硕士学位论文

基于肢端动作信号的癫痫发作识别研究

RESEARCH ON EPILEPTIC SEIZURE RECOGNITION BASED ON EXTREMITY MOVEMENT SIGNAL

李业鸿

哈尔滨工业大学

2020 年 6 月

国内图书分类号：TP181　　学校代码：10213
国际图书分类号：004　　密级：公开

工程硕士学位论文

基于肢端动作信号的癫痫发作识别研究

硕 士 研 究 生：李业鸿
导　　师：梁廷伟 教授
申 请 学 位：工程硕士
学　　科：航天工程
所 在 单 位：航天学院
答 辩 日 期：2020 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP181
U.D.C: 004

Dissertation for the Master Degree in Engineering

RESEARCH ON EPILEPTIC SEIZURE RECOGNITION BASED ON EXTREMITY MOVEMENT SIGNAL

Candidate: Li Yehong
Supervisor: Prof. Liang Tingwei
Academic Degree Applied for: Master of Engineering
Speciality: Astronautical Engineering
Affiliation: School of Astronautics
Date of Defence: June, 2020
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

癫痫病目前是世界上第二大神经系统疾病，如今我国癫痫病患者人数也在逐年增加。癫痫病的发作反复而突然，并且伤害隐患极大，许多家庭因此而不堪重负，急需要一种能够快速有效的识别检测预警方法。传统的癫痫病检测识别是通过脑电信号来进行判断，但脑电信号的检测往往脱离不了病房的医疗设备，所以在空间性要求和便利性上有所局限。近年来，因为基于肢端动作信号的运动姿态识别技术和可穿戴式设备领域的迅猛发展，人们试图通过肢端动作信号来找到这样一种能够快速有效识别癫痫发作的技术方法。在此背景下，本文提出了一种基于肢端动作信号的癫痫发作识别方案。

1. 详细介绍目前关于癫痫病发作识别的国内外研究现状，并对其进行深入分析，阐述本次基于肢端动作信号的癫痫发作识别研究的设计方案与具体流程；

2. 采集数据，并对采集到的原始数据进行数据预处理和特征提取，以此来得到可用的训练集与测试集数据格式；

3. 阐述逻辑回归、支持向量机、随机森林等三种分类算法的理论基础，并依此进行模型构建，通过实验对比分析，得到最适用于本课题的分类器算法；

4. 对得到的实验结果进行对比分析，综合考虑后提出一种符合性能指标要求、快速有效的基于肢端动作信号的癫痫发作识别的最终方案，并对其进行在线测试和有效性验证。

主要设计了由数据采集、数据预处理、特征提取、算法模型构建和在线测试等五个步骤形成的基于肢端动作信号的癫痫发作识别的具体方案。其中使用合成加速度的方式来解决信号方向的不确定性问题，并通过多个实验来验证数据选择、窗口时间大小、数据均衡等对于分类器识别效果的影响，最终得到适合本课题的全套细节参数和具体代码实现。

关键词：肢端动作信号；癫痫发作识别；逻辑回归；支持向量机；随机森林；

Abstract

Epilepsy is now the second largest neurological disease in the world, and the number of epilepsy patients in China is increasing year by year. The seizure of epilepsy recurrent and sudden, and great harm hidden danger. Many families overwhelmed because of this, in dire need of a quick and effective method to detect early warning recognition. The traditional detection and recognition of epilepsy is made by eeg signals, but eeg signals are often not separated from the medical equipment in the ward, so there are limitations in space requirements and convenience. In recent years, with the rapid development of motion gesture recognition based on limb movement signals and wearable devices, it is attempted to find such a technical method that can quickly and effectively identify epileptic seizures through limb movement signals. Based on the background of this subject, this paper proposes a seizure recognition scheme based on limb movement signals.

1. To introduce in detail the current status of epilepsy seizure recognition at home and abroad, and to carry out an in-depth analysis of it. In addition, the design scheme and specific process of this study on epilepsy seizure recognition based on limb movement signal are described.

2. Collect data, and to get the raw data for data preprocessing and feature extraction, in order to get the available training set and testing set data format.

3. The theoretical basis of three classification algorithms, including logistic regression, support vector machine and random forest, is expounded, and the model is constructed accordingly. The classifier algorithm most suitable for this topic is obtained through comparative analysis of experiments.

4. Analysis of experimental results, synthetically considering conforms to the performance index, fast, effective seizures recognition based on limbs movement signal final plan, and carries on the online testing and validation.

In this paper, the specific scheme of seizure recognition based on limb movement signal is designed by data acquisition, data preprocessing, feature extraction, algorithm model construction and online test. The effects of data selection, window time size and data balance on the classifier recognition effect were verified through multiple experiments.

Keywords: Extremity movement signal, Seizure recognition,
Logistic regression, support vector machine, random forest

目 录

第 1 章 绪　论

1.1 课题背景及研究的目的和意义

1.1.1 课题来源

本课题来源于企业需求和现实需要。目前我国有许多家庭正遭受着癫痫病症的困扰。往往一个癫痫病人需要 1-2 个人来进行看护和照顾，这对患者家庭本身就是一个很大的负担。而传统的癫痫检测识别手段，是医生配合脑电医疗设备来进行实时的监控和诊断。这种方式要求患者必须时刻呆在病房里戴着复杂的脑电医疗仪器，很不方便。所以考虑到癫痫病人对新型检测识别方式的巨大需求，本课题聚焦于癫痫发作时的肢体动作表现，通过理论分析和实验验证，基于肢端动作信号来进行几种机器学习算法的训练和对比，综合提出一种基于肢端动作信号的癫痫发作识别方法，并要求能够有效地进行在线识别诊断。

此外，因为考虑到患者的方便性要求，所以本次肢端动作信号是由体积小、方便佩戴的腕带式癫痫检测装置来进行数据采集。本课题所指的肢端动作信号主要包括三轴加速度信号、皮电信号、体温信号和心率信号等。而近年来由于微传感器技术和 MEMS 系统的蓬勃发展，许多可穿戴式设备已经不仅局限于运动领域，开始向医疗器械等领域扩张。例如作为多传感器集合的腕带式癫痫检测装置，不仅能够进行三维加速度和角速度等位置信号的测量，也能够测出皮电、肌电、心率和体温等多种肌体信号。所以在此硬件基础上，完全可以推进本课题相关方向的算法研究。

1.1.2 研究意义

癫痫属于慢性脑功能障碍综合症，它的发病病因复杂难测，主要特征表现为由脑部神经元异常放电所引起的突然性、反复性和短暂性的中枢神经系统功能失常[1]。在癫痫病人发作时，往往还会伴随着暂时性的意识模糊或不受控的剧烈抽搐，所以在此期间如果没有获得行之有效的救助，轻则损害神经，严重者或伤及生命，不容小觑。如今全球约有六千余万人正饱受癫痫疾病的折磨，在中国也有七百万到一千万人身患癫痫[2]。且中国的癫痫病患者与日俱增，每年约增长四十余万人。由于我国人口老龄化的日益严重，癫痫病发病率已呈现逐年上升的趋势，整体情况不容乐观。

癫痫病的发作病因繁杂多样，导致其主要发病机理目前仍没有被完全探明。同时因为脑部神经系统的难以观测，所以在很多情况下无法完全解释其对应症状的结构或代谢的异常。癫痫发作的临床表现症状有很多种，其中主要以三种类型为主：简单部分性发作、复杂部分性发作、全身强直阵挛发作，其中最常见且危害性最大的就是全身强直阵挛发作（generalized tonic-clonic seizures, GTCS）。在 GTCS 发作时，患者本身可以会因为高热不退或由神经元兴奋引起的毒性损伤等对大脑造成永久性伤害，且发作期间易产生电解质紊乱、循环衰竭等其它症状，严重者或伤及性命[3,4]。而除了癫痫自身带来的病痛之外，患者还可能在其突然发病时发生摔伤、晕倒等其它情况，从而对身体健康带来更多的伤害。所以癫痫病反复突然的发作使病人和家庭的正常生活都困苦不堪，很大程度上影响到了日常行为活动和身体健康状况。因此，对于癫痫发病机理的探究，在医学领域与社会生活中都有着重大意义。特别是快速而准确的癫痫发作识别能够帮助患者、家属与医护人员在第一时间内观察到病情发作，从而可以帮助患者脱离险情并及时用药来抑制癫痫小发作，避免 GTCS 的发生。

如今癫痫发作的主要诊断方式是通过记录和读取患者个人详实完整的病历，依此结合病人的临床表现与脑电图记录来诊断识别。常使用脑电图作为检查手段和识别依据，是由于脑电图本身能够记录出癫痫病人发作时的脑部神经元异常放电的活动数据。因此它成为了探索癫痫特征的重要工具，并被广泛应用于包括癫痫诊断、癫痫发作定位、定性研究、癫痫发作预测和癫痫发作控制在内的各项癫痫研究中[5]。目前在医院里的癫痫发作识别，主要仍是由专业医生观察病人临床表现和脑电图异常信号来做出判断。

但实际上通过人工来进行癫痫发作识别其实有诸多弊端。例如，伴随着癫痫病人的观察时间的增长，脑电图的采集数据会越来越多。医生在面对着这海量的脑电图数据时，在判断识别上极其耗费精力且效率相对较低。其次，医生本人也会受到精力下降的影响，在长时间的工作过后进行判断可能会导致误判病例的增加。此外脑电图显示的部分癫痫特征的表现都非常微小，只依据人眼很难区分辨别。而且癫痫发作的形式多样，发作期间常有如肌肉收缩、肢体抽搐、眨眼伪迹等运动的产生来干扰脑电癫痫特征的表现。不同个体的发作形式差异明显，同一患者的多次发作的发作机理和发作形式也不尽相同[6]。所以仅依靠医生个人的主观经验，也常难以进行准确判断。综上可见，进行癫痫发作的自动识别算法研究对于癫痫病的诊断治疗具有重大的现实意义。

近年来，许多科学家致力于通过机器学习和深度学习的方法来建立起基于脑电图信号的癫痫发作自动识别的智能系统。但因为脑电信号本身的高度复杂

性、非线性和非平稳性等特点，使得对于脑电信号的有效分析始终是国内外研究的难点。且因为监测脑电信号的仪器往往多而繁杂，不能满足癫痫病人随时随地的空间适用性和方便性要求。所以本课题把焦点放在癫痫发作时的不受控抽搐的肢端动作特征上，试图通过肢端动作的多个维度信号来进行有效分析和发作识别。通过对手臂多个信号的分析处理与脑电信号的研究相结合，来共同完成对癫痫发病机理的认识和诊断，从而将癫痫病人从繁杂的脑电监测和病房中解放出来，减轻整个患者家庭的负担。

1.2 国内外研究现状

1.2.1 国外研究现状

国际上早在二十世纪七十年代就有了部分对于癫痫发作识别的研究。那时候最早的癫痫发作机制的研究主要是关注医学和生物层面的神经成分，例如神经递质受体或特定载体的研究等[7]。因为癫痫病的发作，本质上还是属于脑部神经异常，所以试图通过脑部神经生物信息来找到识别方法。例如 Niederhoefer 等人[8]就是从原始生物信号里提炼出有效信息来进行癫痫发作的判断。这一类方法的主要思想是试图在生物信号的参数分析中来找到癫痫发作的生物特征，以此来进行识别。

延续这条思路研究下去，科学家们很快就发现了对脑电图信号（Electroencephalogram, EEG）的分析可以更好地来探究癫痫病的发作机理，因为脑电图信号本身就是特定时间戳的脑电生理状况的直接反映。所以科学界就开始对脑电图信号与癫痫发作之间的关系进行了深入的研究。1975 年，Viglione 等人通过对多位患者的 7 次癫痫发作 EEG 数据进行统计分析和对比研究，提取出了部分能够进行有效预测癫痫发作的脑电特征。于是他们基于这些特征设计了癫痫预警电子装置，尽管该装置的确准确预判了部分发作，但同时漏报率和误报率也居高不下。1998 年，Osorio 等人[9]通过时频分析的方法，对大量发作间期以及发作期的癫痫患者脑电信号样本进行实验分析，结果发现其中有 92%的癫痫发作是可以通过时频分析的方法来预测到的，平均预测时间大约为 15.5s。所以依据以上研究可以发现，癫痫发作本身并不突然，它同样也是需要一个长时间的积累才会发病，只是这些特征过于微小而难以观察。实际上，癫痫的发作在脑电图信号上具有可预见性[10]。

所以基于 EEG 的癫痫发作识别实际上是属于模式识别的问题，只不过是具体应用的背景为癫痫发作。因此可以参考模式识别的方法，建立起学习算法

模型来训练发作间期与发作前期的样本[11]，以此达到癫痫发作识别的目的。在2009年，Ghosh Dastidar等人[12]和Subasiursoy和Acharya等人[13]就使用主成分分析法来对癫痫患者的脑电信号进行分类，依次用PCA、ICA和LDA降维的方式来进行癫痫发作识别。而O. Faust[14]等人则是通过基于AR自回归模型的参数估计法来对癫痫病患者的脑电图信号进行功率谱密度的计算和比较分析。2010年，Chisci等人[15]则通过AR模型、最小二乘参数估计器与二进制SVM分类器三者结合，来对脑电信号进行特征提取和模型构建，以此对癫痫发作的发作前期、发作期和发作间期三个阶段进行分类。2012年，Zandi[16]对患者脑电信号图进行移动滑窗后，通过对过零率和过零点的特征分析，以此来研究癫痫发作识别方法，他在测试了561小时的EEG信号之后，得到了88.3%的准确率。2016年，Supriya[17]依据图论的思想来进行癫痫发作研究，通过建立不同核函数的支持向量机与k-近邻等分类学习器来做对比实验分析，发现该方法在发作识别上精度很高，但在预测的效果上精度较低。2017年，Hamad[18]等人使用离散小波变换来对采集来的EEG信号进行特征提取，然后选取差异较大的特征来作为训练集和测试集数据输入给支持向量机进行分类学习，最终模型的分类正确率为95%。与此同时，Sriraam等人[19]对Bern Barcelona数据集进行了EEG信号的特征提取，并经过统计检验和综合考虑之后，最终提取出了21个特征作为支持向量输入SVM中进行训练学习，得到了92.15%的癫痫发作识别准确率。

但正如前面研究意义所述，尽管基于脑电信号的癫痫发作识别方法已经到了一个成熟发展的时期，但是由于脑电信号的获取条件往往对传感器和设备的要求较高，所以不能满足病人癫痫发作实时检测的空间性要求和适应性要求。于是在此需求基础上，随着可穿戴式设备在小型化、智能化、运算速度加快等趋势上的不断发展，在癫痫、帕金森之类的精神病识别领域，众多科学家把目光聚焦于肢端动作信号上，依托可穿戴式设备采集的肢端动作信号来进行检测识别。

例如在辅助医疗帕金森领域，就有一款名叫Kineti Graph的健康智能手环。它通过记录佩戴人的肢端动作信号，来实时监测佩戴人的运动状态，并得到完整的震颤监测报告[20]。这对于帕金森患者来说，不仅可以记录和反映出病情发展，而且能够辅助医生根据该报告进行配合用药、对症治疗。如今这款智能穿戴式健康设备已经取得美国联邦医药管理局的许可，帕金森症病人只要简单地在家里佩戴好KinetiGraph手环就可以持续地监测帕金森病情。KinetiGraph手环如下图1-1所示。

图 1-1 KinetiGraph 手环

而在癫痫发作识别监测领域，美国企业 Empatica 则走在了国际研究的前沿。早在 2007 年时，公司研究小组的 Marieke van Dooren 等人[21]就发现了使用皮肤电活动和运动信号来检测全身性强直-阵挛性癫痫发作的有效性。随后 Empatica 的研究人员不断地改进癫痫发作识别算法，于 2018 年初推出了第一款癫痫发作监测手环 Embrace，被美国 FDA 批准为官方医疗设备[22]。如今已经推出第二代 Embrace 2，如下图 1-2 所示。该产品的主要特色是个人定制化。基于机器学习算法的它可以通过积累用户的运动数据和发病周期，来进行反复训练学习，从而提高识别能力。采集到的个人相关发病数据越多，算法模型的效果和准确率则越高。

图 1-2 Embrace 2

1.2.2 国内研究现状

在癫痫发作识别方法的研究领域，我国起步时间相对较晚，但近年来也在大力推动相关医疗机器学习算法的研究。关于癫痫病的发作特征，早在 1980 年张香桐院士就在《癫痫问答》书中提到，临床治疗时观察癫痫病人的发作，发现在发作前病人会表现出部分特殊的征兆，可以利用该征兆进行癫痫发作的识别与预测。

