

潘潇潇, 郑建立. 基于 DBSCAN 聚类算法的心电图 R 峰检测研究[J]. 智能计算机与应用, 2024, 14(7): 198-202. DOI: 10.20169/j.issn.2095-2163.240731

基于 DBSCAN 聚类算法的心电图 R 峰检测研究

潘潇潇, 郑建立

(上海理工大学 健康科学与工程学院, 上海 200093)

摘要: 本研究致力于提高包含运动噪声、肌电噪声和基线漂移等干扰的心电信号中 R 峰的检测准确性, 为准确估计心率和心率变异性等重要生理参数提供依据。研究通过整合 4 种不同的 R 峰检测算法 Pan_Tompkins、Hamilton、Engzee 和 GQRS 的结果, 并应用 DBSCAN 聚类算法, 用于提升 R 峰的检测性能。实验选用布尔诺理工大学心电图质量数据库作为数据源, 从中提取了 1 000 s 的第二类信号质量样本, 这些样本含有运动伪迹和噪声干扰, 但 R 峰依旧可识别, R 峰定位的误差容忍度设为 100 ms 以内。实验结果表明, DBSCAN 算法显著提高了 R 峰检测的准确率至 87.6%, 灵敏度至 93.1%, 阳性预测值至 93.8%, 以及 F 分数至 0.934, 说明 DBSCAN 算法能有效提升在复杂噪声环境下的 R 峰检测性能, 满足临床及研究需求。

关键词: 心电信号; R 峰检测算法; DBSCAN

中图分类号: R339.4

文献标志码: A

文章编号: 2095-2163(2024)07-0198-05

Research on electrocardiogram R peak detection based on DBSCAN clustering algorithm

PAN Xiaoxiao, ZHENG Jianli

(School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: This study aims to improve the detection accuracy of R-peaks in electrocardiogram signals containing interference such as motion noise, electromyographic noise, and baseline drift, providing a basis for accurately estimating important physiological parameters such as heart rate and heart rate variability. The study aims to improve the detection performance of R peaks by integrating the results of four different R peak detection algorithms, Pan-Tompkins, Hamilton, Engzee, and GQRS, and applying the DBSCAN clustering algorithm. The experiment selected the electrocardiogram quality database of the University of Technology of Brno as the data source, and extracted 1 000 seconds of second-class signal quality samples. These samples contain motion artifacts and noise interference, but the R-peak can still be recognized, and the error tolerance for R-peak positioning is set to be within 100 milliseconds. The experimental results show that the DBSCAN algorithm significantly improves the accuracy of R-peak detection to 87.6%, sensitivity to 93.1%, positive predictive value to 93.8%, and F -score to 0.934, indicating that the DBSCAN algorithm can effectively improve the performance of R-peak detection in complex noise environments, meeting clinical and research needs.

Key words: ECG signal; R-peak detection algorithm; DBSCAN

0 引言

心血管疾病已成为全球的头号死因, 根据世界卫生组织的研究报告, 2021 年, 有 2 050 万人死于心血管疾病^[1]。在临床和健康监测应用中, 通常用心电图 (ECG) 监测和诊断心脏活动, R 波波峰 (R 峰) 是 ECG 信号处理的关键步骤之一。然而, 在运动状态下, 心电信号会受到较为严重的噪声干扰, 包括运

动噪声、肌电噪声和基线漂移等。运动噪声是运动引起的身体震动和电极移动所引入的高频噪声, 使 ECG 信号变得不规则和嘈杂^[2]; 肌电噪声是肌肉活动产生的电信号对心电信号的干扰; 基线漂移是运动状态下, 心电信号的基线可能因肌肉运动导致电极接触问题而漂移。这些噪声因素共同导致了传统的 R 峰检测算法在运动状态下的性能下降^[3]。如何在多种噪声源的干扰下准确识别 R 峰, 对于实现

作者简介: 潘潇潇 (2000-), 男, 硕士研究生, 主要研究方向: 健康物联网。

通讯作者: 郑建立 (1965-), 男, 博士, 副教授, 硕士生导师, 主要研究方向: 医学信息系统与集成技术, 健康物联网。Email: zhengjianli163@163.com

com

收稿日期: 2023-11-14

哈尔滨工业大学主办 ◆ 科技创新与应用

移动健康监测、运动生理学研究 and 心脏疾病诊断起着重要作用^[4]。

本文提出的基于DBSCAN聚类算法的心电图R峰检测是一种两阶段的心电信号处理方法,旨在准确检测心电图中的R峰这一心脏信号中的关键特征。首先,对原始信号用多种R峰检测方法进行检测,再通过DBSCAN聚类算法,将不同的R峰检测结果结合在一起,以降低误检和漏检的风险,提高整体性能。

