Journal of Image and Signal Processing 图像与信号处理, 2020, 9(3), 165-171
Published Online July 2020 in Hans. http://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2020.93020

Hans汉斯

ECG Signal's Classification Algorithm Based on LightGBM

Yuguang Hong[1], Bo Wang[1], Hudi Pan[1], Shengyi Qian[2*]

[1]Ningbo Customs District Technology Center, Ningbo Zhejiang
[2]School of Electronic Information, Hangzhou Dianzi University, Hangzhou Zhejiang
Email: *2762435642@qq.com

Received: Jun. 4th, 2020; accepted: Jun. 26th, 2020; published: Jul. 3rd, 2020

Abstract

Electrocardiogram (ECG) has been widely used in the diagnosis of arrhythmia such as sinus tachycardia, ventricular premature beats and atrial fibrillation, and has shown great clinical application value in the diagnosis and analysis of heart diseases. In order to improve the classification performance of ECG computer-aided diagnosis, an ECG signal classification algorithm based on LightGBM was proposed. The algorithm extracts single heart beat features, heart rhythm volatility features and full waveform features from ECGs to establish a mixed feature set, and uses LightGBM to classify four categories: normal heart beat, atrial fibrillation, other arrhythmia, and noise. Finally, the performance parameter $F1_{nao}$ of the algorithm reached 0.824 on PhysioNet/CinC Challenge dataset, which was better than CART and CatBoost. At the same time, in order to accelerate the speed of ECG feature extraction, this paper screens the key features according to the importance of features to reduce the number of features required for classification, and the feature extraction time is reduced to 17.8% while maintaining the performance of classification.

Keywords

Electrocardiogram, Machine Learning, Signal Processing, LightGBM

基于LightGBM的心电信号分类算法

洪宇光[1]，王　波[1]，潘湖迪[1]，钱升谊[2*]

[1]宁波海关技术中心，浙江　宁波
[2]杭州电子科技大学电子信息学院，浙江　杭州
Email: *2762435642@qq.com

收稿日期：2020年6月4日；录用日期：2020年6月26日；发布日期：2020年7月3日

*通讯作者。

文章引用: 洪宇光, 王波, 潘湖迪, 钱升谊. 基于LightGBM的心电信号分类算法[J]. 图像与信号处理, 2020, 9(3): 165-171.
DOI: 10.12677/jisp.2020.93020

摘　要

心电图(Electrocardiogram, ECG)被广泛应用于窦性心动过速、室性早搏和心房颤动等心律失常诊断中，进而在心脏疾病诊断分析方面展现巨大的临床应用价值。为提升计算机辅助诊断心电图的分类效果，本文提出一种基于LightGBM的心电信号分类算法。该算法从心电图中提取单心拍特征、心律波动性特征以及全波形特征建立混合特征集，并采用LightGBM实现正常心拍、心房颤动、其他心律不齐、噪声四个类别的分类。最终该算法的性能指标 $F1_{nao}$ 在PhysioNet/CinC Challenge数据集上达到0.824，优于CART和CatBoost算法。同时为了加快心电图特征提取的速度，本文根据特征重要性筛选关键特征来减少分类所需的特征数量，在保持分类性能的同时将特征提取时间降为原来的17.8%。

关键词

心电图，机器学习，信号处理，LightGBM

1. 引言

心电图是放置于皮肤上的电极所记录下人体心脏电生理活动的可视化结果，通过心电图展现的心率、S-T 段、P 波、QRS 波群的形态及出现位置等特征能够诊断出窦性心动过速、窦性心律不齐、室性早搏和心房颤动等心律失常[1]。在长期的临床应用中，医生们积累了大量的 ECG 诊断规则[2]，但也正因为心电图分析中存在大量的规则和经验知识的积累，使得心电图分析技能的学习时间长，现有的心电图医生数量不足以应对海量的心电图，医生们需要计算机辅助诊断来提高诊断速度。

