电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

学士学位论文

BACHELOR DISSERTATION

论文题目 基于床垫 BCG 信号的在床状态判别研究

学科专业	生物技术
学　　号	2016000000000
作者姓名	Leereliu
指导教师	曾东

分类号 _____ 密级 _____

UDC [注1] _____

学 位 论 文

_____

_____

姓名

（作者姓名）

指导教师 _____

_____

_____

_____

（姓名、职称、单位名称）

申请学位级别 **学士** 学科专业 _____

提交论文日期 **2000.00.00** 论文答辩日期 **2000.00.00**

学位授予单位和日期 **电子科技大学 2000 年 00 月**

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

Research on XXXXXXXXXX

A Masteral Dissertation Submitted to

University of Electronic Science and Technology of China

Discipline: _____

_____

Author: _____

Supervisor: _____

School: _____

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名：_____ 日期：　　年　　月　　日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日期：　　年　　月　　日

摘　要

本文主要研究的是如何从一段 BCG 信号中判断人体的在床状态，这种状态分为在床、离床两种，其中在床又分为体动和安静状态。本文需要研究如何有效的提取 BCG 信号的特征，对不同状态的 BCG 信号进行区分，利用多种机器学习的方法来实现一个可以准确对床垫 BCG 信号分类的分类算法，并对他们的分类效果进行比对和总结。

本文首先介绍了 BCG 信号采集信息系统的工作原理以及 BCG 信号的生理意义，以确保监督分类学习算法在生理上的可行性。

然后本文研究了 BCG 信号特征的提取，采集了 BCG 信号的有量纲特征和无量纲特征，通过监督学习的方式，将采集到的特征作为输入值，将已标记好的体动数据和安静状态数据采用交叉验证的方式一部分作为训练集，一部分作为验证集。分别使用决策树模型 LightGBM、随机森林、BP 神经网络、SVM 支持向量机、LR 逻辑回归等监督学习分类算法进行训练后对它们进行判定分类，使用统一的 F1score 以及 AUC 作为分类算法评价指标。最后经过比较，基于决策树的 LightGBM 模型对床垫 BCG 信号分类的效果最好。

关键词：BCG; 监督学习; 特征提取; 神经网络;

ABSTRACT

This article mainly studies how to judge the human body's on-bed state from a pre-processed BCG signal. This state is divided into on-bed and off-bed, in which the bed is divided into body movement and quiet state. This article needs to study how to effectively extract the characteristics of BCG signals, distinguish BCG signals in different states, and use a variety of machine learning methods to implement a classification algorithm that can accurately classify BCG signals of mattresses and perform their classification effects. Compare and summarize.

This paper first introduces the working principle of the BCG signal acquisition information system and the physiological significance of the BCG signal to ensure the feasibility of the supervised classification learning algorithm in physiology.

Then this paper studies the extraction of BCG signal features, collects the dimensional and non-dimensional features of the BCG signal, and uses the collected features as input values through supervised learning to use the marked body movement data and the quiet state. The data adopts the cross-validation method as part of the training set and part of the verification set. Supervised learning classification algorithms such as decision tree model, random forest, BP neural network, SVM support vector machine, LR logistic regression, etc. are used to train and classify them, and unified F1score and AUC are used as classification algorithm evaluation indicators. Finally, after comparison, the decision tree-based lightgbm model has the best effect on mattress BCG signal classification.

Keywords: BCG; Supervised learning; Feature extraction; Neural Networks;

目 录

第一章 绪论

1.1 研究背景及意义

健康是每个人生活中的第一要素，睡眠不仅可以缓解疲劳，还可以加快新陈代谢，有研究表面，一天中人在睡眠的时候新陈代谢最为迅速，充足有效的睡眠也可以保证第二天的好心情。早睡早起、足够的深睡眠时长也是放松疲惫一天身体的良药。没有足够的睡眠很难保持良好的精力和心情。

事实上很多疾病同缺乏睡眠相关，缺乏睡眠容易使人产生紧张、焦虑、甚至抑郁的情绪，最为重要的是它会使人的一天昏昏沉沉，做事情没有效率。所以研究睡眠健康有着很重要的意义，它对个人的健康，同时对于社会健康发展也有着积极的作用。有研究证明，睡眠质量与睡眠时的体动有着密切关系，因此对个体进行体动检测并给予干预指导，在养老和康复治疗过程中也有重要的应用价值。

目前，多导睡眠仪[1]可以用来监测睡眠过程中各种症状的参数，但需要在医院或实验室环境中进行。同时多导睡眠仪需要佩戴多个电极板和其他传感器设备，这些设备使用不方便，而且监测费用昂贵。心冲击描记(ballistocardio-gram，BCG)信号是由人的生命体征的活动，如心跳、呼吸、身体移动等信号构成。各种迹象表明，相对于其他信号 BCG 信号具有抽样方法简单，不干扰用户的优势，具有广阔的应用前景。

1.2 本文主要研究内容及结构

本文主要讲了通过相关设备提取到了 BCG 信号后，通过特征提取再通过监督学习方式来从原始 BCG 信号中分类出体动、离床、安静等在床状态，本文尝试了基于决策树的分类算法：lightgbm：同样基于决策树的随机森林、逻辑回归分类算法、SVM 支持向量机、BP 神经网络等算法来做 BCG 信号的分类，并以 F1score、AUC 作为分类算法好坏的评价指标。来避免数据所产生的偏斜类对算法的准确性产生偏差。

本文第一章主要介绍了 BCG 信号的研究背景及意义并概述了本文的主要研究内容及结构。

第二章介绍了 BCG 信号的原理，并介绍了本文所使用的的数据的示例和标签的结果。

第三章分别介绍了本文将要采用的分类算法的原理，分别是 lightgbm、随机森林、逻辑回归、支持向量机、BP 神经网络。

第四章提出了监督学习算法来分类已经标注好了的 BCG 信号。具体思路是对已标记好的 BCG 数据集根据最小标记精度做区间划分，划分为更多更小的数据集。每个数据集均有一个在床状态判别标签，随后对每个区间的数据作特征提取，分别使用 lightgbm、随机森林、逻辑回归、支持向量机、BP 神经网络等模型对提取的特征进行判定分类，将已有的数据作 5 折交叉验证，取 20%做验证集、80%做训练集，采用 F1score 以及 AUC 作为分类算法好坏的评价指标。经过比较 lightgbm 模型在这种时间序列信号的分类中表现最好。

第五章是总结和展望，对本文中所使用的相关特征和各分类算法结果作总结和分析，并指出了现有工作的不足，对未来更好的相关信号的分类工作做了展望。

第二章 BCG 信号概述和 BCG 信号数据示例

2.1 BCG 信号概述

心率是人体最重要的生理指标之一。它的收集方式主要分为接触式和非接触式。接触式的搜集方式就是普通的心电图仪。非接触式的搜集方式有 BCG 信号，它是由心脏跳动时的压力变化引起的。还有 PPG 信号，它是由血流速度和血管中氧气浓度的差异引起的。心冲击图是指使用对仪器敏感的压力变化来捕获由压力变化信号引起的一系列相应的弱体动引起的心跳反应。将该压力变化信号转换为电子信号波形就是 BCG 信号。收集和记录此 BCG 信号的仪器称为 BCG 记录仪。因为 BCG 记录仪具有非接触性、记录方便的特点，随着智能家居、物联网时代的到来，BCG 记录仪在未来有望走进家家户户、逐渐小型化。成为居家生活中必不可少的一个仪器。可以使用智能的 BCG 记录仪来进行心率检测、睡眠状态检测、在床体动状态判别、呼吸频率检测甚至能提前对心脏的疾病发生预警。目前学术界和工业界对 BCG 信号的研究和 BCG 记录仪的小型化正在如火如荼的进行中。伴随着计算机算力的提高和深度学习风潮的来临。对 BCG 的研究也不再局限于传统的信号时域频域相关的分析，背靠机器学习有越来越多的方法加入进来。BCG 信号将作用与并不局限于心率提取，呼吸提取，在床体动判定、心功能检测，睡眠质量分析，心脏病监测等领域。

