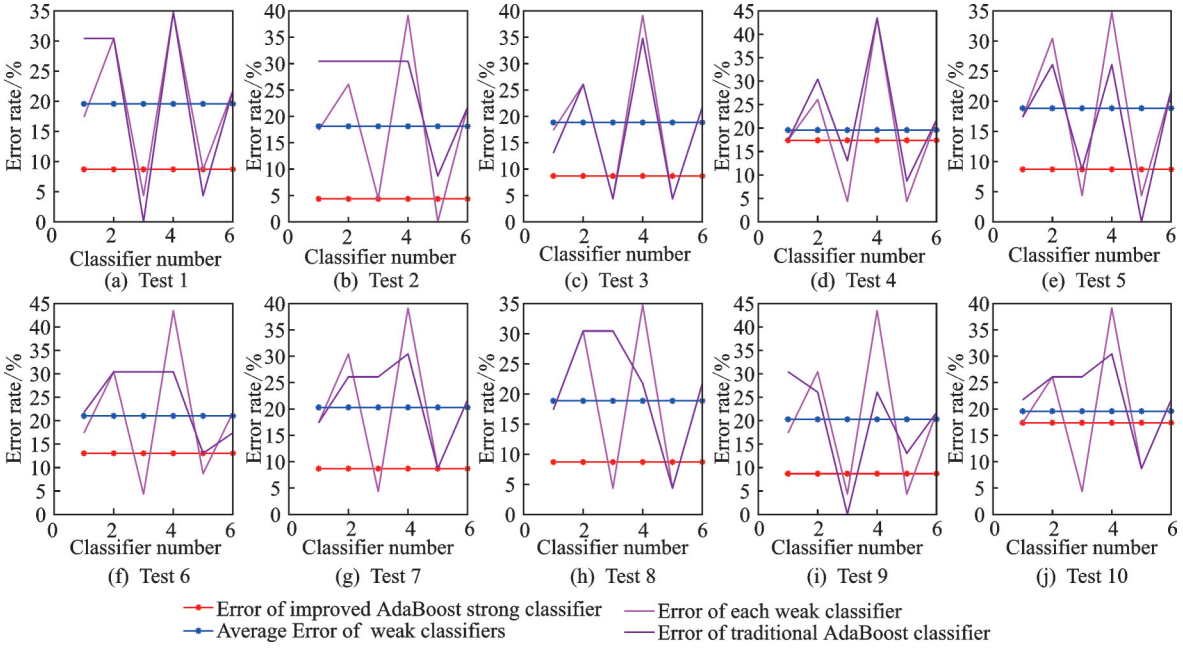

图4 改进型AdaBoost分类器与改进前各子分类器的错误率对比

Fig.4 Comparison of error rate of improved AdaBoost classifier versus pre improved sub classifiers

小;F_1-score为精准率和召回率的调和平均数,是分类问题的一个重要指标,$F1$-score越大,分类器的性能越好。从图中可以看出,改进AdaBoost的肝癌呼气鉴别算法和基于BP算法的传统AdaBoost算法,各项性能明显优越于其他算法。但是,改进AdaBoost肝癌呼气鉴别算法的敏感性和精准率均超过92%,高于其他算法,在正确鉴别肝癌呼气信号方面的优势明显。因此,改进型AdaBoost算法的总体性能较好。另外,基于逻辑回归LR和基于BP神经网络的传统AdaBoost算法的性能也比较高,并且结合图4(见图中横坐标3和5的对应值),发现这两个算法的错误率相对比较低,稳定性也较好。

