DOI:10.3969/j.issn.1005-202X.2021.01.013

医学信号处理与医学仪器

基于机器学习的心音识别分类研究

马晶,蔡文杰,杨利
上海理工大学医疗器械与食品学院,上海 200093

【摘要】心音信号可反映心脏的病理信息,是诊断心脏健康的重要依据之一。本文首先从心音信号提取时频域、梅尔倒谱系数等145个特征作为机器学习的输入数据集,然后在随机森林、LightGBM、XGBoost、GBDT、SVM共5种分类器中选出效果最佳分类器与递归特征消除算法结合进行数据挖掘,找出重要特征集并对其分类效果做比较与分析,最后运用Stacking模型融合方法优化模型。数据挖掘特征子集比同数量特征子集在准确率、召回率、精确率、F1值上分别提高了33.51%、14.54%、20.61%、24.04%;采用LightGBM和SVM模型融合可将F1值提高至92.6%。本文提出了一种有效的心音识别分类方法,挖掘出心音最重要的8个特征,为临床诊断提供参考。

【关键词】机器学习;心音分类;数据挖掘;模型融合
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2021)01-0075-05

Research on heart sounds classification based on machine learning

MA Jing, CAI Wenjie, YANG Li
College of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: The heart sound signals can reflect the pathological information of the heart and is therefore one of the important evidences for the diagnosis of heart diseases. In this paper, 145 features, such as time-frequency domain and Meier cepstrum coefficient, were extracted from heart sound signals as input data sets for machine learning, and then the best classifiers are selected from five classifiers, namely Random Forest, LightGBM, XGBoost, GBDT and SVM. The classifier is combined with the recursive feature elimination algorithm to conduct data mining, in order to find out the important feature set and compare and analyze its classification effect. Finally, the Stacking model fusion method is applied to optimize the model. The data mining feature subset improved the accuracy, recall rate, precision and F1 value by 33.51%, 14.54%, 20.61% and 24.04% respectively over the same number of feature subsets; the fusion of LightGBM and SVM models improved the F1 value to 92.6%. In this paper, an effective heart sound recognition classification method is proposed to mine the eight most important features of heart sounds for clinical diagnosis.

Keywords: machine learning; classification of heart sounds; data mining; model fusion

前 言

受人口老龄化等因素影响,我国心血管疾病的发病率和死亡率近年来快速增长。据国家心血管病中心编撰的《中国心血管病报告2016》显示,心血管病死亡占城乡居民死亡原因的首位,占疾病死亡构成的40%以上,及早诊断与及时治疗是应对这一危机的有效措施[1]。心脏听诊操作简单,但主要靠医生的主观经验来判断病症。在大数据的热潮下,通过机器学习方法来判断心脏是否健康成为了一个备受关注的研究热点。机器学习是人工智能的核心,通过计算机模拟和实现人类的学习行为,利用已有数据寻找规律,并对未知的数据进行预测,目前已广泛应用于图像、音频、语言处理等领域[2]。

机器学习已在心脏信号辅助诊断方面取得了很多成功。Zheng等[3]提出了基于心脏储备指数的混合特征提取方法,并采用最小二乘支持向量机实现慢性心力衰竭计算机辅助智能诊断,诊断准确率、灵敏度和特异性均达到了90%。Hadi等[4]提出了用S-Transform(S变换)技术提取特征,并运用多层感知器网络进行心音病例的分类,达到了很高的正确分类率。Maragoudakis等[5]提出了一种基于马尔科夫链的贝叶斯推理方法,在198个心音信号数据集中已得

【收稿日期】2020-05-15
【基金项目】上海市浦江人才计划项目(15PJ1406100);国家自然科学基金重点项目(31830042)
【作者简介】马晶,硕士研究生,主要研究方向:机器学习,E-mail: 1183794950@qq.com
【通信作者】蔡文杰,博士,副教授,主要研究方向:医学人工智能,E-mail: wenjiecai@aliyun.com