采集 BCG 信号的主要方法有三种：站姿，坐姿和卧姿。由于本文中的 BCG 信号主要用于判断个人在床上的体动状态，因此下面主要介绍卧姿 BCG 信号的采集办法。

卧姿 BCG 信号采集是当前学术界的主要采集办法，也是最科学、最商业的采集方法。卧姿的 BCG 信号采集相比站姿和坐姿采集具有更持久和非接触式的特点，因此已经有很多人研究了 BCG 信号的卧姿采集办法。卧姿 BCG 信号的采集是由测量人在装有采集装置的床垫上进行。在 2005 年，K.Watanabe 提出了一种系统，该系统使用气垫床来监测生理信号，例如心跳，呼吸和体动等。它使用床垫的气垫来感应压力变化，再通过压力传感器的传导来得到 BCG 信号，但是这份信号中包含了不止体动信息，还包含了心率信息、呼吸信息等，需要通过相关的滤波器对信号进行预处理后才能得到所需要的相关生理信息。

本文中的信号源是已经经过滤波后的人体动信号，不再包含无关的数据（例如心率和呼吸），并且已经采用人体动信号的分类标签对数据集进行了标记。

2.2 BCG 信号数据示例

原始的 BCG 数据信号如图 2-1 所示，共由 437265 个采样点组成，总样本数为该数据示例的 40 倍。

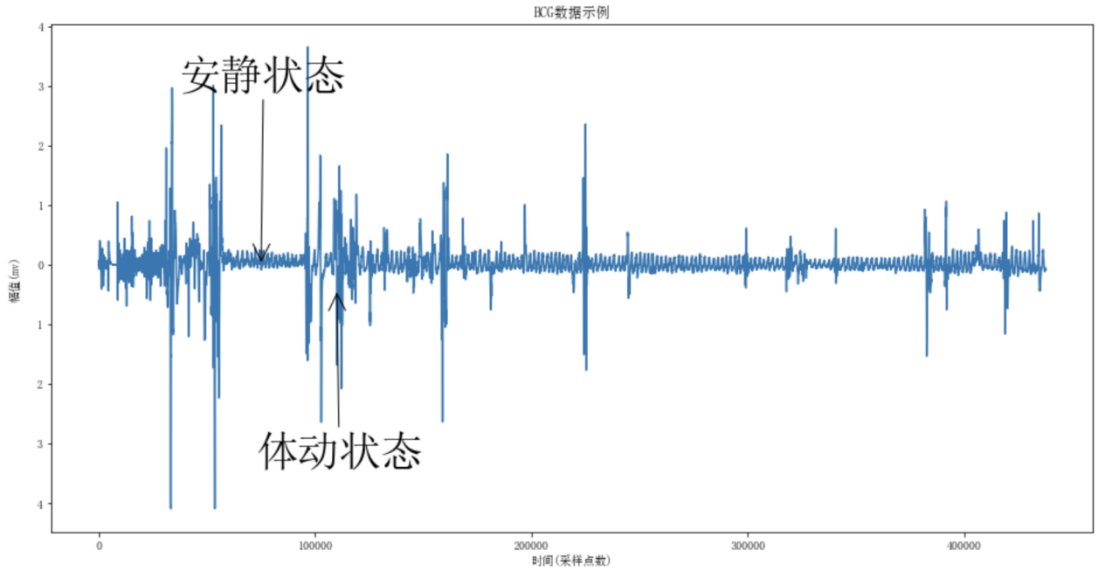

图 2-1 原始的 BCG 数据信号

由上图中可以看出 BCG 的体动数据示例是不平稳的时间序列数据，每隔一段时间会产生波动较大的峰值信号，每隔一段时间又会有一段较为平稳的时间信号。从直观上可以较为方便的看出，波动较大的信号段为体动时的状态，波动较小且序列平稳的信号段为安静时的状态。事实上通过已标记好的标签数据给上图进行标记的结果也较为复合直观判断的结果。标记的采样结果以每 45 个采样点为一个序列，标记其体动状态，安静状态置为 0，体动状态置为 1。下图 2-2 展示了经过标签标记后不同体动状态下采集到的 BCG 信号。

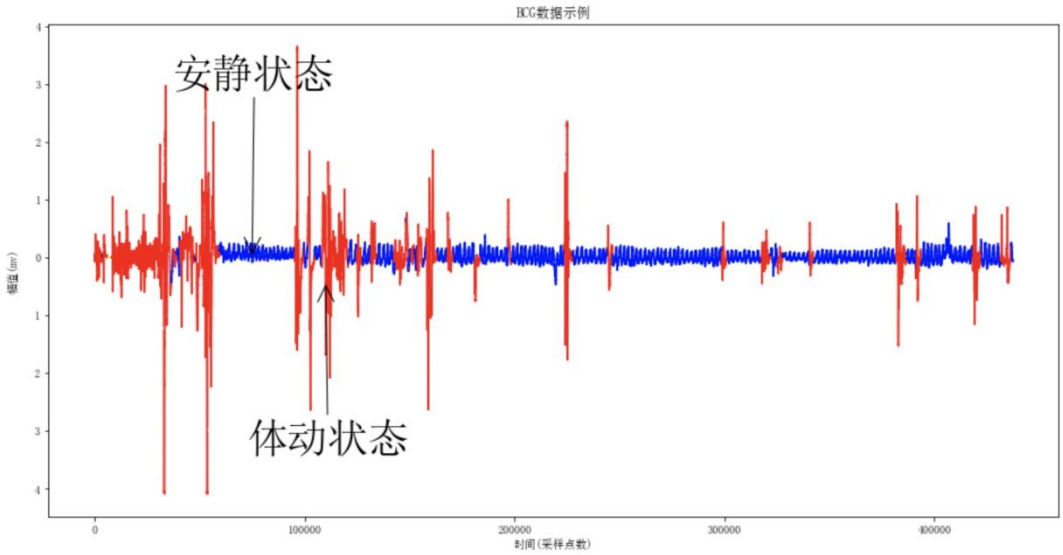

图 2-2 采集到的不同体动状态的 BCG 信号标注结果

将体动的状态在图中绘制为红色，安静的状态绘制为蓝色。较直观的看出，波动较大的数据段为体动状态，波动较小的数据段为安静状态。可以以 45 个采样点为最小的分割点，在提取这 45 个采样点的特征来作为分类的特征。

第三章 分类算法概述

3.1 随机森林

随机森林是由随机构成的决策树组成，随机森林中每一颗决策树都不相互关联，随机森林在处理输入样本时，会让森林[3]里的每一个决策树去判断输入示例样本属于哪一个分类。最后哪一个输出样本的分类统计最多就判断样本分类结果属于哪一类。随机森林既可以分类线性数据，也可以分类离散化的数据。

