ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 4,Jul. 2022,pp. 860－871
DOI:10. 16337/j. 1004-9037. 2022. 04. 014

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

基于启发式集成特征选择的人体活动识别

戴健威，李瑞祥，陈金瑶，乐燕芬，施伟斌

（上海理工大学光电信息与计算机工程学院，上海 200093）

摘　要：针对人为提取的冗余特征集和无关特征集导致可穿戴传感器的人体活动识别分类性能降低的问题，提出一种基于启发式集成特征选择的人体活动识别方法。该方法首先选取了包含功率谱密度（Power spectrum density，PSD）的特征集用于识别易混淆的活动，在此基础上借助皮尔逊系数法（Pearson correlation coefficient，PCC）筛选出低相关的特征子集，然后使用改进的正余弦优化算法（Sine cosine algorithm，SCA）进行特征优化，通过两次特征筛选得到最优特征子集。实验结果表明，在实验室采集的数据集中使用该方法后的特征子集维数为34，识别准确率达到了98.21%。在公开的SCUT-NAA数据集中进行对比实验，特征子集维数为39，低于以往基于该数据集研究方法的特征维数，并且识别准确率达到了96.51%。

关键词：人体活动识别；特征选择；正余弦算法；功率谱密度；可穿戴传感器

中图分类号：TP391　　**文献标志码**：A

Human Activity Recognition Based on Heuristic Integrated Feature Selection

DAI Jianwei，LI Ruixiang，CHEN Jinyao，LE Yanfen，SHI Weibin

（School of Optical-Electrical and Computer Engineering，University of Shanghai for Science and Technology，Shanghai 200093，China）

Abstract：To address the problem that artificially extracted redundant feature sets and irrelevant feature sets lead to the degradation of human activity recognition classification performance of wearable sensor，this paper proposes a human activity recognition method based on heuristic integrated feature selection. The method first selects the feature set containing power spectral density（PSD）for recognizing confusing activities. Then，on this basis，the method screens out the lowly correlated feature subsets with the help of Pearson correlation coefficient（PCC）method，then uses an improved sine cosine algorithm（SCA）for features and obtains the optimal feature subset by screening the feature twice. The experimental results show that the feature subset dimension after using this method in the data set collected in the laboratory is 34，and the recognition accuracy rate reaches 98.21%. In the public SCUT-NAA data set for comparison experiments，the feature subset dimension is 39，lower than the feature dimension of previous research methods，and the recognition accuracy rate reaches 96.51%.

Key words: human activity recognition(HAR); feature selection; sine cosine algorithm(SCA); power spectrum density(PSD); wearable sensor

基金项目：国家自然科学基金（51705324）。
收稿日期：2021-10-09；**修订日期**：2022-01-02

引 言

人体活动识别(Human activity recognition, HAR)是一种典型的模式识别问题,近年来已经成为非常活跃的研究领域。HAR应用于日常生活的方方面面,如医疗监控[1]、人机交互[2]、智能家居[3]等领域。如今,在老人步态分析、儿童行为检测、手臂运动检测、跌倒行为检测[4]等特定检测领域也广受关注,具有广泛的研究和应用价值。

从HAR数据获取的类型来看,研究方法主要分为两类:基于计算机视觉[5]和基于可穿戴传感器[6]的人体活动识别方法。随着嵌入式系统的快速发展,基于可穿戴传感器的HAR在不用获取大量视频图像的情况下,利用传感器获取的数据信息对活动进行分类,既保护了用户的隐私,又降低了计算复杂度,扩展了使用场景[7]。国内外研究者对基于可穿戴传感器的HAR进行了广泛的研究,大多数研究工作仍停留在人工的特征工程结合传统的机器学习算法[8]。人工提取的时、频域特征虽然足以识别基本类别的身体活动或人体姿态,但是面对复杂类别的活动,如何从原始传感器数据中提取鲁棒性强的特征是研究工作的难点。目前深度学习技术可以自动提取特征且在上位机拥有良好的识别性能,但深度学习算法的计算成本较高且需要大量训练数据支撑,不适合实时计算[9]。由此可见,基于可穿戴传感器的HAR仍有许多问题亟待解决,以提高在现实条件下的识别稳定性和准确性。其中一些挑战是[10]:(1)在现实条件下构建一个更便捷的数据采集系统;(2)活动信号的预处理与有效特征的提取;(3)特征选择的优化问题;(4)设计高精度的活动识别系统。可以看出,HAR的性能根据每个阶段使用方法的差异而有所不同。

特征选择是提高HAR性能的一个重要研究方向[11],特征选择的主要目的是降低特征集的空间维数,以提高算法的分类效率并保证分类的准确性。近年来诸多学者开展了与特征选择相关的HAR研究,例如Mohd等[12]对比了随机森林(Random forest, RF)、K近邻(K-nearest neighbor, KNN)、长短期记忆网络(Long short-term memory, LSTM)、支持向量机(Support vector machine, SVM)和多层感知机(Multilayer perceptron, MLP)分类器对人体活动识别分类性能的影响,采用RF和主成分分析(Principal component analysis, PCA)相结合的方法,取得了87.5%的识别精度。但是PCA适用于特征变量之间具有较强相关性的数据,若原始特征变量之间相关性较弱,则不能起到很好的降维作用。相比较之下,Lu等[13]利用S变换来提取特征,然后引入了一种基于监督正则化的鲁棒子空间学习方法,将原始特征空间映射到低维特征空间表示,取得了94.0%的识别精度,该方法虽然可以从原始特征中提取判别特征以提高分类性能,但不能有效筛选并剔除贡献度小的特征。Chanvichet等[14]将卡方滤波、互信息滤波、方差分析滤波和皮尔逊系数滤波的4种特征选择方法进行了研究和比较,提出了结合互信息滤波的极端梯度增强算法(Extreme gradient boosting, XGBoost)的方法,取得了91.22%的识别精度,该方法可以快速地得出特征选择结果,但是未考虑特征之间的相关性,因此可能存在大量线性冗余。从上述HAR特征选择的研究来看,主要是基于过滤式和包裹式的特征选择,无法同时满足降低特征空间维数和保持算法的识别准确率。