到验证，分类效果高于其他分类器。

心音是心脏搏动过程中产生的一种振动信号，能够很好地反映心脏活动、血液流动和心脏的健康情况，特别在诊断血流动力学异常方面比心电图更具优势[6]。心音在疾病诊断中提供初步线索，有助于医生对疾病进行评估，因其方便快捷在临床上被广泛应用[7]。目前将数据挖掘应用到分析心音特征重要性的研究非常少见。本研究将基于多种新型机器学习算法的心音特征进行分类预测，并运用数据挖掘算法分析心音特征的重要性排序，实现了高准确率的心音正异常分类。

1 实验方法

1.1 数据集的准备

1.1.1 获取数据集 实验数据来源于physionet的开源数据库[8]。心音记录（wav格式）共3 153条，其中正常心音记录2 488条、异常心音记录665条。

1.1.2 心音特征提取 特征提取的第一步是心音分割。基于隐半马尔科夫模型的分割算法是目前分割心音效果较好的算法[9]。图1为运用隐半马尔科夫模型分割心音的结果，可看出图中信号共有5个完整的心动周期。

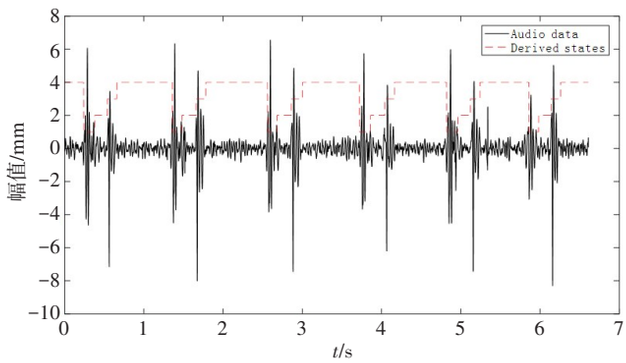

图1 心音信号分割结果

Fig.1 The result ofheart sound signal segmentation

特征提取算法主要有时频域、梅尔频率倒谱系数（Mel Frequency Cepstral Coefficents, MFCC）[10]、连续小波变换。时域分析方法是将信号$x(t)$的能量分布表示为时间t的函数；频域分析方法是通过傅里叶变换获得信号频域及其能量频域分布。信号$x(t)$的小波变换为：

$$WT_X(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}x(t)\psi^*\left(\frac{t-b}{a}\right)\mathrm{d}t \qquad (1)$$

其中，a为伸缩因子，b为平移因子，$\psi(t)$为母小波。本文取单频率正弦Morlet小波为母小波，运用Matlab小波工具箱中的一维连续变换函数cwt对信号进行小波变换。

MFCC表达了一种常用的从语言频率到“感知频率”的对应关系[11]。通常对梅尔频率[12]做以下转换：

$$f_{\mathrm{mel}}=2\ 595\ \log_{10}\left[1+\frac{f}{700}\right] \qquad (2)$$

MFCC特征提取的具体步骤为：对信号进行预加重，帧移位取10 ms，帧长取25 ms对信号进行分帧，加海明窗并进行快速傅里叶变换，通过26个滤波器进行滤波和刻度转换，对取对数能量后的信号做离散余弦变换和均值归一化。

1.1.3 心音特征数据预处理 由于数据样本的异常样本数量要远少于正常样本数量，若数据非线性可分，则会导致在分类过程中判断失误，分类效率严重下降。本研究采用smote人工合成数据的方法扩充异常心音数据[13]。smote算法的实现步骤可描述为：①从所有异常样本中找到异常样本x_m的K个近邻，记为n。②从n中随机抽取一个样本x_n，生成一个范围为(0,1]的随机数δ，按照$x_{m1}=x_m+\delta*(x_n-x_m)$，生成新样本$x_{m1}$。③重复步骤②$i$次，即可生成$i$个新样本。

