



(12)发明专利申请

(10)申请公布号 CN 108030494 A

(43)申请公布日 2018.05.15

(21)申请号 201711088016.2

G06K 9/68(2006.01)

(22)申请日 2017.11.08

(71)申请人 华南理工大学

地址 511458 广东省广州市南沙区环市大道南路25号华工大广州产研院

(72)发明人 崔巍 李耀光

(74)专利代理机构 广州市华学知识产权代理有限公司 44245

代理人 李斌

(51)Int.Cl.

A61B 5/0452(2006.01)

A61B 5/0402(2006.01)

A61B 5/00(2006.01)

G06K 9/00(2006.01)

G06K 9/62(2006.01)

权利要求书1页 说明书5页 附图2页

(54)发明名称

基于交叉验证的心电信号错误标记训练样本识别方法

(57)摘要

本发明公开了一种基于交叉验证的心电信号错误标记训练样本识别方法,包括以下步骤:一、对心电信号采样数据进行预处理,并对预处理后的心电信号采样数据进行特征提取;二、对心电信号采样数据提取出来的特征进行标准化处理,并对标准化后的特征进行PCA降维;三、利用交叉验证和多种机器学习分类算法对心电信号采样数据进行错误标记识别;四、将识别出的带有错误标记的心电信号采样数据从训练集中移除,再训练分类器,并对测试集进行分类。所述方法能够有效地识别出心电信号训练样本中的错误标记样本,进而提高心电信号自动分类的准确率。



1. 一种基于交叉验证的心电信号错误标记训练样本识别方法,其特征在于,所述方法包括以下步骤:

一、对心电信号采样数据进行预处理,并对预处理后的心电信号采样数据进行特征提取;

二、对心电信号采样数据提取出来的特征进行标准化处理,并对标准化后的特征进行PCA降维,将降维后的心电信号采样数据作为训练集;

三、利用交叉验证和多种机器学习分类算法对训练集中的心电信号采样数据进行错误标记识别;

四、将识别出的带有错误标记的心电信号采样数据从训练集中移除,再重新训练分类器,并对测试集进行分类。

2. 根据权利要求1所述的一种基于交叉验证的心电信号错误标记训练样本识别方法,其特征在于,所述对心电信号采样数据进行预处理具体包括:利用中值滤波方法对心电信号采样数据进行处理以去除基线漂移,再对中值滤波后的心电信号采样数据利用小波变换方法去除工频干扰。

3. 根据权利要求1所述的一种基于交叉验证的心电信号错误标记训练样本识别方法,其特征在于,所述对预处理后的心电信号采样数据进行特征提取的步骤具体包括:利用小波变换方法检测预处理后的心电信号采样数据中每个心拍的R波峰值位置,对心电信号采样数据依据R波峰值位置进行心拍分割,并检测出该心拍的Q波起始位置和S波终止位置,进而提取出所需的时域特征,即该心拍的QRS波持续时间、R-R间隔以及平均R-R间隔;同时将每个心拍的数据值统一化处理为300个数据值,以代表每个心拍的形态学特征。

4. 根据权利要求3所述的一种基于交叉验证的心电信号错误标记训练样本识别方法,其特征在于,所述对心电信号采样数据提取出来的特征进行标准化处理的具体过程为:将提取出来的3个时域特征和300个形态学特征共303个特征利用最小最大标准化方法化为0~1之间的数值,从而使得两种不同的特征对分类方案的作用力同趋化。

5. 根据权利要求1所述的一种基于交叉验证的心电信号错误标记训练样本识别方法,其特征在于:所述对标准化后的特征进行PCA降维是将标准化后的特征的维数降到包含原始特征信息的95%的维数。

6. 根据权利要求1所述的一种基于交叉验证的心电信号错误标记训练样本识别方法,其特征在于,所述步骤三的具体过程为:将带有错误标记样本的训练集随机等分为10份,每次选取其中1份作为验证集,剩下9份作为小训练集用于训练多种机器学习分类器,并对验证集进行分类,分类结果和验证集原来的标记进行对比,找出错误标记样本;重复以上过程10次,每次选取的验证集都不同,以保证训练集中所有样本都被验证;找出训练集中所有的错误标记样本后,将其从训练集中移除,得到新的训练集,再重新训练分类器,并在测试集上进行测试。

