**诚信声明**

本人声明：

本人所呈交的毕业设计（论文），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者签名： 日期：

**毕业设计(论文)任务书**

设计(论文)题目：基于小波变换和支持向量机的癫痫脑电信号分类

学院：信息科学与技术学院 专业：自动化实验班 班级：自实1601

学生：陈帅华 指导教师： 宿翀 专业负责人： XXX

1. 设计（论文）的主要任务及目标

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX(小4，宋体)。

2、设计（论文）的主要内容

（1）XXXXXXXXXXXXXXX；

（2）XXXXXXXXXXXXXXXXXX；

（3）撰写毕业论文；

（4）翻译一篇5000汉字左右的外文文献。

3．设计（论文）的主要要求

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX(小4，宋体)。

4．主要参考文献

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX(小4，宋体)。

5．进度安排

|  |  |  |
| --- | --- | --- |
|  | 设计（论文）各阶段名称 | 起止日期 |
| 1 | 调研及查阅文献、翻译外文资料、撰写文献综述、开题报告 | 年月日--月日 |
| 2 | 对其中一种构型进行三维建模，初步进行流场模拟 |  |
| 3 | 流场模拟计算的完善及结果分析 |  |
| 4 | 撰写毕业论文、答辩准备 |  |

**基于小波变换和支持向量机的癫痫脑电信号分类**

**摘要：**XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX（宋体，小四号）

**关键词：**EEG，癫痫，小波变换，支持向量机，K最近邻算法，决策树，集成学习

注：

1、**摘要：（黑体，小四号，加粗）**

2、**关键词：（黑体，四号，加粗）**

3、1.25倍行距，段前0.5行

**Mixing Performance of Different Configurations of New Kneading Disc in a Co-rotating Twin Screw Extruder（Times New Roman，三号，加粗）**

**ABSTRACT：** XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX（Times New Roman，小四号）

**KEY WORDS:** XXXXXXXX，XXXXXXX，XXXXXX，XXXXXX（Times New Roman，四号）

注：

1、**ABSTRACT：（Times New Roman，小四号，加粗）**

2、**KEY WORDS:（Times New Roman，四号，加粗）**

3、1.25倍行距，段前0.5行

4、标题实词首字母大写，关键词中实词首字母大写。

目 录

[前 言 1](#_Toc40623211)

[第1章 绪论 2](#_Toc40623212)

[第1.1节 课题研究的背景与意义 2](#_Toc40623213)

[第1.2节 国内外研究现状 3](#_Toc40623214)

[1.2.1、癫痫脑电信号特征提取研究现状 5](#_Toc40623215)

[1.2.2、癫痫脑电信号分类算法研究现状 5](#_Toc40623216)

[第1.3节 本文的创新点与研究难点 5](#_Toc40623217)

[1.3.1、 本文的创新点 5](#_Toc40623218)

[1.3.1、 本文的研究难点 6](#_Toc40623219)

[第1.4节 本文的主要内容与安排 6](#_Toc40623220)

[第2章 脑电相关知识与癫痫检测 9](#_Toc40623221)

[第2.1节 脑电相关知识与癫痫脑电信号分类 9](#_Toc40623222)

[2.1.1、脑电信号特征与分类 9](#_Toc40623223)

[2.1.2、癫痫脑电信号分类 10](#_Toc40623224)

[第2.2节 脑电信号的采集 11](#_Toc40623225)

[第3章 基于离散小波变换的脑电特征提取 12](#_Toc40623226)

[第3.1节 脑电信号数据来源及介绍 12](#_Toc40623227)

[3.1.1、波恩大学癫痫脑电数据集 12](#_Toc40623228)

[3.1.2、CHB-MIT头皮脑电图数据库脑电数据集 13](#_Toc40623229)

[第3.2节 (癫痫)脑电信号预处理 15](#_Toc40623230)

[3.2.1、波恩大学癫痫脑电数据集 15](#_Toc40623231)

[3.2.2、CHB-MIT头皮脑电图数据库脑电数据集 15](#_Toc40623232)

[第3.3节 小波变换理论介绍 17](#_Toc40623233)

[第3.4节 由离散小波变换进行脑电特征提取 18](#_Toc40623234)

[第3.5节 结果与分析 27](#_Toc40623235)

[第4章 基于多种机器学习算法的癫痫脑电信号分类 27](#_Toc40623236)

[第4.1节 支持向量机 27](#_Toc40623237)

[第4.2节 K最近邻算法 30](#_Toc40623238)

[第4.3节 决策树 32](#_Toc40623239)

[第4.4节 癫痫脑电信号的分类 33](#_Toc40623240)

[第5章 集成学习 39](#_Toc40623241)

[第5.1节 集成学习理论介绍 39](#_Toc40623242)

[第5.2节 集成学习1 40](#_Toc40623243)

[结 论 41](#_Toc40623244)

[参考文献 43](#_Toc40623245)

[致 谢 45](#_Toc40623246)

**注：**

1、每章标题字体（黑体，四号，加粗）

2、每节标题及三级小标题字体（宋体，四号）

3、前言在第1章前。结论、参考文献、致谢不单独成章。

前 言

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX（宋体，小四，固定值22磅，首行缩进2字符）

第1章 绪论

第1.1节 课题研究的背景与意义

人类大脑中有上百亿个脑细胞，神经元与神经胶质细胞又共同组成了脑细胞，人类自身的一切活动过程正是靠大脑进行主导和调节的，以保证人类机体和周围环境的平衡，而这一主导和调节过程又正是依靠人类大脑内由大量神经元互相连接进而构成的极其庞大的神经网络实现的。这一种神经网络结构会在神经冲动出现的时候，使众多神经元被激活开始工作，进而传导电信号，故而大脑是靠通过发送传导电信号进行活动，控制神经的。

当人体处于正常的生理状况时，这一电信号的产生和传导过程是有迹可循的、有限的。但是当人体受到遗传因素的影响或者出现脑结构异常和脑功能的损伤时，可能会造成人体大脑的异常放电，这种放电形式比正常生理状况下的电信号传导距离要长，影响范围要广，同时也不受控制，可能还会造成短暂的大脑功能障碍。而这种异常的放电形式往往会诱发癫痫的产生。