对缺失值的处理方法为：用数据集中缺失值所在的特征（列）的均值来代替缺失值。使用零均值标准化方法对数据集进行归一化处理。

1.2 心音特征分类器工作原理及超参数优化

所用机器学习分类器主要有Light Gradient Boosting Machine（LightGBM）、支持向量机（Support Vector Machine, SVM）、eXtreme Gradient Boosting（XGBoost）、梯度提升决策树（Gradient Boosting Decision Tree, GBDT）、随机森林。本文采用Python语言调用分类器工具包实现该功能。

LightGBM采用梯度提升的方式，将基于学习算法的分类树和回归树进行有效的叠加。SVM建立在统计学习理论的VC理论和结构风险最小化原理基础之上[14]。SVM的主要思想是将训练数据通过一定的函数变化到高维度的空间，在高维空间寻找最优的分类面。XGBoost是华盛顿大学陈天奇对Gradient Boosting Machine算法的C++实现[15]，能够自动利用单机CPU的多核进行并行计算降低计算复杂度。GBDT最早由Friedman提出这个概念。GBDT通过采用加法模型，以及不断减小训练过程中产生的残差来将数据进行分类[15]。随机森林是一个包含多个决策树的分类器，并将它们合并在一起以获得准确率更高的预测[16]。

在实际应用中比较常用的参数优化方法有网格搜索法和随机参数优化法。本文使用随机参数优化法寻找最优参数。

1.3 数据挖掘

数据挖掘是从大量、随机的数据中，提取隐含在其中并且潜在有用信息的过程[17]。本研究采用递归特征消除（Recursive Feature Elimination, RFE）方法进行数据挖掘，其通常和很多分类算法联合使用[18]。RFE的主要流程为：（1）用预处理后的特征训练所选择的分类器；（2）按照最优特征子集进行特征的筛选，保留最优特征。

1.4 分类效果评价标准

在分类学习中，常用的分类评估指标有准确率（Accuracy）、灵敏度（Sensitivity）、精确率（Precision）、F1值。在实验中，TP代表将正常信号预测为正常的总数，TN代表将异常信号预测为异常的总数，FP代表将异常信号预测为正常的总数，FN代表将正常信号预测为异常的总数。

准确率表达式为：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

灵敏度表达式为：

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

精确率表达式为：

$$\text{Precision} = \frac{TP}{TP + FN} \tag{5}$$

F1值表达式为：

$$F1 = \frac{2 \times \text{Accuracy} \times \text{Sensitivity}}{\text{Sensitivity} + \text{Accuracy}} \tag{6}$$

2 结　果

2.1 心音特征提取及数据扩充

心音信号经过滤波、去波峰、降采样等处理后，运用隐半马尔科夫模型分割每个心动周期的S1、收缩期、S2、舒张期，利用时域、频域、MFCC以及小波变换算法共提取145个特征。第1~36个特征为时域特征；第36~72个特征为频域特征；第73~90个特征为MFCC特征；第91~145个特征为小波变换、时间复杂度和光谱复杂度特征，提取的特征数据集的维度为3 153×145。在正常和异常心音特征数据中随机抽取800条作为测试集，其余数据作为训练集和验证集，通过smote算法将异常的训练数据由465条扩充为1 860条，即扩充之后训练集和验证集中正常心音特征数据与异常心音特征数据分别为1 888条和1 860条，比例接近1∶1。表1为在采用相同分类器LightGBM的情况下原始数据集与扩充数据集的分类效果对比，可看出扩充前的各项分类指标都较低，扩充后数据集比原始数据集在分类准确率、灵敏度、精确度、F1值上分别提高了16.05%、24.6%、12.92%、20.58%，可证明数据扩充的必要性和有效性。

表1 扩充数据前后分类效果对比图（%）

Tab.1 Comparison of classification effects before and after data expansion (%)