2015 年，戴若梦采用深度学习算法中的稀疏自编码器 SAE 和卷积神经网

络 CNN 的方法对运动想象脑电信号进行分类[23]，结果表明 SAE 在分类脑电信号时准确率比以往传统方法有所提高，但也花费了更多时长；而 CNN 则是在脑电信号识别的精度和图像识别的精度上有较大差距，准确率表现不好。太原理工大学的周梦妮通过排列熵的特征和支持向量机的方式来探索复杂度和时频分析在此方面的应用，得到了很好的效果[24]；北京邮电大学的郑天依基于小波双谱能量熵和颜色距的特征提取，再利用双子支持向量机来构建算法识别模型，在有限时间内得到了 85%以上的识别准确率[25]。

而在基于肢端动作信号的癫痫发作识别领域，我国目前仍然处于空白。对于肢端动作信号，我国科研人员的研究重点主要还是放在运动状态的识别上。例如华南理工的吴海龙，他基于统计模式的识别算法来完成对游泳状态的检测识别，平均准确率约为 78%[26]。山东大学的黄彬则提出了一种创新的识别跌倒检测算法，他先是巧妙地利用支持向量机来找到疑似摔倒的数据，再将这些疑似数据进行分离处理，最后输入到 k 近邻算法中学习训练[27]。

但近年来随着可穿戴式设备由运动领域向医疗领域的扩张，部分基于肢端动作信号的医疗仪器研发论文也开始出现。2017 年电子科技大学的段豪[28]基于 Android 平台建立了帕金森监测预警系统，他对比了传统 BP 神经网路、CNN、RNN 等三种深度学习算法，最终选用 CNN 卷积神经网络来进行帕金森病症识别，得到了 88.7%的识别准确率，成果斐然。

1.2.3 国内外文献综述

综合国内外的研究人员在癫痫病发作识别与预测方面做的工作来看，癫痫发作状态可以通过分析脑电图 EEG、病人体表皮肤电阻的变化以及发作时全身痉挛所带来的振动等信号来进行识别乃至预测。由于这些信号都是时序信号，所以国内外对它们的研究方法也是大致相同——首先进行时域分析、频域分析、时频域分析、非线性动力学分析、熵等特征构建和选择，然后通过支持向量机、K-近邻、决策树、神经网络、随机森林等机器学习方法来进行分类识别。

且由上述资料我们可以看到，国内外对基于 EEG 的癫痫发作识别和基于肢端动作信号的运动状态识别研究等几个领域都有了一定的基础和发展。在基于脑电信号分析癫痫发作的方面，算法识别准确率已经高达 85%，而且很多研究也证明，癫痫发作前的确存在很多征兆可以用来预测。此外随着诸多基础技术的蓬勃发展，越来越多的可穿戴式设备向医疗监测方面进行扩展，已经出现了很多成熟的产品和技术手段。正因为这几个基础条件已经成熟，所以我们可以预见到，基于肢端动作信号的癫痫发作识别将会是医疗监测识别领域下个阶

段的热点和难点。

在市场需求方面，众多癫痫病人和家属都亟待一款这样的产品来解放他们，以至于病人可以摆脱沉重的脑电监测仪器和病房，家属可以摆脱无时无刻的看护而仅需等待监测系统的预警即可。这块市场需求是很庞大的，目前已知的也仅有 embrace 一家公司有相关产品问世。

所以在技术手段成熟和市场需要的前提下，面对国内相关方面研究仍处于空白的情况，本文把目光聚焦于癫痫发作时的不受控抽搐的肢端动作特征，在借鉴相关运动状态识别研究的基础上，对比几种算法模型，提出一种基于肢端动作信号来精准判断癫痫病发作的识别方法，并能够有效地进行在线识别诊断。

1.3 课题主要研究内容

本文将基于多维度的肢端动作信号来建立起对癫痫发作进行有效识别的在线癫痫诊断系统，主要包括数据采集、数据预处理、特征提取、算法模型构建、在线测试等几个部分，如下图 1-3 所示。

图 1-3 主要研究内容

A. 数据采集

本文拟通过精心设计的基于 STM32F103 芯片及其外围电路构成的腕带式癫痫检测装置来进行数据采集。数据主要是从国内著名癫痫医院采集而来，由知名医生通过专业医用设备来进行配合标注癫痫病大发作 GTCS 的时间段分布，所以数据来源真实有效。

B. 数据预处理

因为采集过程中不可避免的环境热噪声问题和仪器本身带来的精度问题，所以采集的数据必然需要经过一定的滤波降噪处理，并进行定标阈值处理以消除小振动干扰。除此之外，为了后续有效的特征提取，也需要对采集来的数据进行数据选择和数据分割的预处理操作。

C. 特征提取

特征提取对于模式识别尤为重要，将直接影响到算法模型的分类效果。本文首先通过时域分析、频域分析等方式进行特征构建，再利用随机森林模型的特征重要性功能来进行特征选择，最终选定部分高效的特征来表示数据，在保留数据原有规律的同时，减少数据中的不确定因素。

D. 算法模型构建

采用所选择提取出来的特征，进行癫痫发作识别算法的设计。本次研究将对比 Logistic Regression（逻辑回归）、SVM（支持向量机）、Random forest（随机森林）等三种机器学习模型，通过对训练学习的结果分析，综合提出一种能够进行有效在线测试的识别算法。

E. 在线测试

实验分两部分实现。一是还原机器学习模型得到的决策边界，在腕带式癫痫检测装置底层编程，并进行现场测试。二则是在 PC 端进行在线模型判别。

1.4 研究内容及规划

第一章主要讲述了课题来源以及研究意义，对癫痫病症特征研究、基于 EEG 的癫痫发作预测识别以及基于肢端动作信号的机器学习算法等国内外研究现状进行分析，确定了研究内容和具体相关工作，然后对文章各章节的具体内容做出规划。

第二章主要是针对采集来的原始数据信号进行数据预处理，以便获得干净有用的数据。并针对数据信号进行特征相关的分析研究后，进行特征构建和特征选择，为后续的算法模型构建提供高效的数据输入。

第三章主要是介绍三种不同思路、不同应用的机器学习算法，并进行相应的算法模型构建，通过训练学习，对该算法进行系统的实验验证。

第四章主要是对算法的实验结果进行多层次的系统分析，综合提出一种适用于实际生活的快速有效的癫痫发作识别方法，并进行在线诊断测试。

结论主要对基于肢端动作信号的癫痫发作识别方法研究取得的成果进行总结，并对其未来的发展做出一定展望。

第 2 章 数据预处理与特征提取

每个计算机实验都从数据的预处理开始起步。本章首先简单介绍了腕带式癫痫检测装置的硬件基础，然后从本次数据的采集开始，具体描述数据预处理和特征提取的多步操作，最后通过数据集处理，来得到可用于算法模型输入的训练集和测试集等数据格式。

2.1 硬件基础与数据采集

由基于 EEG 的癫痫发作识别方法转型到基于肢端动作信号的癫痫发作识别，其应用的理论依据不同。EEG 脑电信号可以及时地反应出病人脑部神经异常放电的图像，所以这是从癫痫发作的脑部表现来进行识别。而肢端动作信号，则是依据癫痫病人大发作时手臂肢体的不受控抽搐或者全身性强直等表现出发。所以考虑到病人发作时肢端动作的表现，本次肢端信号主要由三轴加速度信号、皮电信号、心率信号、体温信号等多维数据组成。

选取三轴加速度信号，是因为它是常用且重要的位置信号。通过它，我们可以参考论文进行大多数的常见运动状态识别，比如计步、奔跑、上下楼检测等。在本课题中，它是体现病人肢端抽搐活动的最有效的信号之一。使用皮电信号，是参考了 Embrace 公司研究小组的论文结果，他们发现并且实验验证了皮肤电活动对检测癫痫大发作全身强直性表现的有效性。心率和体温信号，则是作为辅助信号来辅助判别，在病人发作时心率和体温较平常状态会有所升高，身体表现出紧张痛苦的特征。

在确定了本次肢端动作信号主要由以上信号组成之后，则开始进行硬件基础的设计和选用。本次课题选用的是基于 Cortex-M3 内核的 STM32F103 芯片及其外围电路设计的腕带式癫痫检测装置。

传感器模块由 STM32F103 芯片进行主要控制。其中加速度传感器选用的是常用的 6 轴运动处理组件传感器，它可准确追踪快速与慢速动作，性能基本上符合本次实验的精度要求，并且以后还可以拓展角速度采集，以实现更高精度的运动姿态识别。

蓝牙传输模块则主要是以 STM32F103 芯片外接蓝牙传感器为主，辅以外围电路进行设计。其主要的目的是确保不丢包的将数据稳定传输到智能手机上或

者 PC 端，以便后续的数据处理和机器学习在线诊断。

而此处为了解决不稳定传感器的异步传输丢包问题，考虑到传感器性能频率大于实验所需要的频率，本课题应用缓存队列重采样的方式，从芯片中划出一定的异步队列对传感器采集来的数据进行缓存操作，输入数据可以是不稳定频率的数据，但输出端输出或者蓝牙传输的是固定采样频率的数据包。

由于考虑到整体设备的功耗，本次数据采集三轴加速度传感器、皮电信号、体温信号、心率信号的采集频率均为低频率采集。不同的传感器采样数据不同，但采样格式保持一致，以三轴加速度为例，每个采样点记录采样时间、三轴加速度数据，采样格式为<X 轴加速度，Y 轴加速度，Z 轴加速度，采样日期，采样时间>，另外几个数据的采样格式为<皮肤体温，皮肤电，心率，采样日期，采样时间>。数据格式示例如表 2-1 所示。

表 2-1 采样数据示例

xAccel	yAccel	zAccel	dataTime	dataTime
0.023	-0.036	-0.018	2018/8/18	0:25:00
0.023	-0.031	-0.019	2018/8/18	0:25:00
0.021	-0.036	-0.007	2018/8/18	0:25:00
0.022	-0.034	-0.014	2018/8/18	0:25:00
0.019	-0.032	-0.007	2018/8/18	0:25:00
0.022	-0.036	-0.018	2018/8/18	0:25:01

skinTem	skinIm	HeartRate	dataTime	dataTime
33.3	5955.2	71.14	2018/8/18	0:25:00
33.3	5947.3	71.02	2018/8/18	0:25:01
33.31	5947	71	2018/8/18	0:25:02
33.29	5948	71	2018/8/18	0:25:03

本次课题的数据采集是在 2018 年 8 月从国内著名癫痫医院采集而来。考虑到病人发作时会伴随的肌肉强直或不受控的抽搐，所以把检测装置紧紧戴在手腕处进行数据采集。采集数据的现场如图 2-1 所示。

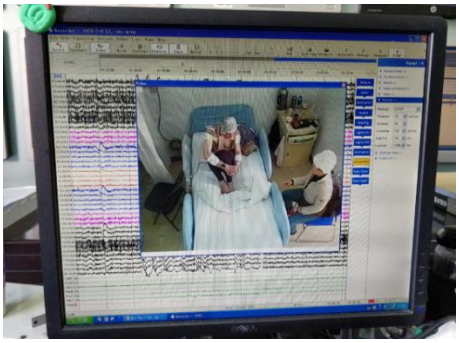

图 2-1 采集数据现场

由图 2-1 我们可以看到，在用腕带式癫痫检测装置进行数据采集时，病人同时也佩戴着专用的医疗仪器进行检测。所以我们本次采集的癫痫大发作 GTCS 的时间段分布，是由知名医生通过专业医用设备来进行配合标注的，数据来源真实可靠。具体发作数据如下图 2-2 所示。

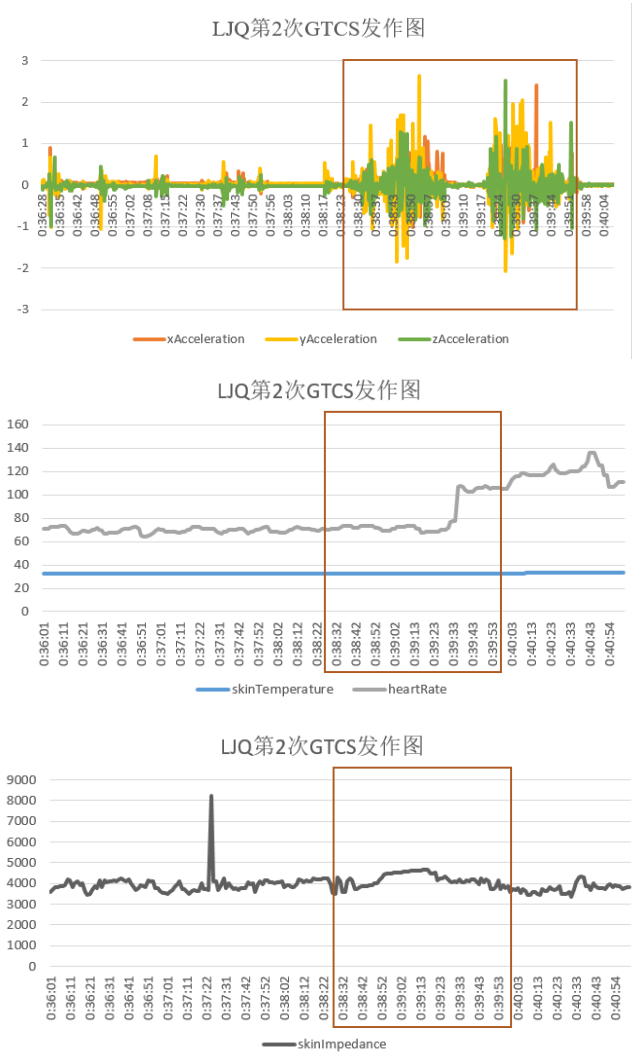

图 2-2 LJQ 第 2 次 GTCS 发作数据

在图 2-2 中，因为各维数据单位指标不同，所以只能把三轴加速度、皮电和心率、皮肤电信号分别进行展示。其中橙色的线框，就是医生配合脑电信号判断病人发作的起始时间段。

由图 2-2 中易知，患者 GTCS 大发作时，三轴加速度的幅度和频率与平常表现都有明显的区别；体温、心率、皮肤电信号等则与平常表现没有明显区分。且在进行后续实验操作中我们发现，体温、心率、皮肤电信号因为表带设计的不合理，大多数病人佩戴检测装置时，装置不紧扣而导致与病人的手腕经常脱离，使得后三维信号跳跃严重，存在明显的失真现象，所以后续的实验预处理

暂不对其进行分析。

本次采集到的数据共有 17 组（人），其中观测到大发作（GTCS）共有 10 组（人），数据特征相对明显的有 22 人次。其余 7 组（人）在观察期间没有发病或者小发作不易识别，所以暂不用作为数据集。所以本次的有效数据仅为 22 次大发作，累计时间约为 10 小时。

2.2 数据预处理

由于腕带式癫痫检测装置的硬件工艺水平限制以及少许环境噪声的影响，本次采集得到的原始信号里不可避免地存在一些干扰噪声和精度损失。为了确保算法训练学习样本的可用性，所以必须对原始数据进行一定量的数据预处理，以提取出干净有效的数据。此外在应用到各机器学习算法之前，我们也需要通过提纯后的数据信号才能提取出稳定高效的特征，因此对原始数据的预处理不可缺少。

2.2.1 初步预处理

为了提纯原始数据，我们现在主要考虑以下两点干扰因素：

（1）由于肢端动作方向的不确定性，导致三轴加速度的各自表征不明显；

（2）由于工艺水平限制和环境因素干扰，数据存在噪声和零漂现象；

我们可以针对这两点干扰因素，分别进行对应预处理。

（1）对三轴加速度进行合成，来提取关键信息

对于腕带式癫痫检测装置来说，因为肢端剧烈运动时手腕的翻转而使得加速度数据的有效轴时刻都在变化，所以单一地提取任一轴的数据都不足够对运动进行完全的表征。

而与此同时，x、y、z 三轴中总会有至少一个轴具有突出的运动特征和相对明显的周期性变化[29]。所以基于此特点，我们使用合成加速度的方式来消除运动有效轴变化的影响，以提取出关键有效的运动信息，计算公式如式(2-1)所示：

$$\mathrm{A} = \mathrm{sign}（A_x * A_y * A_z）\sqrt{A_x^2 + A_y^2 + A_z^2} \tag{2-1}$$