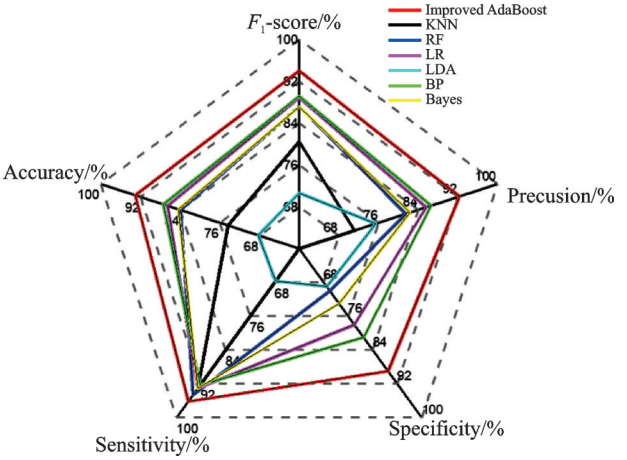

图5 不同肝癌呼气鉴别算法的性能参数对比

Fig.5 Comparison of performance parameters of different liver cancer breath identification algorithms

为了进一步对比以上3种算法分类器性能的优劣,分析计算了在不同特征维度下分类器的性能指标[27]。特征选择是机器学习重要的第一步。特征选择是从候选特征中选出“优秀”的特征。通过特征选择可以达到降维、提升模型效果和性能的效果。一般来说,当特征达到某个数量时,分类器模型的效果达到最优。过多或过少的特征都会引起分类器性能的下降。为了综合对比在不同特征维度下以上3种算法的性能,在此分别将特征维度降至5、10、20、30、40、50、60、70、80、90和100维,分别计算对比各个分类器10次运行后的平均性能,结果如表2所示。从表2可以看到,在特征维度优化为40后,改进型AdaBoost分类器的性能趋于最佳,综合各项性能指标,优于其他传统型AdaBoost分类器。

表2 不同分类器的平均性能

Table 2 Average performance of different classifiers %

分类器	维度	准确率	敏感性	特异性	精准率	F_1-score
Improved AdaBoost	5	76.95	77.69	76	81.53	79.11
Traditional AdaBoost (BP)		76.95	83.07	69	77.81	80.22
Traditional AdaBoost (LR)		66.95	70.00	63	71.63	70.07
Improved AdaBoost	10	82.61	80.76	85	88.56	83.89
Traditional AdaBoost (BP)		82.60	79.23	87	89.03	83.00
Traditional AdaBoost (LR)		71.73	73.07	70	77.07	74.20
Improved AdaBoost	20	86.08	86.92	85	88.76	87.46
Traditional AdaBoost (BP)		85.21	88.46	81	87.6	87.16
Traditional AdaBoost (LR)		86.08	91.53	79	86.03	88.13
Improved AdaBoost	30	86.52	87.69	85	88.88	88.09
Traditional AdaBoost (BP)		87.82	91.53	83	87.64	89.44
Traditional AdaBoost (LR)		85.65	88.46	82	86.87	87.43
Improved AdaBoost	40	93.04	96.15	89	92.31	94.03
Traditional AdaBoost (BP)		86.52	93.07	78	85.58	88.87
Traditional AdaBoost (LR)		87.39	92.30	81	86.60	89.15
Improved AdaBoost	50	88.69	91.53	85	89.76	90.31
Traditional AdaBoost (BP)		87.82	90.76	84	88.73	89.52
Traditional AdaBoost (LR)		89.13	90.00	88	90.93	90.29
Improved AdaBoost	60	86.95	88.46	85	89.18	88.38
Traditional AdaBoost (BP)		84.34	86.92	81	87.01	86.32
Traditional AdaBoost (LR)		85.21	93.84	74	83.22	87.26
Improved AdaBoost	70	89.13	90.00	88	91.05	90.22
Traditional AdaBoost (BP)		81.73	80.00	84	87.52	82.52
Traditional AdaBoost (LR)		90.00	91.53	88	91.29	91.13
Improved AdaBoost	80	86.95	86.92	87	89.91	88.14
Traditional AdaBoost (BP)		82.17	83.07	81	85.22	83.81
Traditional AdaBoost (LR)		85.21	89.23	80	86.40	87.21
Improved AdaBoost	90	85.65	86.92	84	88.35	87.22
Traditional AdaBoost (BP)		85.65	80.00	93	93.90	86.28
Traditional AdaBoost (LR)		83.91	86.15	81	86.49	86.00
Improved AdaBoost	100	83.04	83.07	83	87.36	84.50
Traditional AdaBoost (BP)		82.17	82.30	82	85.82	83.68
Traditional AdaBoost (LR)		87.39	90.76	83	88.56	89.17

