文章编号：2095-2163(2023)06-0013-07 中图分类号：P315.69 文献标志码：A

基于手机加速度异常极值处理的步态身份识别

陈志强[1]，苗敏敏[1,2]，胡文军[1,2]
（1 湖州师范学院 信息工程学院，浙江 湖州 313000；
2 浙江省现代农业资源智慧管理与应用研究重点实验室，浙江 湖州 313000）

摘 要：针对手机加速度信号预处理后仍有小范围异常极值的现象，提出一种基于四分位数去除异常极值的新方法。该方法利用四分位数特性优化信号，削弱了极端值和异常值的影响来提高识别准确率。首先，在福坦莫大学无线数据挖掘实验室开放的公用数据集中筛选步态为步行的数据，将数据经过滤波算法过滤后，采用本文所提方法进行异常值处理，随后进行模板划分以及时域与频域特征的提取，最后实现身份识别。此外，自行采集人体真实步行数据进行实验验证。实验结果表明，在公用数据集和自采数据集上，与未经本文算法处理的数据相比，本文所提方法在支持向量机、BP 神经网络和级联森林 3 种分类模型上均能提高识别准确率，当采用级联森林分类器时，准确率分别达到 99.31%和 99.23%。

关键词：步态身份识别；四分位数；级联森林；手机加速度；支持向量机

Gait identification based on abnormal extremum processing of mobile phone acceleration

CHEN Zhiqiang[1], MIAO Minmin[1,2], HU Wenjun[1,2]
(1 School of Information Engineering, Huzhou University, Huzhou Zhejiang 313000, China; 2 Zhejiang Province Key Laboratory of Smart Management and Application of Modern Agricultural Resources, Huzhou Zhejiang 313000, China)

【Abstract】 Aiming at the phenomenon that mobile phone acceleration signal still has small range abnormal extreme value after preprocessing, a new method based on quartile to remove abnormal extreme value is proposed. This method uses the quartile characteristic to optimize the signal and reduces the influence of extreme value and outlier value to improve the recognition accuracy. Firstly, the gait data of walking is selected from the public dataset provided by the Wireless Data Mining Lab at Fort Hays State University. After the data is filtered by a filtering algorithm, the proposed method is used to remove outliers, followed by template segmentation and extraction of time-domain and frequency-domain features for identity recognition. In addition, real human walking data is collected for experimental verification. The experimental results show that compared with the unprocessed data, the proposed method could improve the recognition accuracy in support vector machine, BP neural network and cascaded forest models on both the public dataset and self-collected dataset. For the cascaded forest classifier, the accuracy reaches 99.31% and 99.23%, respectively.

【Key words】 gait identification; quartile; cascade forest; cell phone acceleration; support vector machine

0 引 言

近年来，智能设备已成为日常必需品，人们每天都在依赖智能设备来完成日常生活中的各种任务，因此对智能手机的安全问题也越来越重视[1-2]。移动智能设备最广泛使用的身份识别技术主要分为 3 大类：第一类是基于 PIN、密码[3]的识别方式，这种方式的缺点是密码容易被遗忘和泄露；第二类是基于人脸、指纹以及虹膜等生物特征[4]的识别方式，这种方式要求被监测对象必须近距离获取信息，同时需要高昂的手机成本；第三类方法是基于三维手势、步态等行为习惯特征[5]的识别方式，该方式具有改变困难、模仿困难等优点，其步态识别是唯一一个在远距离非接触情况下，也能正确识别身份的一种行为习惯特征。如今大多数智能手机都配备了许多内置传感器，例如加速度、陀螺仪等。这些传感器

基金项目：国家自然科学基金(62101189, U20A20228)。

作者简介：陈志强(1995-)，男，硕士研究生，主要研究方向：模式识别；苗敏敏(1989-)，男，博士，讲师，硕士生导师，主要研究方向：模式识别、生物医学信号处理；胡文军(1977-)，男，博士，教授，硕士生导师，主要研究方向：模式识别、数据挖掘。
通讯作者：苗敏敏 Email：02746@zjhu.edu.cn