数据集	准确率	灵敏度	精确率	F1值
原始数据集	74.27	65.8	78.24	69.78
扩充后数据集	90.32	90.4	91.16	90.36

2.2 心音特征分类效果评估

将数据集随机打乱，按照4∶1的比例分配给训练集和验证集。测试集的维度为800×145。实验中分别采用LightGBM、SVM、XGBoost、GBDT、随机森林对数据进行训练和预测。参数优化结果如下：对于LightGBM，learning_rate=0.15，max_depth=60，n_extimators=150，num_leaves=300；对于SVM模型，高斯径向基函数为核函数，C=1，gamma=0.01；对于XGBoost模型，colsample_bytree=0.8，gamma=0.1，learning_rate=0.3，max_depth=10，reg_alpha=0.1，min_child_weight=1，subsample=0.6；对于GBDT，learning_rate=0.1，max_depth=40，n_extimators=200，min_samples_split=70；对于随机森林模型，max_depth=156，max_features=30；n_extimators=120，min_samples_leaf=70。

经过填充均值、归一化、调节最优参数等处理后，各分类器预测后得到的分类效果如表2表示，其中各分类指标结果均为正常与异常两种类别的的平均值。

在分类准确率方面，由表2可看出，在相同的数据集下，LightGBM效果最好，准确率为90.32%，XGBoost次之。由于正样本更被医学上重视，将正样本误判为负样本的代价要远远大于将负样本误判为正样本的代价，因此灵敏度、精确率、F1值往往也是很好的评判标准。每个分类器指标中的准确率、灵敏度、精确率、F1值差距较小，除了随机森林外基本差距在1%左右。结合这4个指标可看出，LightGBM效果最好。

表2 各分类器心音分类效果（%）

Tab.2 Heart sounds classification effects of each classifier (%)

方法	准确率	灵敏度	精确率	F1值
LightGBM	90.32	90.40	91.16	90.36
SVM	86.76	86.80	89.24	86.78
XGBoost	87.40	87.80	89.18	87.60
GBDT	84.18	84.40	85.40	84.29
随机森林	71.28	72.00	79.50	71.64

2.3 数据挖掘结果

根据分类效果的对比可看出LightGBM效果最好，故本文选用LightGBM与RFE结合，选择LightGBM算法作为递归特征消除算法的分类器。通过数据挖掘从145个特征中提取出重要性前8名的特征，分别是第75、73、77、17、78、93、79、52个特征，代表的含义依次为“倒谱系数偏度中心距离”、“倒谱系数最小值”、“信号谱分析系数”、“收缩期和S1振幅比”、“信号谱分析频率”、“小波变换谱系数”、“线性预测系数”、“收缩期加窗傅里叶变换中位数值”。

在数据集划分比例相同、分类器均为LightGBM、调参参数设置相同的情况下，这8个特征构成一个特征子集，分别将145个特征组成的数据子集和从145个特征中随机抽取8个特征组成的特征子集进行比较评估各项指标。对比结果如表3所示，可以看出：数据挖掘特征子集与由145个特征组成的特征集分类效果差不多，而随机抽取的特征子集从准确率、灵敏度、精确率、F1值上看效果都跟前两者有很大差距。数据挖掘的特征子集比随机抽取特征子集分类准确率高0.225 1，提高了33.51%；灵敏度高0.114，提高了14.54%；精确率高0.155，提高了20.61%；F1值高0.173 9，提高了24.04%。以上足以证明数据挖掘的8个特征的重要性。

表3 数据挖掘效果对比图(%)

Tab.3 Comparison chart of data mining (%)

数据集	准确率	灵敏度	精确率	F1值
原数据子集	90.32	90.40	91.16	90.36
数据挖掘特征子集	89.68	89.80	90.69	89.74
随机抽取等量特征子集	67.17	78.40	75.19	72.35