图6进一步给出了更多维度下分类器对测试样本的预测错误率。由图6可以看出,随着特征维度的增加,分类器的性能有所改善,错误率逐渐下降。当特征维度达到40时,分类器的错误率达到最低值,仅为6.96%,F_1-score指标也达到最大;特征维度在50至100之间时,分类器的错误率和F_1-score相对变化缓慢,但性能有所下降;而随着特征维度的继续增加,分类器的性能虽有波动,但始终没有更优于在特征维度为40时的分类器性能,而计算时间和数据量却大大增加。综合来看,对于此次研究,将特征优化为40维度是比较合理的选择。

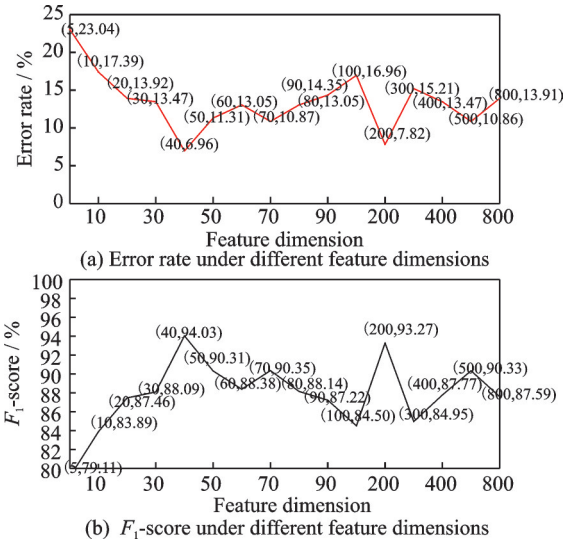

图6 分类器的错误率和F_1-score随特征维度的变化

Fig.6 Change of error rate and F_1-score of classifiers with feature dimension

此外,评价一个分类器的好坏,更多的还要看其泛化性能,而分类器的性能不仅与特征的选择有关,本质上也与训练样本有关。基于不同的训练样本,可得到不同的分类器。为了满足分类器泛化性高的目的,希望找到适用于所有潜在样本的共性特征,并尽量避免过拟合和欠拟合的情况发生。本文提出的改进算法,旨在通过将全部训练集依次送入训练器得到子分类器,避免样本选择造成的过拟合或欠拟合,提升分类器的泛化性能和鲁棒性。为了验证基于本文提出的改进型AdaBoost算法构建的分类器是否具有良好的泛化性能,文中进一步将样本数据随机划分100次,每次以其中80%的样本作为训练集,构建分类器,利用剩余的20%的样本,测试分类模型的性能。每次样本划分随机且独立,因此每次的训练样本和测试样本都将不同,以此来模拟用不同的训练样本构建分类器,并预测不同的测试样本。图7所示在特征维度优化为40后,随机100次测试中分类器的各项性能指标统计情况。从图中看出,在100次相互独立的测试中,改进型AdaBoost分类器的性能存在一定的波动变化,其中,综合指标F_1-score在100次测试中的波动最小,仅为6.59%;特异性的变异系数最大,为13.55%,但也在小于