收稿日期：2022-06-13

支持隐式身份识别技术,能从传感器中捕捉到用户的行为特征。孔菁等人[6]提出采用坐标轴转换算法,让基准坐标系和惯性坐标系重合,提取特征后使用支持向量机算法进行分类识别,识别准确率达到95.5%。Sun 等人[7]提出一种速度自适应步态周期分割方法和个性化阈值生成方法,与基于固定步行速度和恒定阈值的最新技术相比,用户身份识别准确率提高了 21.5%。Hoang Minh Thang 等人[8]采用时域和频域进行实验,采用支持向量机对提取的特征进行分类识别,得到的准确率分别为 79.1% 和92.7%。胡春生等人[9]通过对数据进行特征提取和数据权重分析,构建 BP 神经网络进行训练和匹配识别实验,准确率可达 96.67%。王彬等人[10]建立了步频分布的特征模型,利用相对熵判别用户身份,识别准确率达到 86%。

由于经过滤波后加速度信号中仍然存在小范围的异常极值,为进一步提高识别的准确率,本文提出一种基于四分位数去除异常极值的算法,并采用公用数据集以及自行采集的数据集分别对所提算法进行实验。实验结果表明,在两个数据集上使用本文提出的算法后,准确率均得到提高。

1 公开数据集身份识别方案的设计

1.1 数据获取

采用福坦莫大学无线数据挖掘实验室所提供的公开步态数据集[11]进行实验。该数据集利用手机加速度传感器采集了 36 个不同受试者的步行、慢跑、静坐、站立、上楼和下楼 6 种步态,采样频率为20 HZ。采集过程中手机放在大腿的便携包里,Z 轴指向为前进方向。通过文献[12]发现重力方向上的加速度信号比其他两个方向上的信号更稳定。因此,本文重点分析重力方向的加速度信号,即 Y 轴。

为保证实验数据充足,在 6 种步态中选取步态为步行,且采样点大于 3 000 的时间序列作为实验数据(某个时间序列代表在某一连续时间内采集到的步态加速度数据),共得到 57 个步行时间序列。在各时间序列中分别选取连续的 2 000 个采样点进行实验,其中前1 200个采样点作为训练数据,后 800 个采样点作为测试数据。

由于受试者在步行过程中,并未始终处在稳定步行状态,此时采集到的数据不是步行的真实数据。针对这一现象,本文提出一种基于波峰和波谷方差之和的最小值方法来自适应截取数据。以窗口大小为 2 000,步长为 400 进行滑窗操作,计算 Y 轴方向上各窗口内所有波峰值与波谷值的方差,选取波峰方差与波谷方差之和最小的窗口作为实验数据。某步行时间序列各个窗口波峰方差与波谷方差之和统计如图 1 所示。从图中可以看到,在 1 600-3 600 范围内的数据方差之和最小,表示该范围内的数据较为平稳,因此将该段数据选为稳定步行数据进行实验。

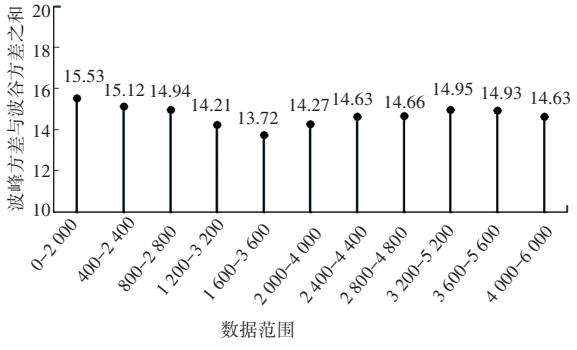


图 1 时间序列选择

Fig. 1 Time series selection

1.2 数据预处理

采用 Savitzky-Golay 滤波[13]方法对数据进行两次平滑处理,两次平滑窗口分别为 7 和 5。考虑到滤波之后仍然会有一些异常极值,因此本文提出一种基于四分位数[14]去除异常极值的算法来提高识别准确率。

四分位数分为下边缘、下四分位数、中位数、上四分位数和上边缘。其中,下四分位数位置记为 Q_1,中位数位置记为 Q_2,上四分位数位置记为 Q_3。

最小估计值公式:

$$\mathrm{Min} = Q_1 - k(Q_3 - Q_1) \tag{1}$$

最大估计值公式:

$$\mathrm{Max} = Q_3 + k(Q_3 - Q_1) \tag{2}$$

其中,k 表示异常值检测因子,设定为 1.5。

当数值大于最大估计值或小于最小估计值时都记为异常值。异常极值去除算法实现步骤如下:

(1)计算信号的所有极大值点,计算如公式(3):

$$x_{i-1} < x_i \& x_i > x_{i+1} \tag{3}$$

其中,x_i 表示当前时刻的采样点,x_{i-1} 和 x_{i+1} 分别是前一时刻和下一时刻的采样点。

(2)使用公式(1)、公式(2)计算所有极大值的最小估计值和最大估计值,若满足公式(4),则被确定为异常极大值。

$$p_i > \mathrm{Max} \mid p_i < \mathrm{Min} \tag{4}$$

其中,p_i 为信号的极大值。

(3)将异常极大值点前一个极大值点到后一个极大值点之间的采样点删除,其余两轴使用与该轴数据相同的起始与结束位置进行删除,以保证删除后的三轴数据在时间上一一对应。

(4)计算去除异常极大值后信号的所有极小值点,使用公式(1)、公式(2)计算所有极小值的最小估计值和最大估计值。若满足公式(5),则被确定为异常极小值。

$$v_i > \mathrm{Max} \mid v_i < \mathrm{Min} \tag{5}$$

其中,v_i 为信号的极小值。

(5)将异常极小值点前一个极小值点到后一个极小值点之间的采样点删除,其余两轴使用与该轴数据相同的起始与结束位置进行删除,以保证删除后的三轴数据在时间上一一对应。得到去除异常极值后的信号。

各步行时间序列的异常波峰箱线与波谷箱线如图 2 所示。去除异常波峰与波谷后的对比图如图 3 所示,图 3(a)和图 3(b)分别表示处理前和处理后的波形图,可以看出使用所提算法后,异常极值已被有效去除。

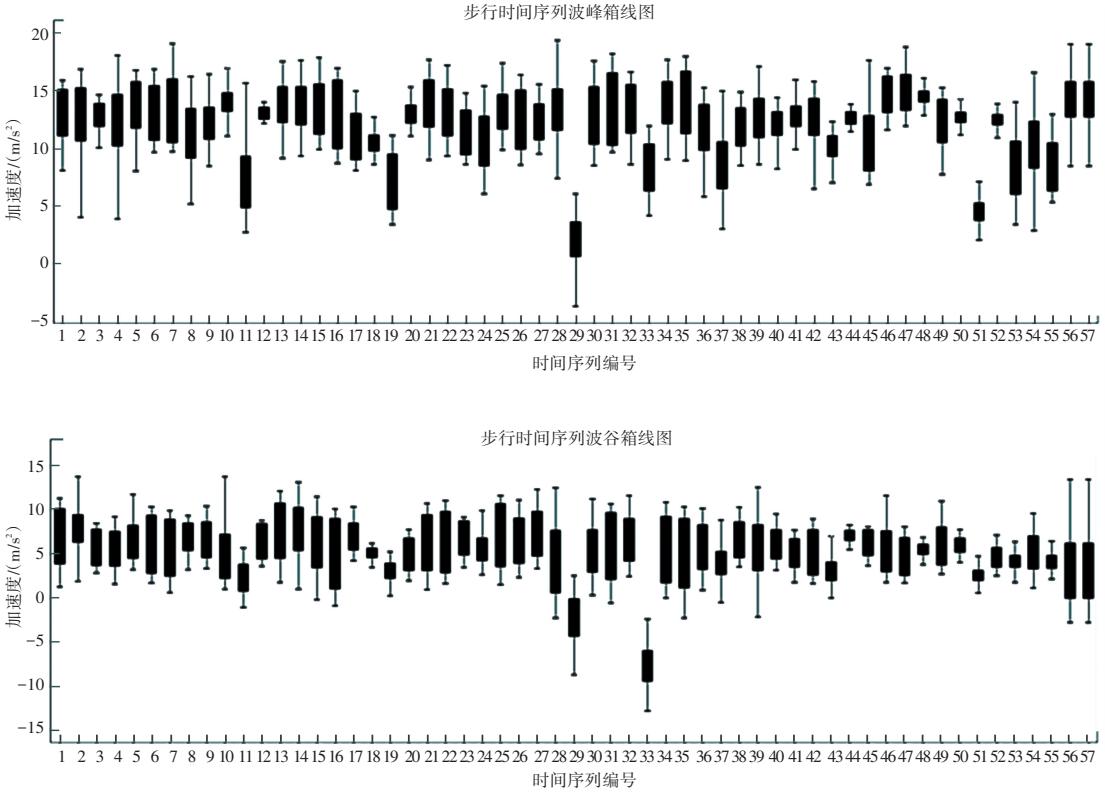
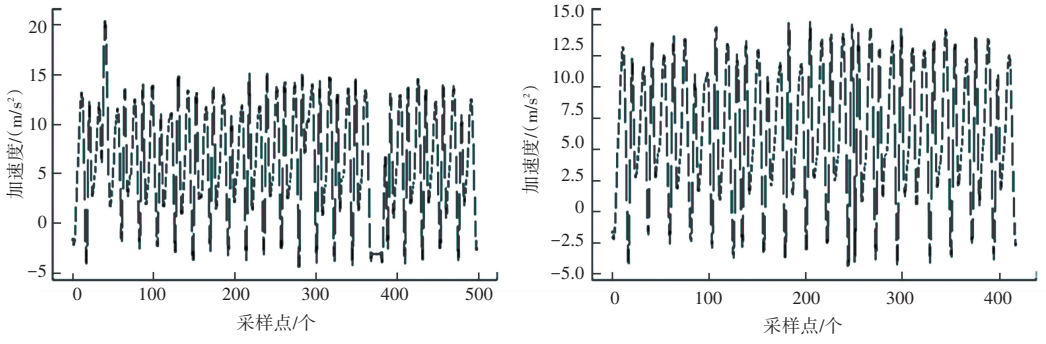