决策树是一个树的结构，其中每个叶子节点代表一种分类判断类别，每个分支代表在大的分类判别下的分类。使用决策树对某个输入特征示例进行判断时，首先从根节点开始，一层一层向下层判断，选择自己的分支节点，直到到达叶子节点来决定示例所属的类别。

下面是随机森林的构造过程：

1. 从所有样本中随机选择 n 个样本作为森林的根节点。每个样本都是一个决策树的根节点。

2. 分支的构建从样本的属性中随机选择 m 个属性，然后采用一定的策略如信息增益等算法来决定一个属性作为该节点的分支属性。

3. 根据 2 步骤来对根节点进行分裂构建，直到节点没有属性来构建分支属性时，一颗决策树就构建完成了。同时构建决策树的过程中不进行剪枝的操作。

4. 根据 1、2、3 的步骤构建大量的决策树，以此来形成一个随机森林。

决策树有很多的优点：

a. 它很难陷入过度拟合的状态。

b. 在许多当前数据集中，与其他算法相比，它具有很大的优势。 两种随机性的引入使得随机森林具有良好的抗噪能力

c. 它可以处理很高维度的特征，或者说数据量很大的特征，并且不需要专门对数据进行正则化。同时适用于离散数据和连续数据。

d. 可以生成一个 Proximities =（pij）矩阵来测量样本之间的相似性：pij = aij / N，aij 表示样本 i 和 j 在随机森林中的同一叶节点中出现的次数，N 表示随机森林数树木。

e. 随机森林使用无偏估计来判断误差。

f. 基于决策树的分类可以对特征的重要性进行排序。

g. 可以检测到特征之间的相互影响结果

h. 实现比较简单，可以很好的支持并行化处理来加快训练速度。

随机森林可用于分类和回归模型，随机森林使用决策树来作为基本单位，在构建决策树时，每个根节点都是随机的，同时每个特征值和分支也是随机的，最后组合这些随机的特征值的决策树来构成随机森林，本文主要使用随机森林来作分类的问题，它可以组合每个决策树的分类结果来构建成最终的分类结果，这种组合式所产生的结果一般要比单一的分类器分类所产生的结果可信度高。虽然随机森林中每个单一的决策树的分类都很差，但是将他们组合成随机森林后效果却出奇的好，好比好多专家从不同角度查看新问题（新输入数据）。最后，每位专家将投票以得出结果。这恰好是群体智能，是经济学中的隐形手，它也是分布式分类系统。每个子领域的专家都使用自己独特的默认知识来对产品进行分类，并确定是否需要生产。随机森林的效果取决于多个分类树是否彼此独立。如果要在不过度适应的情况下继续发展经济（即政府主导的经济增长，但是在遇到新情况后就会出现泡沫），我们需要独立发展，独立选择自己的特征。

3.2 LightGBM

lightgbm 简称 lgb。是一个微软开源的梯度 boosting 框架，使用基于集成学习算法的决策树，传统的提升树模型是利用加模型和前向分布算法实现的，如 XGBoost，GBDT，pGBRT 等。其中 GBDT 采用梯度迭代（Gradient Boosting）即通过之前树得到的残差来更新目标值，其中残差是最小均方损失函数关于预测值的反向梯度。XGBoost 则是对损失函数进行了二阶泰勒展开，使得精度更高。但是它们在决策树寻找最佳分裂点时采用的预排序算法均需要遍历所有样本，在大量数据和特征维度的情况下，时间和空间的消耗都很高。lightgbm 则很好的解决了这些问题。它采用单边梯度抽样算法、直方图算法、互斥特征捆绑算法、基于最大深度的 Leaf-wise 的垂直生长算法、类别特征最优分割、特征并行和数据并行、缓存优化等策略来做到更快的训练速度和更少的内存使用。

下面分别介绍这几种策略。

1. 单边梯度抽样算法

GBDT 算法的梯度大小可以反映样品的质量。 梯度越小，模型拟合越好。 单面梯度采样算法（基于梯度的单边采样，GOSS）使用此信息来对样本进行采样，从而减少了很多对于小梯度的样本，仅需注意高梯度的样本即可。这大大减少了计算量。 GOSS 算法保留具有大梯度的样本，并随机保留具有小梯度的样本。 为了不改变样本的数据分布，为具有小梯度的样本引入常数以在计算增益时进行平衡。一方面，该算法将更多的注意力放在训练不足的样本上，另一方面，它乘以权重，以防止样本过多地影响原始数据的分布。

2. 直方图算法

GBDT 是基于决策树的集成算法。它使用正分配算法，并且在每次迭代中，通过负梯度生成的残差来学习决策树。最耗时的步骤是找到最佳的分区点。一种流行的方法是预排序，其核心是枚举排序后的特征值上的所有可能特征点。另一个改进是直方图算法。它将连续特征值划分为 k 个桶，并在这 k 个点中选择划分点。 k << d，因此内存消耗和训练速度更好，并且实际数据集显示离散分割点对最终精度几乎没有影响，甚至更好。由于决策树本身学习能力较弱，而直方图算法的使用将起到正则化作用，可以有效地防止模型过度拟合。 Lightgbm 正是基于直方图的增强算法。

直方图算法的基本思想是先将连续浮点特征值离散化为 k 个整数，然后构造宽度为 k 的直方图。遍历数据时，离散值用作指示符以累积直方图中的统计数据。遍历数据一次后，直方图将累积所需的统计信息，然后遍历以找到最佳分割点。直方图算法的基本思想是将连续特征离散为 k 个离散特征，同时为统计信息构造一个宽度为 k 的直方图（包含 k 个 bin）。使用直方图算法，我们不需要遍历数据，只需要遍历 k 个 bin 即可找到最佳分割点。我们知道，特征离散化具有许多优点，例如存储方便，计算速度更快，鲁棒性强以及模型更稳定等。在构造叶节点的直方图时，我们还可以通过从相邻叶节点的直方图中减去父节点的直方图来构造，从而将计算量减少一半。在实际操作过程中，我们还可以先计算出直方图小的叶节点，然后利用直方图求和，得到直方图较大的叶节点。

3. 互斥特征捆绑算法

高维特征通常是稀疏的，并且如果两个特征不是完全互斥的（例如仅部分情况），则特征可能是互斥的（例如两个特征不同时为非零值），则可以使用互斥率指示互斥的程度。独占功能捆绑（EFB）指出，如果某些特征被融合和绑定，则可以减少特征数量。为了响应这个想法，我们将遇到两个问题：哪些特征可以绑定在一起？绑定特征后，如何确定特征值？

对于问题一：EFB 算法使用要素和要素之间的关系来构造加权无向图，并将其转换为图着色算法。我们知道图着色是一个 NP-Hard 问题，因此使用贪婪算法来获得近似解。具体步骤如下：

（1）构造一个以顶点为特征，边为两个特征相互排斥的加权无向图；

（2）根据节点的程度降序排列，程度越大，与其他特征的冲突越大；

（3）遍历每个功能部件，将其分配给现有功能部件包或创建新的功能部件包，总体冲突很小。

对于问题二：lightgbm 提供了一种特征合并算法，关键是可以将原始特征与合

并特征分开。假设捆绑中有两个要素值，则 A 取值为[0，10]，B 取值[0，20]。为了确保特征 A 和特征 B 互斥，我们可以在特征 B 上添加偏移量，将其转换为[10，30]，绑定之后的特征值为[0，30]，从而实现特征合并。