癫痫也即民间常叫的“羊癫疯”，这种疾病发作的时候往往会对人体造成很大的伤害，尤其是对大脑造成的损伤，比如会造成人体大脑记忆力的下降和认知能力的下降，而如果该疾病频繁地发作，甚至会造成对大脑永久性的伤害。由世界卫生组织于2018年发布的报道可知，在全球范围内，罹患癫痫疾病的人数已经超过了5000万，而其中绝大一部分都出现在了中低收入国家，而在我们国家中，有900万以上的国人正深深地受到该疾病的困扰，此外癫痫病例增长的速度也十分的可怕，每年都会有60万例的增长[1]。而罹患癫痫之后，不仅患者的整个家庭要承受高昂的治疗费用，还可能会对患者的心理健康和未来的发展造成十分不利的影响，比如患者可能会遭受到社会隔离甚至是污名化，也可能会造成患者较低的教育成绩与糟糕的就业效果。故而针对癫痫的预防和鉴别诊断具有及其重要社会意义和现实意义。

但在实际的诊断过程中，诊断癫痫可能会存在很大的困难，出现这种情况的原因主要是由于有不少其他的疾病可能会展示出同癫痫疾病发作时十分类似的表现和病征，例如昏厥、换气过度综合征（Hyperventilation syndrome，HVS）、偏头疼和发作性睡病等等之类的疾病。目前癫痫的主要检查手段是脑电图（Electroencephalogram，EEG），该检查手段主要是在检查测试的时候，将精密的传感器通过黏黏的导电胶安装在人体的头皮之上，以获取人体大脑细胞互相发送信息的时候所产生的电信号。通过脑电图仪所记录的脑电信号，大脑的活动可以被清晰地呈现出来，若在脑电图监测期间出现癫痫发作的情况，那么脑电图上就会可以看到正常的大脑活动模式发生了改变，并可以看到癫痫发作过程中所独有的特征波。

但是通常由于脑电图本身是非常复杂的，人类的脑电波也是多种多样，而在临床上，主要靠医生在几分钟到几个小时之内测量脑电波，进而找到大脑之中异常脑电活动的依据，并且找出癫痫患者罹患的癫痫发作类型以及起源。但是这种方式十分依赖于医务人员自身的诊断经验和主观标准，进而可能会出现对患者的误诊断和漏诊断的情况，此外长时间采用这种方式会对医务人员的精力造成极大的消耗而且效率也会比较低，这对于越来越多的癫痫患者来说却是远远不够，因而寻找对癫痫脑电信号的识别、特征提取与分类方法对社会，对患者家庭和患者自身而言都有着极其重要的意义！

第1.2节 国内外研究现状

脑电信号具有不平稳和非线性的特性，是一种随着科学技术的进近年来，癫痫脑电信号检测与分类技术迅速发展，支持向量机、隐马尔科夫链、极限学习机、随机森林、深度学习和人工神经网络等方法也被广泛应用**错误!未找到引用源。错误!未找到引用源。错误!未找到引用源。错误!未找到引用源。**。

对于脑电信号特征的提取，针对癫痫脑电和正常脑电的不同特点，国内外研究人员对于脑电信号特征的提取也提出了很多方法。王海玉等人对其进行了总结，例如在时域上的方差分析、直方图分析等，在频域上的频谱分析、全息谱分析等方法**错误!未找到引用源。**，但无论是时频分析还是频域分析，都未充分体现EEG信号的不平稳和非线性的特性。李小兵等人指出时频分析方法充分考虑了信号的不平稳特性，在时间和频率上均有很好的分辨率**错误!未找到引用源。**。但由于人脑由大量神经元构成，组成了一个复杂的非线性系统，因此非线性分析方法也广泛用于脑电的分析**错误!未找到引用源。**，王海玉等人也指出Lyapunov指数、复杂度、样本熵、近似熵等都是目前研究脑电主要的非线性方法**错误!未找到引用源。**。

韩敏等人利用小波变换(WT)计算出每个频带小波系数的标准差、平均功率等指标，并将其作为脑电信号的特征，最后在对正常脑电信号和癫痫脑电信号的分类过程中取得了98.38%的分类精度**错误!未找到引用源。**。此论文最后得到了较高的分类精度，但是脑电信号特征维数较大，达到了22维；且分类过程中消耗的时间也相对更多。Yatindra Kumar等人将脑电信号进行离散小波分解得到各个频带的小波系数后求取了相对小波能量和小波熵作为脑电信号的特征，由于小波熵能衡量信号的有序度和无序度，反映非线性信号潜在的动力学特性，因此对于区分癫痫脑电信号有着很好的效果**错误!未找到引用源。**。Cheol Seung Yoo等人发现在癫痫发作过程中脑电信号的样本熵值下降，在癫痫发作终止时达到最小值，癫痫发作终止后，样本熵迅速恢复到癫痫发作前的数值**错误!未找到引用源。**。Julius Georgiou等人将具有低计算复杂度的排列熵作为自动癫痫检测特征进行了可行性研究，最终对正常脑电和癫痫脑电信号取得了93.55%的最佳平均辨别率**错误!未找到引用源。**。

在对脑电信号进行特征提取后，由分类器根据特征向量进行分类。支持向量机作为一种监督学习算法，在分类和回归问题上被广泛地应用**错误!未找到引用源。错误!未找到引用源。错误!未找到引用源。错误!未找到引用源。**。肖文卿等人将支持向量机用于对正常小鼠和癫痫小鼠分类，并选取径向基核函数，在融合了多种特征向量后，得到了99.80%的分类准确率**错误!未找到引用源。**。吕健和吴艳在各自的研究中分别采用网格搜索等方法对支持向量机中的核函数参数以及惩罚因子进行了优化**错误!未找到引用源。错误!未找到引用源。**，最后的结果也证明了其较好的性能及泛化能力。

而Shahab Abdulla等人在他们的研究中除了使用支持向量机外，也采用了K最邻近算法(KNN)，该算法是机器学习算法中最为简单的一种，其关键在于对k的选取上，Shahab Abdulla等人将k定为1，在此研究中，支持向量机的性能优于KNN**错误!未找到引用源。**，这很可能是由于其k选取的过小，导致整体模型变得复杂，发生了过拟合。在Manjusha等人的研究中，为了解决KNN在一个大的训练集中花费较长时间寻找最近邻域的问题，他们采取了降维方法，但在最后对癫痫危险度分类中，使用KNN得到的准确度，灵敏度，特异性等指标较低，有待提高**错误!未找到引用源。**。