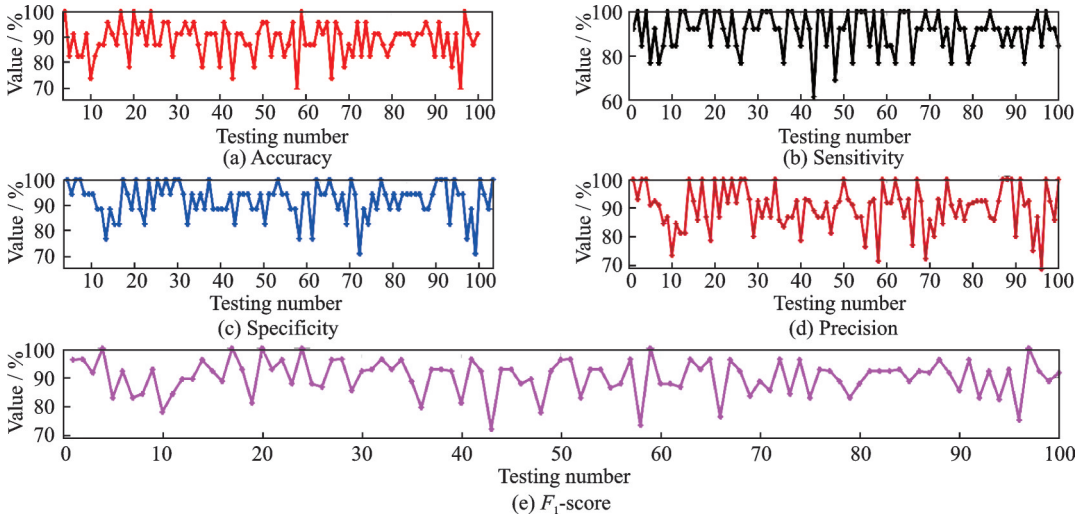

图7 100次随机测试中分类器的性能统计

Fig.7 Performance statistics of classifier in 100 random tests

15%范围内[28]。在100次测试中,5个性能指标数据均在正常波动范围内,由此说明该分类器具有一定的稳定性,改进型AdaBoost算法的鲁棒性和泛化能力较好。

应用本文算法构建肝癌患者鉴别模型,尚有很多因素未充分考虑,如肝癌患者的用药情况、肝癌患者的病程等,另外由于数据量的有限,也将影响分类准确率分类器的泛化性能。在后期的研究中,将进一步考虑这些因素,并尝试探索更适合的特征选择方法,以真正获取潜在于各样本中的共性特征,并构建适合的分类器,得到更稳定的分类结果。

4 结束语

电子鼻是近年来一种新型发展的诊断仪器,可应用于食品检测、环境监测和医疗诊断等领域。该设备通过采集人体呼气中的挥发性气体(VOCs),监测和诊断人体疾病,具有无创、操作简单及检查费用低廉等优点。但由于人体代谢的复杂性和疾病的多样性,以及不同训练样本对分类器性能的影响,电子鼻在疾病检测中一直没有得到广泛的临床应用。有效区分肝癌患者与健康人的呼气数据,是电子鼻诊断肝癌患者的核心工作[29]。

电子鼻模式识别单元主要依靠采集大量的肝癌患者和健康人的呼气数据,探索一种合适的智能算法对呼气数据进行降维有助于提高电子鼻肝癌检测的识别率。集成学习的核心是通过一定的规则生成多个学习器,再采用某种集成策略进行组合,最后综合判断输出的最终结果。本文基于集成强化学习理论,融合Stacking模型和Bagging模型的集成方式,构建多个基分类器组,作为异质子分类器[30],改进AdaBoost算法。同时,由于该算法采用K折交叉划分的方法,依次用不同的数据样本训练出若干基分类器,在一定程度上提升了分类器的鲁棒性和泛化性能,有利于提升肝癌呼气信号鉴别的性能。但本文研究也发现,分类器的性能与特征及其维度有很大关系。为了得到稳定的肝癌呼气鉴别结果,还需进一步在肝癌呼气特征信号的采集、信号的特征提取和优化及多中心样本数据获取等方面做深入的研究。

参考文献:

[1] FERLAY J, COLOMBET M, SOERJOMATARAM I, et al. Cancer statistics for the year 2020: An overview[J]. International Journal of Cancer, 2021, 149(2): 778-789.