4. 具有深度限制的 Leaf-wise 叶子生长策略

Level-wise 算法可以同时拆分同一层的叶子。执行多线程优化很容易。控制模型的复杂性也很容易，并且不容易过拟合。但是实际上，Level-wise 算法是一种低效的算法，因为它不加区别地对待同一层的叶子，这带来了许多不必要的开销，因为实际上许多叶子的分割增益很低并且不需要搜索和分割。

Leaf-wise 算法是一种更有效的策略。它每次从所有当前叶子中找到分裂增益最大的那个，然后分裂，依此类推。因此，与 Level-wise 相比，Leaf-wise 可以在分割数相同的情况下减少更多错误来获得更好的准确性。 Leaf-wise 的缺点是，它可能会生长更深的决策树，从而导致过度拟合。因此，LightGBM 在 Leaf-wise 之上增加了最大深度限制，以防止过度拟合来确保更高的效率。

5. 类别特征最优分割

LightGBM 本机支持类别特征，并使用“多对多”分割将类别特征分为两个子集，以实现类别特征的最佳分割。 基本思想是每次根据训练目标对类别特征进行分类，根据直方图的累加值$\frac{\sum gradient}{\sum hession}$对直方图进行分组，然后在排序后的直方图上找到最佳分割。 此外，LightGBM 还添加了约束的正则化以防止过度拟合。

6. 特征并行

传统的特征并行算法是对数据进行垂直分割，然后使用不同的机器找到不同特征的最优分割点，基于通信集成获得最佳分割点，然后基于通信告知其他机器。传统的特征并行方法有一个很大的缺点：它需要将最终的分割结果通知每台机器，这增加了额外的复杂性（因为数据是垂直分割的，每台机器包含不同的数据，并且分割结果需要通过通信来通知 ）。 LightGBM 不执行垂直数据划分。 每台机器都有训练集的完整数据。 在获得最佳划分方案之后，可以在本地执行划分以减少不必要的通信。

7. 数据并行

传统的数据并行策略主要是对数据进行水平分割，然后在本地构建直方图，然后将其集成到全局直方图中，最后在全局直方图中找到最佳分割点。

这种数据划分有一个很大的缺点：通信开销太大。 如果使用点对点通讯，则机器的通讯开销大约是$O(\#machine * \#feature * \#bin)$。 如果使用集成通信，则通信开销是$O(2 * \#feature * \#bin)$，LightGBM 使用减少分散方法将直方图集成的任务分配给不同的机器，从而降低了通信成本，并通过直方图通信进一步减小了不同

机器之间的差异。

8. 缓存优化

LightGBM 使用的直方图算法本质上对 Cache 友好：

首先，所有功能都使用相同的方法来获取梯度（不同于不同的功能通过不同的索引来获取梯度），只需要对梯度进行排序就可以实现连续访问，大大提高了缓存命中率。

其次，由于不需要将特征存储到样本的索引，因此减少了存储消耗，并且没有“缓存未命中”的问题。

3.3 支持向量机

支持向量机是由 Cortes 和 Vapnik 等人[5]在 1990 年代提出来的，相比于朴素贝叶斯等传统的机器学习算法。它是一种可靠且快速的分类算法。支持向量机被防范应用于小数据量的分类上，并可以与其他机器学习算法结合进行拓展。在支持向量机被提出来的年代，计算机的运算能力还远没有现在强大，正是支持向量机的横空出世很快占有了分类算法的榜首，神经网络算法也因此在那个年代陷入了衰落。支持向量机在分类迭代优化上理论上可以优化到全局最优解，但是神经网络却很容易陷入局部最优解，运算速度上也优于简单的感知机神经网络模型。

支持向量机在本质上是找到一个可以最大化两个类别边距的超平面[6]对目标进行划分。最靠近超平面的采样点称为支持向量，并且具有最大间隔。寻找超平面实质上是一个凸优化问题目的是在区分样本的同时最大化支持向量和超平面之间的距离。在二维数据中，这样的超平面可以是一条直线。但是如果特征数据在二维平面上并不是线性可分的模型如环模型，无法找到一个线性决策边界来划分这两种类别。这种时候可以引入更高的维度，如第三个维度，在三维空间上对特征空间类别进行划分，如环模型的分类在三维空间上的超平面投影到二维平面上就是一个不规则划分类别的圆。

虽然将数据映射到高维空间可以用来划分非线性数据，但是这种映射会使计算量加大，过多的维度也会导致分类算法效率低下以及过拟合。这时候引入支持向量机独有的核函数，它将特征类别的分类转换为特征类别的数量积来进行分类。普通的二维平面的线性分割即使用了线性核的核函数。若要将二维数据上不可分割的数据映射到高维空间上去分割，则需要采用更高维度的核函数去分割。只需要简单的替换一下目标函数的内积的核函数即可。支持向量机分割出来的超平面即称为决策边界，这个边界代表了分类的边界，多分类只需要多次应用二分类的支持向量机模型即为多分类。核函数本质上并不属于支持向量机的一部分，支持向量机主

要用于找出决策边界，核函数可与其他机器学习方法共同使用，如线性分类逻辑回归等。在线性空间寻找决策边界，可以应用线性核、多项式核等。在高维空间寻找超平面的分割，可以应用高斯核等。

3.4 逻辑回归

逻辑回归是一种和线性回归比较相似的算法，但是线性回归和逻辑回归又有本质的区别，线性回归预测数值时，结果将是一个准确的数值，如收益。但是逻辑回归的预测结果将是一个离散数据，可以通过这个预测的离散数据的概率来判断结果的正负性。如狗的图片是否为猫、一封邮件是否为垃圾邮件等。