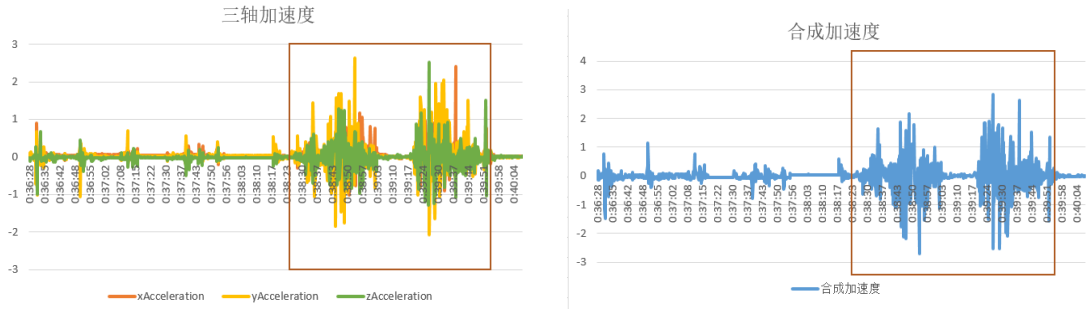

图 2-3 三轴加速度和合成加速度对比

从图 2-3 的对比中我们可以看到，合成加速度的图像不仅包含了原始数据的方向信息，而且对于某一轴峰值等突出运动特点和变化都更加细致地表征。

（2）抑制样本数据噪声—平滑滤波

因为采样环境热噪声与采集过程中手腕异常抖动的原因，原始数据不可避免地会有噪声和零漂现象的存在。在进行了对癫痫发病肢端动作的理论学习之后发现，癫痫发作时肢端动作的频谱能量主要集中在 20Hz 以内，尤其以 10-20Hz 的频谱变化最为明显[30]。

在参考了 Yang J 等人[31]使用阈值和均值滤波来完成数据预处理中的平滑、去噪操作后，本文对采集到的 22 次大发作数据进行了定标统计。通过对样本的统计分析，发现零值附近小振动的数值普遍在 0.05 以下，所以针对 0.05 做了一个阈值过滤操作，把零漂的小振动干扰消除。

随后为了不影响本身数据，考虑到后续滤波程序会移植到检测装置底层芯片上会有一定的计算资源要求，所以本次选用中值滤波的方式，通过 python 库中的 medfilt 中值滤波函数来实现。中值滤波的优势主要是消除椒盐噪声的影响，是图像滤波和机器学习预处理的几个主要的滤波方式之一。本次邻域选择为 5 点，即在 5 个点范围内做中值滤波。滤波前后对比如下图 2-4 所示。

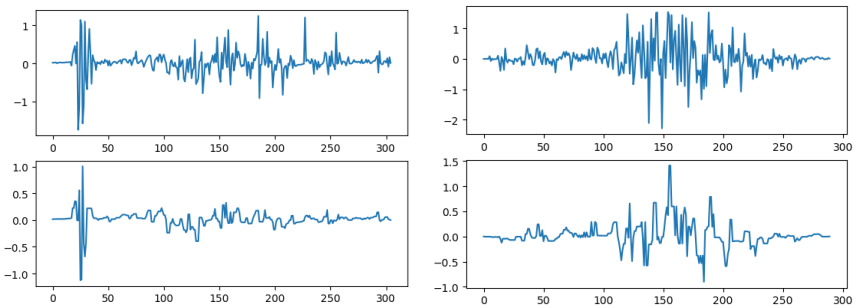

图 2-4 中值滤波前后对比（上图为滤波前，下图为滤波后）

从图 2-4 中不难看出，进行了 5 点中值滤波平滑操作以后，大多数高频噪声都被过滤掉，但整个信号的波形形状却没有改变。这对后续的分割滑窗操作和下一步的特征提取都有着积极的影响。

2.2.2 数据选择与数据分割

（1）数据选择

虽然癫痫病的发作机理尚未完全探明，但是基本上可以确定它属于脑部神经疾病，所以其主要的病理表现是脑部神经信号的异常放电。而本次课题所采用的肢端动作信号，其实是一种癫痫病大发作的附加表现。所以从这一个角度来说，所采集到的发病时间本身就存在由脑到手传递的延迟。此外，本次数据采集的发病时间，是由知名医生通过脑电信号来进行诊断得到标注的，所以也不可避免地存在一定量的判断时间差。所以基于此，我们必须做好发病时间段的数据选择，才能够确保数据的准确性和模型训练效果。

同理，从上图 2-2 中我们可以看到，日常信号中有剧烈波动的片段，也有平缓的片段，分别对应病人运动走动和躺在床上等日常活动。如果我们本次正常数据样本全部选用平缓的片段，则毋庸置疑模型训练效果会很好，但与此同时模型泛化性不强，容易受到运动信号的干扰。

所以为了做好数据选择和验证数据选择对本课题的影响，本课题选用的发病数据段相对原始数据做了一定的调整和优化；选用的正常数据样本设置了一个实验组和两个对照组，实验组的正常数据为波动片段与平缓片段各占比 50%，第一个对照组的正常数据为全部平缓片段，第二个对照组的正常数据为全部波动片段。具体实验结果如第四章 4.2.1 小节所示。

（2）数据分割滑窗

对原始采集的数据进行了初步预处理和数据选择之后，为了适应后面的特征提取和分类器数据格式要求，本课题选择使用滑动窗口的方式对数据进行分割处理。而窗口大小设置是一个重要参数，对算法模型的性能有着重要的影响。

在参考了二十余篇基于 EEG 的癫痫发作识别和基于三轴加速度的运动姿态识别的论文之后，发现一般实验的采样频率都在 20Hz 左右，取最优窗口的点数为 120 点，最优窗口时间为 6 秒，窗口重叠率为 50%。多篇论文中论述，6s 的窗口时间或者 120 点左右为相关窗口长度是最好的识别参数。

考虑到本次实验的采集频率为低频率采集，且要求模型在线测试诊断时间越小越好，所以摒弃 120 点(即 24s)的窗口，选用 6s 作为基本窗口长度，并以此进行对照实验，进行从 3s 到 10s 的窗口长度的效果对比，以便选出本课题自身的最优窗口长度。具体实验结果如第四章 4.2.2 小节所示，图 2-5 为窗口长度 6s 的发病数据图。

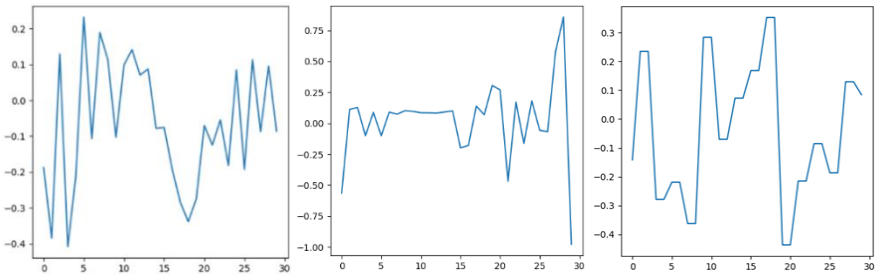

图 2-5 窗口长度为 6s 的发病数据图

除了窗口长度之外，本次实验还需要说明的是窗口重叠率这个参数。因为本次数据采集仅采集到了 22 人次的 GTCS 大发作，相对正常数据来说，发病样本较少。所以采用了重复采样的策略，对于发病数据集来说，进行单点重复采样滑窗，即窗口重叠率为 96%；对于正常数据集来说，则进行 50%重复率滑窗。

综上，本次数据预处理的整体流程图如图 2-6 所示。

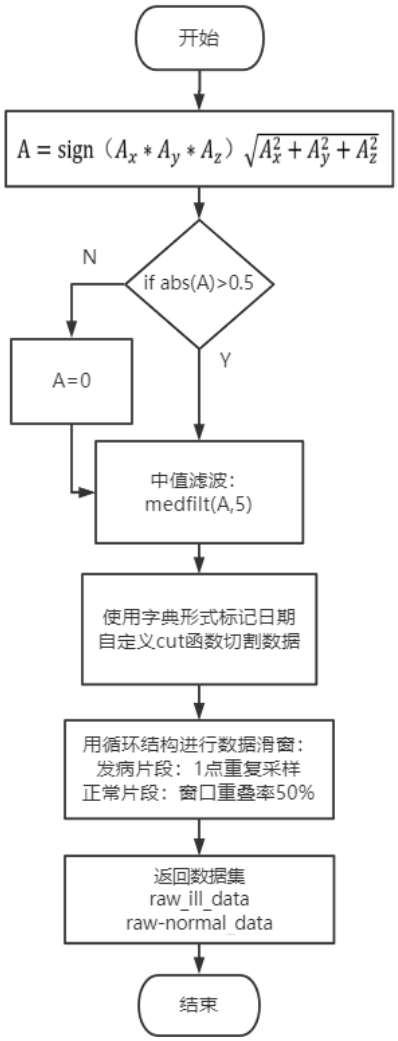

图 2-6 数据预处理流程图

2.3 特征提取

一般来说分类器的识别准确率主要取决于对原始数据的特征表述、分类器算法模型的选择以及学习训练过程等几部分。所以特征提取对于癫痫发作状态识别至关重要，将直接影响到算法模型的识别准确率效果。而构建复杂的机器学习模型，往往也需要通过特征提取来把原始数据进行压缩，在保留数据信息和规律性的前提下减少噪声干扰和数据量。本课题中特征提取主要分为以下两个步骤，特征构建和特征选择。

2.3.1 特征构建

特征构建目前并无固定的工作模式和研究方法，不同的应用背景和分类器选择就需要构建不同的特征，才能得到最优化的识别效果。尽管如此，在参考近年来多篇机器学习和运动姿态识别的相关文献之后，总的来说发现特征构建分别有三个维度、四种方法：三个维度指的是时域分析、频域分析、时频分析；四种方法指的是基于基本统计方法的构建、基于模型的特征构建、基于变换的特征构建、基于分形理论的特征构建[32]。接下来分别从三个维度来进行理论阐述。

时域分析，即是指直接从时间序列数据中提取特征向量，构建出来的特征一般都符合数据的直观表现。其中常用的时域特征有：均值(Mean)、方差或标准差(Std)、相关系数(Corr)、均方根(Rms)、能量(Energy)、平均绝对偏差(Mad)、时域积分(Integration)等。其中的部分特征还具备明确的物理意义，比如平均值代表的是信号中的直流分量、方差刻画出信号的变化程度、四分位距则体现了信号的分散情形。

频域分析，即需要对数据信号先进行傅里叶变换，然后从频域中来构建特征向量。其主要的频率特征有：傅里叶变换系数(Fast Fourier Transform, FFT)、频域熵(Frequence-domain Etropy, PDE)、能量谱密度(Power Spectral Density, PSD)等。

下表 2-2 对近年来多篇机器学习和运动姿态识别的相关文献(如参考文献[33]等)中常用的时域和频域特征进行了统计。由表 2-2 我们不难得知，频域特征的出现次数明显比时域特征要少得多。这主要是由于时域特征较直观地表示出了信号特点，而且直观简单，计算量要求相对较低。

表 2-2 文献中的时域特征与频域特征

参考文献	时域特征							频域特征		
	Mean	Std	Corr	Rms	Energy	Mad	Integration	FFT	FDE	PSD
[34]	√	√	√		√		√	√	√	
[35]	√	√	√	√	√	√				
[36]	√	√								√
[37]		√	√					√		
[38]	√		√		√				√	
[39]	√	√	√		√					
[40]	√	√	√					√	√	
[41]	√	√	√		√					√
[42]	√	√	√	√	√				√	

前面两种方法分别单独地提取出了信号本身的时域特征和频率特征，但并没有很好的结合起来；而时频分析则发挥了这项优势，把两者结合利用并更高效地来表征数据。小波变换是较为常用的时频分析法之一，其中常用的小波基有：Haar 小波基、Daubechies 小波基、Coiflets 小波基等。小波变换的核心思想是将原始信号分解，用一组不同尺度的小波信号来叠加表征原始信号。例如参考文献[43]中就通过对加速度信号高频分量进行了小波分析，提取了不同尺度上的细节系数的平方和、小波包逼近系数的平方和以及标准差和均方根。

而面对相异的问题背景和研究对象，单单按照一个普遍通用的方法来构建特征是不足够的，往往还需要实验人员根据模式识别的指标和要求来设计出更匹配的信号特征。所以除了以上三个维度的特征构建以外，在本次课题的背景下还需扩展的是一种方法和一个指标。本次扩展的特征构建方法是基于模型的特征构建。它指的是用模型来刻画时序信号，再提取模型系数作为特征向量。其常用的模型有：AR 自回归模型、MA 滑动平均模型或两者的组合 ARMA 模型等。在生物电信号领域，比如脑电和肌电信号，一般选用 AR 模型来进行拟合，其识别效果相对于其它两者效果更佳。例如哈尔滨工业大学的彭欣然[41]就曾通过伯格阶数为 4 的 AR 模型作为三轴加速度信号的特征来进行人体姿态识别，获得了 95%的识别准确率。

且由于本次课题研究是基于肢端动作信号对癫痫大发作进行识别，那么可以关注到的是，在室外癫痫发作时往往病人会因为神志不清而跌倒在地。所以基于该课题背景，在特征构建中加入跌倒特征是合理有效的。关于跌倒检测的算法，近年来国内有诸多研究。中国科学技术大学的佟丽娜[44]基于力学量信息来构建了摔倒过程与姿态角、加速度之间的关系，以此建立了数学算法模型，

来进行跌倒检测识别。上海交通大学的孙新香[45]通过三轴加速度的阈值判断和One-class SVM 算法学习训练，设计了一个跌倒探测器，跌倒识别的平均准确率为 88.57%。

综上，本次课题的特征构建主要从时域分析、频域分析、时频分析出发，外加上基于模型的特征和跌倒特征进行辅助，就基本上可以较全面地提取出癫痫发作识别所需要的数据特征。本次实验采用的时域分析的特征主要有：峰峰值、均值、标准差、四分位距、能量、均方根、过零点等。且因为肢端动作信号体现的非线性，所以尝试用多次方的形式来进行非线性拟合，于是对多次方的时域信号也进行了上述特征构建。此外，在参考了西南大学蔡菁[46]通过差分方式提取特征得到了较好的情绪状态识别效果之后，本课题也引入了原始动作信号的差分信号，并对其进行部分时域特征的构建。

至于频域分析，本文先通过 python.np 库中的 rfft 函数实现对数据的快速傅里叶变换，然后再对其傅里叶系数进行特征构建，主要有：峰值频点、平均频点、平均功率频率(MPF)等三个特征。其中平均功率频率的计算公式如式(2-2)所示，P(f)为功率谱：

$$MPF = \int_0^{\infty} fP(f)df / \int_0^{\infty} P(f)\,df \tag{2-2}$$

由于第一次数据采集的采样频率为低频率采集（10Hz 内），而癫痫发作的肢端动作信号主要集中在 10-20Hz 频域段，所以导致高频部分出现了频域混叠现象，整体频域分析效果不理想。功率谱密度图像如图 2-7 所示。

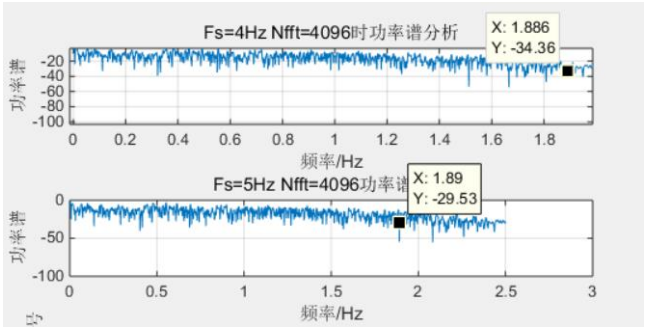

图 2-7 三轴加速度信号的功率谱密度

同理，对于时频分析来说，基于本次数据采样频率稍低的问题，也导致了小波分析效果的不理想，所以综合考虑后本课题没有进行基于小波基的时频特征的构建。此外，对基于模型的特征构建方式来说，因为最终在线测试主要聚焦在腕带式癫痫检测装置上实现，所以其底层芯片对计算资源有一定的限制，考虑到计算量的原因，暂时舍弃这种方式。

且由于本次数据集是在癫痫病院由知名医生配合医疗设备标注采集所得，病人常在病床上进行检测，所以跌倒检测这个指标暂时在数据集中没有体现，

本次课题也不深入研究。

综上，本次实验主要构建的特征共有 25 个，如表 2-3 所示。

表 2-3 特征列表

特征列表				
peak_to_peak	std	mean	energy	iqr
zeronum	d1absmean	d1iqr	d1std	d1peak
d2absmean	d2iqr	d2std	d2peak	x2mean
x2iqr	x2std	x2peak	x3mean	x3iqr
x3std	x3peak	ftpeak	ftmean	MPF

表 2-3 中使用公式如下所示，x_n 表示一个样本数据中第 n 个数：

$$\text{mean} = \frac{\sum x_n}{n} \tag{2-3}$$

$$\text{std} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - mean)^2} \tag{2-4}$$

$$\text{peak\_to\_peak} = \max - \min \tag{2-5}$$

$$\text{iqr} = \text{Q1} - \text{Q3} \tag{2-6}$$

其中 Q1 为第一四分位数，即把窗口数据排序后位于 25%处的数值；Q3 则为位于排序后 75%处的数值。

$$energy = |mean|^2 \tag{2-7}$$

一二阶差分分别如式(2-8) (2-9)所示，其中 X_n 表示数据的第 n 个样本：

$$1d_n = X_{n+1} - X_n \tag{2-8}$$

$$2d_n = X_{n+2} - X_n \tag{2-9}$$

多次方信号的特征，首先是先将数据进行多次方运算，然后再进行特征提取，例如二次方信号平均值即如式(2-10)，以此类推：

$$\text{x2mean} = \frac{\sum x_n^2}{n} \tag{2-10}$$

2.3.2 特征选择

特征选择就是指，通过选出对于解决问题最有用的特征子集来解决无关属性和冗余特征的问题。通常为了改善模型的识别准确率，我们会使用某些方法来对原始特征进行选择。在这个过程中，不需要构建或修改所拥有的特征，只需通过修剪特征就能达到减少噪声和冗余的目的。