为此，各种基于医学经验知识的自动分类算法被提出。Tateno K 和 Glass L 使用 RR 间期和 RR 间期变化速度对心房颤动进行分析诊断[3]；宋莉、孟庆建等人利用小波变换与形态学运算相结合的波形检测算法提取 5 个时域特征、32 个小波域特征和 18 个高阶统计量特征，然后使用支持向量机进行分类[4]；Elhaj F A、Salim N 等人在使用小波分析和 QRS 复合波群检测算法提取线性和非线性波形特征后，使用支持向量机和神经网络进行分类[5]。除此之外，部分学者将特征提取的工作交给卷积神经网络来实现，并由其在完成特征提取后实现心电图分类。颜昊霖、安勇等人利用卷积神经网络在心拍 R 点前后 75 ms 的 QRS 波群上对心电图特征进行自动提取，从而减少医学先验知识需求，但是对于需要观察较长时间段心电信号的病种这些模型无法充分提取时间相关特征导致分类效果不好[6]；Kiranyaz S、Ince T 等人使用一维卷积神经网络对长时段心电信号进行特征提取并分类，但每个病人需单独训练一个模型，不适用于临床心电图诊断场景[7]。

考虑到心电信号的特征种类较多并且无法预知哪些特征对于心电信号分类比较重要，本文提出基于 LightGBM 的心电信号分类算法，在包含 314 维的混合特征集中训练出分类模型，由模型根据信息增益自行选择利于分类的特征作为分类标准。另一方面，随着移动医疗的快速发展，便携设备实现心电信号实时分类的需求也随之出现，这要求心电信号分类算法具有较快的分类速度，为此本文根据特征重要性从 314 维混合特征中筛选出较关键的 64 维特征来减少特征提取耗费的时间，进而加快算法分类速度。

2. 算法框架

本文所提出的基于 LightGBM 的心电信号分类算法流程如图 1 所示。原始心电信号由于基线漂移、工频干扰、肌电干扰噪声的影响易出现失真不利于分类[8] [9]，所以先使用 3~45 Hz 的带通滤波器对原始心电信号进行简单的去噪处理，去除大部分噪声影响；接着对去噪后的心电信号进行特征提取，得到总计 314 维的混合特征，包括单心拍特征、心律波动性特征以及全波形特征；以上特征作为基于 LightGBM 的分类器的输入，最终得到心电信号的类别标签：正常心拍、心房颤动、其他心律不齐以及噪声。

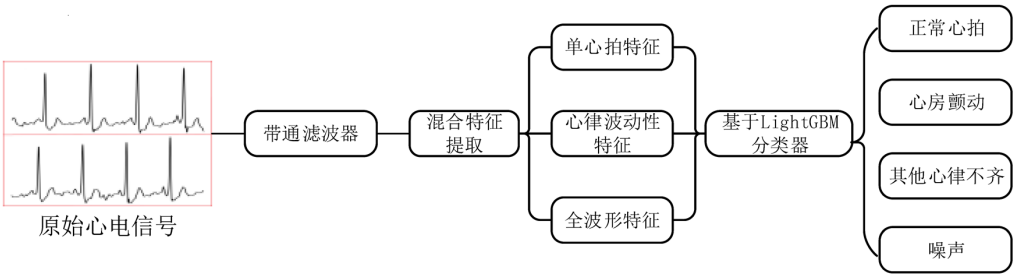

Figure 1. The flow of ECG signal's classification algorithm based on LightGBM
图 1. 基于 LightGBM 的心电信号分类算法流程

3. 算法介绍

3.1. LightGBM 算法

LightGBM 是由微软提供的针对梯度提升决策树(Gradient Boosting Decision Tree, GBDT)的改进算法，具有内存使用低、训练速度快和准确率高等优点[10]。LightGBM 算法抛弃了传统 GBDT 算法按层生长(level-wise)的决策树生长策略，采用带有深度限制的按叶子生长(leaf-wise)策略。Leaf-wise 策略分裂叶子节点的过程如图 2 所示，每次从当前所有叶子中找到分裂增益最大的一个叶子，然后分裂，如此循环。这样做可以在相同分裂次数情况下得到误差更低，性能更好的决策树，但是也容易长出深度较深的决策树因其过拟合，所以会增加一个决策树深度限制来防治过拟合的出现。