人工神经网络是基于人类大脑神经元结构和功能活动创立的数学模型，它以自身独有的非线性自适应信息处理能力被广泛应用，Yatindra Kumar等人在对癫痫脑电信号的分类中使用了BP神经网络，其结构为：输入层4个神经元，隐藏层6个神经元，输出层2个神经元，在对正常脑电和癫痫脑电的分类中取得了95%的分类准确率**错误!未找到引用源。**，但是该方法在处理大容量癫痫样本数据时，收敛速度会极为缓慢，对神经网络计算能力要求也十分苛刻，目前只能在最佳性能与计算复杂度间折衷考虑。

1.2.1、癫痫脑电信号特征提取研究现状

分为两部分，

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX [5]。（宋体，小四，固定值22磅，首行缩进2字符）

XXXXX[9-12]。

1.2.2、癫痫脑电信号分类算法研究现状

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX [12,16,18,20]。

第1.3节 本文的创新点与研究难点

* + 1. 本文的创新点

(1)采用两个数据集进行了研究验证

在癫痫脑电信号的特征提取与分类这一领域的过往研究之中，有很大一部分的研究都是针对波恩大学癫痫脑电数据集这一单一数据集进行的。但是考虑到如果仅仅针对该数据集进行信号的预处理、特征提取和信号的分类，那么最终得到的结论本身就不具备普遍性。本文考虑到这一点后，除去下载的波恩大学癫痫脑电数据集之外，又下载了波士顿儿童医院的CHB-MIT头皮脑电图数据库中的一部分数据集进行分析研究。本文通过对这两个数据集的研究进而使得最后得到的结论具备了一定的普遍性，对研究癫痫脑电信号的分类有一定的借鉴意义和实际应用意义。

(2)将集成学习方法应用于癫痫脑电信号分类，提高分类效果

本文将集成学习的方法应用在癫痫脑电信号分类这一领域之中，通过使用集成学习中的stacking方法，本文将若干个单一分类算法作为基学习器使用数据集进行训练，之后将通过基学习器得到的分类结果重新作为新的训练集来训练另一个新的分类算法，该分类算法得到的结果就作为最终的分类结果。

通过使用该算法，可以充分利用多个分类器进而得到更好的分类效果，同时提高了整体的泛化能力，很大程度上弥补了前人研究中使用单独使用某一模型时所带来的缺点。

* + 1. 本文的研究难点

1. 对CHB-MIT头皮脑电图数据库数据的读取和预处理

由于该数据库中的数据均为edf格式，而该格式的数据文件不像txt或者csv文件那样可以直接进行读取，而是要在找到合适的api之后，然后调用再进行读取。针对该问题，经过一番调查、研究与学习之后，发现eeglab工具箱功能十分强大，它可以对数据进行读取、可视化、滤波、独立成分分析和消除伪影等等一系列操作，因此本文决定下载eeglab工具箱，在matlab环境之下，对edf格式的数据进行读取和预处理操作。

第1.4节 本文的主要内容与安排

本文的主要研究内容是通过小波变换来对脑电数据集进行特征的提取，然后使用支持向量机(三种机器学习算法)根据特征向量/特征矩阵来对癫痫脑电信号进行分类。详细的研究内容安排如下：

第一章中介绍了该课题的研究背景与意义，并从癫痫脑电信号的特征提取和分类算法两方面对国内外研究现状进行了介绍，然后提出了本研究中的创新点与研究难点，最后对本研究主要进行的工作和工作的安排进行了介绍。

第二章则是对脑电信号的相关知识包括脑电信号的特征，脑电信号的采集方式(10-20系统)，癫痫脑电信号的特征波等内容进行了介绍。

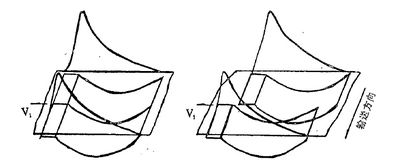
第三章的主要内容则是使用离散小波变换对脑电信号进行特征的提取。首先本文对要使用到的两个数据集进行了细致的介绍，并对其相应进行了预处理以放方便后续的特征提取。然后本文对小波变换的理论进行了介绍，并确定使用离散小波变换来进行特征提取，之后通过比较三种典型的小波基，确定以’db4’为小波基。在进行了五层离散小波变换得到小波系数之后，计算得出其标准差和样本熵，并将其作为分类的特征向量。

第四章的主要内容是使用支持向量机、K最近邻算法和决策树三种机器学习算法根据单一的标准差特征、单一的样本熵特征、融合后的标准差特征和融合后的样本熵特征分别对波恩大学癫痫脑电数据集和CHB-MIT数据集进行了四种情况的二分类和病例chb01、病例chb03癫痫发作间期与发作期的分类，并使用准确度、灵敏度和特异度这三个评价指标对分类效果进行了评价。

第五章的内容则是本文的创新点，为了提高分类器的分类效果，本文采用了集成学习中的stacking方法组成了两层分类模型，其中第一层基础模型由支持向量机和K最近邻算法两个算法构成，第二层元模型是决策树算法。将第四章中得到的融合后的特征矩阵作为这两层分类模型的输入，同样也进行第四章中所有情况的分类，最后也同样使用准确度、灵敏度和特异度这三个评价指标对分类效果进行评价。

最后在结论中，本文对所进行的癫痫脑电信号分类工作进行了总结，分析了研究中存在的不足之处，并对未来的研究提出了展望。

如图1·1所示为捏合盘元件的输送特性图。



(a) 单个捏合盘 (b) 一组捏合盘

图1·1 捏合盘元件输送特性（宋体，五号）

论文中公式表示方式：

 （1·1）

 （1·2）

（1·2）表示第1章中第2个公式

**论文中表格表示方式**：

表1·1 四种不同捏合块构型下流道的物理参数（宋体，五号，表格内文字皆用五号）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 捏合块  元件 | 导程/ | 螺杆外径/ | 机筒内径/ | 捏合盘厚度/ | 错列角/° | 中心距/ | 计算域长度/ |
| NKBLB |  |  |  |  |  |  |  |
| NKBRB |  |  |  |  |  |  |  |
| NKBLF |  |  |  |  |  |  |  |