1 研究现状

心室兴奋所产生的左、右心室除极电位和时间的变化(QRS波群),为尖锐的上下波动,持续时间为0.06~0.1 s,在标准肢体导联中,ECG上显示的向上偏移(R峰)是主波,只有准确检测到R峰后,才能正确定位出心房去极化(P波)和心室复极化的晚期阶段(T波),进一步计算心率、峰峰间期(RR间期)等心电图特征^[5]。因此,R峰检测是心率变异性分析、心血管疾病诊断等应用的首要步骤。

随着可穿戴式心电检测设备的迅速发展,信号质量波动较大,对采集到的ECG信号进行质量评估成为临床诊断前不可或缺的环节^[6]。目前对ECG信号进行质量评估就是根据采集到的ECG信号受噪声污染的严重程度将信号分为3个质量类别:第一类,所有重要波形即P波、T波和QRS复合体都清晰可见,且可以可靠地检测到这些波形的起点和偏移;第二类,噪声水平增加,心电图中的重要点不清楚,但QRS波群可见,R峰清晰可见,信号能够可靠地进行QRS波群或R峰检测;第三类,QRS波群不能可靠地检测,信号不适合任何分析^[7]。

当前QRS波和R峰检测算法很多,主要包括差分阈值法、滤波器法、模板匹配法、小波变换法和自适应检测法等。Pan-Tompkins差分阈值法,因其实时性和广泛的适用性而广为使用,尤其适合处理运动状态下的ECG信号,但其在强噪声环境中的准确性可能会受到影响^[8];Hamilton滤波器算法,特别适用于低信噪比环境,但对于高频噪声和基线漂移的抗干扰能力较弱^[9];模板匹配法通过匹配已知的QRS波和R峰模板来定位R峰,适用于具有明显周期性和规律性的心电信号^[10];SWT(Stroke Width Transform)小波变换法算法利用小波系数的极值点来检测R峰,在噪声抵抗方面表现良好,但在低信噪比信号中表现不稳定^[11]。Engzee算法能够适应QRS波形的多样性,但在低质量的ECG信号中可能

导致误检或漏检^[12];QRS自适应法算法(Gqpost-QRS detector and post-processor),能够适应不同类型的心电信号,但在处理多峰信号时出现重复检测的问题^[13]。尽管结合多种检测算法可以提高检测效果,但在高噪声环境下,这些检测算法的效果仍面临挑战^[14]。

聚类算法是一种根据数据点的相似性分组成不同的簇的方法^[15],通过将相似的数据点放在同一簇中,有助于揭示数据内在的结构,提供有关数据分布和关联性的洞察。在运动心电信号的检测算法中,聚类方法将信号数据分成不同的组,从而帮助识别和理解不同的心电信号模式。这些聚类方法包括K-means(K均值)聚类、均值漂移聚类、基于密度的聚类算法(DBSCAN)以及使用高斯混合模型(GMM)的最大期望(EM)聚类等^[16]。聚类方法在处理复杂噪声环境下的运动心电信号时能够提供有价值的信息,从而帮助改进信号分析和检测性能。

2 相关方法

2.1 候选R峰数据集

使用Pan-Tompkins、Hamilton、QRS和Engzee4种算法对R峰探测及其位置标定如图1所示,通过预实验评估各个算法在不同信号质量和噪声水平情况下的性能,Pan-Tompkins算法和Hamilton算法在灵敏度(SEN)、真阳率(PP)、F分数(F-score)3个指标上均有相对较好表现,但是准确率(ACC)较低。Engzee算法在真阳率较好,但是在准确率和灵敏度较低,QRS算法在灵敏度上有较为优异的表现,但是在准确率、真阳率、F分数均较低。综上所述,为保留各个算法优点,补齐算法缺点,4种算法分别对样本进行R峰检测,合并形成候选R峰数据集。