特征选择算法可能会用到评分方法来进行排名和选择特征，评分指标可能是相关性或者其它确定特征重要性的指标。而更深入有效的方法是依据试错的方式，筛选得到最优特征子集。

根据特征选择的形式，可以把特征选择方法主要分为 3 种：Filter、Wrapper、

Embedded。Filter 即为过滤法，是根据发散性或者相关性等指标来对每个特征打分，通过设置阈值或者待选择阈值的个数来选取特征。Wrapper 即为包装法，根据目标函数，每次选择若干特征，或者排除若干特征。Embedded 即为嵌入法，是指先建立某种机器学习模型来学习训练，获取每个特征的权值系数或贡献值，再将权值按从大到小来排序选取特征。Embedded 嵌入法有些类似于 Filter，区别在于嵌入法需要通过训练来确定特征的优劣。

本次实验选用 Embedded 嵌入法来进行特征选择。其中使用的是基于树模型的特征选择法，利用随机森林算法模型本身的特征重要性打分的特性来进行评估和选择。得到的特征重要性结果如下图 2-8 所示。

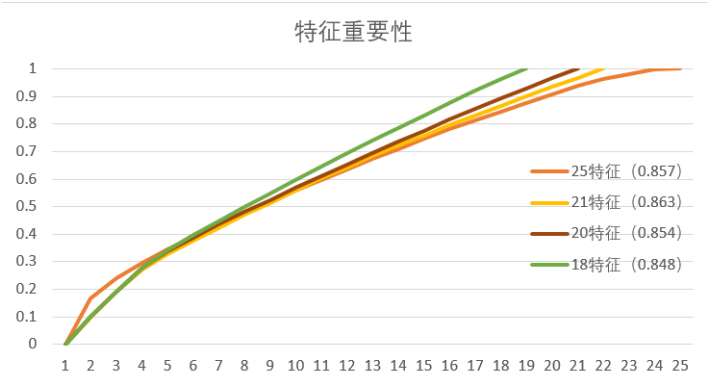

图 2-8 特征重要性结果图

由图 2-8 中我们可以看出，在识别准确率为 85.7%的随机森林模型训练学习下，排序在前列的 22 个特征就已经具备了 95%的重要性贡献度。其中重要性排前五的特征是 x2mean(0.0888)、iqr(0.079)、x3iqr (0.0713)、std(0.0558)、d1std(0.0509)，重要性排倒数前五的是 x2std(0.03)、MPF(0.027)、ftmean(0.0178)、ftpeak(0.0165)、zeronum(0.0014)。所以再删除了后面四个加和占比不足 5%重要性的特征，再次进行学习训练，识别准确率提高到了 86.3%。以此思想迭代类推，不断的删除和增添特征进行学习训练，依照识别准确率和特征维度进行选择，综合考虑选出最优的特征子集。

综上，本次实验选择 21 个特征来作为最终的特征集进行后续的处理和学习训练。选择后的特征列表如表 2-4 所示。

表 2-4 选择后的特征列表

最优特征				
x3iqr	iqr	x2mean	d1absmean	d2absmean
std	d1std	x2iqr	d2peak	d2std
d1iqr	mean	x3mean	d1peak	peak_to_peak
energy	x3std	d2iqr	x3peak	x2std
x2peak				

2.4 数据集处理

将采集来的数据进行了上述两个步骤之后，得到的仅是特征集。这离算法模型要求的输入格式还有所偏差，所以必须进行数据集处理，才能得到最后可用于输入的数据集。

在算法模型的学习训练中，最终目标始终是将训练好的模型应用到现实生活中去检测，期望训练好的模型可以于真实数据上获取优秀的预测结果。也就是说，最终的目的是减小训练好的模型的泛化误差，并较好的用于现实生活中。所以这就需要把数据集切分成训练集和测试集——用训练集的数据来进行学习训练，再将测试集上的误差当作最终算法模型在用于真实数据时的泛化误差来进行估计。

由前 2.1 可知，本次有效数据集为 22 人次，考虑到后面在线测试所需的数据，本次实验以 20 人次的数据特征集来进行数据集切分。依照普遍惯例“二八原则”，将 80%的数据作为训练集，20%的数据作为测试集，两者相互独立。数据集划分流程图如下图 2-9 所示。

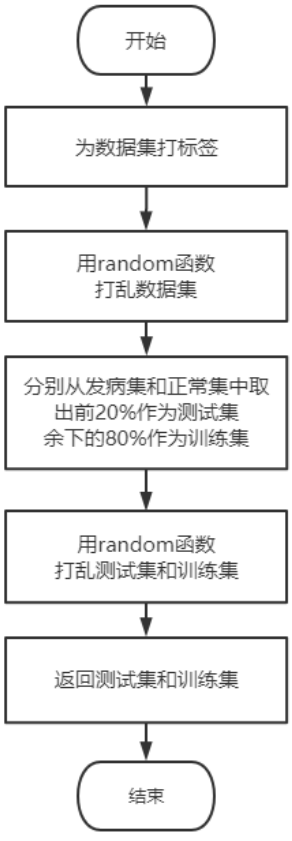

图 2-9 数据集划分流程图

实现数据集划分的方式有很多，可以通过自行设计步骤，也可以通过 sklearn 库的 model 函数进行切分。本次实验考虑易用性，所以进行自行编程设计，具体的数据集处理步骤如下：首先，为发病数据和正常数据添上标签，发病数据为正样本，标 1；正常数据为负样本，标 0。随后为了训练集和测试集的相互独立，用随机函数来打乱数据集。然后测试集取前 20%的数据，训练集则取剩余的 80%数据，再进行一次随机打乱，确保独立性。最后返回得到训练集和测试集，进行文件保存。

但是这里有一个问题是，本次正常数据比发病数据要多得多。也就是说，不论是测试集或训练集，负样本数据比正样本数据要多，整体正负样本数据分布不均衡，所以这可能会带来识别效果偏差的现象，例如识别负样本的精度高而识别正样本的精度低。

所以为了测试数据分布均衡与否对本课题的影响，本次实验分数据均衡和数据不均衡两种取出方式进行对照实验。具体结果见第四章 4.2.3 小节。

2.5 本章小结

本章主要介绍了数据采集、数据预处理、特征提取和数据集处理的具体流程步骤。通过将采集来的数据进行简单处理、数据选择、数据分割、特征提取以及后续的数据集处理之后，才可以得到最终能用于算法模型输入的数据格式。

第 3 章 机器学习算法模型构建

在得到了可输入的训练集和测试集之后，接下来就是进行算法模型的构建和训练学习了。随着计算机科学的日益发展，机器学习模型愈来愈多，相关理论也愈加深奥。如何在众多机器学习模型中选出相对符合课题要求的算法模型，是一个难题。本章从课题面对的数据是否线性可分的问题出发，选用了逻辑回归、支持向量机、深度学习等三种算法，在阐述了相关算法基础后进行模型的搭建和学习训练。

3.1 问题分析

本次课题研究的主要是基于肢端动作信号的癫痫发作识别，并构建算法模型进行训练学习以达到在线测试诊断的目的。而在选用算法模型时，数据的可分性是一个值得考虑的问题。

首先我们得判断出该数据问题是线性问题还是非线性问题，这样才能更好的依据其特点来选用算法模型。通常来说，确定数据集是否线性可分的方式有三种：一是把正负样本的特征一一画出来，通过直观地来看是否线性可分；二是使用最优间隔或感知机等简单线性算法做验证看效果，因为该两种方法只能处理线性可分的情况；三是首先使用 quickhull 算法得到数据的凸包，再通过 sweepline 算法来判断凸包的边界是否相交，若边界不相交/分离则线性可分。本次实验主要通过前两种方式来进行判别。部分特征散点图如图 3-1、图 3-2 所示。

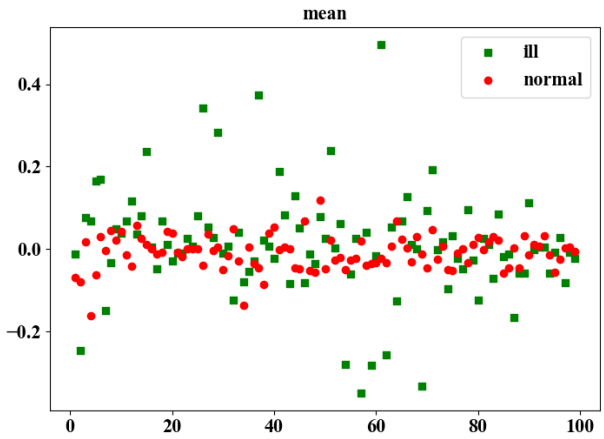

图 3-1 mean 特征散点图

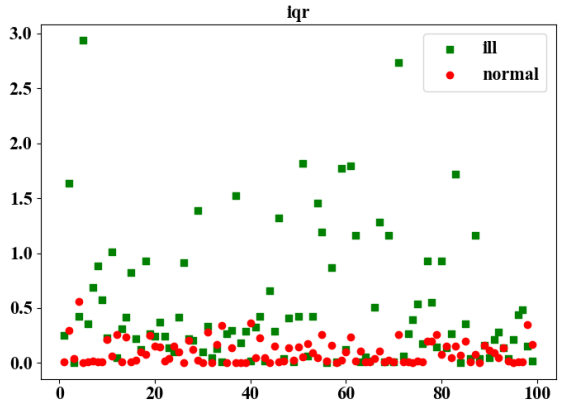

图 3-2 iqr 特征散点图

从图 3-1 和图 3-2 中我们可以看到，ill 和 normal 这两个正负样本特征图像并没有一个很明显的线性平面将他们区分，多数正样本点和负样本点还是相互混杂在一起。

第二种方式则是通过使用线性 svm 和 rbf 核 svm 的识别效果来进行对比，如第四章所示，线性的识别效果比 rbf 核非线性模型的准确率低了 20 个百分点左右。综上，无论是从原始数据的振动信号（如图 2-5）来看，还是从特征散点图、线性模型的识别效果对比等来看，本次数据表现主要为线性不可分。所以后续考虑算法模型选用时，以考虑非线性模型为主。

3.2 逻辑回归

逻辑回归（Logistic Regression, LR）虽然被称作回归，但与线性回归不同，它其实是一种分类器算法，常用来估计某种事物的可能性，解决二分类问题。逻辑回归的本质思想是假设数据服从 Logistic 分布，然后使用极大似然估计来做参数的估计[47]。Logistic 分布是连续分布，其分布函数和密度函数分别为：

$$F(x) = P(X \le x) = \frac{1}{1+e^{-(x-\mu)/\gamma}} \tag{3-1}$$

$$f(x) = F'(X \le x) = \frac{1}{\gamma(1+e^{-(x-\mu)/\gamma})^2} \tag{3-2}$$

其中，μ表示位置参数，$\gamma > 0$为形状参数，具体图像特征如下图 3-3 所示：

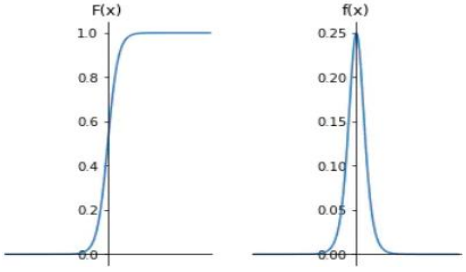

图 3-3 logistic 分布的分布函数 F(x)和密度函数 f(x)

逻辑回归的优点在于可解释性很强，且分类时计算量较小，耗费的计算资源低，能够较好地得到样本的概率分数。缺点主要是不能较好地处理大量多类特征或变量的问题，容易欠拟合。

3.2.1 逻辑回归算法基础

逻辑回归的表达式是以线性回归为基础的。线性回归的一般形式为：

$$y = wx + b，\ y \in [-\infty, +\infty] \tag{3-3}$$

逻辑回归在其基础上，将线性回归得到的 y 带入到 Sigmoid 函数的 t（如式 3-4）中，即可得到逻辑回归的一般模型方程（如式 3-5）：

$$S(t) = \frac{1}{1+e^{-t}} \tag{3-4}$$

$$P(y=1|x;w,b) = p(w,b) = \frac{1}{1+e^{(aX+b)}}，\ p \in [0,1] \tag{3-5}$$

所以通过机器学习训练好了一组权值 w 和截距 b 以后，只需要把接下来的数据样本 X 代入到上式(3-5)中，就可以得到预测值 p。得到了 p 值，我们就能判断出输入数据是属于哪个类别。例如假设分类的阈值为 0.5，则 h>0.5 时即为 1 分类，反之则为 0 分类。即决策函数如(3-6)所示，阈值可以按照现实需要来自行调节。

$$y^* = 1，\ \text{if}\ P(y=1|x;w,b) > 0.5 \tag{3-6}$$

那么如何通过一组采集到的真实样本，来训练出参数 w/b 的值呢？此处就引入损失函数的方法。由前决策函数(3-6)，可以推出单个样本发生的概率式为：

$$P(y_i|x_i) = p^{y_i}(1-p)^{1-y_i} \tag{3-7}$$

式(3-7)表明当 y=1 时，概率为 p；当 y=0，概率为 1-p，y 代表样本的分类标签。那么当我们采集到 N 个样本时，合事件发生的总概率即等于每一个事件发生的概率相乘，即有：

$$P_{总} = P(y_1|x_1)P(y_2|x_2)\ldots\ldots P(y_N|x_N) = \prod_{n=1}^{N} p^{y_n}(1-p)^{1-y_n} \tag{3-8}$$

为了方便后续运算，取对数的形式得到损失函数：

$$F(w) = -ln\left(P_{总}\right) = -\sum_{n=1}^{N}(y_n\, ln(p) + (1-y_n)\, ln(1-p)) \tag{3-9}$$

其中$p = p(w,b)$，见式(3-5)。从式(3-9)的推导中我们可以看出，损失函数的值就等于合事件发生总概率的值的负数，所以说当我们选取的某个 w/b 的值刚好使得总概率$P_{总}$取得最大、损失函数 F(w)最小时，这个 w/b 就是我们要寻找的最优参数值，这就是最大似然估计的思想。

所以我们可以用损失函数来衡量我们当前模型的输出结果跟实际的分类结果之间的差距。且基于最大似然估计的思想，把问题转化成了如何找到一个 w^*

使得损失函数 F(w)取得最小值的问题，即有：

$$w^* = arg\min_{w} F(w) \tag{3-10}$$

求解最值问题的方法有很多，逻辑回归算法模型中常用的是梯度下降法和牛顿法。梯度下降法原理图如图 3-4 所示。

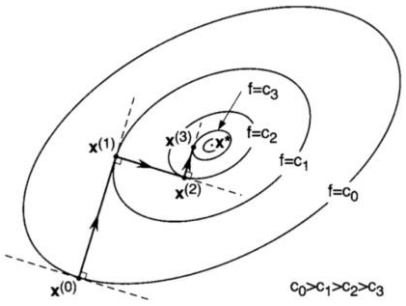

图 3-4 梯度下降法

从式(3-9)我们可以看出，逻辑回归的损失函数 F(w)是一个连续的凸函数，所以它只会有一个全局最优点而不存在局部最优。正因为逻辑回归的这个特点，所以可以使用梯度下降法的收敛来快速找到最优解。

对(3-9)做梯度运算，其中变量只有 w/b：

$$\nabla \mathrm{F(w)} = \frac{\partial \mathrm{F(w)}}{\partial \mathrm{w}} = \sum_{\mathrm{n=1}}^{\mathrm{N}} (\mathrm{p} - \mathrm{y_n})\,\mathrm{x_n} \tag{3-11}$$

梯度下降法的主要方法是先随机初始化 w_0，然后给定步长ρ，通过不断地修正$w_t \to w_{t+1}$，从而得到取得最大值的点。也就是说不断进行迭代、更新参数直至$\|F(w_{t+1}) - F(w_t)\|$的值小于阈值或者到达最大迭代次数即止。迭代公式见式(3-12):

$$w_{t+1} = w_t + \rho \nabla F(w) \tag{3-12}$$

综上，按照以上过程计算，即可获得一组 n 维的 w 和 b。w，又被称为权重向量，它的每一个维度值，都代表了这个维度的特征对最终分类结果的贡献值。即如果为正数，值愈大越说明了这个特征对于正分类的起的识别作用愈明显；反之亦然。而对于截距 b 来说，则在一定程度上说明了正负两个类别的判定难易情况，例如 b>0 则代表该样本集更容易被分成正类，以此类推。也正是因为逻辑回归模型得到的 w/b 可以部分对每个特征的重要程度进行量化，所以才说它有很强的可解释性。