在实现方面，逻辑回归仅将 Sigmoid 函数添加到线性回归的计算结果中，并将数值结果转换为 0 到 1 之间的概率（Sigmoid 函数图像如下）

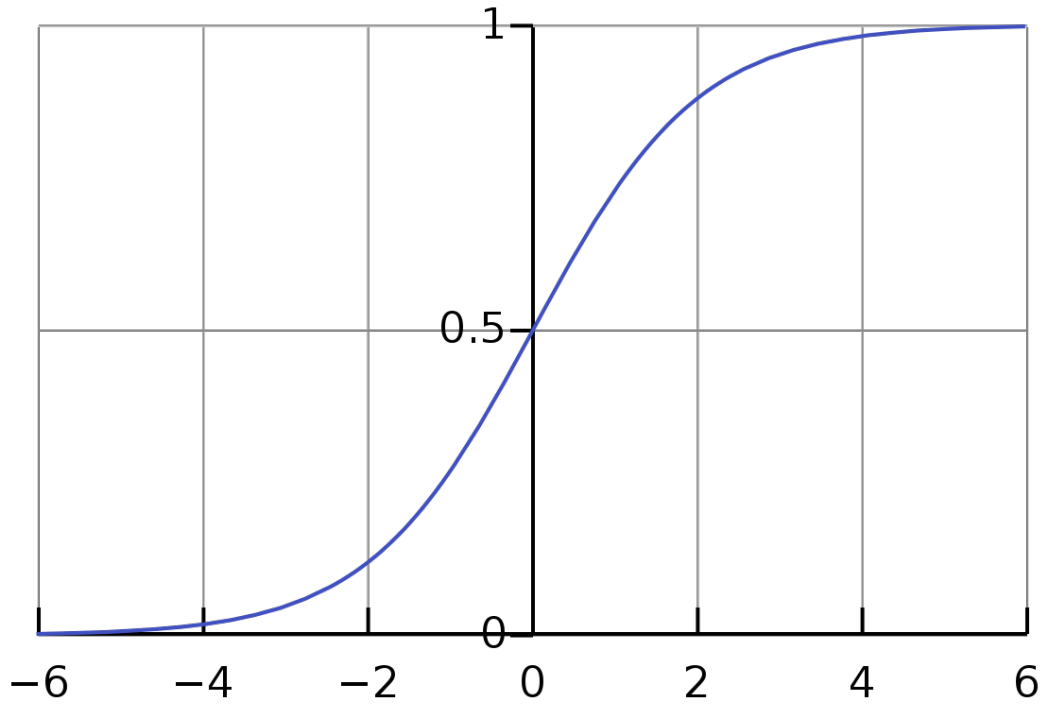

图 3-1 Sigmoid 图像

我们就可以基于该概率进行预测，例如，如果概率大于 0.5，则表明电子邮件为垃圾邮件，直观上，逻辑回归画出了一条分类线，请参见下图 3-2。

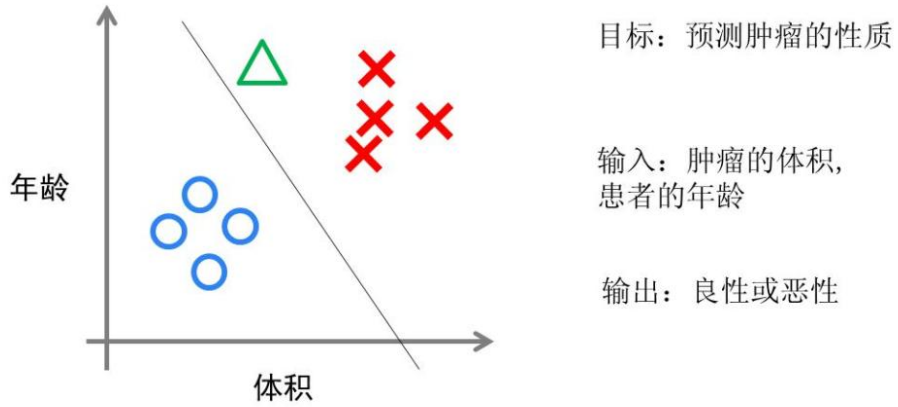

图 3-2 肿瘤患者数据分类

假设有一组肿瘤患者的数据。这些患者的一些肿瘤是良性的（图中蓝色点），而某些则是恶性的（图中红色点）。在这里，肿瘤的红色和蓝色可以称为数据的标签。每组数据都包含两个特征：患者的年龄和肿瘤的体积。将这些特征和标签绘制到二维空间上就形成了如图 3-2 所示的数据。

当有绿点时，应该判断肿瘤是恶性还是良性的？根据红色和蓝色点，训练了逻辑回归模型，这是图中的分类线。此时，绿色点出现在分类线的左侧，因此判断其标记应为红色，这意味着它属于恶性肿瘤。

通过逻辑回归算法绘制的分类线基本上是线性的（也存在具有非线性分类线的逻辑回归，但是此类模型在处理大量数据时效率非常低），这意味着逻辑回归在非线性分类上的表达能力十分不足，需要寻求别的分类算法。

3.5 神经网络

人工神经网络是诞生于上世纪 80 年代的算法，它在诞生之初非常的流行，但是受限于当时的计算算力的限制，被运算速度更快的支持向量机等传统的机器学习模型击败。随着最近几年 GPU 算力的提高，深度学习的大火，神经网络又走进了人们的视野中。

人工神经网络最初诞生于生物界的学者对人大脑的研究机制之中。人的神经细胞有用来接收传入信息的树突和有着很多轴突末梢的轴突，其中轴突连接着下一个神经元的树突，在神经冲突传导来临之时上一个神经元的轴突末端释放神经递质到下一个神经元的树突上以此来完成一次神经冲动的传递。在 Hubel-Wiesel 实验中，有学者研究了猫的视觉分析机制。

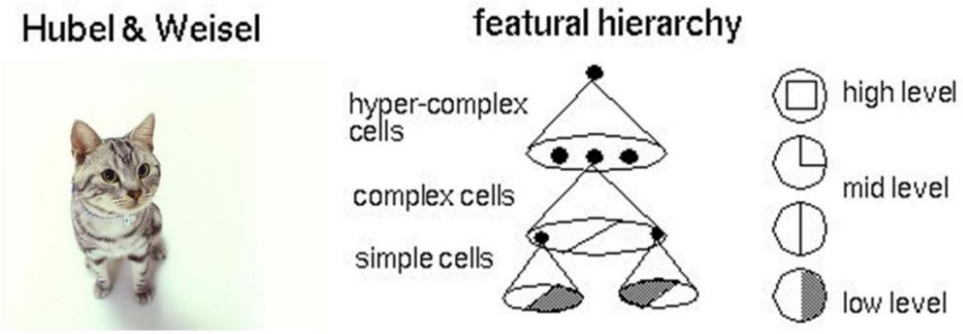

图 3-3 猫的视觉分析机制

在猫的神经元之中，最先看到的是是一个正方形，下一层神经元中看到的是一个折线，再下一层神经元之中看到的一条直线，最后一层神经元则看到的是黑白两个表面。可以看出图像在猫的神经元中被分解成了很多细小的细节分散进不同的神经元中分别处理，但是最终得出的结论是这是一个正方形。人们根据这种大脑视觉识别的机制发明了人工神经网络。

图 3-4 是一个简单的两层感知机的模型，它的入口处的节点集合称为输入层，中间的节点集合称为隐藏层，最后的节点称为输出层。输入层负责特征的输入，中间每个节点线段都连接下一层的节点，每条线段都代表着权重值，中间的隐藏层的输入经过激活函数处理后再乘以隐藏层到输出节点的权重值得到输出节点的输入值，输出节点再经与真实值（即标签值）的比对后（选择一个损失函数）反馈到前面的节点，如此反复循环，不断的拟合到最优值，这个过程称为神经网络的训练，神经网络正向传播完毕后反向回来跳转中间节点的值的算法称为 BP 算法。正是 BP 算法的提出，使得神经网络风靡一时。