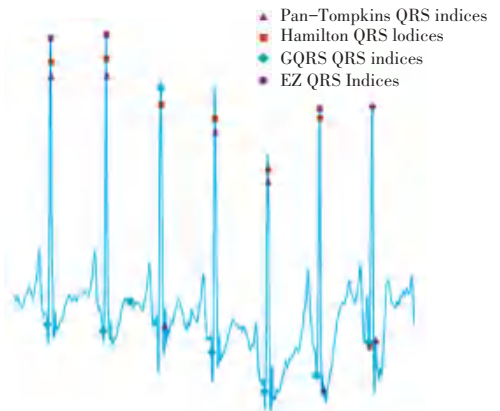


图1 四种算法对R峰探测及其位置标定
Fig. 1 Four algorithms for R peak detection and position annotation

2.2 DBSCAN 算法

DBSCAN (Density - Based Spatial Clustering of Applications with Noise) 聚类算法用于在数据集中发现具有高密度的区域,并将其划分为簇。与传统的 K-means 等聚类算法不同,DBSCAN 不需要预先指定簇的数量,可以自动识别数据中的不同簇,同时还能有效处理噪声点。DBSCAN 的核心思想是根据数据点的密度将其分为核心点 (Core Points)、边界点 (Edge Points) 和噪声点 (Noise Point),通过这种方式实现聚类^[17]。

采用 DBSCAN 聚类算法对候选 R 峰数据集进行聚类,算法描述如下:

输入 D: 候选 R 峰数据集

E 邻域: 选定的 R 峰范围

MinPts: 阈值范围内 R 峰个数

输出 R 峰检测聚类结果 R

方法: 标记所有 R 峰数据为 unvisited

Do

 随机选择一个 unvisited 的 R 峰数据 p

 标记 p 为 visited

 If p 在 E 邻域内至少有 MinPts 个 R 峰数据

 创建一个新簇 C, 并把 p 添加到 C;

 令 N 为 p 的 E 邻域中的 R 峰数据集

 For N 中每个点 p'

 If p' 是 unvisited

 标记 p' 为 visited

 If p 的 E 邻域至少有 MinPts 个 R 峰数据, 把这些数据添加到 N

 If p' 还不是任何簇的成员, 把 p' 添加到 C

 End for

 将 C 的平均值添加到 R

 Else 标记 p 为噪声

Until 没有标记为 unvisited 的 R 峰数据

3 实验验证

3.1 数据集介绍

本研究使用了 PhysioNet 2020 年发布的布尔诺理工大学心电图质量公开数据集 (BUT QDB)^[18], 记录数据受试者年龄范围从 21~83 岁, 平均年龄为 41 岁, 中位年龄为 37 岁, 自 2018 年 8 月至 2019 年 10 月期间在正常生活条件下进行的数据采集, 数据集包含来自 15 名受试者的 18 个单导联 ECG 的长期记录。

数据集中 ECG 信号根据质量被专家分为 3 个类别: 类别 1 为清晰可见的显著波形; 类别 2 为噪音增加但 QRS 波清晰可见; 类别 3 为不适合分析的信号如图 2 所示。

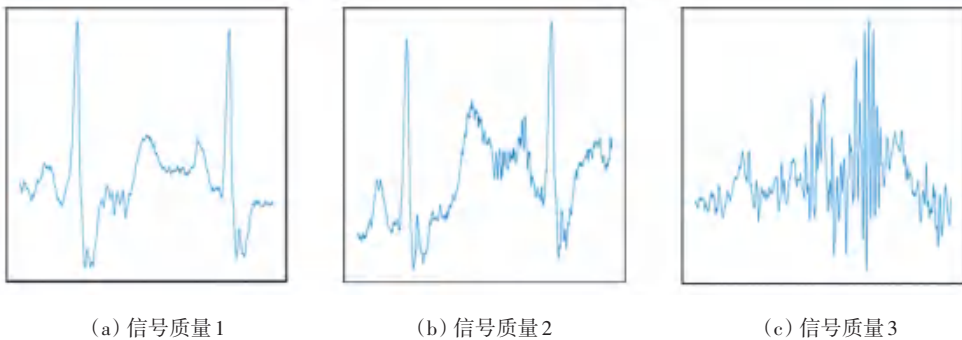


图 2 数据集 ECG 信号质量分类

Fig. 2 Dataset ECG signal quality classification

根据信号质量进行分类后, 测试数据集从其中 11 位受试者中提取, 从这些受试者的数据中切割出经过专家标注为信号质量 2 的 20 组样本, 每组包含 50 s 的心电信号, 总计 1 000 s。为保证数据的独立性, 同一志愿者提供的不同样本之间的时间间隔需超过 600 s, 见表 1。