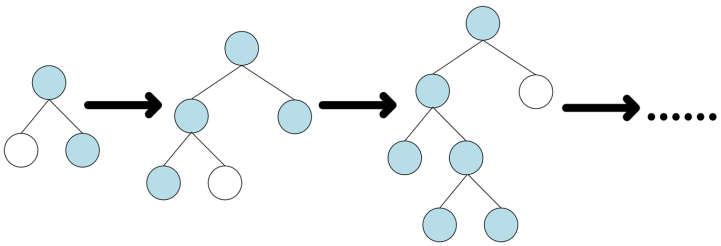

Figure 2. The leaf nodes' splitting process of leaf-wise strategy
图 2. Leaf-wise 策略叶子节点分裂过程

此外，LightGBM 相较于传统 GBDT 算法更加高效主要是因为采用直方图算法(Histogram Algorithm)和单边梯度采样算法(Grandient-based One-Side Sampling, GOSS)。直方图算法的思想是将连续浮点型的特征值用 K 个整数来离散化表示，并为样本集合的每一个特征用宽度为 K 的直方图来统计数据信息，最后直方图里的信息会被用来计算分裂增益从而选出增益最大的特征和分割点。特征直方图化的过程如图 3 所示，这样做可以使得 LightGBM 在计算分裂增益上的时间开销从 O (样本数 × 特征数)减少到 O (K × 特征数)，大幅提升算法训练效率。

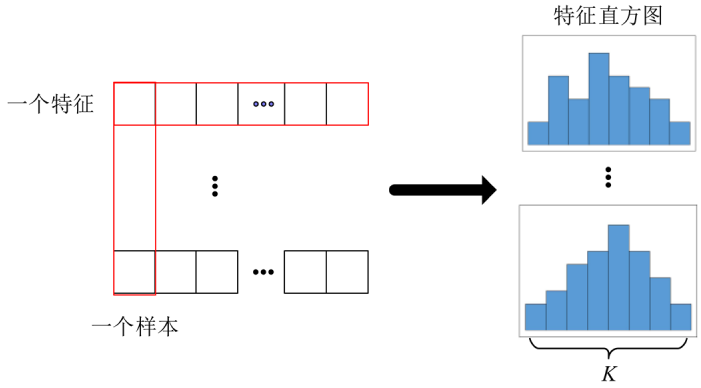

Figure 3. The process of feature histogram
图 3. 特征直方图化

GOSS 算法的改进在于仅使用部分训练样本来估算整个训练样本的分裂增益，从而提升 LightGBM 的训练速度。假设存在一个训练样本集 $S=\{x_1,x_2,\cdots,x_n\}$，特征维度为m。每次梯度迭代中，训练样本集 S 对模型输出的负向梯度记为 $\{g_1,g_2,\cdots,g_n\}$。GOSS 算法会先对训练样本集 S 的梯度绝对值进行降序排列，然后选取前 $a\times100\%$ 梯度绝对值较大的样本作为一个新的子集 A；剩余的后 $(1-a)\times100\%$ 梯度绝对值较小的样本集合记为 A^c，并从中随机采样 $b\times|A^c|$ 个样本作为子集 B；最后使用样本集合 $A\cup B$ 来计算信息增益估计值 $\tilde{V}_j(d)$，并根据该增益值来进行分割点的选择。$\tilde{V}_j(d)$ 的定义如公式(1)所示：

$$\tilde{V}_j(d)=\frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_l}g_i+\frac{1-a}{b}\sum_{x_i\in B_l}g_i\right)^2}{n_l^j(d)}+\frac{\left(\sum_{x_i\in A_r}g_i+\frac{1-a}{b}\sum_{x_i\in B_r}g_i\right)^2}{n_l^j(d)}\right) \tag{1}$$