3.2.2 逻辑回归模型的学习训练

本次逻辑回归模型主要通过使用 python 语言来进行学习训练。首先是把输入来的测试集和训练集的特征与标签进行分离，以便于后续的学习训练和结果预测。然后再选用库中封装的逻辑回归模型，考虑到之前的算法基础来适当的

设定超参数。在确定下模型和参数后，输入训练集的特征和标签进行学习训练，得到一个训练好的参数固定的模型。再把测试集的特征输入到模型中获得预测结果，用其与真实标签相比较。比对结果可通过混淆矩阵来进行表征，如若不满足指标要求，则重新调节超参数再次训练，直到调参达到要求或者最优效果。最后返回决策边界的参数以及保存模型，以待第四章在线测试诊断所用。逻辑回归模型的具体构建训练流程如图 3-5 所示。

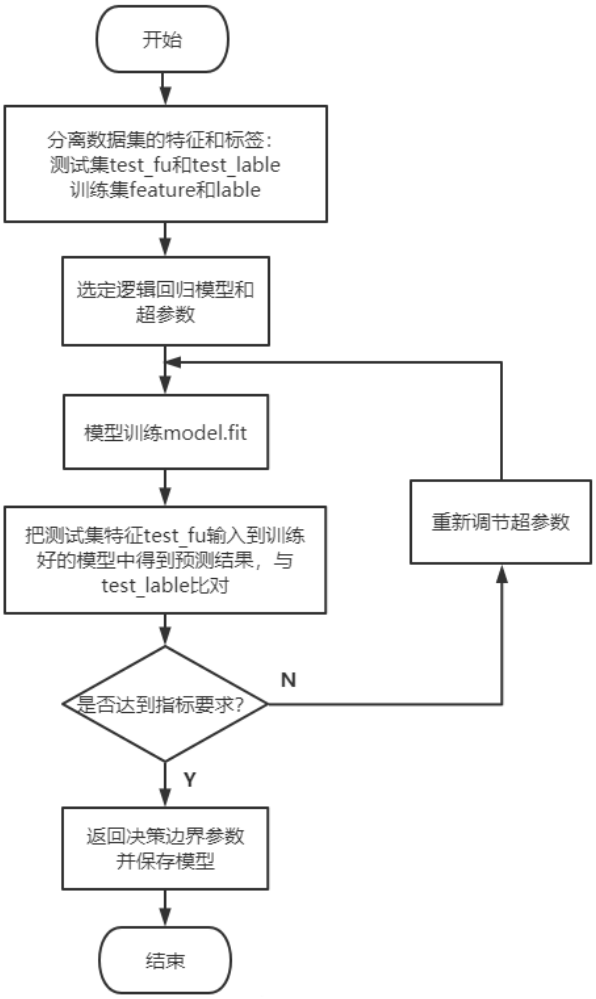

图 3-5 逻辑回归模型的构建训练流程图

在上述构建和学习训练中需要特别注意的是超参数的设置。逻辑回归模型共有 14 个超参数。在本次实验中比较重要的是 penalty、tol、c、solver、max_iter 等这几个超参。Penalty，是用来指定惩罚项正则化的规范，本次实验中默认为 L2 规范。超参数 c，为正则化系数 λ 的倒数，数值越小则正则化越强，即对模型的惩罚和约束程度越强。这两个超参数主要是用于设置正则化。

而 tol、solver、max_iter 等三个则关系到逻辑回归的迭代求解。solver，即为选择算法优化参数，有五个可选项——liblinear、lbfgs、newton-cg、sag、saga。

选择 lbfgs、newton-cg，即使用牛顿法进行损失函数最小求解。sag、saga 即表示选用随机梯度下降法。本次实验考虑到数据样本小和二分类的问题，选用 liblinear 即坐标轴下降法来迭代优化损失函数。max_iter，用来设定算法收敛的最大迭代次数，默认值为 10。tol，用来设定停止求解的标准，默认值为 1e-4。

综上，在 penalty、tol、max_iter 等三个参数选择默认值也合适的基础上，本次主要调节的超参数为 c 和 solver。

3.3 支持向量机

支持向量机(support vector machine, SVM)是一种在分类和回归分析中分析数据的监督式学习模型与相关学习算法，最早是由 Vapnik 等科学家在 1995 年提出。模型的基本定义是特征空间上间隔最大的线性分类器，学习策略为间隔最大化，即找到最优超平面使得特征空间上的两个类别距离最远[48]。如图 3-6 所示，当+1 和-1 这两类平面离得最远时，$\vec{w} * \vec{x} + b = 0$即为要找的最优超平面。

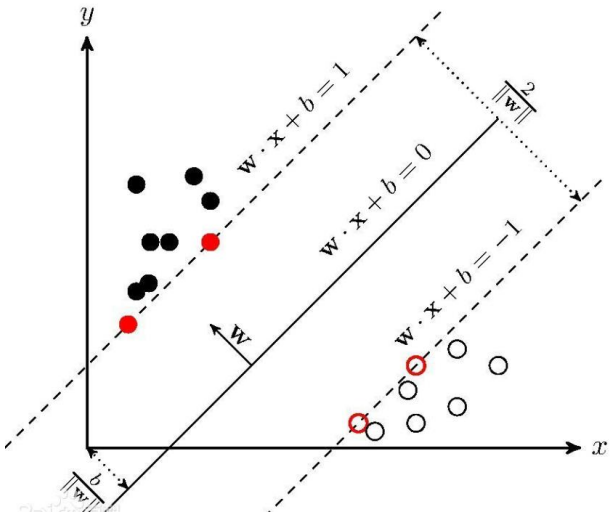

图 3-6 支持向量机原理说明

与逻辑回归相同，支持向量机也常用来解决二分类问题。但两者不同的是，逻辑回归的损失函数是 log loss，而 SVM 使用的是 hinge loss；SVM 关注的是支持向量，关注类别边缘的关键信息，而逻辑回归则是进行全局优化。正因为支持向量机的运算较复杂，所以常用于小样本数据处理。也恰好本次实验数据样本不大，故选用支持向量机来做训练学习和模型对比。

3.3.1 支持向量机算法基础

在以下算法基础说明中，x 代表样本特征（矢量），y 为结果标签，下标 i 表

示第 i 个样本数据，即数据格式为$(x_i，\ y_i)$。

从图 3-6 中可知，我们可以将 SVM 的分类函数用$f(x) = w^T x + b$来表示，即把样本数据 x 代入 f(x)中，若 f(x)<0 则判别样本为-1，反之亦然。这说明了通过观察$w * x + b$的符号与类别标签 y 的符号是否一致就可以检测出分类识别的效果。

且既然 SVM 的学习策略是间隔最大化，那么就应该对间隔进行定义，以及找到最大化等优化问题的解法。而在超平面$\vec{w} * \vec{x} + b = 0$确定的情况下，$|w * x + b|$即表示点 x 到超平面的间隔距离。所以此处引入函数间隔$\hat{\gamma}$的概念：超平面(w，b)关于样本数据 T 中所有样本点$(x_i，\ y_i)$的间隔距离最小值，即为超平面(w，b)关于训练数据集 T 的函数间隔，见式(3-13)：

$$\hat{\gamma} = min\ \hat{\gamma_i} = min|w * x_i + b|, (i = 1, \dots, n) \tag{3-13}$$

但这样定义的函数间隔有问题，当(w，b)倍增的时候，函数间隔$\hat{\gamma}$也在倍增，但超平面却没有改变。所以我们得引入几何间隔$\tilde{\gamma}$作为真正的点到超平面的距离，见式(3-14)：

$$\tilde{\gamma} = \frac{\hat{\gamma}}{\|w\|} \tag{3-14}$$

式中，$\|w\| = \sqrt{\sum_{i=1}^{n} w^2}$。此时的几何间隔$\tilde{\gamma}$才是要找的间隔最大化中所说的“间隔”。几何间隔$\tilde{\gamma}$只随着超平面的变动而变动。

依此，则使得间隔分类最大化的目标函数可以定义为$\max \tilde{\gamma}$，再满足一些约束条件，例如函数间隔的定义$\hat{\gamma_i} > \hat{\gamma}$等即可。由于令目标函数中的$\hat{\gamma}$等于 1 既可以方便推导和优化，又对目标函数本身不影响，所以把$\hat{\gamma}$替换。且求$\frac{1}{\|w\|}$的最大值相当于求$\frac{1}{2}\|w\|^2$的最小值，两者等价，所以最终间隔最大化的目标函数问题可转化为如式(3-15)所示：

$$\min \frac{1}{2}\|w\|^2 \tag{3-15}$$
$$s.t., |w * x_i + b| \geq 1, (i = 1, \dots, n)$$

从式(3-15)可以看出，目标函数为二次，约束条件是线性的，所以实际上是一个凸二次规划的问题。又因为这个问题的特殊结构，所以本次利用拉格朗日对偶性，即通过求解与原问题等价的对偶问题来得到原问题的最优解。这就是线性可分条件下支持向量机的对偶解法，如此也方便引入核函数从而推广到解决非线性问题。使用拉格朗日法求解，加入拉格朗日乘子α之后，目标函数的对偶问题为：

$$\min_{w,b} \theta(w) = \max_{\alpha_i \geq 0} \min_{w,b} L(w, b, \alpha) \tag{3-16}$$

其中$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i(|w * x_i + b| - 1)$。

目标函数问题转化为式(3-16)后，按照对偶问题求解的三步骤：固定α，使得 L 函数关于 w 和 b 最小化，即令偏导$\partial L/\partial w$和$\partial L/\partial b$等于 0；再求对α的极大值，最后通过 SMO 算法求解出拉格朗日乘子α。最后解得 w/b 为：

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{3-17}$$

$$b = -\frac{\max_{i:y_i=-1} w^T x_i + \min_{i:y_i=1} w^T x_i}{2} \tag{3-18}$$

目标函数优化求解已完毕，现在找到了使得超平面间隔最大化的 w 和 b。将式(3-17)的 w 代入分类函数$f(x) = w^T x + b$中可得：

$$f(x) = (\sum_{i=1}^{n} \alpha_i y_i x_i)^T x + b = \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \tag{3-19}$$

式中$\langle x_i, x \rangle$为 x 与x_i的内积。

且对式(3-19)进行说明，所有非支持向量（即两个类平面上的样本点）所对应的系数α都等于 0，所以对于新点 x 的预测实际上只要关注少量的“支持向量”即可，而不用关心全局数据。

但目前解到此处也只能应对如图 3-5 中的线性情况，而面对非线性分类的问题，还需要引入核函数来扩展解决。从式(3-19)可以看出，对于新样本点 x 的预测，由于α、b、x_i、y_i经过训练数据集学习后已经固定，所以只需要计算$\langle x_i, x \rangle$内积即可。这一点很重要，因为只有在此基础上，后续才能在此处引入核函数而完成非线性分类问题求解的推广。

实际上现实生活中绝大部分环境数据和课题数据都不是线性可分的。在这种情况下想找到满足间隔最大化条件的超平面很难。SVM 对这种非线性分类情况的扩展解决方法是，通过使用核函数 k<x_i,x>，来将低维数据映射到高维空间中，解决低维空间中线性不可分的问题。这种思路如图 3-7 所示，图中一维数据在二维空间中无法划分，但简单映射到三维空间则分离超平面很容易就能找到。

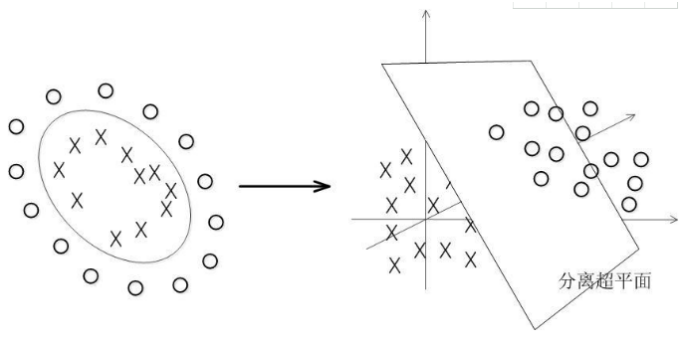

图 3-7 高维映射解决低维线性不可分的现象

通过核函数进行高维空间映射的方便之处在于，无需苦苦找寻具体的映射函数，而是在特征空间中直接计算内积，在(3-19)的基础上建立一个非线性学习

器。核函数，即对所有的 x、z，都满足$\mathrm{K(x,z)} = \langle\varphi(x)\cdot\varphi(z)\rangle$，其中$\varphi$指代从样本数据 x 到内积特征空间 F 的映射。将核函数代入到式(3-19)中，得到非线性的分类函数有：

$$\mathrm{f(x)} = \sum_{i=1}^{n}\alpha_i y_i K\langle x_i, x\rangle + b \tag{3-20}$$

通常使用的核函数有：

高斯核函数为：

$$\mathrm{K(x,z)} = \exp\left(-\frac{\|x-z\|^2}{2\delta^2}\right) = \exp(-\gamma\|x-z\|^2) \tag{3-21}$$

多项式核函数有：

$$\mathrm{K(x,z)} = (\mathrm{x}\cdot\mathrm{z}+1)^{\mathrm{p}} \tag{3-22}$$

线性核函数有：

$$K(x,z) = x\cdot z \tag{3-23}$$

最后为增强 SVM 算法的学习效果，消除 outlier 偏离正常位置的支持向量噪声点(如下图 3-8 所示)，所以在原算法基础上引入松弛变量θ和惩罚因子 C，使得模型对于部分噪声点和错分样本点具备一定的容忍度。因此，把目标函数优化问题(3-15)转化为下式(3-24)，如此便能较为完整的搭建起 SVM 算法：

$$\min_{w,b,\theta}\frac{1}{2}\|w\|^2 + c\sum_{i=1}^{N}\theta_i \tag{3-24}$$

$$\mathrm{st.}\, y_i(w\cdot x_i + b) \geq 1-\theta_i, \theta_i \geq 0$$

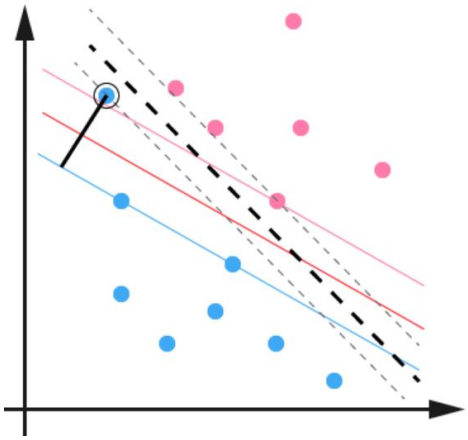

图 3-8 outlier 噪点对超平面的影响

3.3.2 支持向量机模型的学习训练

本次支持向量机模型主要通过使用 python 语言来进行构建。虽然各自算法不同，但是大部分流程步骤与逻辑回归模型的搭建相同。首先仍是分离出训练集和测试集的特征x_i和标签y_i。然后选定所使用的 SVM 模型来进行搭建，并选用调节合适的超参数。之后再把训练好的模型中带到测试集中进行泛化测试，

满足性能指标要求即可，不满足则重新调节超参数或者特征集，直到满足要求或者达到最优性能为止。最后依据式(3-20)输出分类决策函数以及决策边界，保存模型，以待第四章在线测试中使用。支持向量机模型的具体构建训练流程如图 3-9 所示。

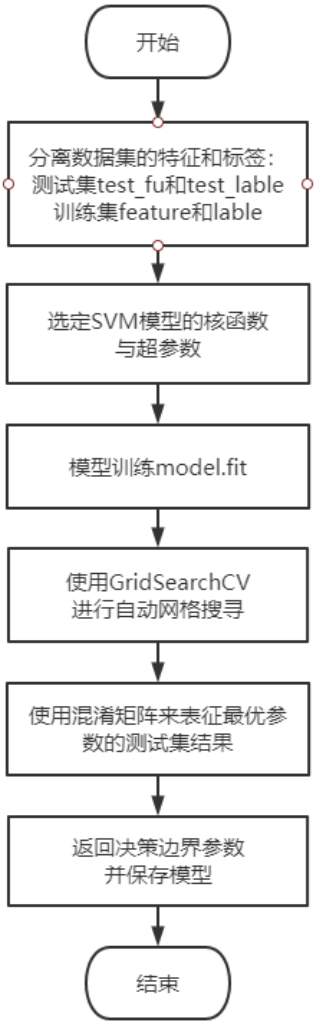

图 3-9 SVM 模型的构建训练流程图

尽管搭建步骤大致相同，但支持向量机模型的学习训练中有几点细节很重要，也与上述不同。一是对核函数的使用，由于本次数据集是自身采集的，对于数据分布是否线性可分（见 3.1 节）还需要用线性模型和非线性模型的效果来对比验证。所以本次核函数的选用分为线性核和 rbf 核两种来进行对比实验。