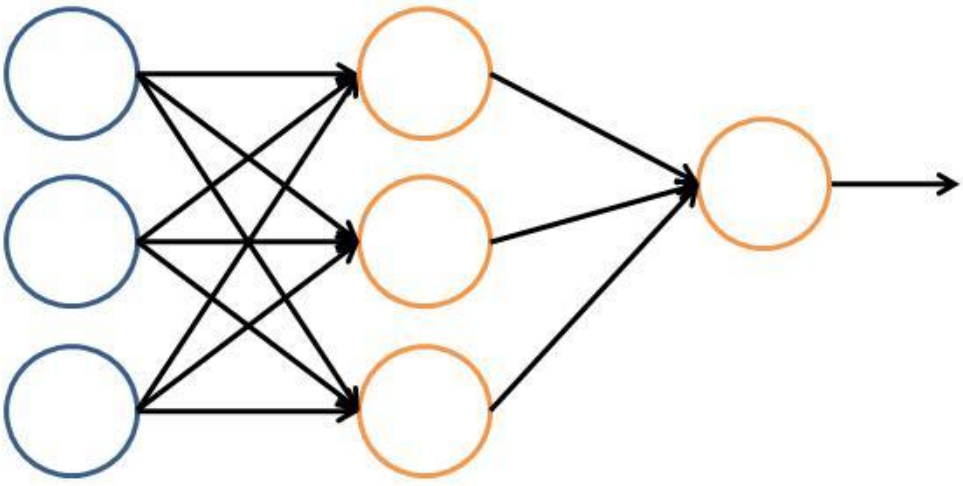

图 3-4 神经网络示意图

在隐藏层中，每个节点都有一个激活函数，这个激活函数通常为方便训练的线性函数，如 sigmoid、tanh 等。其中若使用 sigmoid 函数，函数求导后，值将向 0、1 靠近，基于 sigmoid 函数的这种性质，神经网络将可用于分类模型。

但是神经网络在提出后的 1990 年代还是遇到了困难，它尽管使用了 BP 算法，但是基于当时的计算能力的限制，训练耗时严重，且中间隐藏层的节点数需要调参，甚至训练结果最后会拟合到局部最优解而不是全局最优解。所以当不用调参、能迅速拟合到全局最优解的支持向量机模型出现后，神经网络的研究进入了冰河期。

在本文中，使用了最简单的双层感知器模型。 27 个输入层节点对应于 27 个提取的特征，100 个隐藏层节点，2 个输出节点代表 BCG 信号的两个在床体动状态。

第四章 基于监督学习的方式分类床垫 BCG 信号在床状态

4.1 信号的序列构建

对于已经得到的 BCG 信号，每个数据示例的采样点数均为 47w 采样点数左右，已标注的数据以 45 个采样点为一个区间，每个区间均有在床状态的标注，以此每个小区间作最小分类单元。

4.2 特征提取

特征提取是传统机器学习中最重要的任务之一。重要的特征对训练的准确率有着事半功倍的效果。基于 BCG 信号的时间序列特点，提取了相关的统计学、信号相关的特征。提取的特征有

（1）最大值：在一个域上取得最大值的点的函数值。

（2）最小值：在一个域上函数取得最小值的点的函数值。

（3）平均值：一组数据之和除以这组数据个数。

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i \quad (4\text{-}1)$$

（4）方差：描述随机变量的离散程度，取每个随机变量距期望的平方的平均值。

$$\delta^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 \quad (4\text{-}2)$$

（5）标准差：方差的正平方根即为标准差，同样描述随机变量的离散程度。

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2} \quad (4\text{-}3)$$

（6）平均绝对偏差：表示各个变量值之间差异程度。指各个变量值同平均数的绝对值的算术平均数。

$$X_{mad} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{X}| \quad (4\text{-}4)$$

（7）峰度（Kurtosis）：表示概率密度分布曲线在平均值处峰值高低的特征数。反映了波形峰部的尖度。

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\delta^4} \quad (4\text{-}$$

5）

（8）偏度：偏度衡量实数随机变量概率分布的不对称性。

$$\alpha = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\delta^3} \quad (4-6)$$

（9）峰值：峰值指相应的时间间隔如周期内，变化的电流、电压或功率等波形数据的最大值。

$$X_{peak} = max(|x_i|) \quad (4-7)$$

（10）整流平均值：指信号绝对值的平均值，也是信号经过全波整流后的平均值。

$$X_{arv} = \frac{1}{n}\sum_{i=1}^{n}|x_i| \quad (4-8)$$

（11）均方根值：是二次方的广义平均数的表达式。

$$X_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \quad (4-9)$$

（12）波形因子：是相同功率的直流讯号和原交流讯号整流后平均值的比值。

$$S = X_{rms}/X_{arv} \quad (4-10)$$

（13）峰值因子：为波形的振幅再除以波形 RMS(time-averaged)所得到的值。

$$C = X_{peak}/X_{rms} \quad (4-11)$$

（14）脉冲因子[7]：是信号峰值与整流平均值（绝对值的平均值）的比值。

$$I = X_{peak}/X_{arv} \quad (4-12)$$

（15）过零点数：指一个信号的符号变化的个数，如信号从正数变成负数，或反过来。

差分是指以时间为主轴统计维度的序列中，以下一个数值减去上一个数值，当间距相等时，就称为一阶差分，一阶差分对于不平稳的离散数据往往有奇效，因为差分的作用就是减轻数据之间的不规律波动，使其波动曲线平稳。因此又提取了一阶差分后的最大值、最小值、平均值、方差、标准差、平均绝对误差、峰度、偏度作特征。

傅里叶变换是指将一个时域的连续信号，转换为一个在频域的连续信号。要做的研究内容是将 BCG 信号分类为体动、安静、离床等在床状态，目前所提取的特

征均为时域内的特征，不妨将这些连续的时域的信号经过傅里叶变换后变为频域的信号，也许在床状态也和信号频率相关。因此也提取了快速傅里叶变换后的方差、平均值、最大值、最小值特征。

4.3 序列分类

提取了相关特征后需要对已有的数据进行序列分类，在本课题中使用了随机森林、lightgbm、支持向量机、逻辑回归、BP 神经网络这几个常用的分类模型做分类，在训练过程中通过采用五折交叉验证的方式并通过计算模型预测结果的 F1score、AUC 值来评价各个分类器的性能。将所有已标记好的数据随机分成五份，（这里具体在代码中得设置一个相同的随机种子，为了避免对照分类算法时出现偏差。）其中每份数据之间都互斥。共进行 5 轮训练，每轮训练都只使用其中的 4 份做训练数据，另外的 1 份做验证数据，每轮的验证数据集都不相同。使用 5 轮训练结果的评价指标的平均值作为该分类器性能的评价指标，通过该方法来尽能力的消除随机因素的影响

在验证分类算法的评价指标上，如果使用准确率作为评价指标，需要考虑偏斜类的影响即大量的样本倾向了某一类型。在本研究使用的数据中即产生了偏斜类，在床数据远远多于离床数据，安静状态数据远远多于体动状态数据。在这种情况下，单纯的使用误差不再能完善的评价模型的好坏了。

因此我们定义阳性(Positive)为: 表示正样本。当预测和实际结果都为正样本时，表示真阳性(True Positive); 而预测为正样本，实际为负样本时表示假阳性(False Positive)。

阴性(Negative)为: 表示负样本。当预测和实际结果都为负样本时，表示真阴性(True Negative); 而预测为负样本，实际为正样本时表示假阴性(False Negative)。

定义查准率(Precision)为:

$$\text{Precision} = \frac{\text{TurePos}}{\text{PredicatedPos}} = \frac{\text{TruePos}}{\text{TruePos+FalsePos}} \quad (4\text{-}13)$$

要提高查准率，就需要降低假阳性出现的频次。

定义召回率(Recall)为：

$$\text{Recall} = \frac{\text{TruePos}}{\text{ActualPos}} = \frac{\text{TruePos}}{\text{TruePos+FalseNeg}} \quad (4\text{-}14)$$