[2] 袁放, 马青松, 宋彬. 肝癌的影像学研究进展及其在多学科诊疗中的应用[J]. 中国普外基础与临床杂志, 2021, 28(3): 297-302.
YUAN Fang, MA Qingsong, SONG Bin. Progress in imaging studies of hepatocellular carcinoma and its application in multidisciplinary diagnosis and treatment[J].Chinese Journal of Bases and Clinics in General Surgery, 2021, 28(3): 297-302.

[3] HAGSTRM H, THIELE M, SHARMA R, et al. Risk of cancer in biopsy-proven alcohol-related liver disease: A population-based cohort study of 3410 persons[J]. Clinical Gastroenterology and Hepatology, 2022, 20: 910-929.

[4] 焦红彬.AFP在原发性肝癌不同因素分层诊断中的作用研究[D].唐山:华北理工大学,2020.
JIAO Hongbin. Role of AFP in stratified diagnosis of different factors of primary hepatic cancer[D]. Tangshan: North China University of Science and Technology, 2020.

[5] SONG Peipei, XIA Jufeng, INAGAKI Y. Controversies regarding and perspectives on clinical utility of biomarkers in hepatocellular carcinoma[J]. World Journal of Gastroenterology, 2016, 22(1): 262-274.

[6] MAZZONE P J. Analysis of volatile organic compounds in the exhaled breath for the diagnosis of lung cancer[J]. Journal of Thoracic Oncology, 2008, 3(7): 774-780.

[7] OAKLEY-GIRVAN I, DAVIS S W. Breath based volatile organic compounds in the detection of breast, lung, and colorectal cancers: A systematic review[J]. Cancer Biomarkers, 2017, 21(1): 29-39.

[8] KE Yan, ZHANG David. A novel breath analysis system for diabetes diagnosis[J]. IEEE, 2012: 166-170.

[9] GERMANESE D, COLANTONIO S, D ACUNTO M, et al. An e-nose for the monitoring of severe liver impairment: A

preliminary study[J]. Sensors, 2019,19(17): 3656.

[10] KITIYAKARA T, REDMOND S, UNWANATHAM N, et al. The detection of hepatocellular carcinoma (HCC) from patients' breath using canine scent detection: Proof-of-concept study[J]. Hepatology,2016,64(1): 222A-223A.

[11] 秦涛.肝癌患者呼气中挥发性标志物的定量分析研究与呼气诊断函数模型的建立[D].合肥:安徽医科大学,2009.
QIN Tao. Quantitative analysis of volatile markers in breath of hepatocellular carcinoma patients and the establishment of the breath diagnostic function models of the cancer[D]. Hefei: Anhui Medical University,2009.

[12] WIJAYA D R, AFIANTI F, ARIFIANTO A, et al. Ensemble machine learning approach for electronic nose signal processing[J]. Sensing and Bio-Sensing Research, 2022,36: 100495.

[13] KRESNAWATY I, MULYATNI A S, ERIS D D, et al. Electronic nose for early detection of basal stem rot caused by Ganoderma in oil palm[J]. IOP Conference Series Earth and Environmental Science, 2020,468: 012029.

[14] HENDRICK H, HIDAYAT R, HORNG G J, et al. Non-invasive method for tuberculosis exhaled breath classification using electronic nose[J]. IEEE Sensors Journal, 2021, 21(9): 11184-11191.

[15] 刘蕊.基于Stacking集成学习算法的骨龄自动评估研究[D].重庆:重庆医科大学,2020.
LIU Rui. The research on automatic bone age assessment based on ensemble learning of stacking[D]. Chongqing: Chongqing Medical University,2020.

[16] 赵乐,麦范金,张兴旺.多特征融合的Voting-SRM情感分类研究[J].小型微型计算机系统,2019,40(11): 2269-2273.
ZHAO Le, MAI Fanjin, ZHANG Xingwang. Voting-SRM sentiment classification based on multi-feature fusion[J]. Journal of Chinese Mini-Micro Computer Systems,2019,40(11): 2269-2273.

[17] AN T K, KIM M H. A new diverse AdaBoost classifier[C]//Proceedings of 2010 International Conference on Artificial Intelligence & Computational Intelligence. [S.l.]: [s.n.], 2010:359-363.