而在支持向量机模型的搭建过程中，主要有 14 个超参数，选用了线性核函数和 rbf 核函数以后，主要的是惩罚因子 C、gamma、tol 等三个。惩罚因子 C，数值越小则对模型的惩罚和约束程度越强。gamma 与 rbf 核相关，与式(3-21)中γ成正相关。tol 表示停止训练的误差值大小，默认为 1e-3，在本次实验中默认

值可用。

二是由上可知，若使用 rbf 核函数，则至少有两个超参数需要进行调节——高斯系数γ以及惩罚因子 C。所以面对多个超参数的调节优化，像逻辑回归那般手动调节是很难找到全局最优的，故本次使用自动搜寻网格法来进行超参数调节，其中主要通过 python 中的 GridSearchCV 函数来进行实现。

网格搜寻是指，在选定的参数范围内按照步长依次调整参数来找到测试集上精度最高的参数，它在调参的三种方法中精度损失最小、优化效果最好。且 GridSearchCV 是最常用的网格搜寻法，因为它既能实现参数网格搜寻又能进行交叉验证。所以本次使用 GridSearchCV 函数来实现对(gamma，c)的网格搜索与交叉验证。其中交叉验证次数设定为 3，以 f1 分数、综合准确率作为衡量指标来进行调节。

3.4 随机森林

随机森林(Random Forest, RF)是属于 bagging 类型的一种优越的集成学习算法，最早由 1995 年 Leo Breiman 等科学家提出。它的核心思想是通过训练多颗决策树来进行投票，然后依据投票结果的众数来得到最终决策分类[49]，如图 3-10 所示。基于此，所以它和上述两种机器学习算法不同的是，它不是个体学习器，而是集成学习，对多个基学习器进行集成决策的多分类器系统。

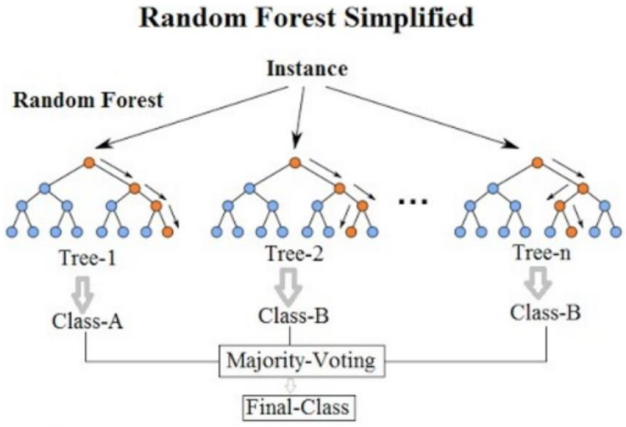

图 3-10 随机森林算法简单原理图

它的优点是不容易过拟合，计算开销小、训练速度快，且可以做成并行化方法。哪怕面对高维度的数据学习，不用进行特征选择也能得到很好的分类效果。此外，由于各个基分类器的相互独立、训练样本的随机切分，所以它能检测出各特征之间的相互影响，最终给出各特征的特征重要性值。随机森林在很多现实任务的应用中也体现出了很好的性能效果。

本次实验选用随机森林算法模型，一是因为它可以输出特征重要性数值的特点，可用于特征选择；二是因为它属于集成学习方法来解决非线性问题，可以通过它与 rbf 核函数的 SVM 算法的实验结果对比，来体会出个体学习器与集成学习器的各自特点和优劣。

3.4.1 随机森林算法基础

对于随机森林算法进行理解和说明，则必须从随机森林的本身意义出发，通过“随机”和“森林”两方面来分别阐述。“随机”主要是指对样本的随机取样和特征子空间的随机选择。而“森林”则主要是说明该算法是基于多颗决策树分类器集成而来。所以以下便通过这两方面进行详细探讨。

（1）森林

决策树是随机森林的最基本组成要素，它是一种树状结构的直观模型。它类似于我们人类对数据思考然后进行分类的过程，它会对一系列的数据进行统计和询问，直到最终得出一个预测类别为止。具体原理结构如图 3-11 所示，它的每个内部节点都表示出对于一个特征属性的判断，每个分支输出一个判断结果，每个叶节点都显示一种分类结果。

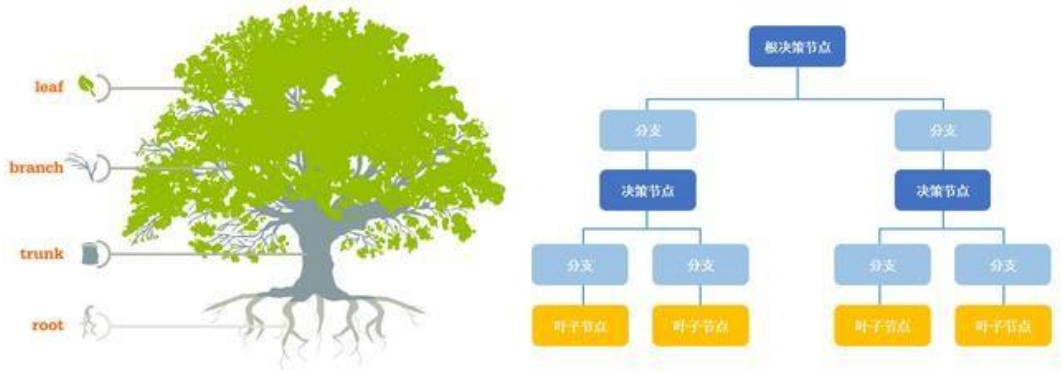

图 3-11 决策树简单原理图

决策树的生成中最关键的是两个步骤，一是节点的分裂，即当一个节点所代表的特征属性无法给出判断时，要将该节点分为多个子节点。第二步就是分类阈值的确定。常见的决策树分类算法有 ID3、C4.5 和 CART。

由于随机森林中默认使用的是 CART 二叉树，所以对 CART 决策树进行深入说明。CART 只能将一个父节点分为两个子节点，并依据基尼系数来决定如何分裂。

基尼系数公式有：

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{3-25}$$

$$Gini_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Gini(D_j) \tag{3-26}$$

$$\Delta\mathrm{Gini(A)} = \mathrm{Gini(D)} - \mathrm{Gini_A(D)} \tag{3-27}$$

其中$p_i = \frac{|C_{i,D}|}{|D|}$代表在数据集 D 中属于$C_i$类别的数据条的比例。如果数据集总体 D 内包含的类别越纷乱，基尼系数则越大。CART 二叉树主要通过基尼系数的增量$\Delta\mathrm{Gini(A)}$最大的属性来作为分裂指标。

单颗决策树由 CART 来分裂生成，但森林的树与普通决策树不同的是，森林里的每棵树都尽最大程度的生长，并且没有剪枝过程。而多颗相互独立的决策树生成后，则通过投票法的结合策略来集成为森林。投票法，即为少数服从多数，聚合每棵树的预测分类结果，然后将得票最多的结果作为最终的森林预测结果。这样集成多颗决策树的森林，可以提升模型的泛化能力，降低糟糕局部极小点的风险。

（2）随机

“随机”主要是指对样本的随机取样和特征子空间的随机选择。这个可以分为两个步骤来实现。假设数据集 S 中共有 n 个独立的样本，每个样本中包含有 M 个特征属性，则有：

1）每棵树从集合 S 中随机且有放回的抽取一个样本，重复抽取 n 次，生成新的训练集集合 S*；

2）从新集合 S*中随机选取 m 个属性（m<M）；

第一步中随机抽取一次，集合 S*中不包含第 i 个样本$x_i(\mathrm{i} = 1, \ldots, \mathrm{n})$的概率为：

$$p = (1 - \frac{1}{n})^n \tag{3-28}$$

当 n 趋向于∞时即有：

$$\lim_{n\to\infty} p = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e^{-1} = 0.368 \tag{3-29}$$

这种随机取样的方式被称为 Boostrap 自助重采样。通过这样的抽取方式，可以使得每个树的训练集既不相同，又包含重复的训练样本。从式(3-29)可知，新集合 S*尽管样本数量也为 n，但当 n 很大时，去除里面的重复样本，新集合 S*中共包含了原集合 S 中 63.2%的样本。另外剩余的 36.8%的样本数据，可以作为袋外数据来对模型进行评估。袋外错误率是随机森林泛化误差的一个无偏估计，随机森林模型可以依据此进行内部评估，选出最优的 m，而无须进行复杂的 k 折交叉验证。

这种基于 Boostrap 重采样得到新训练样本，并且基于每个新训练样本训练一个弱分类器，再通过结合策略集成得到强分类器的方法被称为 Bagging。其具

体结构如下图 3-12 所示。随机森林是属于 bagging 类型中优越的一种集成学习算法。

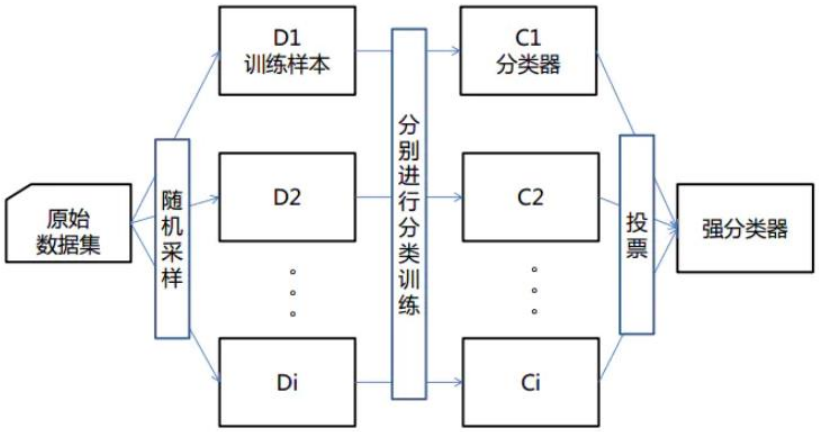

图 3-12 Bagging 结构

而第二步中对于特征子空间的随机选择，即是随机森林算法在 Bagging 结构的基础上进行优化。如此则可以使得森林里每个决策树的节点都能够等概率地随机引入地某个特征，再通过其特征信息来对节点进行拆分，从而选出最优特征。这一步也就是为什么随机森林具有特征重要性评估的原因所在。

综上，通过随机划分数据，并训练每颗决策树后进行投票集成，就能够得到基础的随机森林模型。

3.4.2 随机森林的学习训练

在 3.4.1 的算法基础上，本次实验依此进行了随机森林的模型搭建。首先是对训练集样本的 Boostrap 重采样，生成 K 个子训练集。然后再通过每个子训练集来生成对应的单颗 CART 二叉树，并且从子训练集中的 M 个属性中抽取 m 个属性作为节点的分裂属性，按照 m 个属性中最好的分裂方式来分裂节点。m 值维持不变，通过袋外错误率来进行评估。其中的每棵树都完整生长，不进行剪枝。训练完之后，代入测试集样本，得到每颗决策树的分类结果，采用投票的方式，将众数结果作为随机森林最终的分类结果输出。

在模型搭建完以后，后续考虑对模型的学习训练。随机森林的学习训练与之前上述 SVM 支持向量机的训练流程大体相仿，其中不一样的细节主要是对于超参数的选用以及随机森林不能够还原出决策函数。具体算法流程图如 3-13 所示。

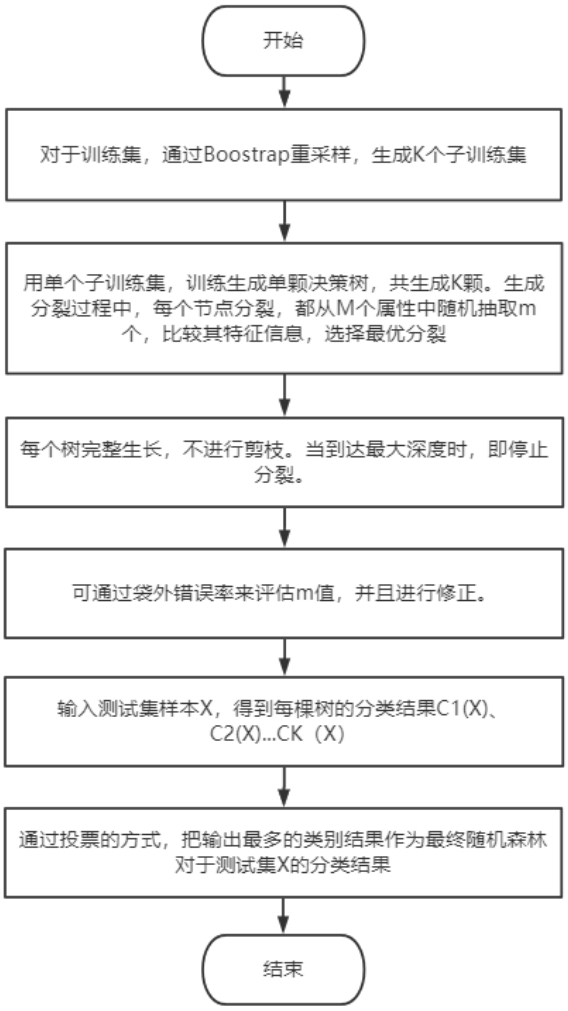

图 3-13 随机森林算法流程图

随机森林模型主要的超参数有：n_estimators、oob_score、criterion、max_feature、max_depth、min_samples_split、max_leaf_nodes。其中属于 Bagging 框架的参数是 n_estimators、oob_score、criterion。n_estimators，即为最大的弱学习器的个数，即进行 Boostrap 重采样的 K 值。oob_score，即图 3-12 中的第四步，确定是否通过袋外样本来评估模型的好坏，可设置为 Ture。Criterion，即 CART 决策树分裂时对特征的评价指标，默认为基尼系数。

属于 RF 决策树的超参数有 max_feature、max_depth、min_samples_split、max_leaf_nodes。max_feature，即 RF 节点分裂时考虑的最大特征，也就是对应前文所说的 m 值。默认为 sqrt，意味着划分时最多考虑 $m = \sqrt{M}$ 个特征。max_depth，即为最大树深，控制着每棵树的最大生长。min_samples_split，为叶子节点最少样本数；max_leaf_nodes 为最大叶子节点数，这两个值加以限制，则可以防止部分过拟合，默认为 None。

由上可知，面对本次实验的样本小特征少的数据特点，上述部分超参数选择默认使用是可以符合要求的，例如 max_feature 默认为 sqrt。重点需要调节的超参数主要为 n_estimators 和 max_depth。这两个超参数的调节，本次实验选用 GridSearchCV 函数来进行网格搜寻和交叉验证，其中交叉验证次数设定为 3，以 f1 分数作为衡量指标来进行调节。

此外，需要特别说明的是，与前两种机器学习方法不同，随机森林模型无法给出明确的决策函数和决策边界。因为这种 bagging 框架的随机森林算法，在提升分类准确率、防止过拟合的同时，以失去部分解释性来作为代价。所以在后续的在线测试章节中，随机森林模型主要通过保存模型的方法，在 PC 端进行在线测试。

3.5 本章小结

本章首先对实验数据是否线性可分的问题进行了分析，然后主要阐述了逻辑回归、支持向量机、随机森林等三种机器学习算法的理论基础，并基于各自理论完成了三种模型的搭建和训练学习。

第 4 章 实验结果分析与在线测试

本章首先介绍了衡量实验结果的重要指标，然后对本课题进行的窗口点数实验、数据均衡实验、分类器效果对比实验、个体识别效果实验等进行结果分析和探讨，最终给出了在线测试方案并进行实验验证，完成基于肢端动作信号的癫痫发作的识别。

4.1 实验说明

本次实验基于 VS code 编辑器下的 python 环境进行实验。实验用机的处理器为英特尔 i5-4210H cpu@2.90GHz，内存(RAM)为 8GB，操作系统为 window10 x64。

本次实验的数据来源于自身采集。可用的数据共有 22 组，其中 20 组作为数据集，依据二八惯例，80%的数据作为训练集，20%的数据作为测试集。剩余两组数据作为独立组，用于评估本章的在线测试结果。

评估本次实验的重要指标主要是准确率、误报率和漏报率。实验结果输出主要通过混淆矩阵的形式来进行表征。混淆矩阵，也称为误差矩阵，是评估分类器模型准确度中最基本、最直观的方法，如下图 4-1 所示。

		预测值	
		0	1
真实值	0	True negative(TN)	False positive(FP)
	1	False negative(FN)	True positive(TP)

图 4-1 混淆矩阵

由图 4-1 可知，混淆矩阵中包含了 TP、FP、FN、TN 等四个基础指标。其中，TP(Ture Postive, 真阳性)表示的是正样本被正确预测为正样本的个数；FP(False Postive, 假阳性)表示的是负样本被错误预测为正样本的数量；TN(Ture Negative, 真阴性)表示的是负样本被正确预测为负样本的数量；FN(False Negative, 假阴性)表示的是正样本被错误判断为负样本的个数。通过列表列写这四个指标，就形成了基本的混淆矩阵。