7. 根据权利要求6所述的一种基于交叉验证的心电信号错误标记训练样本识别方法,其特征在于:所述多种机器学习分类器采用多种机器学习分类算法训练得到,包括朴素贝叶斯算法、线性判别分析、最近邻法则、支持向量机以及决策树算法,当5种算法中有4种以上认为某个样本为错误标记样本的话,模型判断该样本为错误标记样本,其中,模型对样本的判断能够重复多次,以增加可靠性。

基于交叉验证的心电信号错误标记训练样本识别方法

技术领域

[0001] 本发明涉及生物医学信息处理领域,具体涉及一种基于交叉验证的心电信号错误标记训练样本识别方法。

背景技术

[0002] 近年来,心脏疾病已经越来越成为威胁人类生命安全的关键问题,如何能够尽早地检测出心脏疾病是解决该问题的核心所在,因为心脏疾病很容易引发猝死,这样的情况造成了心脏疾病的检测往往比治疗更加迫在眉睫。随着计算机技术的发展,利用机器学习算法对心电信号进行自动分析并做出诊断已经得到了广泛的应用,但在心电信号的自动分析中,错误标记样本是广泛存在的,其原因可能是因为专家的诊断失误,也有可能是在预处理或特征提取等阶段出现误差等等。这些错误标记样本会对分类器产生负面影响,影响分类器的建立,进而降低分类的准确率,使得实际临床辅助诊断可靠性下降。

[0003] 在机器学习领域,解决训练中样本错误标记问题的办法一般有两种:(1) 识别出这些错误标记的样本并将其移出训练集;(2) 设计能够抵抗这些错误标记干扰的更好的分类器。本发明基于交叉验证的心电信号错误标记训练样本识别方法采用方法(1),并结合时域特征和形态学特征对心拍的特点进行识别,其中时域特征可以反映心拍最基本的心率特点,形态学特征则详细反映了心拍电特性的全貌。时域特征和形态学特征的结合可以较为准确地表示一个心拍的关键特点。

发明内容

[0004] 本发明的目的是针对现有技术的不足,提供了一种基于交叉验证的心电信号错误标记样本识别方法,所述方法利用交叉验证的方法和多种机器学习算法对心电信号中的错误标记样本进行识别,并将识别出来的错误标记样本移除,进而提高了心电信号自动分类的准确率。

[0005] 本发明的目的可以通过如下技术方案实现:

[0006] 一种基于交叉验证的心电信号错误标记训练样本识别方法,所述方法包括以下步骤:

[0007] 一、对心电信号采样数据进行预处理,并对预处理后的心电信号采样数据进行特征提取;

[0008] 二、对心电信号采样数据提取出来的特征进行标准化处理,并对标准化后的特征进行PCA降维,将降维后的心电信号采样数据作为训练集;

[0009] 三、利用交叉验证和多种机器学习分类算法对训练集中的心电信号采样数据进行错误标记识别;

[0010] 四、将识别出的带有错误标记的心电信号采样数据从训练集中移除,再重新训练分类器,并对测试集进行分类。

[0011] 进一步地,所述对心电信号采样数据进行预处理具体包括:利用中值滤波方法对

心电信号采样数据进行处理以去除基线漂移,再对中值滤波后的心电信号采样数据利用小波变换方法去除工频干扰。即心电信号采样数据序列中一点的值用该点的一个邻域中各点的中值代替,从而达到去除基线漂移的目的,对中值滤波后的心电信号采样数据进行小波变换,使得50Hz/60Hz的工频干扰和有用心电信号的主要能量落在不同尺度上,从而达到去除工频干扰的目的。

[0012] 进一步地,所述对预处理后的心电信号采样数据进行特征提取的步骤具体包括:利用小波变换方法检测预处理后的心电信号采样数据中每个心拍的R波峰值位置,对心电信号采样数据依据R波峰值位置进行心拍分割,并检测出该心拍的Q波起始位置和S波终止位置,进而提取出所需的时域特征,即该心拍的QRS波持续时间、R-R间隔(该心拍的R波峰值点和上一个心拍R波峰值点之间的时间间隔)以及平均R-R间隔(该心拍之前的10个心拍的平均R-R间隔);同时将每个心拍的数据值统一化处理为300个数据值,以代表每个心拍的形态学特征。

[0013] 进一步地,所述对心电信号采样数据提取出来的特征进行标准化处理的具体过程为:将提取出来的3个时域特征和300个形态学特征共303个特征利用最小最大标准化方法化为0~1之间的数值,从而使得两种不同的特征对分类方案的作用力同趋化。