基于以上分析,本文提出一种基于启发式集成特征选择的人体活动识别方法。首先,该方法针对相似活动易混淆问题,提取了包含功率谱密度的特征集。其次,为了提高特征选择的有效性,采用皮尔逊系数法将提取的特征进行重要性排序,剔除重要性低的特征。最后,将筛选后的特征使用改进的正余弦优化算法结合本文设置的权重系数进一步进行特征选择,筛选后的特征子集能够准确地判断出人体活动类别。本文与前人研究的不同之处在于提出了一种过滤式与包装式特征选择相结合的方法,同时兼顾了特征维数和识别准确率两方面的考虑。

1 启发式集成特征选择的人体活动识别模型

1.1 模型结构

以往基于可穿戴传感器的人体活动识别研究中，一般的实验步骤为数据采集与预处理、特征提取、特征选择、训练分类器模型、模型评估等[10]。为了有效去除人工选取的冗余特征，同时保证分类模型的准确性，本文提出了一种基于启发式集成特征选择的人体活动识别方法，算法的总体流程如图1所示。

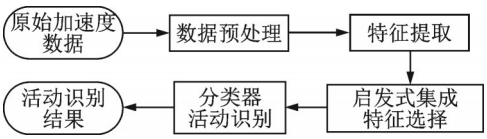

图1 启发式集成特征选择的人体活动识别流程图

Fig.1 Human activity flow chart of heuristic integrated feature selection

1.2 数据采集与预处理

在数据采集过程中，需要保证传感器的佩戴位置和方向是固定的，以避免传感器的不平衡带来数据质量的下降。研究发现[15]，当佩戴单个传感器位于腰间或胫骨位置时，活动分类表现较胸部、前臂、头部和大腿等位置时更优。因此，本文设计了可穿戴惯性传感器模块加开发板的硬件系统，佩戴在受试者的腰间用来采集人体活动的信号数据。数据采集的硬件结构如图2所示，传感器采集模块主要采用STM32F4超低功耗微控制器、MPU6050六轴惯性传感器和存储模块组成。数据采集过程由MPU6050传感器通过I2C接口将采集到的加速度计运动数据传输到STM32F4的闪存中，并以逗号分隔的格式存储，以供进一步离线的模型训练和分类。

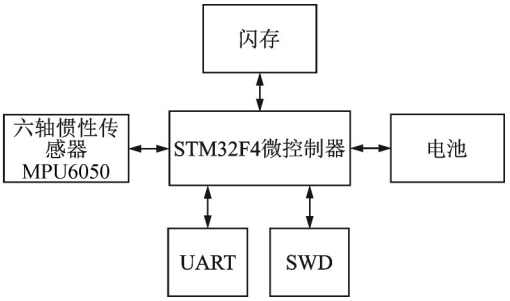

图2 数据采集硬件结构图

Fig.2 Data acquisition hardware structure diagram

利用上述硬件采集系统，本文自采集的数据集共收集了8名志愿者(性别：5名男生和3名女生。年龄：22～28岁。体重：45～85 kg。身高：155～190 cm)的6种日常活动数据(躺、站立、行走、倒走、上楼和下楼)，每名志愿者的6种活动各有10 240个样本。本实验中MPU6050传感器的采样频率设置为50 Hz。每名志愿者需要对每种类型的活动执行两次，每次持续时间为3 min。结合前人的研究工作来看[16]，窗口间隔2 s左右为识别速度和准确率的最佳平衡。因此，结合本实验室采集的数据进行预处理，采用窗口大小为2.56 s的滑动窗口，进行50%重合的数据分割。使用两个巴特沃斯滤波器，分别对高于25 Hz和低于0.2 Hz的噪声数据进行滤波。

考虑到所提出方法的普适性，采用公开的SCUT-NAA数据集进行对比实验，该数据集包含44名志愿者(34名男性和10名女性)的10种日常活动，包含骑车、上下楼、跳跃、行走等多类活动，每名志愿者的10种活动各有12 667个样本。加速度计的采样频率为100 Hz，采用窗口大小为2 s以及50%重合的数据分割。考虑到识别系统的泛化性能，数据集均采用70%的数据作为训练集，30%的数据作为测试集。

1.3 特征提取

原始加速度信号可以作为分类算法的输入，但是在这种情况下，活动识别的准确率就会不尽如人意。因此，从加速度信号中提取有效特征是至关重要的。本文提取了加速度数据的时域特征和频域特征两大类，如常见的时、频域特征统计量[17]：最大值(MAX)、最小值(MIN)、均方根(RMS)和绝对中位

差(MAD)等。表1列出了本文提取的9种、共51维的活动特征统计量,以构成人体活动识别的特征集。

上述特征都是基于统计性质来度量的,对于统计性质相似的特征,其识别性能未能达到令人满意的效果。例如,上、下楼梯活动具有相似的均值和方差。功率谱密度(Power spectrum density, PSD)[18]能够反映随机信号振动的功率对于频率的分布密度,它显示了作为频率函数的能量强度。功率谱密度分析可用来研究随机振动信号在单位频带内的功率分布情况,有助于活动特性的模拟分析。基于功率谱密度的这些特点,本文在特征集中加入了功率谱密度这一特征。