图 2　各步行时间序列的异常波峰箱线图与波谷箱线图

Fig. 2　Abnormal wave peak and trough boxplots of each walking time series



(a) 波峰波谷处理前　　(b) 波峰波谷处理后

图 3　去除异常波峰与波谷前后加速度信号对比

Fig. 3　Comparison of acceleration signal before and after removing abnormal wave peak and trough

1.3 模板划分

以每 8s 的步行数据作为步行模板。由于采样频率为 20 Hz,因此一个步行模板包含 160 个采样点。统计步行时间序列中 Y 轴方向上加速度数据的波峰位置信息,以每个波峰位置作为起始点,后 160 个采样点作为步行模板。

为了得到更多的步行模板,当正向截取结束后,再从最后一个波峰位置开始逆向截取,X 轴与 Z 轴使用与 Y 轴的波峰位置进行数据截取。模板截取如图 4 所示。

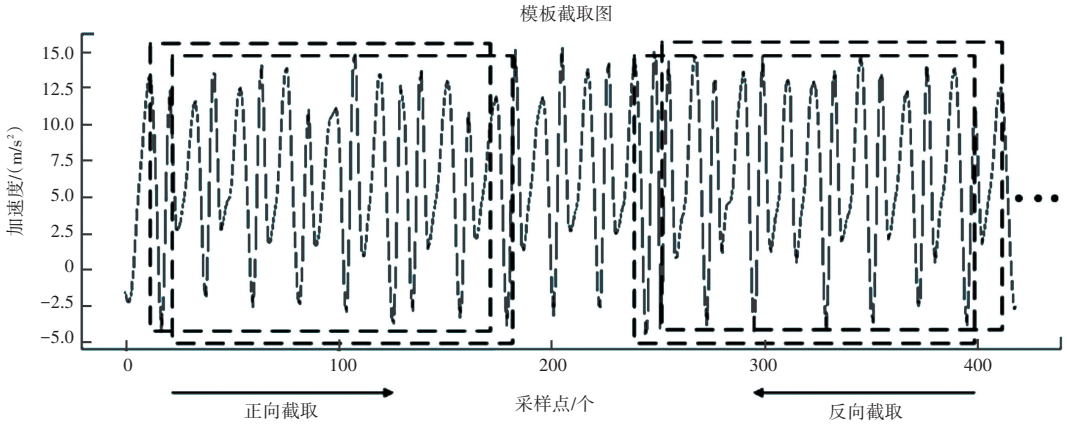


图 4 模板截取

Fig. 4 Template to intercept

1.4 特征提取

根据文献[15-17],从时域和频域两方面对步态模板进行特征提取。下面描述这些特征,括号中注明的是每个特征类型生成的特征数量。

(1)时域特征

平均值(3):X 轴、Y 轴、Z 轴的平均值。

标准差(3):X 轴、Y 轴、Z 轴的标准差。

相关系数(3):X 轴和 Y 轴、X 轴和 Z 轴、Y 轴和 Z 轴的相关系数。

三轴加速度合成标量最大值(1):X 轴、Y 轴、Z 轴加速度平方和的平方根的最大值。

M 为三轴加速度合成标量最大值,计算公式如式(6):