[0014] 进一步地,所述对标准化后的特征进行PCA降维是将标准化后的特征的维数降到包含原始特征信息的95%的维数。

[0015] 进一步地,所述步骤三的具体过程为:将带有错误标记样本的训练集随机等分为10份,每次选取其中1份作为验证集,剩下9份作为小训练集用于训练多种机器学习分类器,并对验证集进行分类,分类结果和验证集原来的标记进行对比,找出错误标记样本;重复以上过程10次,每次选取的验证集都不同,以保证训练集中所有样本都被验证;找出训练集中所有的错误标记样本后,将其从训练集中移除,得到新的训练集,再重新训练分类器,并在测试集上进行测试。

[0016] 进一步地,所述多种机器学习分类器采用多种机器学习分类算法训练得到,包括朴素贝叶斯算法、线性判别分析、最近邻法则、支持向量机以及决策树算法,当5种算法中有4种以上认为某个样本为错误标记样本的话,模型判断该样本为错误标记样本,其中,模型对样本的判断能够重复多次,以增加可靠性。

[0017] 本发明与现有技术相比,具有如下优点和有益效果:

[0018] 1、本发明采用多种机器学习算法对包含错误标记样本的心电信号训练集进行交叉验证,识别率很高,误识率较低,能够有效地将训练集中的错误标记样本剔除,从而提高心电信号分类准确率。

[0019] 2、本发明基于交叉验证的心电信号错误标记训练样本识别方法改善了由于训练集中存在错误标记样本而导致分类器分类准确率下降的问题,在错误标记样本占比不高于20%的情况下,可以使得分类准确率回升到和没有错误标记样本非常接近的水平。

附图说明

[0020] 图1为本发明实施例基于交叉验证的心电信号错误标记样本识别方法的流程图。

[0021] 图2为本发明实施例正常心电周期波形示意图。

[0022] 图3为本发明实施例利用交叉验证和多种机器学习分类算法对训练集中的心电信

号采样数据进行错误标记识别的具体示意图。

具体实施方式

[0023] 下面结合实施例及附图对本发明作进一步详细的描述,但本发明的实施方式不限于此。

[0024] 实施例:

[0025] 本实施例提供了一种基于交叉验证的心电信号错误标记样本识别方法,所述方法的流程图如图1所示,包括以下步骤:

[0026] 一、对心电信号采样数据进行预处理,并对预处理后的心电信号采样数据进行特征提取;

[0027] 其中预处理具体包括:利用中值滤波方法对心电信号采样数据进行处理以去除基线漂移,再对中值滤波后的心电信号采样数据利用小波变换方法去除工频干扰。即心电信号采样数据序列中一点的值用该点的一个邻域中各点值的中值代替,从而达到去除基线漂移的目的,对中值滤波后的心电信号采样数据进行小波变换,使得50Hz/60Hz的工频干扰和有用心电信号的主要能量落在不同尺度上,从而到达去除工频干扰的目的。对预处理后的心电信号采样数据进行特征提取的步骤具体包括:利用小波变换方法检测预处理后的心电信号采样数据中每个心拍的R波峰值位置,对心电信号采样数据依据R波峰值位置进行心拍分割,并检测出该心拍的Q波起始位置和S波终止位置,如图2所示,进而提取出所需的时域特征,即该心拍的QRS波持续时间、R-R间隔(该心拍的R波峰值点和上一个心拍R波峰值点之间的时间间隔)以及平均R-R间隔(该心拍之前的10个心拍的平均R-R间隔);同时将每个心拍的数据值统一化处理为300个数据值,以代表每个心拍的形态学特征。

[0028] 二、对心电信号采样数据提取出来的特征进行标准化处理,并对标准化后的特征进行PCA降维,将降维后的心电信号采样数据作为训练集;

[0029] 其中,进行标准化处理的具体过程为:将提取出来的3个时域特征和300个形态学特征共303个特征利用最小最大标准化方法化为0~1之间的数值,从而使得两种不同的特征对分类方案的作用力同趋化。所述对标准化后的特征进行PCA降维是将标准化后的特征的维数降到包含原始特征信息的95%的维数。

[0030] 三、利用交叉验证和多种机器学习分类算法对训练集中的心电信号采样数据进行错误标记识别;