表1 特征提取列表

Table 1 Feature extraction list

类型	特征	维数
时域+频域	最大值(MAX)	6
时域+频域	最小值(MIN)	6
时域+频域	均方根(RMS)	6
时域+频域	绝对中位差(MAD)	6
时域+频域	自相关系数(AC)	6
时域+频域	标准差(STD)	6
时域+频域	四分位距(IQR)	6
时域+频域	过零率(ZCR)	6
仅频域	功率谱密度(PSD)	3

1.4 启发式集成特征选择

在特征提取之后,最重要的工作是选择有效特征和去除冗余特征。特征选择的普遍做法是选取包含所有重要信息的特征子集,由于学习任务不同,特征选择的方法不是固定不变的。因此,本文针对可穿戴传感器的人体活动识别领域提出了一种启发式集成特征选择算法,以适应在不同特征提取及分类器下的识别任务。

1.4.1 皮尔逊系数法

PCC通常用来衡量两个变量之间相关性的大小,它决定了两个变量之间线性相关的强度。两组随机变量$\boldsymbol{X}$:$\{X_1, X_2, \cdots, X_n\}$和$\boldsymbol{Y}$:$\{Y_1, Y_2, \cdots, Y_n\}$的相关性可以表示为它们的协方差除以它们标准差的乘积

$$P_{XY}=\frac{\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})}{\sigma_X\sigma_Y} \tag{1}$$

式中,总体协方差表示为

$$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=\frac{\sum_{i=1}^{n}\left(\boldsymbol{X}_i-E(\boldsymbol{X})\right)\left(Y_i-E(\boldsymbol{Y})\right)}{n} \tag{2}$$

总体均值表示为

$$E(\boldsymbol{X})=\frac{\sum_{i=1}^{n}X_i}{n} \tag{3}$$

$$E(\boldsymbol{Y})=\frac{\sum_{i=1}^{n}Y_i}{n} \tag{4}$$

标准差表示为

$$\sigma_X=\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-E(\boldsymbol{X})\right)^2}{n}} \tag{5}$$

$$\sigma_Y=\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i-E(\boldsymbol{Y})\right)^2}{n}} \tag{6}$$

当P_{XY}为0时，表示对应变量之间没有相关性，当P_{XY}为1时，表示对应变量之间完全正相关，当P_{XY}为－1时，表示对应变量之间完全负相关。

皮尔逊系数法作为初步筛选的目标是获得一组特征之间彼此低相关的特征子集，则该特征子集存在大量线性冗余的可能性大大降低，并且对于后续算法的分类性能有着重大影响。本文设置的皮尔逊系数法特征筛选的得分方程为

$$\mathrm{score}(\boldsymbol{X})=-\sum_{1}^{\dim}P_{XY}(\boldsymbol{X},\boldsymbol{X}_{\dim}) \tag{7}$$

式中dim为特征维数。

1.4.2 正余弦优化算法

SCA是Mirjalili[19]在2016年提出的一种基于数学性质的新型随机优化算法，SCA首先初始化多个随机候选解，然后利用正余弦函数模型，使得这些解朝着最优解方向或反向波动，利用多个随机变量和自适应变量来计算当前解的所在位置，从而可以搜索空间中的不同区域。正余弦优化算法寻优过程是根据正、余弦函数并结合随机因子来更新当前解在每一维度的值，正余弦函数模型的更新方程为

$$X_i^{t+1}=X_i^t+r_1\times\sin r_2\times\left|r_3p_i^t-X_i^t\right| \tag{8}$$

$$X_i^{t+1}=X_i^t+r_1\times\cos r_2\times\left|r_3p_i^t-X_i^t\right| \tag{9}$$

式中：X_i^t为第t次迭代时，当前解在第i维中的位置；r_1、r_2、r_3为随机数；p_i为最优解在第i维中的位置。

通常将上述两个方程组合起来使用，有

$$X_i^{t+1}=\begin{cases}X_i^t+r_1\times\sin r_2\times\left|r_3p_i^t-X_i^t\right| & r_4<0.5\\ X_i^t+r_1\times\cos r_2\times\left|r_3p_i^t-X_i^t\right| & r_4>0.5\end{cases} \tag{10}$$

$$r_1=a-t\frac{a}{T} \tag{11}$$

式中：a为一个常数，t为当前迭代次数，T为最大迭代次数，参数r_1决定了下一位置区域的移动方向，该区域可能位于当前解和最优解之内或之外；参数$r_2\in[0,2\pi]$决定了当前解朝向或者远离最优解的距离；参数$r_3\in[0,2]$为最优解的随机权重，以便随机地强调($r_3>1$)或者弱化($r_3<1$)最优解在定义候选解移动距离时的影响效果；参数$r_4\in[0,1]$之间的随机数，参数r_4用于等概率地在正余弦之间进行切换。

本文优化算法的适应度函数结合了准确率与特征维数两方面的考虑，目标是得到使识别准确率高且特征维数少的子集，考虑到这一目的，设置适应度函数方程以评估该特征子集，有

$$\mathrm{Fitness}=\alpha\times\mathrm{acc}+(1-\alpha)\times(\mathrm{num}_{\mathrm{feature}}-\mathrm{num}_{\mathrm{agent}})/\mathrm{num}_{\mathrm{feature}} \tag{12}$$

式中：$\alpha\in(0,1)$为权重值；acc为子集的分类准确率；$\mathrm{num}_{\mathrm{feature}}$为特征全集维数；$\mathrm{num}_{\mathrm{agent}}$为特征子集维数。