公式(1)中，$A_l=\{x_i\in A:x_{ij}\le d\}$，$A_r=\{x_i\in A:x_{ij}>d\}$，$B_l=\{x_i\in B:x_{ij}\le d\}$，$B_r=\{x_i\in B:x_{ij}>d\}$，分别表示子集 A 中样本在第 j 个特征上特征值小于等于 d 的样本集合，大于 d 的样本集合以及子集 B 中样本在第 j 个特征上特征值小于等于 d 的样本集合，大于 d 的样本集合；而乘上系数 $(1-a)/b$ 是为了让子集 B 的梯度和尽可能接近 A^c。GOSS 算法的这种抽样方式不仅可以减少计算资源的消耗和提升训练速度，还可以在不损失过多训练精度的同时提高模型的泛化能力[10] [11]。

3.2. 混合特征提取

建立一个区分度高的特征集对心电信号分类来说十分重要。本文对心电信号采用 3~45 Hz 的带通滤波器进行去噪后，提取出总计 314 维度的混合特征用于模型的训练和测试。314 维的混合特征包括 140 维的单心拍特征、126 维的心律波动性特征以及 48 维的全波形特征三种。

单心拍特征的提取是首先对一份心电信号进行心拍检测生成单心拍模板，然后通过计算单心拍模板和所有单心拍的相关系数来区分好、坏两类心拍(坏心拍主要是由病变或者噪声叠加引起)。单心拍特征分为两部分：一部分是好心拍的 P、QRS、T 波的能量、幅值、时间间期以及相关系数的最大值、最小值、方差，Fisher 信息等时间序列指标；另一部分是坏心拍占比以及坏心拍和模板的相关系数的中值、方差等时间序列指标。

心律波动性特征的提取是先根据每份心电信号的 R 峰值点位置计算出 RR 间期；并由此计算出心率，RR 间期变化速度以及 RR 间期变化加速度；最后对心率、RR 间期、RR 间期变化速度以及加速度统计最大值、最小值、方差、Hjorth 参数、信息熵等时间序列指标，并将这些指标作为心律波动性特征。

全波形特征的提取是先对整段心电信号进行最大值、最小值、均值、中值、方差、峰态系数的计算，然后对心电信号进行连续小波变换并统计小波系数的信息熵，分形维数等时间序列指标，最后将这些指标作为全波形特征。

4. 性能试验

4.1. 实验数据集

本文所使用的心电图均来自 2017 PhysioNet/CinC Challenge 数据库，该数据库是 AliveCor 公司提供给 PhysioNet 举办的心脏病学计算竞赛的。数据库的心电图都是使用 AliveCor 设备采集的，主要用于心房颤动识别算法的训练和性能测试。数据库分为训练集和测试集两部分，测试集有 3658 份心电图信号，但是不对外公开；训练集有 8528 份心电信号，其中正常记录 5154 份，心房颤动 771 份，其他心律不齐 2557 份，噪声 46 份，数据分布信息如表 1 所示。

Table 1. The distribution information 2017 PhysioNet/CinC Challenge dataset training set
表 1. 2017 PhysioNet/CinC Challenge 数据集训练集分布信息

类别	数量	时长(s)				
		平均值	标准差	最大值	中位数	最小值
窦性心律	5154	31.9	10.0	61.0	30.0	9.0
心房颤动	771	31.6	12.5	60.0	30.0	10.0
其他心律	2557	34.1	11.8	60.9	30.0	9.1
噪声	46	27.1	9.0	60.0	30.0	10.2
总计	8528	32.5	10.9	61.0	30.0	9.0