[0031] 具体过程如图3所示,将带有错误标记样本的训练集随机等分成10份,每次取其中1份作为验证集,其余9份作为小训练集用于训练机器学习分类器,并对验证集进行分类,将分类结果对比验证集样本原始标记,判断验证集样本是否为错误标记样本。重复以上过程10次,每次选取的验证集都不同,这样训练集中所有样本都被验证。找出训练集中的所有错误标记样本后,将其从训练集中移除,得到新的训练集,再训练分类器,并在测试集上进行测试。

[0032] 为了增强错误标记样本识别的可靠性,利用小训练集训练分类器时采用了多种机器学习算法,分别是朴素贝叶斯算法、最近邻法则、线性判别分析、支持向量机和决策树算法,将多个分类器分别对验证集进行分类并将结果和原始标记进行对比,只有当5个分类器中有4个以上分类器都判断一个样本是错误标记的,才将该样本从训练集中剔除。进一步,

可以将上述过程重复多次,增强错误标记识别的准确性。

[0033] 四、将识别出的带有错误标记的心电信号采样数据从训练集中移除,再训练分类器,并对测试集进行分类。

[0034] 采用来自MIT-BIH心律失常数据库中的20个记录实验数据对本实施例的基于交叉验证的心电信号错误标记样本识别方法进行验证,经过预处理、特征提取、标准化和PCA降维后,通过人为更改样本标记的方法引入比例分别为5%、10%、20%、30%、40%的错误标记样本。在各个比例错误比较样本的训练集上,利用5种机器学习算法和交叉验证识别出错误标记样本,将这些样本剔除出训练集,再重新训练真正的分类器(朴素贝叶斯、最近邻法则、线性判别分析),并在测试集上进行测试,得出分类准确率。

[0035] 以一个不引入任何错误标记的训练集直接训练分类器并在测试集上进行测试作为对照,来评估本实施例的技术方案在识别错误标记样本并提升心电信号分类准确率上的有效性。其中表1为不引入任何错误标记的训练集直接训练分类器并在测试集上进行测试的错误标记样本识别情况表,表2为采用朴素贝叶斯算法训练分类器并在测试集上进行测试的错误标记样本识别情况表,表3为采用最近邻法则训练分类器并在测试集上进行测试的错误标记样本识别情况表,表4为采用线性判别分析训练分类器并在测试集上进行测试的错误标记样本识别情况表:

[0036]

误标记比例	实际误标记数量	识别误标记数量	识别实际误标记数量	识别率	误识率
5%	255	334	239	93.73%	37%
10%	510	572	474	93%	19.22%
20%	1020	912	825	81%	8.53%
30%	1530	1344	1216	79%	8.37%
40%	2040	1453	1207	59%	11.18%

[0037] 表1

[0038]

误标记比例	不处理	完全去除误标记	交叉验证去除误标记
0	73.30%	--	--
5%	71.96%	73.13%	75.09%
10%	69.33%	73.20%	75.53%
20%	62.01%	73.04%	74.15%
30%	51.85%	73.06%	72.34%
40%	38.77%	72.74%	57.87%

[0039] 表2

[0040]

误标记比例	不处理	完全去除误标记	交叉验证去除误标记
0	97.50%	--	--
5%	96.60%	97.29%	97.25%
10%	94.97%	97.20%	96.80%
20%	87.84%	97.12%	95.53%
30%	77.48%	96.96%	92.22%

40%	64.16%	96.86%	78.15%
-----	--------	--------	--------

[0041] 表3

	误标记比例	不处理	完全去除误标记	交叉验证去除误标记
	0	74.50%	--	--
[0042]	5%	73.88%	74.27%	74.38%
	10%	73.41%	74.35%	74.23%
	20%	71.27%	74.62%	73.82%
	30%	63.78%	74.35%	70.95%
[0043]	40%	48.89%	74.40%	58.27%

[0044] 表4

[0045] 从中可以看出,当错误标记样本比例不超过20%时,本实施例的技术方案可以使得心电信号分类准确率回升到和无错误标记样本条件下同一水平;当错误标记样本比例为30%时,分类准确率比无错误标记条件下稍低;当错误标记样本比例达到40%,虽然分类准确率相比无错误标记条件下已有不小差距,但相比于不处理的情况还是有很大的回升。

[0046] 以上所述,仅为本发明专利较佳的实施例,但本发明专利的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明专利所公开的范围内,根据本发明专利的技术方案及其发明专利构思加以等同替换或改变,都属于本发明专利的保护范围。