续表1.1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| NKBRF |  |  |  |  |  |  |  |
| NKBRF |  |  |  |  |  |  |  |

第2章 脑电相关知识与癫痫检测

第2.1节 脑电相关知识与癫痫脑电信号分类

2.1.1、脑电信号特征与分类

脑电信号是由大脑神经活动产生的一种自发性电位活动，该电位活动缓慢、具有节律性并且持续。但是该信号较为微弱，其幅值一般在5~200之间，频率范围则一般在0.5~100Hz。当根据频率的不同对其脑电信号分类时，其可以分为四大类，具体如表2·1所示，而这四类典型的脑电信号波形图如图2·1所示。

表2·1 脑电信号的四种分类

|  |  |  |  |
| --- | --- | --- | --- |
| 种类 | 频率 | 幅值 | 出现情况 |
| 波 | 0.5~4Hz | 20~200 | 常见于成人和儿童睡眠时 |
| 波 | 4~8Hz | 100~150 | 一般在大脑系统活动处于抑制状态下会出现 |
| 波 | 8~13Hz | 20~100 | 一般出现在成人大脑处于安静的状态时 |
| 波 | 13~30Hz | 5~20 | 当中枢神经神经系统强烈活动或紧张时出现 |

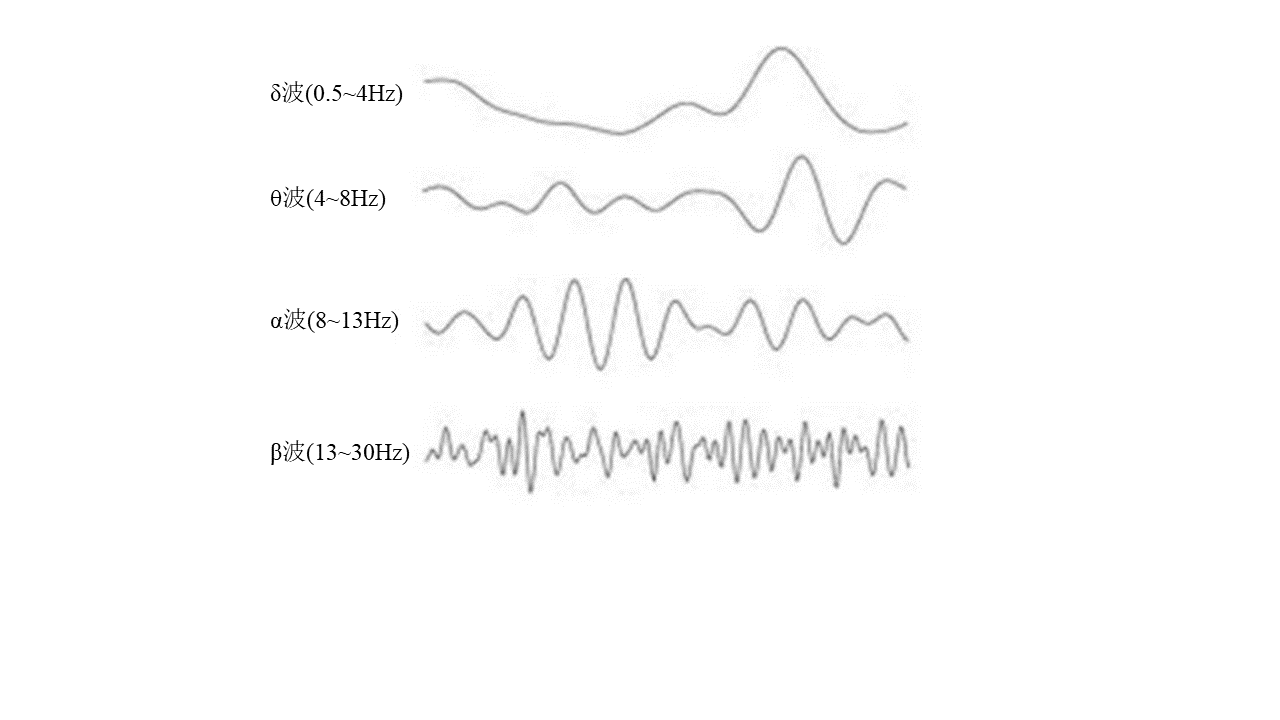


图2·1 脑电信号四种典型波形图

2.1.2、癫痫脑电信号分类

不同于正常状态下的脑电信号，癫痫脑电信号是当大脑神经元异常放电时出现，

表现出无规律性、发复性和急促性等特性。目前癫痫的发作大致可以分为三类：局灶性癫痫发作、全身性癫痫发作和位置的发作性癫痫。当癫痫发作时会出现一系列在与正常脑电信号波形显著不同的特征波，比如尖波、棘波、棘-慢复合波和尖-慢复合波等。图2·2对这些典型的癫痫特征波进行了展示，而表2·2则对图2·2中前四种典型的典型特征波进行了介绍。