要提高召回率，需要降低假阴性出现的频次。

理想状况下，我们希望能同时具有高准确率和高召回率，因此引入 F1score:

$$F_1Score = 2\frac{PR}{P+R} \quad (4\text{-}15)$$

从公式中也可以看出，只有查准率和召回率两者都高时，F1score 的值才会高。同时我们以 FP 假阳性为 x 轴，TP 真阳性为 y 轴构建一条曲线，如下图所示

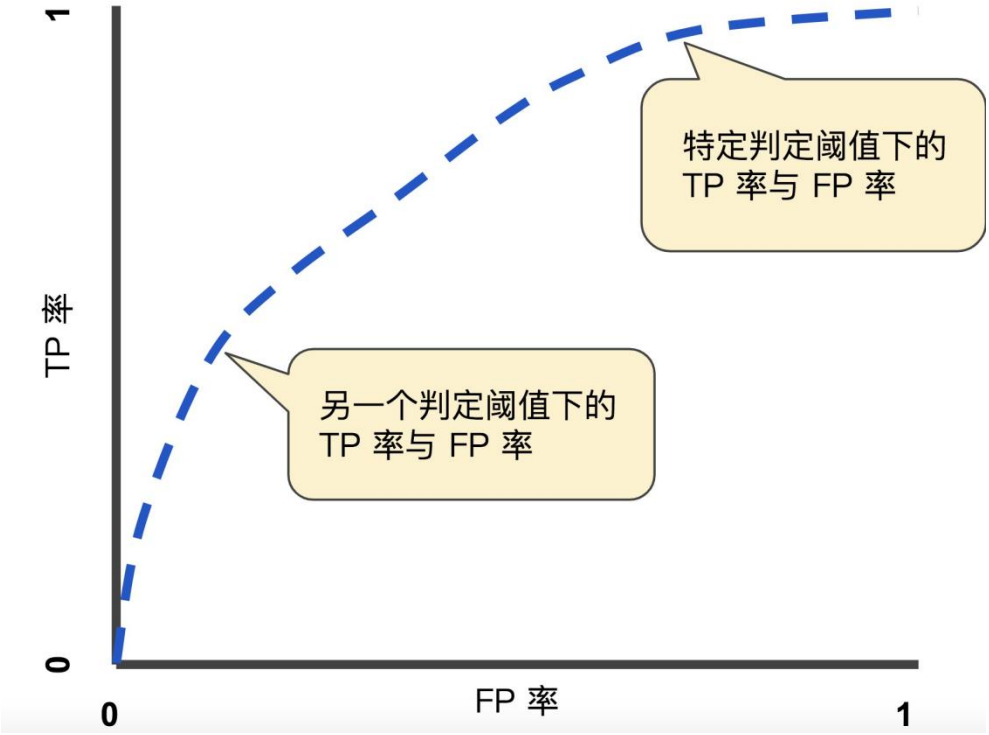

图 4-1 ROC 曲线

将曲线下的面积 Area under curve 称之为 AUC。该值也能很好的避免偏斜类的影响。因此本文同时采用 F1score 和 AUC 来判断这些分类算法的好坏。

4.4 结果分析

表 4.1 分类算法结果分析

分类算法	AUC	F1score	重要特征
lightgbm	0.929384	0.73301	diff_kurt、diff_skew、diff_var
逻辑回归	0.899912	0.66064	-
随机森林	0.917277	0.56742	diff_mad、diff_std、peak

支持向量机（使用高斯核）	0.811969	0.64981	-
BP 神经网络	0.799876	0.59634	-

从结果上可以看出基于决策树的 lightgbm 和随机森林算法拟合结果要优于逻辑回归、支持向量机、bp 神经网络。而且基于树的集成算法有一个很好的特性，在模型训练结束后可以输出模型所使用的特征的相对重要度，上表中已将 lightgbm 和随机森林模型训练结束后所使用的特征的相对重要度最高的三个结果输出了出来。从重要特征中我们可以看出对原有序列作一阶差分后提取的特征相对于原始序列上提取的特征更为重要。lightgbm 的重要特征分别为一阶差分后的峰度、一阶差分后的偏度、一阶差分后的方差。随机森林的重要特征分别为一阶差分后的平均绝对偏差、一阶差分后的标准差、峰值。而不基于决策树的算法要普遍若于基于决策树的算法，这里推测为不基于决策树的算法将所有特征的初始权置为相同来训练，削弱了重要特征的影响力，导致最后的预测结果没有决策树相关的算法好。逻辑回归和 BP 神经网络只是简单的对特征和结果作了一个二元线性分类的拟合，使用高斯核的支持向量机效果较差的原因推测为在高维空间仍没有找到较好的超平面分割点，固分类结果较差。

第五章 总结和展望

5.1 总结

BCG 信号由于其无接触式判断心率、呼吸、体动等数据的特点，被工业级和学术界广泛研究。本文使用了传统的机器学习方法来对床垫的 BCG 信号进行了分类和判别。

本文第一章介绍了该课题的研究背景，因为 BCG 信号的非接触的特点，被广泛研究。它可以用来分析睡眠质量也可以用来检测心脏的相关疾病。本文首先提出通过机器学习的办法来分析 BCG 信号的体动信息。为床垫 BCG 体动信号的分类提供了一个新的思路。

本文的第二章重点讲述了 BCG 信号产生的原理及其生理意义。为后文采用监督式机器学习方法做分类做铺垫。证明机器学习方法在 BCG 信号这种时间序列上的可行性。同时本章还介绍了本文所使用的数据示例和标签标记后的数据结果。

本文的第三章分别介绍了本文所使用的算法随机森林、lightgbm、逻辑回归、支持向量机、BP 神经网络的算法原理，为后文使用这些算法作 BCG 床垫信号的分类训练做铺垫。

本文的第四章是为使用监督学习的方式判断体动非体动方式而采集特征。在这一章中，通过介绍如何划分子序列，如何采集特征，在序列分类中分别采用了 lightgbm、逻辑回归、随机森林、支持向量机、BP 神经网络对采集的特征在子序列上进行提取后训练，最后得到的结果是 lightgbm 的训练效果最好，其中相对重要的特征为 diff_kurt、diff_skew、diff_var 即一阶差分后的偏度，一阶差分后的峰度、一阶差分后的方差，根据 BCG 信号的特点也不难解释为什么这几个特征相对重要。因为差分的作用就是减轻数据之间的不规律波动，使其波动曲线平稳，而偏度、峰度、方差这几个特征又恰好是描述一段信号平稳性所必须的。

5.2 展望

本文虽然使用几种机器学习的方法对 BCG 床垫信号在床状态进行了判别，但是由于时间和学术水平的限制，课题中仍有很多地方可以优化和改进。如：

在使用传统的监督式机器学习算法中，最重要的一环就是特征的构建和寻找重要特征，我们可以从多个角度，多门学科中来挖掘更多更有效的特征来用于分类在床状态。但这需要更多的信号、统计相关的更深层次的知识，希望后面有望能找

出更具影响力的特征。

在算法的分类过程中，我使用了五种分类器进行训练对比，但是每一种分类器都有其特有的调参方法，后续希望可以使用更多的调参方式来提高训练的准确率。

在本文中使用的分类器均为传统的提取特征再使用分类器进行训练的做法，并未使用深度学习的方式对 BCG 信号的在床状态进行判别，深度学习不需要对特征进行提取，只需要选好所需要的网络构建网络模型在进行训练即可，避免了提取特征的痛苦，像本文中的基于时间序列的信号数据就比较适合使用 RNN 或者 LSTM 循环神经网络来进行建模和训练，可能比传统的提取特征的方式效果更为出色。