1.4.3 改进的正余弦优化算法

为了提高标准SCA的收敛精度和速度，本文对更新方程进行改进，由1.4.2节可以看出，r_1是正余弦算法的关键参数，影响着全局搜索和局部开发，但是标准的SCA中r_1是线性变化的，这不利于算法的全局搜索。因此，对SCA的部分参数r_1的收敛方式进行改进，采用抛物线函数作为参数r_1的更新策略

$$r_1=a\left(1-t\frac{1}{T}\right)^2 \tag{13}$$

1.4.4 启发式集成特征选择

PCC是一种过滤式的、运行效率很高的相关性特征选择方法，可以筛选出相关性较小的特征，以减少特征冗余。而改进的SCA可以针对设定的适应度函数筛选出最优的特征子集且达到较高的准确率，但单独使用计算开销较前者更大。因此结合二者的优、缺点，本文提出的启发式集成特征选择(Heuris-

tic integration feature selection, HIFS)给出了更优的组合解决方案。算法流程图如图3所示,具体的算法实现步骤描述如表2所示。

表2 HIFS算法实现步骤

Table 2 HIFS algorithm implementation steps

输入:原始特征集
输出:优化后的特征子集
算法1:皮尔逊系数法特征选择
(1) 初始化皮尔逊相关性系数矩阵:**PCC** = zeros(dim,dim);
(2) 遍历计算每个特征变量之间的PCC(式(1));
(3) 得到每个特征变量的皮尔逊系数得分,并按照得分降序排名(式(7));
(4) 设置特征数量阈值$\gamma\in[0.6,0.8]$,根据γ剔除得分排名靠后的特征。
算法2:改进正余弦优化算法的特征选择
(1) 输入皮尔逊系数法筛选后的特征,初始化(更新)种群位置及各项参数;
(2) 计算当前解的适应度函数值,保留当前最优候选解(式(12));
(3) 更新参数(r_1,r_2,r_3,r_4)及当前解X_i^t的位置(式(10));
(4) 判断是否达到设定迭代次数,不满足则返回步骤(1);
(5) 返回最优解子集。

1.4.5 算法的时间复杂度分析

假设算法的最大迭代次数为T,种群规模为N,特征维数为dim,初步剔除的特征维数为K,皮尔逊系数法的时间复杂度为$O(\dim^2)$,标准SCA的时间复杂度为$O(T\times N\times\dim^2)$,故HIFS的时间复杂度为$O(\dim^2+T\times N\times(\dim-K)^2)$,不包含常数项的HIFS时间复杂度为$O((T\times N+1)\times\dim^2-2\times K\times T\times N\times\dim)$,$2\times K\times T\times N$相对于dim的数据规模更大。由此可见,HIFS总体时间复杂度较SCA更低,提高了标准SCA的收敛速度,并且能有效剔除线性冗余特征。

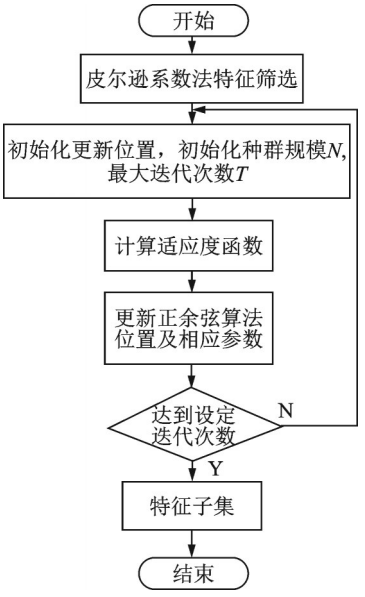

图3 启发式集成特征选择流程图

Fig.3 Flow chart of heuristic integrated feature selection

2 实验结果及分析

2.1 评价指标

本文采用人体活动识别中常用的评价指标进行定量分析,如准确率(Accuracy)、查准率(Precision)、查全率(Recall)和F_1分数,4个指标越高说明所构建系统的性能就越好。表3为真实情况和预测结果的分类混淆矩阵示例。

准确率表示每个样本被正确分类的概率,表达式为

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{14}$$

精度表示预测结果为正例中正确的概率,表达式为

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{15}$$

召回率表示正确分类的正例概率，表达式为

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{16}$$

F_1分数表示精度和召回率的调和平均数，是将二者结合为单一度量的方法，表达式为

$$F_1=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{17}$$

表3 分类结果混淆矩阵

Table 3 Confusion matrix of classification results

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

2.2 实验环境与参数设置

在本节中，采用PyCharm2020.2.3进行实验，运行环境为64位Windows10操作系统，处理器类型为Intel Core i5-8300H，机器学习框架为scikit-learn1.0.1。模型训练过程中的分类算法的超参数如表4所示。改进正余弦优化算法的共有参数设置：种群规模$N=10$，最大迭代次数$T=20$。

表4 分类算法的超参数列表

Table 4 Hyperparameter list of classification algorithm

超参数名称	超参数值	超参数名称	超参数值
RF_n_estimators	120	SVM_kernel	“rbf”
RF_min_samples_split	2	LightGbm_num_leaves	100
RF_min_samples_leaf	1	LightGbm_learning_rate	0.2
SVM_gamma	0.1	LightGbm_n_estimators	200
SVM_C	1	LightGbm_boosting_type	“gbdt”