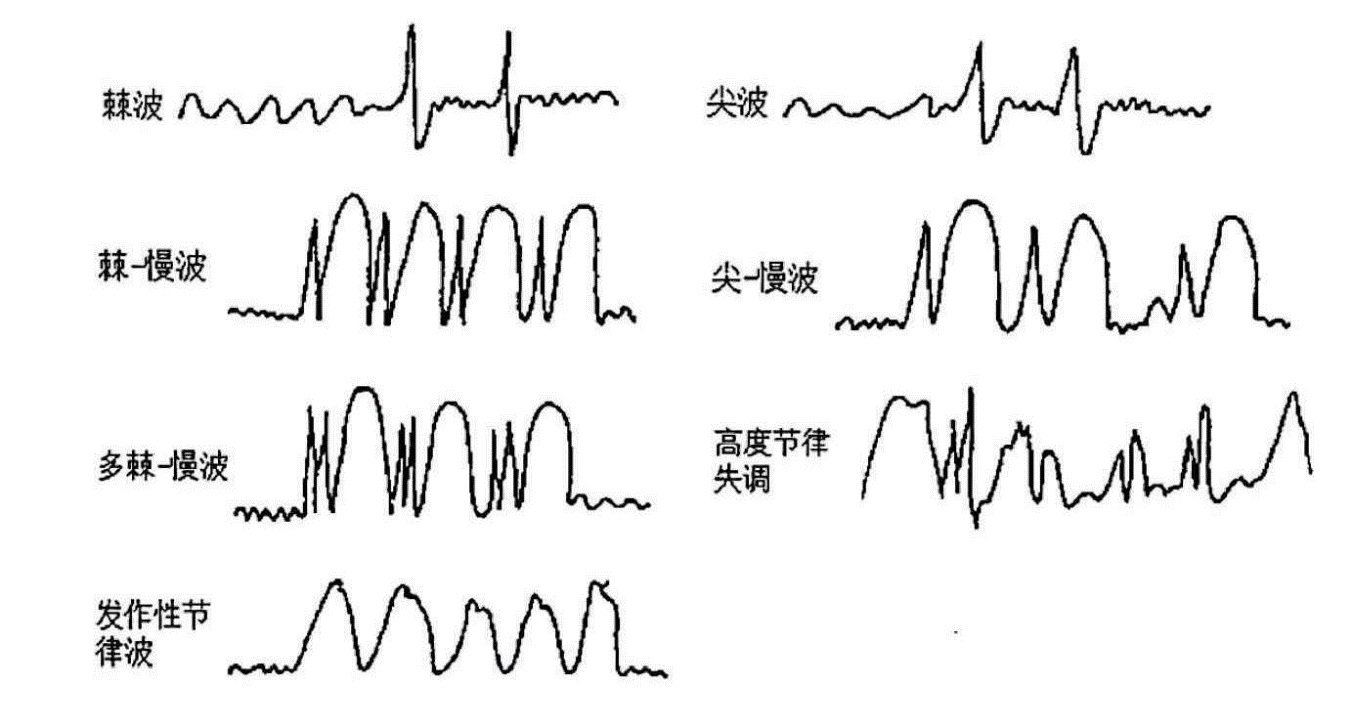


图2·2 几种典型的癫痫特征波

表2·2 四种典型癫痫特征波介绍

|  |  |  |  |
| --- | --- | --- | --- |
| 特征波类型 | 频率 | 幅值 | 特征 |
| 尖波 | 12.5 Hz  以内 | 多在100以上 | 形似棘状，是痫样波形最具特征性的表现之一 |
| 棘波 | 5~12.5Hz | 100~200 | 尖波呈尖峰样，是常见的痫样特征波之一 |
| 棘-慢复合波 | 2~12.5Hz | 150~300 | 由一个棘波与一个慢波组成，波幅一般较高 |
| 尖-慢复合波 | 1.5~2.5Hz/  4~6Hz | 100以上 | 由一个尖波与一慢波组成，可连续性或弥漫性出现出现形式多样 |

第2.2节 脑电信号的采集

目前脑电图(Electroencephalography,EEG)是脑电信号的采集主要方式，由记录脑电图时电极不同的放置位置，可以将脑电图分为三类：头皮脑电图、皮层皮质脑电图和深层脑电图。

通过上文中对脑电信号的介绍，我们可以了解到它的幅值一般都比较低，通常在几百左右，因此如果脑电采集过程中受到外界环境的干扰，很容易对后续的脑电信号处理过程形成极大的阻碍，因此在采集过程中要尽量保证采集环境中没有较强的电场，仪器设备要接地，同时要保证被采集者保持一种平和的状态，不能有剧烈的情绪起伏。

现如今脑电信号的采集主要依照“国际脑电图学会”公布的10-20系统，该系统规定了电极在大脑上16或19个特定的位置，以俯视图中大脑中心线为界，采取左奇右偶的标号方式，具体的位置如图2·3所示，表2·3对图2·3中电极分布信息进行了详细的解释。