2.4 模型融合

为了提高模型的性能，采用Stacking模型融合方法将所用的5种模型进行5折交叉验证输出预测结果，然后将其合并为新特征并使用新模型加以训练得到最终结果。图2展示了不同数量模型融合与分类指标F1值的关系图，分别取LightGBM、XGBoost、SVM、GBDT、RandomForest这5个模型的第1个、前2个、前3个、前4个、前5个模型做模型融合。实验结果表明，LightGBM、XGBoost和SVM模型融合的效果最好，准确率、灵敏度、精确率和F1值分别为92.6%、92.6%、92.7%、92.6%。

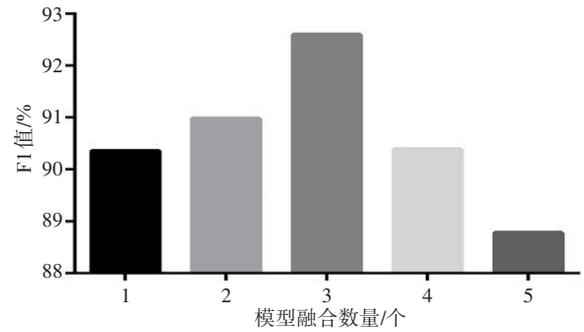

图2 模型数量与分类指标F1值关系图

Fig.2 Relationship between number of models and F1 value

3 讨 论

近年来，智能辅助诊断成为了不少学者研究的热点，在特征分类和图像分类方面有很好的应用。在实际训练时，为了提高模型的鲁棒性和泛化能力，本研究所收集的数据来源于physionet的公开数据集，并从中剔除了信号。在心音特征分类诊断过程中，采取了多种方法解决分类不平衡问题，发现smote人工合成数据能够有效扩充异常信号；对比各分类器的分类效果，发现相比于传统机器学习方法，新兴机器学习方法如lightGBM在保持各项分类指标的同时大大提高了工作效率，在实际应用中是一个很好的优点。

通过数据挖掘的分类效果对比可看出，数据挖掘出的特征子集在各项指标中都表现很好。145个特征组成的特征集虽然在准确率、灵敏度、精确率、F1值这4个指标上略高于数据挖掘特征子集，但是数据量却是后者的19倍。在临床实际应用中，不影响分类效果的前提下，提取的特征数目由145个减少到8个，这不仅大大减少了特征提取所需要的时间，也提高了机器学习训练和测试的效率，可以在相同的时间内进行更多的预测。数据挖掘出来的8个特征，包含MFCC、时频域分析等，对于诊断心血管疾病来说是重要的指标，具有较强的临床参考价值。

运用Stacking模型融合方法将多种分类器结合以提高模型性能，是有效提高分类效果的方法之一，但并不是分类器越多越好，若某个单独分类器本身性能不佳则会影响整个模型的效果。本研究结合了多种心音特征类型进行心音诊断，目的是充分利用心音的信息以提高分类准确率，运用Stacking模型融合方法准确率可达到92.6%。尽管本文的初步研究取得良好的心音分类效果，但physionet的心音数据来源于国外多家医院，不同品牌数字听诊器和不同环境下采集的心音信号可能会夹杂一定的干扰信息，应进一步进行广泛的研究和测试。

4 结 论

本研究通过将机器学习与递归特征消除算法相

结合挖掘到其中最重要的8个特征,在提高模型运行效率的同时也从某些程度上代表了诊断心血管疾病的参考指标,并采用Stacking模型融合方法将心音分类准确率提高至92.6%,该方法具有潜在的临床应用价值。

【参考文献】

[1] 陈伟伟, 高润霖, 刘力生, 等.《中国心血管病报告2016》概要[J]. 中国循环杂志, 2017, 32(6): 521-530.
CHEN W W, GAO R L, LIU L S, et al. China cardiovascular disease report 2016[J]. Chinese Circulation Journal, 2017, 32(6): 521-530.

[2] BAUMGARTNER D, SERPEN G. A design heuristic for hybrid classification ensembles in machine learning[J]. Intelligent Data Analysis, 2012, 16(2): 233-246.