HIFS是皮尔逊系数法和改进正余弦优化算法的组合解决方案，根据准确率的权重值α和皮尔逊特征数量的阈值γ设置组合参数。在自采集数据集和SCUT-NAA数据集中实验发现，α取值过小使得所选特征子集的维数大幅减少，但伴随着准确率的降低，γ取值过大会导致所选特征子集可能存在线性冗余，均与本文目标相反，因此α的取值范围设置为$[0.7,1)$，γ的取值范围设置为$[0.6,0.8]$。设置的权重值与数量阈值通过多次实验以获得性能更优的权重组合，最终选取$\gamma=0.8$、$\alpha=0.9$作为本实验的权重设置。表5是不同权重设置对实验结果影响的展示。

2.3 验证功率谱密度特征的有效性

为了分析功率谱密度特征给模型分类结果带来的影响，将自采集的数据集和公开的SCUT-NAA数据集[20]针对加速度计的x轴数据进行功率谱密度分析，其实验分析均采用穿戴位置位于腰间的传感器数据。

图4(a)为自采集数据集的功率谱密度图，图4(b)为SCUT-NAA数据集的功率谱密度图。谱峰的相对振幅与信号的振荡形状密切相关。从图4(a)可以看出，下楼和上楼活动的功率谱密度波形相较于躺活动变化较大；而图4(b)中，下楼和跳活动的功率谱密度波形变化较大，骑车活动的功率谱密度波形变化较小；此外，上楼活动的功率谱密度比下楼的功率谱密度小，这意味着上楼活动比下楼活动更平稳。

针对人体活动易混淆问题，对加入功率谱密度特征前后的活动识别性能进行对比实验，图5(a)表示自采集数据集中未加入功率谱密度特征的识别混淆矩阵，图5(c)表示SCUT-NAA数据集中未加入功率谱密度特征的识别混淆矩阵，横坐标表示预测类别，纵坐标表示真实类别。从图5可以看出，加入功率谱密度特征后基于自采集数据集和SCUT-NAA数据集总体平均识别率分别由97.1%提升到98.2%、95.1%提升到96.5%。并且对于SCUT-NAA数据集中上楼和下楼易混淆问题，上、下楼活动识别率分别提升3.3%和1.3%，这就验证了针对活动易混淆问题功率谱密度特征的有效性。

表5 不同权重设置对实验结果的影响

Table 5 Influence of different weight settings on experimental results

γ	α	自采集数据集					SCUT-NAA数据集				
		准确率/%	查准率/%	查全率/%	F_1	特征维数	准确率/%	查准率/%	查全率/%	F_1	特征维数
0.6	0.70	96.89	96.91	96.91	96.91	**33**	94.32	94.32	94.32	94.32	**36**
0.6	0.80	97.16	97.16	97.16	97.16	35	94.41	94.43	94.41	94.41	38
0.6	0.90	97.30	97.28	97.28	97.29	36	94.55	94.56	94.56	94.56	39
0.6	0.99	97.28	97.28	97.27	97.28	39	94.51	94.51	94.51	94.52	40
0.7	0.70	97.02	97.01	97.01	97.01	39	94.49	94.49	94.49	94.49	38
0.7	0.80	97.41	97.41	97.41	97.41	41	94.73	94.74	97.74	94.74	39
0.7	0.90	97.53	97.52	97.53	97.53	38	95.02	95.03	95.03	95.02	40
0.7	0.99	98.14	98.14	98.14	98.14	42	94.91	94.91	94.92	94.91	42
0.8	0.70	97.90	97.93	97.91	97.91	33	94.77	94.77	94.79	94.78	38
0.8	0.80	97.40	97.40	97.40	97.40	35	95.63	95.60	95.62	95.63	39
0.8	**0.90**	**98.21**	**98.21**	**98.21**	**98.22**	34	**96.51**	**96.51**	**96.53**	**96.51**	39
0.8	0.99	98.19	98.19	98.21	98.19	37	96.48	96.47	96.48	96.48	43

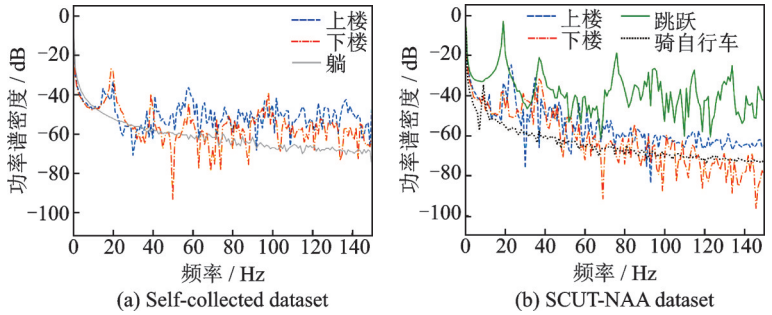

图4 自采集数据集和SCUT-NAA数据集的功率谱密度图

Fig.4 Power spectrum density diagram of self-collected dataset and SCUT-NAA dataset

2.4 HIFS的有效性验证

2.4.1 常见特征选择方法的性能对比

特征的初步筛选对于启发式特征选择尤为重要,本文选取了机器学习中常见的过滤式特征选择方法进行对比实验。从图6可以看出,模型的分类性能在特征维数增加的初期以较快速度提高,当特征维数超过了某一阈值,模型性能将不再升高,甚至逐渐下降;Relief特征选择的总体性能不佳;PCA在低维阶段筛选特征时有较好的表现,但在高维阶段筛选特征时,识别性能略低于系数法特征选择;斯皮尔曼系数法(Spearman's correlation coefficient, SCC)与PCC在特征选择的识别性能上相似,然而二者计算变量相关性程度的方法不同,在计算效率上PCC优于SCC。因此,本文选择PCC作为HIFS的初步筛选。