$$M = \max\left(\sqrt{x_i^{\ 2} + y_i^2 + z_i^2}\right) \qquad (6)$$

波峰平均值(3):X 轴、Y 轴、Z 轴所有波峰的平均值。

波谷平均值(3):X 轴、Y 轴、Z 轴所有波谷的平均值。

直方图(10):重力方向上的加速度轴(Y 轴)数据中的最大值和最小值,相减的差除以 10 的结果作为间隔,算出每个间隔里点的个数所占的百分比。

(2)频域特征

直流分量(3):X 轴、Y 轴、Z 轴经快速傅里叶变换后频率为 0 的分量。

1.5 级联森林模型

级联森林是深度森林的一部分[18],每个级联层包括两个随机森林和两个完全随机森林,每个决策器包含 100 棵决策树。完全随机森林中,每棵树随机选择一个特征作为分裂点,然后一直增长,直到每个叶子节点细分到只有一个类别或者不多于 10 个样本。随机森林中,每棵树选取 $\sqrt{n}$ 个特征(n 为特征维度),再通过 gini 系数[19]筛选分裂节点。级联森林分类原理如图 5 所示。

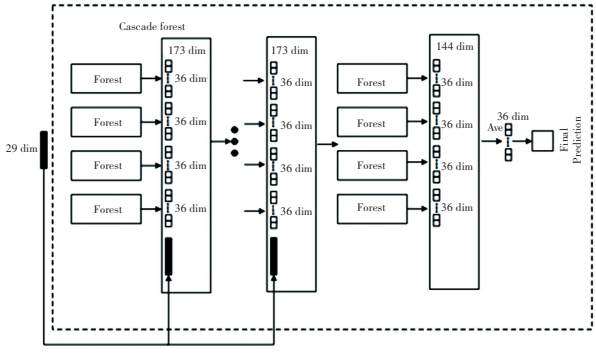


图 5 级联森林

Fig. 5 Cascade forest

首先,将样本特征提取后的 29 维特征分别输入两个随机森林和两个完全随机森林中,每个森林得到一个 36 维概率类向量,将 4 个 36 维类向量与原始特征拼接成一个 173 维数据,并作为下一层的输入,以此类推。最后输出 4 个 36 维数据,将这 4 个 36 维数据中的每一维取平均值,得到一个新 36 维数据,取 36 维数据中结果最大的类别作为最终预测。

2　实验结果与分析

在各时间序列中分别选取连续的 2 000 个采样点进行实验，其中前 1 200 个采样点作为训练数据，后 800 个采样点作为测试数据。将训练集经过预处理、模板划分共得到 14 050 个样本，对样本特征提取后进入模型训练。测试集经过预处理、模板划分共得到 8 651 个样本。

为了验证级联森林分类算法在身份识别中的有效性，分别使用支持向量机和 BP 神经网络模型[9]进行对比。其中，BP 神经网络输入层节点为 29，隐藏层节点为 30，输出层节点为 36，得到一个 29-30-36 的 BP 神经网络模型。同时为了验证本文提出的异常峰值去除算法的有效性，设计实验将未经算法处理的样本与经算法处理的样本在各识别分类算法中训练，在各分类算法中得到的准确率见表 1。

表 1　各分类算法准确率对比

Tab. 1　Comparison of the accuracy of several classification algorithms

No	分类算法	是否使用四分位数去除异常峰值算法	准确率/%	时间(训练+测试)/s
1	SVM	否	90.30	16.53
2	SVM	是	91.45	16.31
3	BP 神经网络	否	89.92	12.53
4	BP 神经网络	是	91.58	10.08
5	级联森林	否	99.19	6.86
6	级联森林	是	99.31	6.69