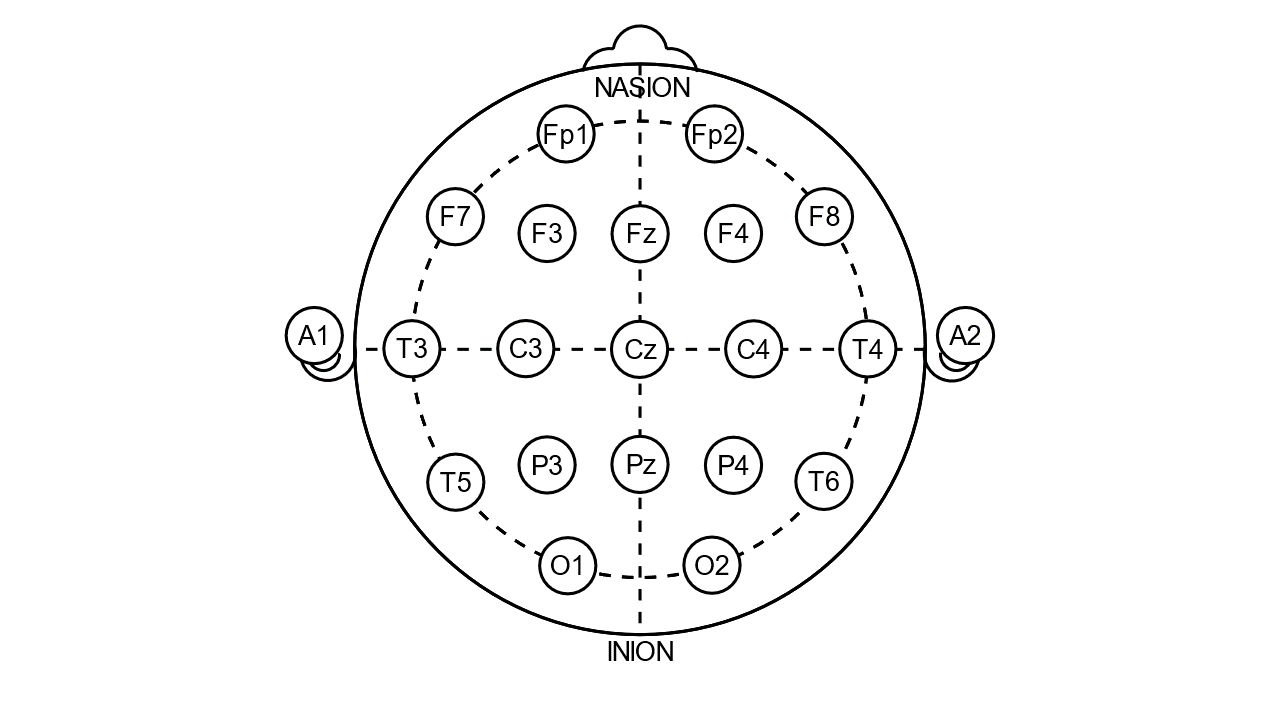


图2·3 10-20国际标准电极方案

表2·3 10-20系统电极分布信息

|  |  |  |
| --- | --- | --- |
| 电极位置 | 英文名称 | 电极标号 |
| 前额 | Frontal Pole | Fp1、Fp2 |
| 额 | Frontal | F3、Fz、F4 |
| 中央 | Central | C3、Cz、C4 |
| 顶 | Parietal | P3、Pz、P4 |
| 枕 | Occipital | O1、O2 |
| 侧额 | Inferior Frontal | F7、F8 |
| 颞 | Temporal | T3、T4 |
| 后颞 | Posterior Temporal | T5、T6 |

第3章 基于离散小波变换的脑电特征提取

第3.1节 脑电信号数据来源及介绍

本研究中所采用的EEG脑电数据集共有两个，分别是来源于波恩大学癫痫脑电数据库的数据集和来源于波士顿儿童医院的CHB-MIT头皮脑电图数据库，这两个数据集均是开源的，并且均可以直接在互联网上下载到，这两个数据集的下载地址分别为：<http://qcsdn.com/q/a/3118833.html>和<http://physionet.mit.edu/physiobank/database/chbmit/>。

3.1.1、波恩大学癫痫脑电数据集

对于来自于波恩大学癫痫脑电数据库的数据集，其一共由五组数据构成，分别用英文字母A～E表示。这五组数据均为从志愿者身上采取得到的，而采集方式则是使用上文中提及的国际标准的10-20系统来安置电极，对志愿者的脑电数据进行记录。这五组数据可以分为两部分：从五位健康志愿者身上采集到的两组数据(包括A、B两组)和从五位患有癫痫的志愿者身上采集到的三组数据(包括C、D和E三组)。而这两部分又分别是志愿者处于不同的状态下的时候，使用脑电图仪在志愿者大脑的不同部位记录采集到的，具体介绍详见下方表3·1所示。在A～E五组数据中，每一组数据中又分别均有100条单通道脑电记录，格式则为txt，其采样频率为173.61Hz，每一条脑电记录持续的时间则为23.6s，共有4097个数据点。值得注意的是，这些数据已经提前经过了带通滤波处理(其中滤波通带设置为0.53Hz~40Hz)，也已经经由专业人员进行人工伪迹的去除。

表3·1 波恩大学癫痫脑电数据集概述

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 组别 | A | B | C | D | E |
| 志愿者  身体状况 | 健康 | 健康 | 癫痫 | 癫痫 | 癫痫 |
| 志愿者状态 | 睁眼 | 闭眼 | 发作间期 | 发作间期 | 发作期 |
| 脑电信号  采集部位 | 大脑  表面 | 大脑  表面 | 与病灶相对脑半球的海马区 | 病灶区 | 病灶区 |

在matlab2018开发环境之下，针对A~E这五组数据中每一组的第一条脑电记录，进行脑电数据的读取与图像的绘制，读取之后得到的脑电信号时序图如图3·1所示。图3·1极其明确地展示了各组志愿者的脑电信号波形是怎么样的，通过观察这一张图可以发现A组和B组这两组的波形相似度较高，而C、D和E这三组则各不相同，和正常志愿者相比，可以发现C、D和E这三组的脑电信号时序图中存在着许多异常的波形，与正常人的脑电信号差异较大，这也为后续的脑电信号研究分类工作奠定了成功的基础。

通过研究图3·1之后，本研究将对癫痫脑电信号的分类工作共分为四种情况，其中第一种情况是针对A/E两组进行二分类，第二种情况是针对B/E两组进行二分类，第三种情况是针对C/E两组进行二分类，第四种情况则是对D/E两组进行二分类。