[18] 李朋飞,于洪.基于属性约简的自采样集成分类方法[J].数据采集与处理,2021,36(3): 498-508.
LI Pengfei, YU Hong. Self-sampling ensemble classification method based on attribute reduction[J]. Journal of Data Acquisition and Processing,2021,36(3): 498-508.

[19] YANG D H, LEE H J, LIM D J. Rolexboost: A rotation-based boosting algorithm with adaptive loss function[J]. IEEE Access, 2020, 8: 41037-41044.

[20] 郝丽俊,黄钢.基于电子鼻的呼气无创肝癌检测方法研究[J].传感器与微系统,2020,39(4): 46-48.
HAO Lijun, HUANG Gang. Non-invasive detection of liver cancer by respiratory based on electronic nose[J]. Transducer and Microsystem Technologies, 2020,39(4): 46-48.

[21] HAO L J, ZHANG M, HUANG G. Feature optimization of exhaled breath signals based on pearson-BPSO[J]. Mobile Information Systems, 2021, 2021(7): 1-9.

[22] GHOSH P, AZAM S, JONKMAN M, et al. Efficient prediction of cardiovascular disease using machine learning algorithms with Relief and LASSO feature selection techniques[J]. IEEE Access, 2021, 9: 19304-19326.

[23] GUPTA V, MITTAL M. KNN and PCA classifier with autoregressive modelling during different ECG signal interpretation[J]. Procedia Computer Science, 2018, 125: 18-24.

[24] GERHARDT N, SCHWOLOW S, ROHN S, et al. Quality assessment of olive oils based on temperature-ramped HS-GC-IMS and sensory evaluation: Comparison of different processing approaches by LDA, KNN, and SVM[J]. Food Chemistry, 2019, 278(25): 720-728.

[25] YOO W, FERENCE B A, COTE M L, et al. A comparison of logistic regression, logic regression, classification tree, and random forests to identify effective gene-gene and gene-environmental interactions[J]. International Journal of Applied Science and Technology, 2012, 2(7): 268-280.

[26] MAHABUB A. A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers[J]. SN Applied Sciences, 2020, 2(4): 1-9.

[27] 秦锋,杨波,程泽凯.分类器性能评价标准研究[J].计算机技术与发展,2006,16(10): 85-88.
QIN Feng, YANG Bo, CHENG Zekai. Research on measure criteria in evaluating classification performance[J]. Computer Technology and Development,2006,16(10): 85-88.

[28] 李玲玲,宋玉芳,史奕.赤子爱胜蚓(Eisenia fetid)EROD活性测定的HPLC法的建立[J].沈阳大学学报:自然科学版,2021,

33(1): 33-40.
LI Lingling, SONG Yufang, SHI Yi. Establishment of EROD activity determination in earthworm (eisenia fetida) by HPLC[J]. Journal of Shenyang University(Natural Science), 2021, 33(1): 33-40.

[29] ANDREAS V, VICO B, REMATE R, et al. Smelling renal dysfunction via electronic nose[J]. Annal of Biomedical Engineering, 2005, 33(5): 656-660.

[30] 韩亮,杨婷,蒲秀娟,等. 用于阿尔茨海默症分类的模糊逻辑特征选择和异质集成学习方法[J]. 电子与信息学报, 2021, 43(11): 3319-3326.
HAN Liang, YANG Ting, PU Xiujuan, et al. Method on Alzheimer's disease classification utilizing fuzzy logic feature selection and heterogeneous ensemble learning[J]. Journal of Electronics & Information Technology, 2021, 43(11): 3319-3326.

作者简介:

郝丽俊(1981-),女,讲师,研究方向:医学信号处理及医学智能化,E-mail:sunnyshm@163.com。

黄钢(1961-),**通信作者**,男,教授,主任医师,研究方向:分子探针、肿瘤影像组学、生物医学工程,E-mail: huanggang@sumhs.edu.cn。

(编辑:王静)