从表 1 中可以看出，使用异常峰值去除算法后，各分类算法的准确率均有提高，平均提高了 0.98%。从各分类算法可以看出，级联森林分类算法的准确率明显高于支持向量机和 BP 神经网络，分类效果更佳；同时在运行时间方面(训练+测试)也表明了级联森林算法模型训练的高效率以及可扩展性。

3　实验验证

为了验证公开数据集实验方案的有效性，本文自行采集人体真实步行数据进行实验验证。使用自行开发的手机 APP 采集受试者行走时加速度传感器的数据，采样频率为 14 Hz，使用一阶低通滤波器去除重力的影响。在公开数据集中，手机放在大腿的便携包里，加速度传感器的 Y 轴平行于重力加速度方向，而在自采数据集中，手持手机行走时的加速度传感器的 Z 轴平行于重力加速度方向，如图 6(a)所示。因此，在验证实验中以 Z 轴数据为基准进行实验。

所选的受试者均为在校本科生和研究生，身高为 155~180 cm，体重 45~80 kg，年龄 20~25 岁。本次实验共采集了 11 名受试者的步行数据，要求受试者在走廊从规定的起始点步行至结束点，记为第一次步行实验数据 name1，再从结束点步行至起始点，记为第二次步行实验数据 name2。如此反复，每名受试者采集 5 次，采集到的数据以 txt 文本文件记录，以“姓名+编号”命名，实验步行场景如图 6(b)所示。将每名受试者的 5 次步行数据样本中，编号为 1、2、3 的步行数据作为训练集，编号为 4、5 的步行数据样本作为测试集。由于每次步行数据的采样点在 350-450 之间，因此以步长为 30 且窗口为 250 选择有效数据段，将训练集选择后的数据以 1.2 节的方法进行预处理，特征提取后使用分类算法训练。测试集数据使用与训练集数据同样的数据选择方式、预处理和特征提取操作，最后输入到模型中进行身份识别。

(a) 手机方向　　(b) 实验场地

图 6　实验采集场地和手持手机图

Fig. 6　The site for collecting data and picture of handheld phone

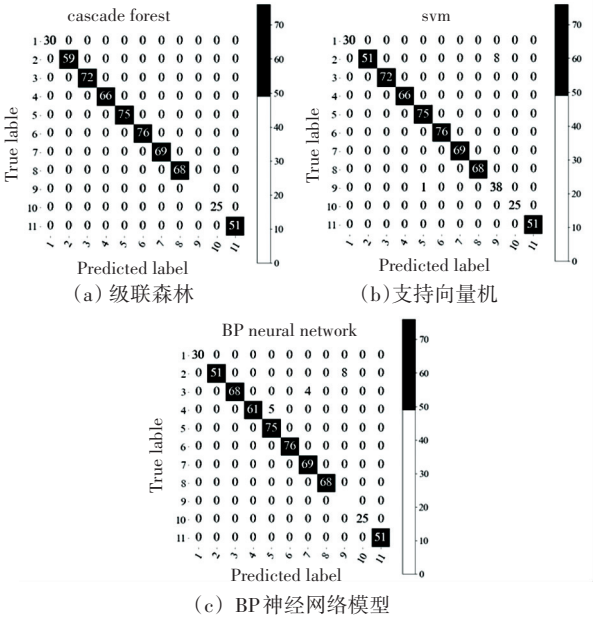
选择的训练样本数据和测试样本数据得到的结果可能存在偶然性，因此对 5 次步行数据样本轮流选取 3 次步行数据样本作为训练数据进行实验，剩余 2 次步行数据作为测试数据，即 $C_5^3 = 10$ 组。使用支持向量机、BP 神经网络和级联森林分类算法在各样本组的准确率见表 2。其中，训练数据样本编号为 1、2、3 在级联森林、支持向量机和 BP 神经网络

模型的混淆矩阵分别如图 7(a)、图 7(b)和图 7(c)所示。

表 2 三种分类算法在各样本组的准确率

Tab. 2 The accuracy of three classification algorithms in various sample groups %