具体如表3·2所示。

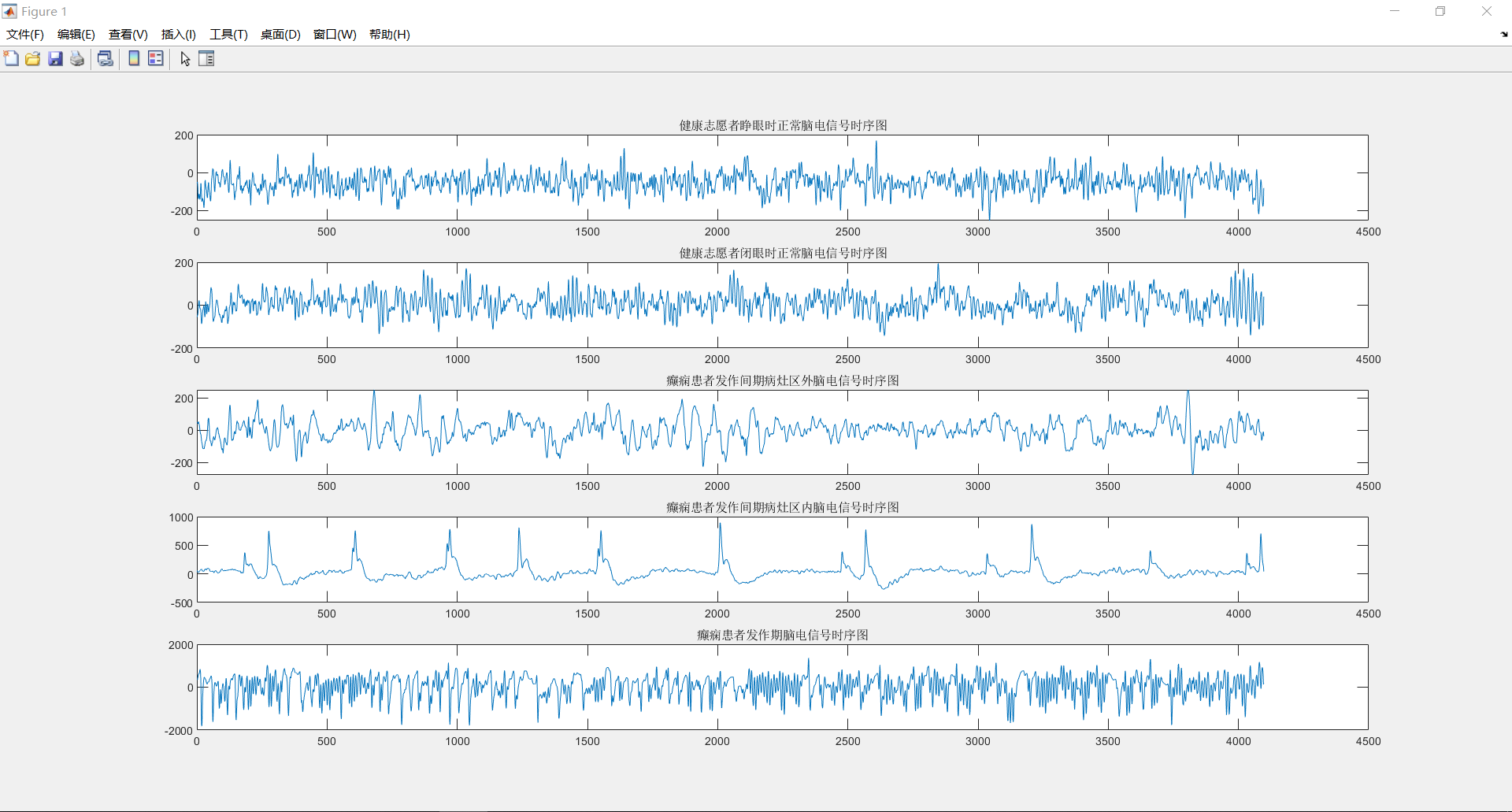


图3·1 A~E五组志愿者脑电信号时序图

表3·2 波恩大学癫痫脑电数据集四种分类任务

|  |  |  |
| --- | --- | --- |
| 情况 | 志愿者临床状态 | EEG子集 |
| 1 | 正常 vs 发作期 | A vs E |
| 2 | 正常 vs 发作期 | B vs E |
| 3 | 发作间期 vs 发作期 | C vs E |
| 4 | 发作间期 vs 发作期 | D vs E |

3.1.2、CHB-MIT头皮脑电图数据库脑电数据集

对于来源于波士顿儿童医院的CHB-MIT头皮脑电图数据库的EEG脑电数据集，其则是由儿科患者的脑电图记录所构成，这些儿科患者均罹患有难治疗性癫痫。据了解，该数据库中共包含有22位受试者的23个病例记录(chb01~chb24，其中病例chb01在1年半之后重新进行了脑电图记录，并记为了chb21；而病例chb24则是在2010年12月份的时候添加到了该数据库当中，该病例的性别和年龄并未记录在其中，而是仅仅记录了该病例的癫痫发作情况)，该22位受试者则包括5例男性和17例女性，其年龄分别为3岁到22岁和1岁半到19岁。

该数据库中的数据和波恩大学癫痫脑电数据库的数据集一样，均根据国际标准的10-20系统来安置电极，对患者的脑电数据进行记录，这些所有采集到的数据均为以256Hz的采样率以16位分辨率进行采样。其中每一个案例均包含有若干个.edf数据文件(从最少的9个到最多的42个不等)，其中EDF(European Data Format)则为一种在生理信号(如脑电信号和心电信号等)领域中常用的一种标准数据格式，它可以对多通道数据进行存储，并且能够使得每一个通道的信号有不一样的采样频率。在大多数情况之下，每个.edf文件均正好包含有病患长达一个小时的数字化EEG信号，只有对少数的数据文件，其才会记录四个小时的EEG脑电信号，例如病例chb01有chb01\_01.edf~chb01\_46.edf共46个数据文件，每一个数据文件中均记录了病例chb01一个小时的EEG信号。这些数据文件中，大多数有23个脑电信号通道，只有少部分有24个或者26个通道。

本研究在matlab2018环境下使用eeglab工具箱对病例chb01下的数据文件chb01\_03进行了读取并对脑电图像进行了绘制,具体如图3·2所示。

在本研究中，采用该数据集的目的是为了对针对波恩大学癫痫脑电数据库的数据集提出的癫痫脑电信号分类方法进行验证以得到具有普遍性的结论，因此在本研究中仅选取部分病例的部分发作间期和发作期的脑电数据进行研究。经过研究，本研究选取病例chb01和chb03的FT9-FT10通道，并选取若干个记录有癫痫发作的数据文件中的发作间期和发作期各一部分时间(每次记录起始时间均为0s)来进行研究，详细情况如表3·3和表3·4所示(以下选取的数据文件，数据记录时间均为1个小时)。