2.4.2 与其他优化算法的性能对比

将HIFS与其他一些新的启发式优化算法——正余弦优化算法、粒子群优化算法(Particle swarm optimization, PSO)[21]、布谷鸟搜索算法(Cuckoo search, CS)[22]、均衡优化器(Equilibrium optimizer, EO)[23]进行性能对比。表6是在自采集数据集和SCUT-NAA数据集中应用不同优化算法的测试结果,从准确率、特征维数和运行时间3个方面对不同优化算法进行评价。从表6可知,HIFS在自采集数据集和SUCT-NAA数据集中

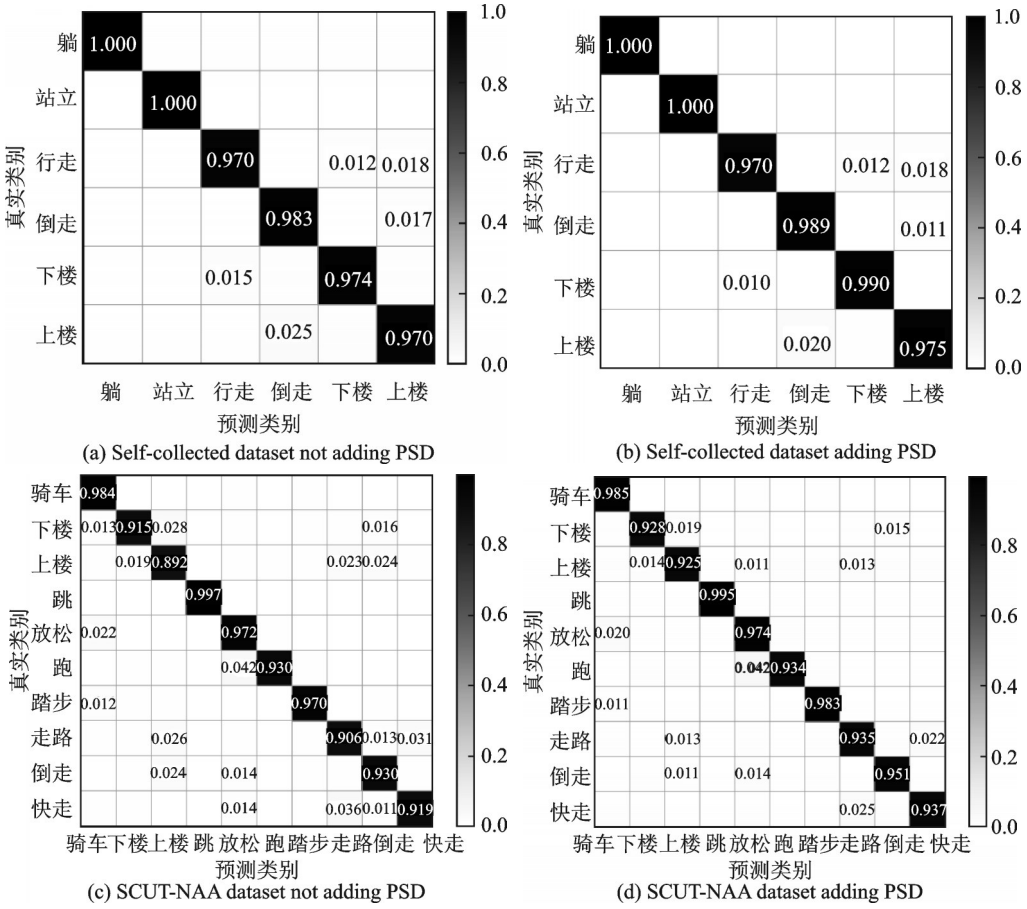

图5 功率谱密度特征对人体活动识别的影响

Fig.5 Influence of PSD on human activity recognition

相较于其他优化算法都有最高的准确率,并且优化过程的计算开销最低。另外,HIFS在自采集数据集结合RF算法和在SCUT-NAA数据集结合轻量梯度提升机(Light gradient boosting machine,LightGBM)算法得到的特征维数最少。例如,将本文提出的方法应用于LightGBM分类算法在自采集数据集中识别准确率达到了98.21%,在SCUT-NAA数据集中识别准确达到了96.51%,寻优精度高于对比的其他4种优化算法,并且寻优的计算开销最低,证明了HIFS具有较好的寻优精度与收敛速度。

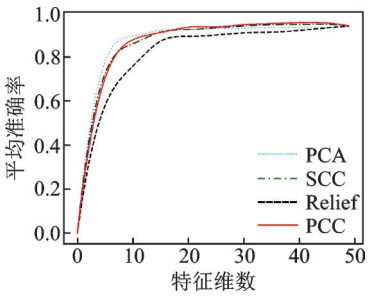

图6 常见特征选择方法性能对比

Fig.6 Performance comparison of common feature selection methods

由式(11)可知,适应度函数值越大表示特征选择的效果越优,即识别准确率高、特征维数少。图7是在自采集数据集和SCUT-NAA数据集上使用HIFS和不同智能优化算法的适应度函数收敛曲线图。从图7(a,b)中可以看出,CS算法和EO算法在两个数据集中适度函数值分别达到0.987和0.962、0.980和0.966,但二者在寻优过程中最优解的范围相对集中,说明这两种算法在本实验中的全局探索能力不强,在迭代20次的情况下未能找到全局最优解;PSO算法相较于