4.2. 评价指标

本文主要使用 F_1 度量作为主要性能指标，定义为基于召回率 R 和查准率 P 的调和平均：

$$\frac{1}{F_1}=\frac{1}{2}\cdot\left(\frac{1}{P}+\frac{1}{R}\right) \tag{2}$$

对于试验中分类类别为四类的情况，定义正常类别 F_{1n}、心房颤动类别 F_{1a}、其他类别 F_{1o}、噪声类别 F_{1p} 以及最终模型性能评估指标 $F1_{nao}$ 计算公式如下， $F1_{nao}$ 的计算方式与 PhysioNet/CinC Challenge 最终采用的方式相同，即不包含样本数噪声类别：

$$F_{1n}=\frac{2\times N_n}{\sum N+\sum n} \tag{3}$$

$$F_{1a}=\frac{2\times A_a}{\sum A+\sum a} \tag{4}$$

$$F_{1o}=\frac{2\times O_o}{\sum O+\sum o} \tag{5}$$

$$F_{1p}=\frac{2\times P_p}{\sum P+\sum p} \tag{6}$$

$$F1_{nao}=\frac{F_{1n}+F_{1a}+F_{1o}}{3} \tag{7}$$

其中 N_n 表示模型正确分类的正常心电图，$\sum N$ 表示所有标注为正常的心电图数量，$\sum n$ 表示所有模型分类为正常的心电图数量，其他三类表示方式与此相同。

4.3. 算法分类精度实验

由于 PhysioNet/CinC Challenge 还未公开测试集，本文在训练集上进行 5 折交叉验证来评估算法性能。除了本文提到的 LightGBM 算法，还实现了基于 CART 和 CatBoost 算法的心电信号分类模型，并将三者进行对比，结果如表 2 所示。从表 2 中可以看出，基于 LightGBM 的心电信号分类模型的 $F1_{nao}$ 是最高的，而且每一类的 $F1$ 分数也是最高的。基于 CatBoost 的心电信号分类模型在各类的分类性能上都略次于基于 LightGBM 的心电信号分类模型。以上两个模型的分类性能都明显高于基于 CART 的心电信号分类模型，这主要得益于梯度提升算法。梯度提升算法可以使模型在每一次迭代中都会构建一个新的学习器沿着梯度反方向降低损失，从而弥补当前模型的不足。综上所述，基于 LightGBM 的心电信号分类模型在 314 维的混合特征集上的表现最好。

Table 2. The comparison of model performance's cross validation results
表 2. 模型性能交叉验证结果对比

模型算法	$F1_n$	$F1_a$	$F1_o$	$F1_p$	$F1_{nao}$
CART	0.831	0.701	0.614	0.433	0.715
CatBoost	0.893	0.822	0.727	0.5909	0.814
LightGBM	0.900	0.829	0.743	0.620	0.824

4.4. 算法分类速度提升实验

在本文所提出的算法中，主要的时间消耗在于特征提取上，所以要想提升算法的分类速度首要是降低特征提取的耗时，其最直接的方法是减少特征维数，这也是本文所要研究的内容。混合特征在基于 LightGBM 的心电信号分类算法中前六大特征及其重要性如图 4 所示，其重要性的衡量依据是特征在基于 LightGBM 的分类器中被用于叶子节点分裂的次数，图中数值代表各特征被使用次数和单个特征最大被使用次数的比值。从图 4 中可以看出，P 波的相关特征占了三个，这说明 P 波相关特征在该模型的分类中起到很大的作用，这和传统医学常用 P 波作为诊断依据相符合。

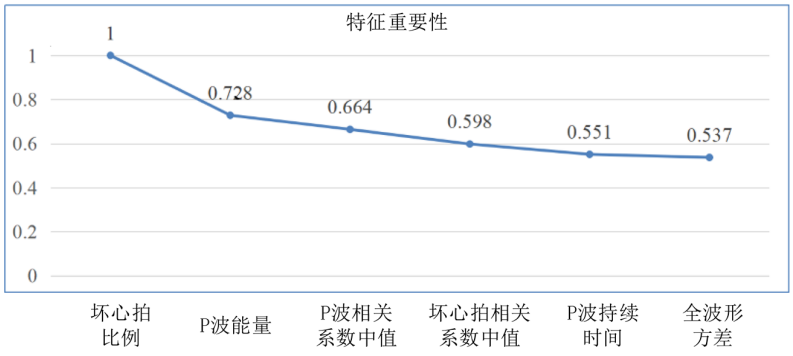

Figure 4. The first six features of mixed features and their importance
图 4. 混合特征的六大特征及其重要性