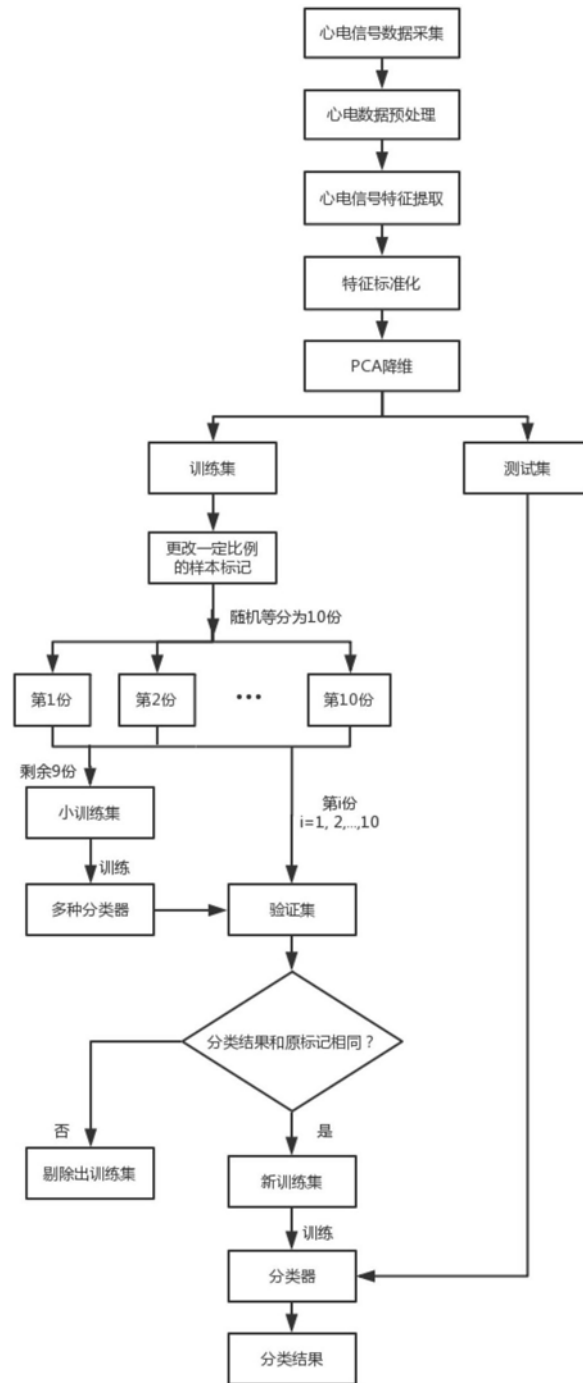


图1

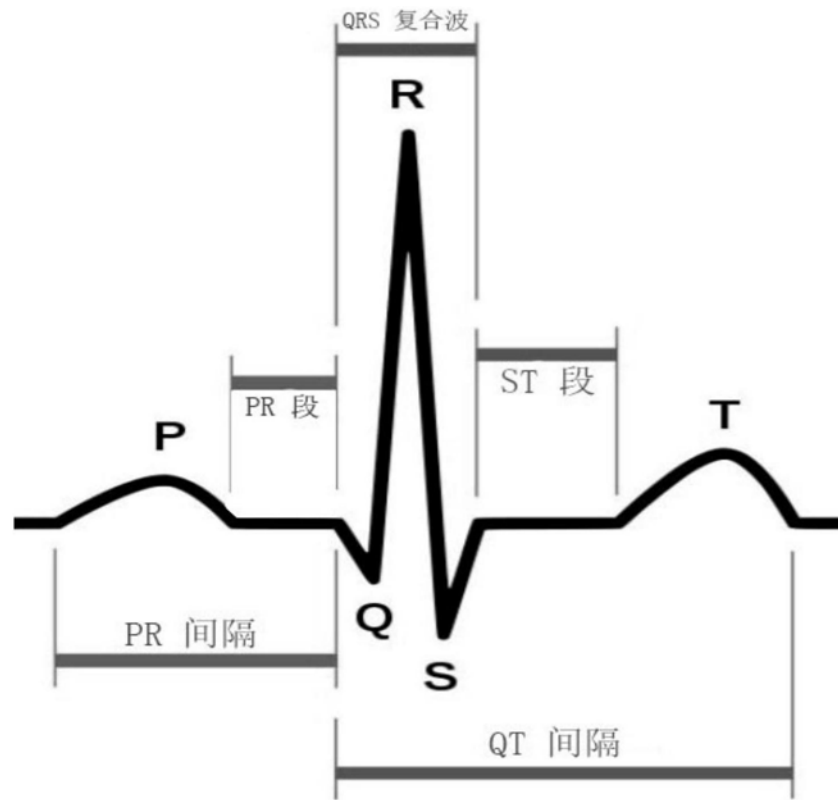


图2

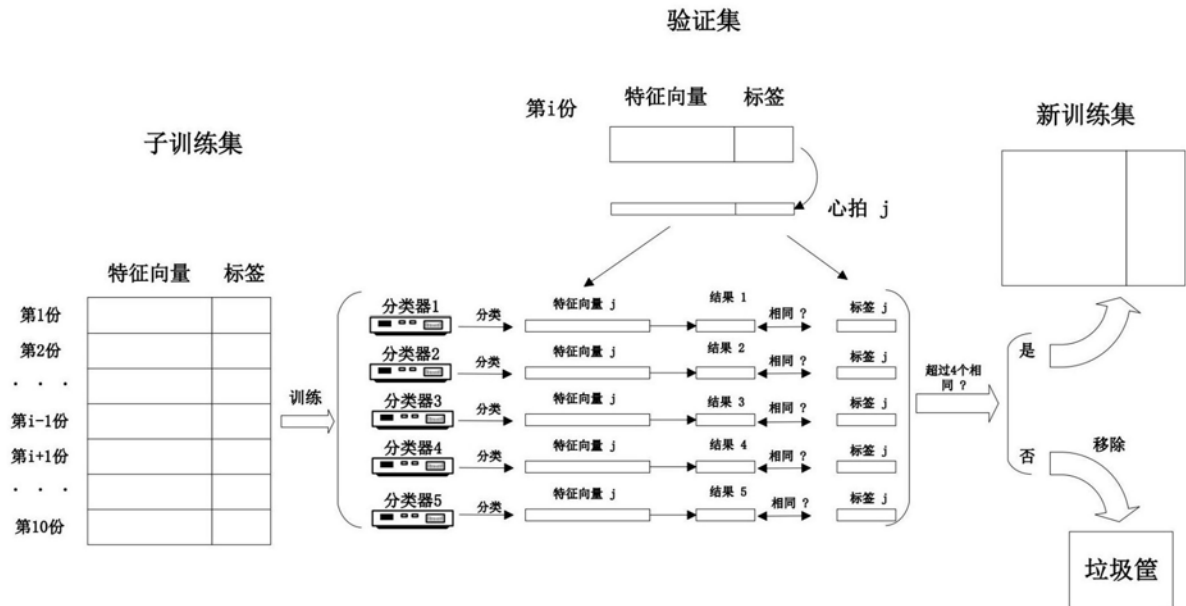


图3

专利名称(译)	基于交叉验证的心电信号错误标记训练样本识别方法		
公开(公告)号	CN108030494A	公开(公告)日	2018-05-15
申请号	CN2017111088016.2	申请日	2017-11-08
[标]申请(专利权)人(译)	华南理工大学		
申请(专利权)人(译)	华南理工大学		
当前申请(专利权)人(译)	华南理工大学		
[标]发明人	崔巍 李耀光		
发明人	崔巍 李耀光		
IPC分类号	A61B5/0452 A61B5/0402 A61B5/00 G06K9/00 G06K9/62 G06K9/68		
CPC分类号	A61B5/0402 A61B5/0452 A61B5/7267 G06K9/00516 G06K9/0053 G06K9/00543 G06K9/6247 G06K9/685 G06K2009/6871		
代理人(译)	李斌		
其他公开文献	CN108030494B		
外部链接	Espacenet SIPO		

摘要(译)

本发明公开了一种基于交叉验证的心电信号错误标记训练样本识别方法，包括以下步骤：一、对心电信号采样数据进行预处理，并对预处理后的心电信号采样数据进行特征提取；二、对心电信号采样数据提取出来的特征进行标准化处理，并对标准化后的特征进行PCA降维；三、利用交叉验证和多种机器学习分类算法对心电信号采样数据进行错误标记识别；四、将识别出的带有错误标记的心电信号采样数据从训练集中移除，再训练分类器，并对测试集进行分类。所述方法能够有效地识别出心电信号训练样本中的错误标记样本，进而提高心电信号自动分类的准确率。