3.2 实验流程及环境

将心电图测试数据集放入选中的 4 种分类器进行 R 峰检测, 得到候选 R 峰数据集, 保存成新的数

据文件, 再通过 DBSCAN 聚类算法, 将候选 R 峰进行聚类, 得到聚类结果, 与专家标注的 R 峰位置进行对比, 根据心电图临床的波形生理特性, 将误差设定在 100 ms 以内, 采用敏感度 (SEN)、真阳性率 (PP)、准确率 (ACC) 和 F 分数 (F - Score) 4 个评价指标评价模型效果。

实验环境: Python3.6.4, VSCode 集成开发环境, 操作系统为 Windows11, 英特尔的 i7-11700k_CPU, 主频 3.2 GHz, 2×8 G_DDR4 内存。

表 1 测试样本抽样表
Table 1 Test sample sampling table

志愿者编号	抽取时间样本	志愿者编号	抽取时间样本
100001	2 209 s_2 322 s 的前 50 s	111001	31 311 s_31 365 s 的前 50 s
100001	27 223 s_28 596 s 的前 50 s	111001	40 281 s_40 333 s 的前 50 s
103001	58 701 s_58 800 s 的前 50 s	113001	28 800 s_28 965 s 的前 50 s
103003	29 458 s_29 504 s 的前 50 s	114001	29 456 s_29 521 s 的前 50 s
104001	58 174 s_58 358 s 的前 50 s	115001	58 088 s_58 167 s 的前 50 s
105001	48 631 s_48 738 s 的前 50 s	118001	57 917 s_58 001 s 的前 50 s
105001	53 826 s_53 913 s 的前 50 s	122001	58 422 s_58 477 s 的前 50 s
111001	194s_288 s 的前 50 s	124 001	28 800 s_29011 s 的前 50 s
111001	15 463 s_15 537 s 的前 50 s	124 001	34 295 s_34 372 s 的前 50 s
111001	20 859 s_21 133 s 的前 50 s	124 001	58 116 s_58 167 s 的前 50 s

3.3 结果分析

测试数据集 R 峰范围为 0.06~0.1 s, 算法检测结果在专家人工标记位置的左右各 50 ms 内, 为准确识别到 R 峰。对每个样本分别计算正确检测到的 R 峰 (真阳性 TP)、未正确检测到的 R 峰 (真阴性 TN)、错误检测到的 R 峰 (假阳性 FP) 和未检测到的 R 峰的数量 (假阴性 FN)。采用敏感度 (SEN)、真阳性率 (PP)、准确率 (ACC) 和 F 分数 (F -Score) 4 个评价指标评估该算法的性能。

灵敏度 (SEN): SEN 是一个评估度量, 用正确检测到的 R 峰 (TP) 的数量与所有真实 R 峰的比值来表示, 如公式 (1):

$$SEN = \frac{TP}{TP + FN} \quad (1)$$

真阳性率 (PP): PP 给出了一个用正确检测到的 R 峰数 (TP) 与检测到的 R 峰的总数的比值计算出来的值。如公式 (2):

$$PP = \frac{TP}{TP + FP} \quad (2)$$

准确率 (ACC): ACC 是一个评估度量, 是正确检测到的 R 峰数与所有 R 峰总数的比值, 计算如公式 (3):

$$ACC = \frac{TP}{TP + FP + FN} \quad (3)$$

F -Score 分数能同时考虑 SEN 和 PP 这两个数值, 平衡地反映这个算法的精确度。 F -Score 越高说明模型敏感度和真阳性率越高, 模型越好如公式 (4):

$$F - Score = \frac{2 \times SEN \times PP}{SEN + PP} \quad (4)$$

参与融合的算法和 DBSCAN 算法结果对比如

图 3 所示, 相比用于参与聚类的 4 种单一算法, DBSCAN 算法在准确率上有较大提升, 相较于 QRS 算法准确率从 45.5% 提升到 87.6%, 相较于 Hamilton 算法, 4 种指标均有 20% 以上的提升, 相较于 Engzee 算法, 准确率和敏感度均有 20% 以上提升, 达到聚类目的, 提高整体性能。