[3] ZHENG Y N, GUO X M, QIN J, et al. Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics[J]. Comput Methods Programs Biomed, 2015, 122(3): 372-383.

[4] HADI H M, MASHOR M Y, SUBOH M Z, et al. Classification of heart sound based on s-transform and neural network[C]. International Conference on Information Sciences Signal Processing and Their Applications. IEEE, 2010: 189-192.

[5] MARAGOUDAKIS M, LOUKIS E. Heart sound screening in real-time assistive environments through MCMC Bayesian data mining[J]. Universal Access in the Information Society, 2014, 13: 73-88.

[6] 张家亮, 江洪, 阙大顺, 等. 基于小波的心音信号分析及其特征提取[J]. 电脑与信息技术, 2011, 19(1): 17-20.
ZHANG J L, JIANG H, QUE D S, et al. Heart sound signal analysis and feature extraction based on wavelet[J]. Computer and Information Technology, 2011, 19(1): 17-20.

[7] SUN S P, WANG H B, JIANG Z W, et al. Segmentation-based heart sound feature extraction combined with classifier models for a VSD diagnosis system[J]. Expert Syst Appl, 2014, 41(4): 1769-1780.

[8] LIU C Y, SPRINGER D, LI Q, et al. An open access database for the evaluation of heart sound algorithms[J]. Physiol Meas, 2016, 37(12): 2181-2213.

[9] SPRINGER D B, TARASSENKO L, CLIFFORD G D. Logistic regression-HSMM-based heart sound segmentation[J]. IEEE Trans Biomed Eng, 2016, 63(4): 822-832.

[10] 胡峰松, 张璇. 基于梅尔频率倒谱系数与翻转梅尔频率倒谱系数的说话人识别方法[J]. 计算机应用, 2012, 32(9): 2542-2544.
HU F S, ZHANG X. Speaker recognition method based on Mel frequency cepstrum coefficient and inverted Mel frequency cepstrum coefficient[J]. Journal of Computer Applications, 2012, 32(9): 2542-2544.

[11] KAY E, AGARWAL A. Drop connected neural networks trained on time-frequency and inter-beat features for classifying heart sounds[J]. Physiol Meas, 2017, 38(8): 1645-1657.

[12] EL-SAYED A A, MAHMOOD M A, MEGUID N A, et al. Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE)[C]//2015 Third World Conference on Complex Systems (WCCS). Marrakech, Morocco: IEEE, 2016: 1-5.

[13] 李喆, 吕卫, 闵行, 等. 机器学习在乳腺肿瘤分类检测中的应用研究[J]. 计算机工程与科学, 2016, 38(11): 2303-2309.
LI Z, LÜ W, MIN H, et al. Application of machine learning algorithms in breast tumor detection[J]. Computer Engineering & Science, 2016, 38(11): 2303-2309.

[14] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data. ACM, 2016: 785-794.

[15] SON J, JUNG I, PARK K, et al. Tracking-by-segmentation with online gradient boosting decision tree [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2016: 3056-3064.

[16] 屠睿博, 陈中华, 王洪凯. 基于随机森林算法的小鼠micro-CT影像中骨骼关节特征点定位[J]. 中国生物医学工程学报, 2017, 36(3): 257-266.
TU R B, CHEN Z H, WANG H K. Bone joints localization in mouse micro-CT images using random forests algorithm[J]. Chinese Journal of Biomedical Engineering, 2017, 36(3): 257-266.

[17] 陈鸿俊. 基于数据挖掘技术的移动互联网业务研究[J]. 计算机与数字工程, 2017, 45(8): 1597-1600.
CHEN H J. Research on mobile internet service based on data mining technology[J]. Computer & Digital Engineering, 2017, 45(8): 1597-1600.

[18] GRANITTO P M, FURLANELLO C, BIASIOLI F, et al. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products[J]. Chemometr Intell Lab Syst, 2006, 83(2): 83-90.

(编辑:薛泽玲)