表6 不同优化算法测试结果

Table 6 Test results of different optimization algorithms

分类算法	特征维数	优化算法	自采集数据集				SCUT-NAA数据集			
			准确率/%	优化后特征维数	优化后准确率/%	优化运行时间/s	准确率/%	优化后特征维数	优化后准确率/%	优化运行时间/s
RF	51	HIFS	96.61	38	98.08	238.8	93.35	38	96.03	382.5
		SCA		42	97.86	357.5		41	95.87	492.2
		PSO		43	97.68	325.1		39	95.01	436.5
		CS		37	97.56	444.5		37	94.93	616.4
		EO		38	97.95	340.6		39	95.61	472.4
SVM	51	HIFS	92.31	38	94.28	132.1	92.21	36	94.93	256.3
		SCA		40	94.21	176		41	94.31	301.2
		PSO		37	94.15	135.1		38	93.97	284.5
		CS		37	92.3	211.6		35	92.71	463.3
		EO		36	93.63	145.3		37	93.16	296.1
LightGBM	51	HIFS	96.92	34	98.21	290.8	94.69	39	96.51	647.4
		SCA		40	98.15	307.4		40	95.63	680.4
		PSO		30	97.69	305.3		39	95.55	669.2
		CS		45	98.2	466.7		41	95.21	792.8
		EO		36	98.2	312.5		42	95.52	657.9

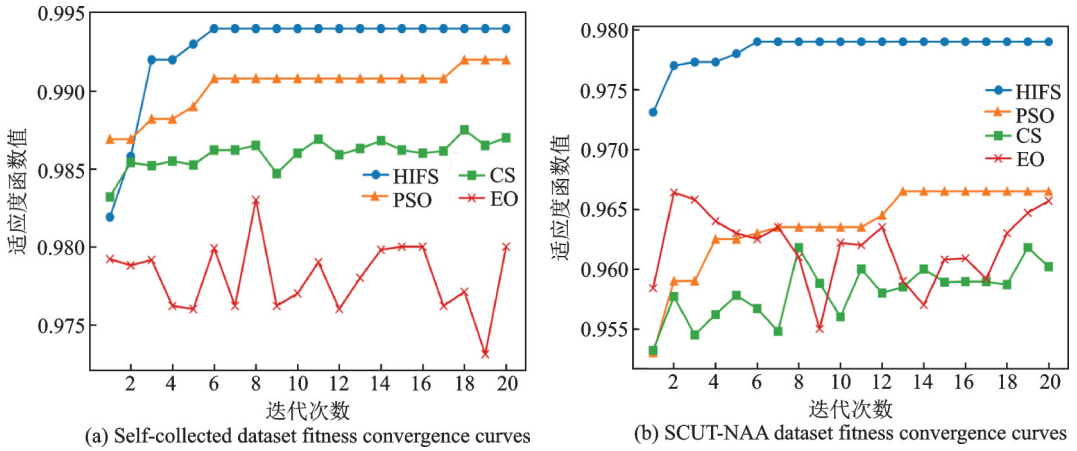

图7 不同群智能算法的适应度函数收敛曲线图

Fig.7 Convergence curves of fitness functions for different swarm intelligence algorithms

CS算法和EO算法有较强的全局探索能力,随着迭代次数增加逐渐找到全局最优解,寻优效果较好;直观可见,HIFS相较于其他智能优化算法具有更快的寻优速度,在迭代前期已经达到了较高的适应度值,并且收敛精度也高于其他智能优化算法,验证了本文方法的有效性。

2.4.3 与基于SCUT-NAA数据集前人工作的比较

在活动识别领域公开的SCUT-NAA数据集上,将本文的方法与前人的工作方法相比较,从表7可以看出,本文提出的基于启发式集成特征选择的方法在SCUT-NAA数据集上达到了比前人更优的性能,验证了本方法的有效性。

表7 基于SCUT-NAA数据集的研究对比

Table 7 Research comparison of SCUT-NAA dataset

研究文献	特征维数	分类算法	准确率/%
文献[24]	43	KNN	89.1
文献[25]	91	SVM	91.2
文献[13]		SVM	94.0
文献[26]	210	RF	93.0
本文	39	LightGBM	96.5

3 结束语

为了去除冗余特征,筛选出最优的特征子集,提高人体活动识别的分类性能,本文提出了一种基于启发式集成特征选择的方法。首先,对于易混淆活动的识别问题,提取了包含功率谱密度的特征集。然后,通过皮尔逊系数法进行初步特征筛选,将初步筛选后的特征作为输入,使用改进的正余弦优化算法结合分类算法进一步特征筛选。最终,发现本文提出的方法在对特征选择后的特征子集与原始特征集相比维数有明显的降低,并且保证了较高的识别准确率。

尽管本文针对基于可穿戴传感器的人体活动识别中特征选择的设计取得了一些进步,但是识别系统很大程度上依托于性能优异的机器学习分类算法。下一步将继续探索更先进的分类算法与本方法相结合,以提高识别系统准确率和计算效率,将筛选后的特征子集应用于实时检测人体活动中,以满足实际应用的需求。

参考文献:

[1] SHRIVASTAVA R, PANDEY M. Human fall detection using efficient kernel and eccentric approach[J]. International Journal of E-Health and Medical Communications, 2021, 12(1): 62-80.

[2] AHMAD J, MARIA M, ABDUL S. Multi-features descriptors for human activity tracking and recognition in indoor-outdoor environments[C]// Proceedings of International Conference on Applied Sciences and Technology. Islamabad: IEEE, 2019: 371-376.

[3] XU H, PAN Y, LI J, et al. Activity recognition method for home-based elderly care service based on random forest and activity similarity[J]. IEEE Access, 2019, 7(7): 16217-16225.