因为混淆矩阵和四个基础指标本身统计的是数量，相对于我们想要的性能

指标仍有所差别，所以在此基础上又延伸出了多个二级指标，如准确率、召回率、F1 值以及本次实验所需要的误报率和漏报率等。

准确率，即预测准确的样本数占所有样本数中的比例，见式(4-1):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4-1}$$

召回率，即真阳性样本数占全部正样本的比例，如式(4-2)所示：

$$Recall = \frac{TP}{TP+FN} \tag{4-2}$$

精确率，即真阳性样本数占全部预测为正样本数的比例，见式(4-3):

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4-3}$$

F1 值，主要通过召回率和精确率综合运算得来，因为往往精确率和召回率是一对矛盾的度量。所以通过 F1 值来进行折中评估模型，例如本次实验使用自动网格搜寻法就以 F1 值作为主要指标。其具体计算公式见式(4-4)：

$$\text{F1} = \frac{2Recall \cdot Precision}{Recall+Precision} \tag{4-4}$$

误报率，即假阳性样本占总样本数的比例，如式(4-5)所示：

$$False = \frac{FP}{TP+TN+FP+FN} \tag{4-5}$$

漏报率，即假阴性样本占总样本数的比例，见式(4-6)：

$$\text{miss} = \frac{FN}{TP+TN+FP+FN} \tag{4-6}$$

实验中得到的实际输出结果如下图 4-2 所示。此为 25 个特征的逻辑回归结果图。其中右上角的矩阵即为混淆矩阵。输出的结果中包含了准确率、精确率、召回率和 F1 值。对于漏报率和误报率，则需要通过混淆矩阵来另行计算。

```
the window30 21feature logistic result is:
[[2343   97]
 [ 340 2100]]
              precision    recall  f1-score   support

           0       0.87      0.96      0.91      2440
           1       0.96      0.86      0.91      2440

    accuracy                           0.91      4880
   macro avg       0.91      0.91      0.91      4880
weighted avg       0.91      0.91      0.91      4880
```

图 4-2 实际输出结果

本实验中最看重的性能指标为准确率、误报率和漏报率，其余二级指标作为辅助指标来帮助评估。此次基于肢端动作信号的癫痫发作识别检测装置的性能指标的基本要求为准确率在 80%以上，且在误报率和漏报率之中更偏向于降低漏报率。

4.2 实验结果分析

4.2.1 数据选择对比实验

如前文 2.2.2 数据选择小节所言，在采集到的发病样本中，需要剔除肢端动作信号延迟于脑部异常放电的部分样本，来确保正样本的可用性。而对于日常活动样本，则需要进行平稳片段和剧烈运动片段、平稳运动片段等三组对比实验，才能够选出更加符合于病人日常活动表现的负样本。

所以本次数据选择对比实验的正常组的数据中，相较于原始数据做了一定的调整和优化——选用的正常数据样本设置了一个实验组和两个对照组，实验组的正常数据为平稳运动片段，即波动片段与平缓片段各占比 50%，第一个对照组的正常数据为全部平稳片段，第二个对照组的正常数据为全部剧烈运动片段。实验中，采集数据的窗口时间为 6s，特征集特征有 21 个（见表 2-3），并进行了数据样本均衡处理。

具体结果如表 4-1 所示，表中衡量实验效果好坏的指标为准确率。

表 4-1 数据选择对比实验结果

	逻辑回归	线性 svm	Rbf 核 svm	随机森林	平均准确率
平稳片段	83.0%	83.2%	89.2%	90.0%	86.4%
平稳运动片段	69.2%	69.5%	80.9%	86.3%	76.5%
剧烈运动片段	65.1%	66.4%	78.5%	84.8%	73.7%

由表 4-1 可知，用平稳片段作为日常样本的分类效果是最优的，平均准确率高达 86.4%，其中随机森林算法的准确率达到了 90%。而当日常样本掺杂了 50%的运动片段以后，平均准确率立即降了十个百分点，降至 76.5%。三者中剧烈运动片段的识别效果最差，平均准确率为 73.7%。

以上结果可以说明，日常的剧烈运动对于基于肢端动作信号的癫痫发作识别存在一定的干扰，但平均准确率仍在 75%左右，其中随机森林模型表现较为突出。

而由于平稳运动片段与剧烈运动片段的准确率仅相差三个百分点，其相比于平稳片段低了十个百分点，所以可以认为平稳运动片段也一定程度上表征了剧烈运动信号对于本次实验的干扰。考虑到平稳运动片段更贴近于日常生活情况，有平稳静态，也有剧烈运动状态，所以后续的日常数据样本选择，都选择平稳运动片段来进行实验。

4.2.2 窗口大小对识别效果的影响

窗口大小的设置，对于数据分割滑窗以及后续的算法模型识别效果来说，都有着至关重要的影响。因为窗口大小这个参数本身表示的是单个样本的点数和信息量——窗口越大，含点数越多，则单个样本自身携带的信息量也就愈多。但与此同时，窗口也不是愈大愈好，因为人体各个运动的周期特性不同，当窗口过大时，则可能窗口本身包含了多种运动模式，从而导致单运动状态的识别效果变差。因此对本课题的应用方向来做一次窗口大小对比实验是很有必要的。

在参考多篇基于三轴加速度的运动姿态识别的相关文献后，发现 6s 的窗口时间或者 120 点左右为相关窗口长度是最优识别参数。考虑到本次采集频率仅为 5Hz，且性能指标上对于识别时间有要求，所以本次窗口大小对比实验围绕着 6s 的窗口时间来进行对照组设置，进行了从 1s 到 10s 的窗口大小的分类结果对比，其具体结果如下表 4-2 所示，表中衡量分类器的指标为准确率。实验中，数据选择的日常样本为平稳运动片段，特征集的特征有 21 个（见表 2-3），并进行了数据样本均衡处理。图 4-3 为窗口大小与分类准确率关系的折线图。

表 4-2 窗口大小对比实验结果

窗口大小	逻辑回归	线性 SVM	Rbf 核 SVM	随机森林	平均准确率
1s	65.4%	65.0%	68.5%	70.5%	67.4%
2s	66.8%	66.8%	71.7%	77.6%	70.7%
3s	66.6%	66.4%	74.9%	79.8%	72.0%
4s	67.0%	67.2%	77.8%	83.1%	73.8%
5s	67.7%	67.7%	80.6%	84.9%	75.2%
6s	69.2%	69.5%	80.9%	86.3%	76.5%
7s	69.1%	68.8%	81.8%	86.7%	76.6%
8s	70.1%	70.7%	83.8%	87.9%	78.1%
9s	69.8%	70.4%	84.1%	88.3%	78.1%
10s	70.2%	70.2%	85.5%	89.2%	78.8%

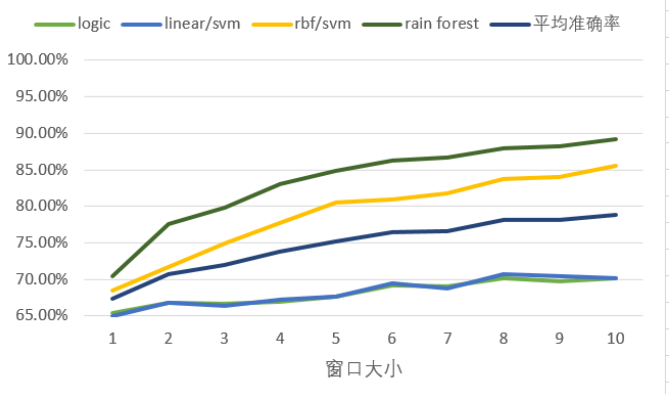

图 4-3 窗口大小与分类准确率关系的折线图

由表 4-2 与图 4-3 可知，本次各个算法模型的准确率随着窗口时间的增长而提高，10s 时达到准确率峰值，平均准确率为 78.8%，其中表现最好的随机森林模型的准确率为 89.2%。出现这个现象的原因预估有两点：一是因为随着窗口时间的延长，窗口点数增加，信息量在增加，所以准确率提高；二是因为本次采样频率稍低，5Hz 的采样频率，10s 的窗口长度仅包含了 50 个点，所以可能没有涵盖完整个癫痫发作的小周期，也并没有达到癫痫识别的阈值点数，所以不像参考文献[26]和[50]中那般，时间窗口增大到一定程度后就递减下降。这也就说明后续应该提高信号的采集频率，才能更加有效地进行癫痫发作识别。

从表 4-2 中可以看出，窗口时间从 5s 到 6s 的改变，平均准确率就提高了 1.3%，而 6s 与 10s 的平均准确率的差别也仅为 2.3%，所以可以看出 5s 到 6s 的识别效果提升明显。此外，当窗口时间分别为 6s 与 7s 时，两者的平均准确率几乎没有差别，甚至窗口为 6s 时的逻辑回归模型与线性 svm 分类器的识别分类效果均强于 7s 时的表现。综上，考虑到本次实验性能指标中对于识别时间的要求，所以后续的算法实验均以 6s 作为窗口时间长度来进行数据滑窗分割。

4.2.3 数据均衡对识别效果的影响

如 2.4 小节所言，本次采集到的数据中，正常样本数据远比发病数据要多，这也就导致了无论是测试集或训练集，负样本数据都比正样本数据多，整体正负样本的数据分布不均衡。这可能会带来识别效果偏差的现象，因此本课题进行了数据均衡对比实验，来验证一下数据均衡对于识别效果的影响。

本次数据均衡对比实验主要通过逻辑回归模型和随机森林模型的识别效果来进行体现。实验中，数据选择的日常样本为平稳运动片段，采集数据的窗口时间为 6s，特征集的特征有 21 个（见表 2-3）。具体结果如表 4-3 与表 4-4 所示。

表 4-3 数据均衡对逻辑回归识别效果的影响

	准确率	正样本召回率	负样本召回率
数据均衡	69.2%	65.7%	72.1%
数据不均衡	70.5%	60.8%	78.6%

表 4-4 数据均衡对随机森林识别效果的影响

	准确率	正样本召回率	负样本召回率
数据均衡	86.3%	87.4%	84.3%
数据不均衡	87.1%	86.7%	87.3%

由表 4-3 与表 4-4 可知，是否进行数据均衡操作，对于各分类器识别的准

确率来说影响都不大，在 2%的影响范围内。但是数据均衡与否，对于正负样本的召回率则影响甚远。就逻辑回归模型来看，数据均衡的正样本召回率比数据不均衡的要高出 4.9%，而其负样本召回率则比数据不均衡的要低 6.5%。出现这个现象是因为前言所说的负样本个数远比正样本个数多，所以导致如果不进行数据均衡操作，则模型训练时会更偏向于负样本训练，使得负样本召回率高而正样本召回率低。

综上，考虑到本次实验性能指标更专注于漏报率的改善，而漏报率与正样本召回率呈负相关关系，所以后续的算法实验中均进行数据均衡操作，以确保降低漏报率。

4.2.4 分类器识别效果对比

由于目前国内对基于肢端动作信号的癫痫发作识别的研究上仍处于空白，所以未能够存在一个适用于肢端动作信号来进行癫痫发作识别的算法标准。本次实验结合第三章的算法模型构建，针对采集来的肢端动作信号，进行一系列实验对比分析，从逻辑回归、支持向量机、随机森林等三种分类算法中找出最优的算法模型。

首先，尝试着不进行特征提取而使用原始数据信号来进行学习训练，看一下原始数据本身是否存在明显的区分关系。实验中，数据选择的日常样本为平稳运动片段，采集数据的窗口时间为 6s，已进行数据均衡操作。具体结果如表 4-5 所示。

表 4-5 基于原始数据的不同分类算法结果

算法	准确率
逻辑回归	52.66%
线性 SVM	38.92%
Rbf 核 SVM	66.36%
随机森林	70.98%

由表 4-5 可知，线性模型的表现相比非线性模型的表现差得多，所以这相对印证了 3.1 节本次数据是否线性可分的结论。此外，基于原始数据的分类器效果表现远不尽人意，离本次实验需要的性能指标还有一定距离。所以这也说明了原始数据本身的规律性不强，需要我们进行特征提取，在保留数据信息和规律性的前提下减少噪声的干扰。但可以看到的是，随机森林算法相对于其它三种算法有着明显的准确率优势。

然后，进行本次标准规格的分类器算法对比实验，来选出适配本课题的最优分类器算法。实验中，数据选择的日常样本为平稳运动片段，采集数据的窗

口时间为 6s，已进行数据均衡操作，特征集的特征有 21 个（见表 2-3）。具体结果如表 4-6 所示，表中使用性能指标要求的准确率、漏报率、误报率来对算法模型分类效果进行评估。

表 4-6 标准实验的不同分类算法结果

算法模型	准确率	漏报率	误报率
逻辑回归	69.21%	16.34%	14.45%
线性 SVM	69.47%	14.86%	15.67%
Rbf 核 SVM	80.93%	6.78%	12.29%
随机森林	86.30%	5.13%	8.57%

由表 4-6 的实验结果可得，随机森林算法模型对于基于肢端动作信号的癫痫发作识别问题的分类效果最佳，准确率为 86.3%，且误报率仅为 5.13%。此外，Rbf 核函数 SVM 也明显优于除随机森林之外的其余算法，准确率为 80.93%。而从表 4-5 与表 4-6 的对比中，则可以看出特征提取对于整体算法模型性能的提升效果。

综上，考虑到本课题的性能指标要求以及随机森林算法的优越表现，本课题选择随机森林算法模型作为本次实验的最优分类器算法。

4.2.5 个人识别效果实验

之前的实验都是使用整体训练集和测试集来进行验证的，其中发病样本共 8134 条，正常样本为 10156 条。但考虑到每个人的动作幅度和频率不同等因素，所以癫痫发作检测识别也存在有个性化的需求。因此本次实验进行了个人识别效果的对比实验，验证对于不同样的个人，各分类器的表现。

实验设置实验组和对照组，实验组的被测试人员日常活动幅度小，偏静态；而对照组的被测试人员则活泼好动，日常活动活跃。两组各有三人，每人的正常样本数和发病样本数均为 600 条左右。实验中采集数据的窗口时间为 6s，已进行数据均衡操作，特征集的特征有 21 个（见表 2-3）。具体实验记录分别如表 4-7 与表 4-8 所示，表中使用准确率作为衡量指标。

表 4-7 不同个人的癫痫发作识别效果(实验组)

	逻辑回归	线性 SVM	Rbf 核 SVM	随机森林	平均准确率
ZLH	82.3%	82.6%	88%	89.7%	85.6%
WHR	86.8%	86.9%	88.8%	90.1%	88.1%
DX	86.7%	85.5%	87.8%	89.7%	87.4%

表 4-8 不同个人的癫痫发作识别效果(对照组)

	逻辑回归	线性 SVM	Rbf 核 SVM	随机森林	平均准确率
CXF	65.4%	63.5%	82.4%	85.2%	74.1%
LJQ	77.6%	77.8%	83.9%	88.8%	82.0%
ZJH	85.0%	86.3%	87.4%	88.3%	86.7%

从表 4-7 与表 4-8 可知，对于不同情况的个人来说，识别准确率的确存在着些许差异，实验中最大的平均准确率差别为 14%。但与此同时可以看到，对于不同个人的癫痫发作识别效果都还算优异，相比总体的分类识别效果（见表 4-6）来说，都有一定程度的提升。除 CXF 一组以外，平均准确率均在 82%以上，其中随机森林算法的表现最优，准确率均在 85%以上。

综上，后续对于不同个体的癫痫发作的在线识别可采用随机森林算法进行训练学习，在确保个性化需求的同时，也能保证极高的准确率。

4.3 在线测试

由前文可知，符合本课题基于肢端动作信号的癫痫发作识别最好的方案是，窗口采样时间为 6 点，进行数据选择和数据均衡处理，利用整理过后的 21 个特征进行随机森林算法模型的构建，并以此进行分类识别。所以在此基础上，本文对在线测试的方案进行探讨，以满足课题要求。

由于本课题不只局限于实验室验证阶段，还要探讨算法的真实应用和项目的现实落地。所以考虑到腕带式癫痫检测装置底层芯片的计算资源，本次在线测试分为两套方案进行：一是通过还原出分类器的决策边界，然后在底层芯片中编写出决策函数，以此进行在线识别；二是通过底层芯片的蓝牙功能，将数据实时传输到 PC 端服务器上，通过 PC 端建立好的算法模型来进行在线识别。

4.3.1 基于决策边界的在线测试

从第三章算法模型构建中可知，随机森林模型作为 Bagging 结构的集成学习算法，类似于黑箱训练，很难还原出具体的决策函数而只能保存模型的参数在电脑端。至于其余三种算法，则均可以还原出具体的决策函数与决策边界。例如逻辑回归模型与线性 SVM，都可以通过返回系数矩阵 w 与截距 b，来得到决策函数 f(x)=wx+b。