致　谢

本科四年匆匆而逝，回望过去四年的校园时光，历历在目，在这段最难忘的青春岁月里，有幸结识了生命中难以忘怀的老师、同学和朋友们，使我从一个踌躇、彷徨的高中毕业男孩，成长为荷载专业知识技能“腹有诗书气自华”的阳光大学生。面对未来的种种不确定，内心淡定、平和，但充满自信。因为四年的锤炼让我能够认识自己，我有能力去赢得未来，让专业技能更好的释放，实现自己的人生价值和社会抱负。

我首先要感谢我的家人，因为疫情在家感谢父母在疫情期间的操劳，让我能衣食无忧。感谢你们的双手，鼓励，让我的个性得到张扬，并推动我在自由的大学里快速找到自己的位置！

还要感谢敬爱的老师，亲爱的同学们，让我不为过去的四年留下太多的遗憾；

还有工作室的小伙伴们，新老朋友们，帮我打开窗户，让我感受到了大学里不一样的精彩！

感谢我的室友们，一段真诚浇灌的感情，几个五湖四海的兄弟，书写了我青春中最靓丽的色彩，感谢一路有你们。

最后，感谢所有在我的人生路上关心我、帮助我的人，谢谢有你们！

参考文献

[1] 闫瑞阳. 基于机器学习的 BCG 信号心率提取方法研究. 电子科技大学 2018

[2] 黄鑫. 压电复合材料性能参数预测. 兰州理工大学硕士论文 2007

[3] 董师师. 随机森林理论浅析. 集成技术 2015

[4] 张万库. 详解随机抽样. 中学生数理化：高一版 2012

[5] 乐明明. 数据挖掘分类算法的研究和应用. 电子科技大学硕士学位论文 2017

[6] 王立国. 非线性支持向量机判别阈值的设置. 黑龙江大学自然科学学报 2011

[7] 葛怡然. 光电脉冲信号的时域和频域分析与测量方法研究. 吉林大学；葛怡然硕士论文 2010

外文资料原文

Definition and Physiological Relevance

Imperative to the discussion of the physiologic origin of the ballistocardiogram is the clarification of what constitutes a ballistocardiogram in the first place. As mentioned in the introduction, mechanical motions due to cardiac and hemodynamic events have been recorded from multiple locations, with multiple types of sensors (position, velocity, and acceleration), leading to a confusing number of techniques and signals, sometimes related, sometimes not. This multitude of methods has certainly contributed to blurring the field in the past, and care should be taken not to repeat this situation.

The ballistocardiogram is defined as the reaction (displacement, velocity or accele ation) of the whole body resulting from cardiac ejection of blood. Consequently, it is an integration of multiple forces related to movements of blood inside the heart, inside the arteries (primarily the aorta), and movement of the heart itself. It is inherently a 3D signal, although most measurement techniques focus on the longitudinal, head-to-toe component. Its interpretation has been rendered more difficult by the fact that the signal is dependent on the measurement method. Early on, an effort was made to standardize the measurement techniques and signal labeling in order to help comparison and dissemination of data. In the case of the classic, Starr-based longitudinal BCG, there is a general agreement that the early peak is related to the motion of the heart early in systole, and that the main IJK complex is related to the ventricular ejection and aortic flow. There is less agreement on the later waves. While the BCGs of healthy people can be rather well interpreted in light of physiologic events, BCGs of patients with cardiovascular diseases tend to be more difficult to interpret because of the complex interplay of the various internal forces. As a result, interpretation of abnormal BCG has been mostly based on experimental data, and largely qualitative. Since the early work on interpretation, research was aimed at refining the understanding of the signals, using various models and transfer functions, but it did not fundamentally improve the situation. Modern imaging and simulation tools, however, may offer interesting new approaches.

In collaboration with the Cardiovascular Biomechanics Research Laboratory at Stanford University, we started using Computational Fluid Dynamics (CFD) to quantitatively relate BCG signals to hemodynamics. Using Computer Tomography (CT)

models of aortas (where it is believed most of the force related to J wave is generated), we computed the forces at the fluid-solid interface. These forces would be transferred to the whole body through the tight coupling of the aorta to the spine. An example of simulation for a case of aortic coarctation is presented in Figure 3. Of particular relevance is the two-fold decrease in generated force (projection over the longitudinal axis), and the magnitude of the force post- operation – approximately 2 N – similar to normal, measured BCGs. This model is still limited, but points to a novel research direction that can potentially augment the understanding of the BCG signals. We are now looking to extend this model to include lower limbs (in order to provide adequate simulation of pulse wave reflections), and a realistic coupling to body tissues, which will help modeling the mass-spring-damper response of the whole body.

外文资料译文

定义和生理相关性

对心动描记图的生理起源的讨论的必要性是首先要明确心动描记图的构成。如引言中所述，由于心脏和血液动力学事件而产生的机械运动已从多个位置记录下来，并具有多种类型的传感器(位置，速度和加速度)，导致令人困惑的技术和信号数量，有时是相关的，有时却不是。过去，多种方法无疑导致了该领域的模糊化，应注意不要重复这种情况。

心动描记图被定义为由于心脏的血液喷射而引起的整个身体的反应(位移，速度或加速度)。因此，它是与心脏内部，动脉内部(主要是主动脉)内部的血液运动以及心脏本身的运动有关的多种力量的综合。尽管大多数测量技术都专注于纵向，从头到脚的分量，但它本质上是 3D 信号。由于信号取决于测量方法，因此其解释变得更加困难。早期，人们努力标准化测量技术和信号标记，以帮助比较和传播数据。对于基于 Starr 的经典纵向 BCG，人们普遍认为，早期峰值(H)与心脏收缩早期的心脏运动有关，而主要的 IJK 复合体 与心室射血和主动脉血流有关。在随后的浪潮中，人们的共识较少。尽管根据生理事件可以很好地解释健康人的 BCG 信号，但由于各种内在因素的复杂相互作用，心血管疾病患者的 BCG 信号往往更难以解释。结果，异常 BCG 的解释主要是基于实验数据，并且在很大程度上是定性的。自从早期的解释工作以来，研究旨在使用各种模型和传递函数来完善对信号的理解，但并没有从根本上改善这种情况。但是，现代的成像和模拟工具与斯坦福大学心血管生物力学研究实验室合作，我们开始使用计算流体动力学(CFD)将 BCG 信号与血液动力学定量相关。使用主动脉的计算机断层扫描(CT)模型(据信，其中大多数与 J 波有关的力都产生了)，我们计算了流固界面处的力。这些力将通过主动脉与脊柱的紧密连接传递到整个身体。图 3 给出了一个主动脉缩窄情况的模拟示例。与之特别相关的是所产生的力下降了两倍(纵向轴线上的投影)，并且术后力的大小–大约 2 N –与正常的 BCG。这个模型仍然有限，但指向了一个新颖的研究方向，该方向可能会增强对 BCG 信号的理解。现在，我们正在寻求将该模型扩展到包括下肢(以提供对脉搏波反射的充分模拟)以及与身体组织的逼真的耦合，这将有助于对整个身体的质量弹簧-阻尼器响应进行建模。