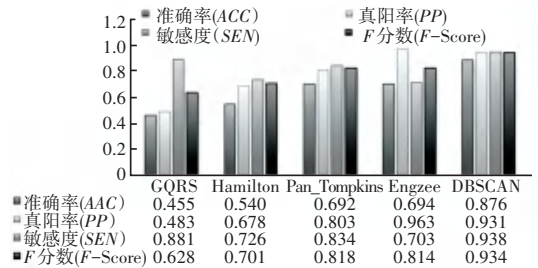


图 3 参与融合的算法和 DBSCAN 算法结果对比

Fig. 3 Comparison of results between algorithms participating in fusion and DBSCAN algorithm

4 结束语

本文算法能够有效降低运动引起的信号失真和噪声对 R 峰位置的误导, 从而提高了 R 峰检测的性能。与单一算法进行比较的结果表明, 本文算法在准确性和鲁棒性方面优于单一峰值检测算法, 有望在心脏健康监测和移动健康应用中发挥重要作用。未来可进一步优化算法以提高性能, 并将其集成到实际医疗设备和应用中。

参考文献

[1] 李小辉, 魏思昂, 陈树嘉, 等. 中国城乡居民饮食结构改变与心血管疾病相关性分析[J]. 中国食物与营养, 2023, 29(2): 81-84, 89.
[2] 马帅, 侯世科, 樊毫军, 等. 智能心电信号监测设备研究现状[J].

- 医疗卫生装备,2020,41(11):95-99.
- [3] 杨承金,聂春燕,车敏诗,等. 心电信号去噪效果的评估与分析[J]. 计算机工程与应用,2022,58(1):300-312.
- [4] 郝婧宇,南格丽,吴水才. 胎儿心电信号检测算法的研究进展[J]. 医疗卫生装备,2021,42(6):90-96,102.
- [5] 杨承金,聂春燕,车敏诗,等. 心电信号去噪效果的评估与分析[J]. 计算机工程与应用,2022,58(1):300-312.
- [6] 余新艳,王志新,郎立国,等. 单导联可穿戴远程心电监测设备在有症状心律失常患者中的应用模式探讨[J]. 实用心电学杂志,2021,30(1):46-49.
- [7] 王帅,赵钟瑶,张翔宇,等. 穿戴式心电信号质量的三分类评估方法[J]. 中国生物医学工程学报,2020,39(5):550-556.
- [8] FARIHA M A Z, IKEURA R, HAYAKAWA S, et al. Analysis of Pan-Tompkins algorithm performance with noisy ECG signals[J]. Journal of Physics: Conference Series. IOP Publishing, 2020, 1532(1): 012022.
- [9] LAITALA J, JIANG M, SYRJÄLÄ E, et al. Robust ECG R-peak detection using LSTM[C]//Proceedings of the 35th Annual ACM Symposium on Applied Computing. IEEE, 2020: 1104-1111.
- [10] 方震,白忠瑞,陈贤祥,等. 基于压电陶瓷传感器的非接触式精准逐拍心率提取方法研究[J]. 电子与信息学报,2021,43(5):1472-1479.
- [11] ZHANG H, HE R, DAI H, et al. SS-SWT and SI-CNN: An atrial fibrillation detection framework for time-frequency ECG signal[J]. Journal of Healthcare Engineering, 2020, 2020(1): 7526825.
- [12] RODRIGUES T, SAMOUTPHONH S, SILVA H, et al. A low-complexity r-peak detection algorithm with adaptive thresholding for wearable devices [C]//Proceedings of the 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 1-8.
- [13] ANUHYA A V, KOLLURU V R, PATJOSHI R K. A deep learning approach for detection and classification of QRS contours using single-lead ECG[J]. International Journal of Pharmaceutical Research, 2020, 12(2):75-91.
- [14] 马建红,段豪,韩颖. 心电信号中特征波分割方法研究综述[J]. 郑州大学学报(理学版),2023,55(2):79-87.
- [15] 陈叶旺,申莲莲,钟才明,等. 密度峰值聚类算法综述[J]. 计算机研究与发展,2020,57(2):378-394.
- [16] 赵嘉,陈磊,吴润秀,等. K近邻和加权相似性的密度峰值聚类算法[J]. 控制理论与应用,2022,39(12):2349-2357.
- [17] 王光,林国宇. 改进的自适应参数DBSCAN聚类算法[J]. 计算机工程与应用,2020,56(14):45-51.
- [18] NEMCOVA A, SMISEK R, OPRAVILOVÁ K, et al. Brno university of technology ECG quality database (but QDB) [J]. PhysioNet, 2020, 101: e215-e220.