[4] 彭玉青,高晴晴,刘楠楠,等.基于多特征融合的跌倒行为识别与研究[J].数据采集与处理, 2016, 31(5): 890-902.
PENG Yuqing, GAO Qingqing, LIU Nannan, et al. Fall behavior recognition based on multi-feature fusion[J]. Journal of Data Acquisition and Processing, 2016, 31(5): 890-902.

[5] 朱艳,李曙生,谢忠志.基于FCM聚类和卷积神经网络的跌倒识别算法[J].数据采集与处理,2021,36(4): 746-755.
ZHU Yan, LI Shusheng, XIE Zhongzhi. Fall recognition algorithm based on fcm clustering and convolutional neural network [J]. Journal of Data Acquisition and Processing, 2021, 36(4): 746-755.

[6] KARIM S, BURNEY S M A, MAHMOOD N. An analysis of human activities recognition using smartwatches dataset[J]. International Journal of Advanced Computer Science and Applications, 2020, 11(12): 334-339.

[7] HASSAN M M, UDDIN M, MOHAMED A, et al. A robust human activity recognition system using smartphone sensors and deep learning[J]. Future Generation Computer Systems—The International Journal of Escience, 2018, 81(1): 307-313.

[8] THIEN H T, HUA C H, NGUYEN A T, et al. Physical activity recognition with statistical-deep fusion model using multiple sensory data for smart health[J]. IEEE Internet of Things Journal, 2021, 8(3): 1533-1543.

[9] SUN Luqian, ZHAO Yuyuan. Human activity recognition using time series pattern recognition model-based on tsfresh features [C]// Proceedings of 2021 International Wireless Communications and Mobile Computing. Harbin: IEEE, 2021: 1035-1040.

[10] BULLING A, BLANKE U, SCHIELE B. A tutorial on human activity recognition using body worn inertial sensors[J]. ACM Computing Surveys, 2014, 46(3): 1-33.

[11] LGNATOV A. Real-time human activity recognition from accelerometer data using convolutional neural networks[J]. Applied Soft Computing, 2018, 62(1): 915-922.
[12] MOHD S H B, NAZMUS S P, PROTAP K S, et al. Sensor-based human activity recognition: A comparative study of machine learning techniques[C]// Proceedings of 2020 2nd International Conference on Advanced Information and Communication Technology. Dhaka: IEEE, 2020: 286-290.
[13] LU W, FAN F G, CHU J H, et al. Wearable computing for internet of things: A discriminant approach for human activity recognition[J]. IEEE Internet of Things Journal, 2019, 6(2): 2749-2759.
[14] CHANVICHET V, THITIPHOOM T, VIRINYA W, et al. Comparison of feature selection and classification for human activity and fall recognition using smartphone sensors[C]// Proceedings of 2021 Joint International Conference on Digital Arts. Chaam: IEEE, 2021: 170-173.
[15] ISAH A, SOPHIA B. Deep human activity recognition with localization of wearable sensors[J]. IEEE Access, 2020, 8(1): 155060-155070.
[16] BANOS O, GALVEZ J M, DAMAS M, et al. Window size impact in human activity recognition[J]. Sensors, 2014, 14(4): 6474-6499.
[17] DANG L M, MIN K, WANG H X, et al. Sensor-based and vision-based human activity recognition: A comprehensive survey [J]. Pattern Recognition, 2020, 108(1): 1-24.
[18] CHELLI A, PATZOLD M. A machine learning approach for fall detection and daily living activity recognition[J]. IEEE Access, 2019, 7(1): 38670-38687.
[19] MIRJALILI S. A sine cosine algorithm for solving optimization problems[J]. Knowledge Based Systems, 2016, 96(5): 120-133.
[20] XUE Y, JIN L. A naturalistic 3D acceleration-based activity dataset & benchmark evaluations[C]// Proceedings of IEEE International Conference on Systems Man & Cybernetics. Istanbul: IEEE, 2010: 4081-4085.
[21] KENNEDY J, EBERHART R. Particle swarm optimization[C]// Proceedings of International Conference on Neural Networks. Perth: IEEE, 1995: 1942-1948.
[21] YANG X, DEB S. Cuckoo search via levy flights[C]// Proceedings of World Congress on Nature & Biologically Inspired Computing. [S.l.]: IEEE, 2009: 210-214.
[23] FARAMARZI A, HEIDARINEJAD M, STEPHENS B, et al. Equilibrium optimizer: A novel optimization algorithm[J]. Knowledge-Based Systems, 2020, 191(1): 1-33.
[24] TAO D, JIN L, WANG Y, et al. Rank preserving discriminant analysis for human behavior recognition on wireless sensor networks[J]. IEEE Trans on Industrial Informatics, 2014, 10(1): 813-823.
[25] VANRELL S R, MILONE D H, RUFINER H L. Assessment of homomorphic analysis of human activity recognition from acceleration signals[J]. IEEE J Biomed Health Inform, 2018, 22(4): 1001-1010.
[26] LI R, LI H, SHI W. Human activity recognition based on LPA[J]. Multimedia Tools and Applications, 2020, 79(41): 31069-31086.

作者简介:

戴健威(1998-),男,硕士研究生,研究方向:人体活动识别,E-mail:963953440@qq.com。

李瑞祥(1967-),**通信作者**,男,讲师,研究方向:无线传感器网络应用,E-mail:lrx@usst.edu.cn。

陈金瑶(1999-),女,硕士研究生,研究方向:人体活动识别。

乐燕芬(1978-),女,副教授,研究方向:无线传感器网络定位。

施伟斌(1967-),男,副教授,研究方向:无线传感器网络协议。

(编辑:夏道家)