根据以上所述的特征重要性，本文从 314 维混合特征中筛选出被基于 LightGBM 的分类器使用次数最多的 64 维特征作为新的特征集用于重新训练基于 LightGBM 的分类器。特征筛选前后所训练出的基于

LightGBM 的分类器分类性能和特征提取耗时对比如表 3 所示，从表中可以看出特征筛选后所训练出的基于 LightGBM 的分类器分类性能和筛选前相差微小，特征提取耗时均值减少到 0.439 s，约为筛选前平均耗时的 17.8%，提速效果显著。

Table 3. The performance comparison before and after feature selection
表 3. 特征筛选前后性能对比

		特征筛选前	特征筛选后
特征维数		314	64
5 折交叉验证平均 $F1_{nao}$		0.824	0.820
特征提取耗时(s)	均值	2.465	0.439
	方差	0.476	0.073
	最大值	4.021	0.645
	最小值	1.599	0.293

5. 结论

本文提出的基于 LightGBM 的心电信号分类算法在从 PhysioNet/CinC Challenge 数据库提取的 314 维混合特征集上 $F1_{nao}$ 达到 0.824，优于使用相同特征集的 CART 和 CatBoost 算法。考虑到 314 维混合特征存在冗余会影响心电信号分类算法的速度，本文根据特征重要性重新筛选出前 64 维关键特征作为分类依据，实验证明特征筛选后算法的分类性能基本不变，特征提取速度降为原先的 17.8%，效果显著。

参考文献

[1] Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., *et al*. (2017) Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks.

[2] 卢喜烈. 现代心电图诊断大全[M]. 北京: 科学技术文献出版社, 1997.

[3] Tateno, K. and Glass, L. (2001) Automatic Detection of Atrial Fibrillation Using the Coefficient of Variation and Density Histograms of RR and deltaRR Intervals. *Medical & Biological Engineering & Computing*, **39**, 664-671. https://doi.org/10.1007/BF02345439

[4] 宋莉, 孟庆建, 张光玉, 等. 基于波形特征和 SVM 的心电信号自动分类方法研究[J]. 中国医学物理学杂志, 2010(4): 2043-2046.

[5] Elhaj, F.A., Salim, N., Harris, A.R., *et al*. (2016) Arrhythmia Recognition and Classification Using Combined Linear and Nonlinear Features of ECG Signals. *Computer Methods & Programs in Biomedicine*, **127**, 52-63. https://doi.org/10.1016/j.cmpb.2015.12.024

[6] 颜昊霖, 安勇, 王宏飞, 等. 基于卷积神经网络的心电特征提取[J]. 计算机工程与设计, 2017(4): 1024-1028.

[7] Kiranyaz, S., Ince, T. and Gabbouj, M. (2016) Real-Time Patient-Specific ECG Classification by 1D Convolutional Neural Networks. *IEEE Transactions on Biomedical Engineering*, **63**, 664-675. https://doi.org/10.1109/TBME.2015.2468589

[8] 王润. 心电信号的预处理算法分析[J]. 现代计算机(专业版), 2018(7): 33-36.

[9] Barhatte, A.S., Ghongade, R. and Tekale, S.V. (2016) Noise Analysis of ECG Signal Using Fast ICA. 2016 *Conference on Advances in Signal Processing* (*CASP*) *Cummins College of Engineering for Women*, Pune, 9-11 Jun 2016, 1-5. https://doi.org/10.1109/CASP.2016.7746149

[10] Ke, G., Meng, Q., Finley, T.W., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Neural Information Processing Systems*, Long Beach, CA, 4-9 December 2017, 3149-3157.

[11] Zhou, Z.-H. (2012) Ensemble Methods: Foundations and Algorithms. CRC Press, New York, 99-118. https://doi.org/10.1201/b12207