此处需要说明的是，原本还原决策边界的在线测试，是打算在得到决策边界以后，编程在底层端芯片上，利用腕带式癫痫检测装置去医院进行实时检测的。但由于现实因素的影响，暂时不能够按预想的实验进行。所以本次决策边

界的在线测试，使用 python 来进行编程还原，实验所用的数据集为本次采集数据中预留的两组数据，该数据没有被分类器学习和训练。具体结果如下表 4-9 所示。

表 4-9 基于决策边界的在线测试结果

	准确率	漏报率	误报率
逻辑回归	68.83%	15.67%	15.50%
线性 SVM	68.98%	14.23%	16.79%
Rbf 核 SVM	87.07%	4.97%	7.96%

从表 4-9 中可知，基于决策边界的在线测试结果与 4.2.4 中的分类算法结果相近，其中表现最好的是 Rbf 核 SVM，在确保 87.07%的准确率的同时，将漏报率降到了 4.97%，完全能够满足本次在线测试的性能指标要求。此外，逻辑回归与线性 SVM 的表现一般，准确率为 69%，并且漏报率在 15%左右，相对来说效果较差，没有满足要求。

其中需要额外考虑的是 Rbf 核 SVM 算法，因为它的决策边界本身是通过将低维数据映射到高维空间后通过各个支持向量来拟合而成，所以本次它的决策边界共涉及到了 5793 个支持向量，见式 3-20。考虑到还原出它的决策边界需要进行的指数运算和累加和等运算，底层芯片的计算资源可能很难满足。

4.3.2 基于调用模型的在线测试

从第三章算法模型构建的各算法流程图中可以看到，各算法模型在搭建、训练学习完之后，都会有一步模型保存。这主要是通过 python 的 joblib 库中函数来进行实现。所以在保存了模型，得到.model 文件之后，本文提出了基于调用模型的在线测试方案。方案实现的流程图如下图 4-4 所示。

由图 4-4 中可知，首先得确保蓝牙功能的稳定通信，收到传送过来的窗口时间为 6s、重叠率为 50%的肢端动作数据包。然后通过相应的特征提取，才能够得到可输入于模型的特征集。最后就是调用.model 文件来进行在线判断，输出结果。

同上 4.3.1，本次基于调用模型的在线测试实验所用的数据集为本次采集数据中预留的两组数据，具体的实验结果如下表 4-10 所示。

表 4-10 基于调用模型的在线测试结果

	准确率	漏报率	误报率
Rbf 核函数	86.58%	5.68%	7.74%
随机森林	88.83%	4.63%	6.54%

从表 4-10 中可知，基于调用模型的在线测试识别表现优异，尤其是随机森

林算法，准确率高达 88.83%，且漏报率低至 4.63%，用时仅不到 1s。所以可以预想到，当蓝牙功能能够进行稳定通信不丢包时，基于调用模型的在线测试是最优的检测识别方案。而后续的算法模型升级，可通过定时的学习训练来进行更新。

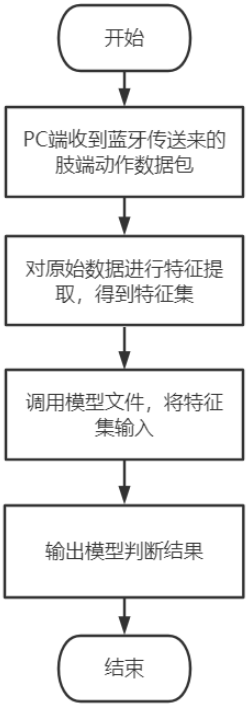

图 4-4 基于调用模型的在线测试流程图

4.4 本章小节

本章首先介绍了本次的实验条件和衡量实验效果的各个指标，然后分别进行了数据选择对比实验、窗口大小对分类器识别影响实验、数据均衡对比实验、各分类器识别对比实验、个人识别效果实验等五个实验，得到了适配本课题的最佳识别方案——窗口采样时间为 6 点，进行数据选择和数据均衡处理，利用整理过后的 21 个特征进行随机森林算法模型的构建，并以此进行分类识别。除此之外，本章还进行了两种不同方案的在线测试，对比分析其优劣并结合现实情况进行选用。

结　论

由于我国癫痫病患者人数逐年增加，越来越多的家庭遭受了癫痫疾病的困扰，所以对于快速有效的癫痫发作检测识别技术有着日益迫切的需求。但传统的基于脑电信号 EEG 的识别诊断方式由于受到空间的局限性而不能推广到日常生活中使用。在此背景下，本文提出了一种基于肢端动作信号来进行癫痫发作识别的系统设计方案，并对其中每个步骤都进行了理论论证和具体实现，通过实验结果展示来进行了有效性验证。本文主要成果分为以下四个部分：

1.利用课题组设计的腕带式癫痫检测装置完成了对病人原始数据的采集，并针对采集来的数据提出了阈值消除小振动干扰、平滑滤波、合成加速度等数据预处理方法，再进行数据选择和数据分割滑窗等处理手段后得到可用的数据格式。其中针对数据选择对课题的影响进行了实验论证，并通过实验得出了适用于本课题的最优窗口时间大小。

2.对适用于肢端动作信号的癫痫发作识别问题的最优特征集进行了理论论证和实际特征提取，并通过数据集处理得到训练集与测试集。针对特征多而繁杂等问题采用基于随机森林特征重要性的方案来进行特征选择，对比不同特征集的分类准确率，来得到适用于本课题最佳的特征集。同时，进行了数据均衡对比实验，验证了数据均衡对于本课题分类效果的影响。

3.研究逻辑回归、支持向量机、随机森林等三种算法，构建不同的算法模型并对比其识别准确率、漏报率、误报率等，综合考虑得到基于肢端动作信号的癫痫发作识别效果最佳的分类算法。

4.综合以上得到的系统流程步骤，设计并对比了两种在线测试方案，进行功能测试和性能指标试验后，提出了面对不同硬件条件时基于肢端动作信号的癫痫发作的具体检测识别方案。

此外对于本课题如今取得的成果，后续可进行深入研究与扩展的方向有：

（1）扩展采集来的数据，使原始数据更多样化。比如提高采样频率，增加单个样本的数据量；比如采集到癫痫病人的日常数据，而不仅是病床上的数据，则可以引入跌倒检测等多个癫痫发作并发征兆的特征，从而提高识别准确率。

（2）在线测试方案上，模型系数更新可以考虑在线训练和在线更新的算法；

（3）扩展更多病症的识别检测，比如帕金森状态识别等。

参考文献

[1] Gotman J. Automatic recognition of epileptic seizures in the EEG[J]. Electroencephalography & Clinical Neurophysiology, 1982, 52(5): 530-540.
[2] Epilepsy Foundation of America. About Epilepsy: The Basics[OL]. Epilepsy Foundation, 2014-03-19. http://www.epilepsy.com/learn/about-epilepsy-basics.
[3] 刘莉莉. 基于神经网络的癫痫发作预测[D]. 北京：北京工业大学, 2018.
[4] Gaitatzis A, Trimble MR, Sander JW. The psychiatric comorbidity of epilepsy[J]. Acta Neurologica Scandinavica, 2004, 110:207–20.
[5] 杜沛冬. 结合卷积神经网络和随机森林的癫痫自动监测[D]. 济南：山东大学, 2018.
[6] Snyder David E., Echauz Javier, Grimes David B., et al. The statistics of a practical seizure warning system. Journal of Neural Engineering[J], 2008. 5(4): 392.
[7] 孙涛, 王峰. 神经外科与癫痫[M]. 人民军医出版社, 2015.
[8] Niederhoefer C, Gollas F, Chernihovskyi A, et al. Detection of seizure precursors in the EEG with cellular neural networks [J]. Epilepsia, 2004, 45(1):245-245.
[9] Osorio I, Frei M G, and Wilkinson S B, Real-time automated detection and quantitative analysis of seizures and short-term prediction of clinical onset. Epilepsia[J], 1998. 39(6): 615-627.
[10] Chu Hyunho, Chung Chun Kee, Jeong Woorim, et al., Predicting epileptic seizures from scalp EEG based on attractor state analysis. Computer Methods & Programs in Biomedicine[J], 2017. 143(C): 75-87.
[11] 祁玉, 基于脑电的癫痫预警及预警—抑制诊疗系统关键技术研究[D], 浙江, 浙江大学, 2015.
[12] Ghosh-Dastidar S, Adeli H, Dadmehr N. Principal Component Analysis-Enhanced Cosine Radial Basis Function Neural Network for Robust Epilepsy and Seizure Detection[J]. IEEE Transactions on Biomedical Engineering, 2008, 55(2):512-518.
[13] Subasi A, Ismail Gursoy M. EEG signal classification using PCA, ICA, LDA and support vector machines[J]. Expert Systems with Applications, 2010, 37(12):8659-8666.
[14] Faust O, Acharya R U, Allen A R, et al. Analysis of EEG signals during epileptic and alcoholic states using AR modeling techniques[J]. Irbm , 2008, 29(1):44-52.

[15] Chisci L, Mavino A, Perferi G, et al. Real-time epileptic seizure prediction using AR models and support vector machines[J]. IEEE transactions on bio-medical engineering, 2010, 57(5):1124.
[16] Zandi A S, Tafreshi R, Javidan M, et al. Predicting temporal lobe epileptic seizures based on zero-crossing interval analysis in scalp EEG[C] // Engineering in Medicine and Biology Society. IEEE, 2010:5537-5540.
[17] Supriya S, Siuly S, Wang H, et al. Weighted Visibility Graph With Complex Network Features in the Detection of Epilepsy[J]. IEEE Access, 2016, 4(99):6554-6566.
[18] Hamad Asmaa, Houssein Essam H., Hassanien Aboul Ella, et al. A Hybrid EEG Signals Classification Approach Based on Grey Wolf Optimizer Enhanced SVMs for Epileptic Detection. in International Conference on Advanced Intelligent Systems and Informatics[C]. 2017. Cairo.
[19] Sriraam N. and Raghu S., Classification of Focal and Non Focal Epileptic Seizures Using Multi-Features and SVM Classifier. Journal of Medical Systems[J], 2017. 41(10): 160.
[20] Maxwell A. Global commercialization of the Parkinson's KinetiGraph[J]. Australasian Biotechnology, 2014, 24(2):45-46.
[21] Van Dooren, De Vries, J. J., Janssen, J. H. (2012). Emotional sweating across the body:comparing 16 different skin conductance measurements locations. Physiology & Behavior, 106(2), 298-304. doi:10.1016/j.physbeh.2012.01.020.
[22] Francesco Onorati, Chiara Caborni, et al. Performance of a wrist-worn multimodal seizure detection system for more than a year in real-life settings. Conference: 13th European congress on Epileptology, At Vienna.
[23] 戴若梦. 基于深度学习的运动想象脑电分类[D]. 北京：北京理工大学, 2015.
[24] 周梦妮. 复杂度和时频分析在癫痫脑电信号诊断与发作预测中的应用研究[D]. 太原：太原理工大学, 2019.
[25] 郑天依. 基于机器学习的癫痫及精神异常脑电信息识别研究[D]. 北京：北京邮电大学, 2019.
[26] 吴海龙. 基于加速度传感器的游泳监测系统分析与实现[D]. 广州：华南理工大学, 2018.
[27] 黄彬. 智能空间中人的行为识别与理解[D]. 济南：山东大学, 2010.
[28] 段豪. 基于 Android 的健康监测应用研究与实现[D]. 电子科技大学, 2017.
[29] 魏芬, 邓海琴. 基于加速度传感器的运动步数检测算法研究[J]. 电子器件, 2016, 39.

[30] Francesco Onorati, Giulia Regalia, et al. Multicenter clinical assessment of improved wearable multimodal convulsive seizure detectors. Epilepsia [J], 2017 Nov, 58(11):1870-1879. doi: 10.1111/epi.13899. Epub 2017 Oct 4.

[31] Wang S., Yang J., and Chen N., et.al. Human activity recognition with user-free accelerometers in the sensor Networks [C]. IEEE Int. Conf. Neural Networks and Brain, 2005, 2:1212-1217.

[32] 林珠, 刑延. 数据挖掘中适用于分类的时序数据特征提取方法[J]. 计算机系统应用, 2012, 21.

[33] 薛洋. 基于单个加速度传感器的人体运动模式识别[D]. 广州：华南理工大学, 2011.

[34] Lester Jonathan, Choudhury Tanzeem, and Borriello Gaetano. A practical approach to recognizing physical activities [C]. LNCS, 2006, 3968:1-16

[35] Yang J.-Y., Wang J.-S., and Chen Y.-P. Using acceleration measurements for activity recognition: an effective learning algorithm for constructing neural classifiers [J]. Pattern Recognition Letters, 2008, 29(16):2213-2220.

[36] 何振宇. 基于三轴加速度传感器的人体运动识别研究[D]. 广州：华南理工大学, 2009.

[37] Choudhury T., Borriello G., Consolvo S., et al. The mobile sensing platform: an embedded system for capturing and recognizing activities [C]. IEEE Pervasive Magazine Special Issue on Activity-Based Computing, April, 2008.

[38] Wu Jia-Hui, Pan Gang, and Zhang Da-Qing, et al. Gesture recognition with a 3-D accelerometer [J]. Ubiquitous Intelligence and Computing, 2009, 5585(6):25-38.

[39] Maguire Domnic, Frisby Richard. Comparison of feature classification algorithms for activity recognition based on accelerometer and heart rate data [C]. The 9th. IT &T Conference, 2009:1-8.

[40] 李景辉. 基于多传感器信息融合的人体姿态识别研究[D]. 济南：山东大学, 2014:20-25

[41] 彭欣然. 基于多传感器的人体姿态识别系统[D]. 哈尔滨：哈尔滨工业大学, 2017:32-43

[42] Li Ming, Rozgic Viktor, and Thatte Gautam, et al. Multimodal physical activity recognition by fusing temporal and cepstral information [J]. IEEE Trans. on Neural Systems and Rehabilitation Engineering, 2010, 18(4):1-10.

[43]. Wang N,Ambikairajah E, Lovell N H, et al. Accelerometry based classification of walking patterns using time-frequency analysis[C]. Engineering in Medicine and Biology Society, 29th Annual International Conference of the IEEE, 2007: 4899-

4902.
[44] 佟丽娜. 基于力学量信息获取系统的人体摔倒过程识别方法研究[D]. 合肥：中国科学技术大学, 2011.
[45] 孙新香. 基于三轴加速度传感器的跌倒检测技术的研究与应用[D]. 上海：上海交通大学, 2008.
[46] 蔡菁. 皮肤电反应信号在情感状态识别中的研究[D]. 重庆：西南大学, 2010.
[47] 凤河. 有限注意与 A 股市场股价回归预测—基于 SVM 与 Logistic 的比较研究[D]. 西安：西安理工大学, 2019.
[48] 李航. 统计学习方法[M]. Beijing: China Machine Press, 2012:364-370
[49] 马骊. 随机森林算法的优化改进研究[D]. 广州：暨南大学, 2016.
[50] 张贵昌. 穿戴式运动状态识别及心率监测的研究. 贵阳：贵州大学, 2019.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于肢端动作信号的癫痫发作识别研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：李业鸿　　日期：2020 年 6 月 16 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：李业鸿　　日期：2020 年 6 月 16 日

导师签名：　　日期：2020 年 6 月 16 日

致 谢

本文是在导师梁廷伟教授的悉心指导下完成的。梁老师从论文开题、中期乃至最后论文完善的整个过程中，都给了我耐心而细致的教导。回首过去的两年读书生涯，梁老师不仅在学术科研中为我指引了方向，在生活中也常常对我们学生关怀备至。所以值此论文完成之际，谨向梁老师致以衷心的感谢和崇高的敬意！

此外，还要感谢金显吉老师和吴字宇师兄在本论文工作上给予我的帮助和指导，感谢韩凯歌同学、孔得慧同学在理论知识和代码实现中给予的指导和帮助。

感谢飞行器与控制研究所提供的良好科研环境，感谢研究所里所有给予我支持和帮助的老师、同学和朋友们。

最后感谢乔晓妍同学在我科研路上的陪伴和情感支持，每当我打算放弃或意欲退缩的时候，是她给了我坚持下去的力量。也感谢我的家人们，正是因为你们的无私奉献和默默陪伴，才让我能在科研路上这么潜心地走下去。

光阴荏苒，白驹过隙，不知不觉间两年的研究生学习生活、六年的哈工大生涯就这么过去了。回望过去在工大的六年，我从一个踌躇满志、懵懵懂懂的青春少年成长为现在成熟稳重的研究生。在过去的六年里，我曾在图书馆里奋战过，也曾在宿舍里迷惘过，那些在哈工大里求学的无声岁月里保存了太多舍不得的人和事。这段求学时光也将是我人生中永远难忘的回忆。感谢你们，感谢百年哈工大，让我们有缘相遇。