训练数据编号	测试数据编号	识别方法		
		SVM	BP 神经网络	级联森林
1,2,3	4,5	98.57	97.30	100
1,2,4	3,5	94.85	95.94	100
1,2,5	3,4	97.30	96.98	99.37
1,3,4	2,5	95.74	96.53	96.68
1,3,5	2,4	98.87	96.28	98.87
1,4,5	2,3	95.14	97.99	98.16
2,3,4	1,5	98.45	98.76	100
2,3,5	1,4	98.90	100	99.53
2,4,5	1,3	99.84	99.84	100
3,4,5	1,2	99.21	99.20	99.68
平均准确率		97.69	97.88	99.23



(a) 级联森林 (b)支持向量机 (c) BP神经网络模型

图 7 三种分类识别算法的混淆矩阵

Fig. 7 Confusion matrix of three classification algorithms

为了进一步验证所提算法的有效性,将经异常极值去除算法处理过的数据与未处理过的数据在级联森林模型中对比,得到的 10 组不同训练集和测试集的准确率如图 8 所示。

由图 8 中可见,经异常峰值去除算法处理后的平均准确率为 99.23%,未经处理的准确率为 98.81%,准确率提升了 0.42%,再次验证了此方法对于提升识别准确率有较好的效果。

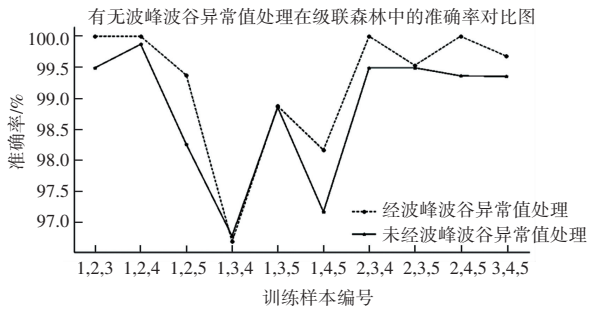


图 8 有无波峰波谷异常值处理在级联森林中的准确率

Fig. 8 Comparison of accuracy with and without peaks and troughs abnormal value processing in cascaded forest

4 结束语

针对现阶段身份识别准确率低的问题,在充分分析加速度信号之后,考虑到信号中会出现小部分异常极值的现象,提出了基于四分位数去除异常极值算法来进一步提高识别准确率。实验结果表明,在公开数据集和自采数据集上,经过算法处理后的识别准确率均得到进一步提高,并验证了级联森林分类算法在身份识别领域有较好的实际价值。

参考文献

[1] 迟有鹏. 可穿戴设备数据安全与隐私保护研究[J]. 电子技术与软件工程,2019(8):196.

[2] NEAL T J, WOODARD D L. Surveying biometric authentication for mobile device security[J]. Journal of Pattern Recognition Research,2016,1(74-110):4.

[3] SPOLAOR R, LI Q Q, MONARO M, et al. Biometric Authentication Methods on Smartphones: A Survey[J]. PsychNol-ogy Journal, 2016,14(2).

[4] JOSEPH L MIL, SHRIVASTAVA P, KAUSHIK A, et al. Methods to identify facial detection in deep learning through the use of real time training datasets management [J]. EF - FLATOUNIA Multidisciplinary Journal, 2021,55(2):1-34.

[5] FILIPI GONÇALVES DOS SANTOS C, OLIVEIRA D S, A. PASSOS L, et al. Gait recognition based on deep learning: A survey [J]. ACM Computing Surveys (CSUR),2022,55(2):1-34.

[6] 孔菁,郭渊博,刘春辉,等. 基于智能手机运动传感器的步态特征身份识别方法[J]. 计算机应用,2019,39(6):1747-1752.

[7] SUN F, MAO C, FAN X, et al. Accelerometer-based speed adaptive gait authentication method for wearable I oT devices[J]. IEEE Internet of Things Journal,2018,6(1):820-830.

[8] THANG H M, VIET V Q, THUC N D, et al. Gait identification using accelerometer on mobile phone[C]//2012 International Conference on Control, Automation and Information Sciences (ICCAIS).IEEE,2012:344-348.

[9] 胡春生,王德,赵汇东. 人体步态特征数据分析和人物身份识别方法研究[J]. 计算机应用研究,2020,37(S2):129-132.

[10] 王彬,付雄,王俊昌. 可穿戴计算中基于步频分布的身份识别研究[J]. 计算机应用与软件,2020,37(10):188-193.

(下转第 29 页)