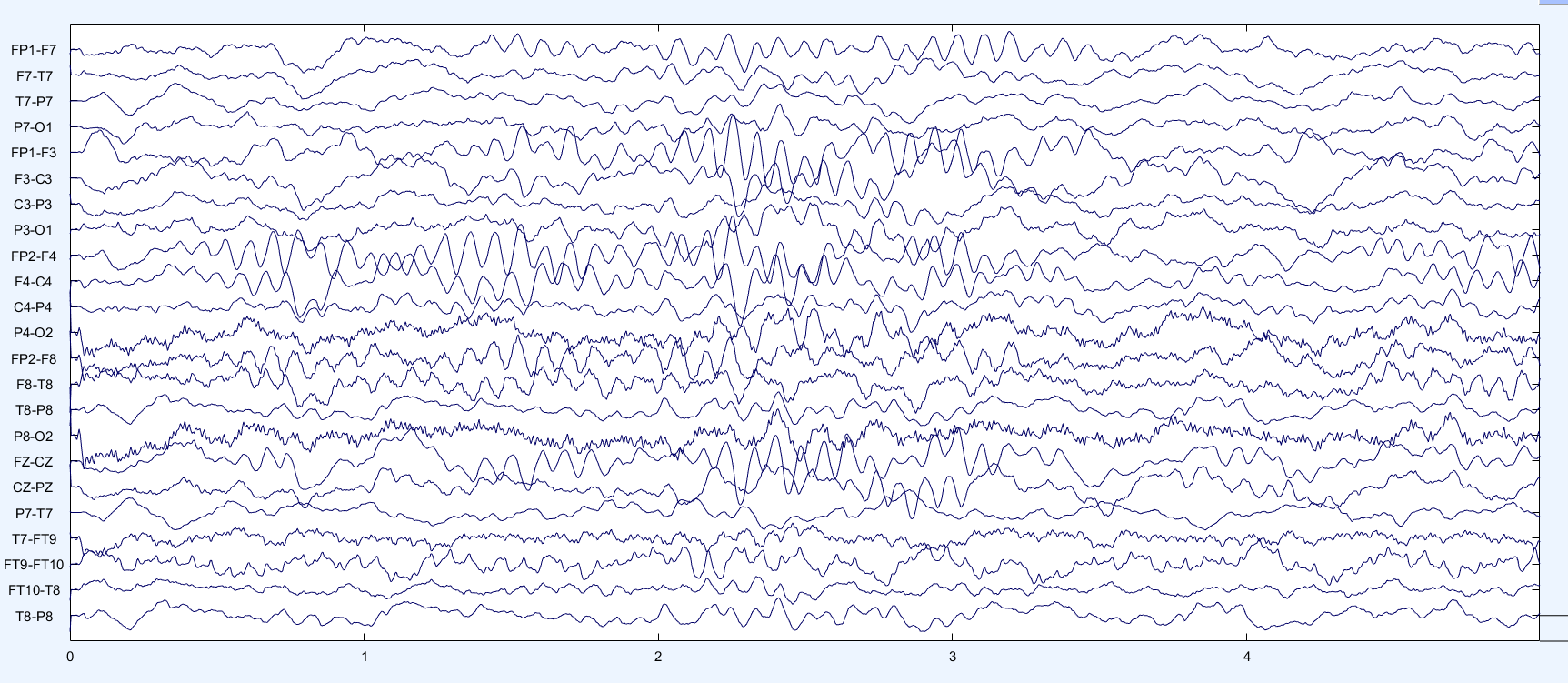


图3·2 chb01\_03脑电信号时序图

表3·3 病例chb01选取的数据文件及时间

|  |  |  |
| --- | --- | --- |
| 病例数据文件 | 选取发作间期时间(s) | 选择发作期时间(s) |
| chb01\_03 | 100~130 | 3001~3031 |
| chb01\_15 | 1000~1030 | 1737~1767 |
| chb01\_16 | 500~530 | 1025~1055 |
| chb01\_18 | 200~230 | 1750~1780 |
| chb01\_21 | 500~560 | 342~402 |
| chb01\_26 | 100~160 | 1882~1942 |
| 共计 | 240 | 240 |

表3·4 病例chb03选取的数据文件及时间

|  |  |  |
| --- | --- | --- |
| 病例数据文件 | 选取发作间期时间(s) | 选择发作期时间(s) |
| chb03\_01 | 2000~2040 | 368~408 |
| chb03\_02 | 150~180 | 748~778 |
| chb03\_03 | 2000~2030 | 452~482 |
| chb03\_04 | 300~340 | 2168~2208 |
| chb03\_34 | 520~560 | 1985~2025 |
| chb03\_35 | 1000~1030 | 2608~2638 |
| chb03\_36 | 100~130 | 1736~1766 |
| 共计 | 240 | 240 |

第3.2节 (癫痫)脑电信号预处理

3.2.1、波恩大学癫痫脑电数据集

对于该数据集，由于该数据集已经经过了通带为0.53Hz~40Hz的带通滤波和由专业人员进行的人工伪迹处理，因此本研究中不再进行额外的数据操作。

对于A~E五组数据中每一组中的每一条单通道脑电记录，本文按照时间顺序，以1024个数据点为一小组，那么一条单通道脑电记录就可以划分为四个小组，A~E这五组脑电数据则总共可以划分为2000个小组。在接下来的分析研究中，本文以一小组为单位进行分析操作。

3.2.2、CHB-MIT头皮脑电图数据库脑电数据集

对于该数据集，由于其还未经过任何处理，因此在通过使用eeglab工具箱将选中的病例chb01和chb03的数据文件读取之时，直接按照表3·3和表3·4中所选定的时间范围进行数据的读取。在进行了数据读取操作之后，本研究通过eeglab工具箱自带的数字滤波器对这些数据文件进行通带为0.1Hz~60Hz的带通滤波操作。通过这一操作，本研究达到了对所选数据文件滤除60Hz工频干扰的目的。

在下一步中，本研究直接将除了FT9-FT10这一通道外的其他通道清除，经过这两步的操作之后可以得到若干个期望的格式为edf的单通道数据文件，这些数据文件所包含的时间范围同样是表3·3和表3·4中需要的。

之后本研究以chb01\_03的发作间期和发作期的数据为例，通过使用EDFbrowser软件对这两部分的数据进行了读取，读取得到的脑电信号时序图分别如图3·3和图3·4所示。通过观察，显然可以发现这两张时序图的特征波和脑电信号幅值存在着较大的差异，这与波恩大学癫痫脑电数据库的数据集经由matlab读取后得到的D和E两组的癫痫脑电信号时序图有异曲同工之处。

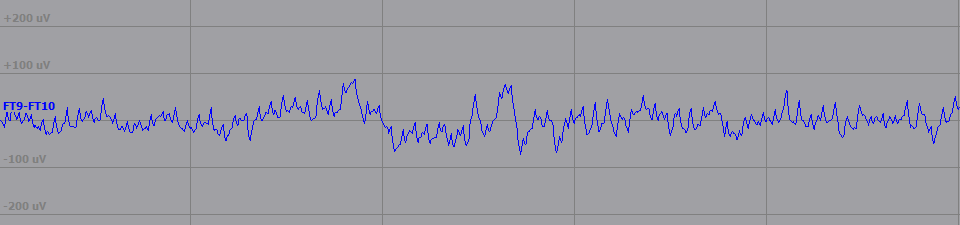


图3·3 chb01\_03发作间期脑电信号时序图

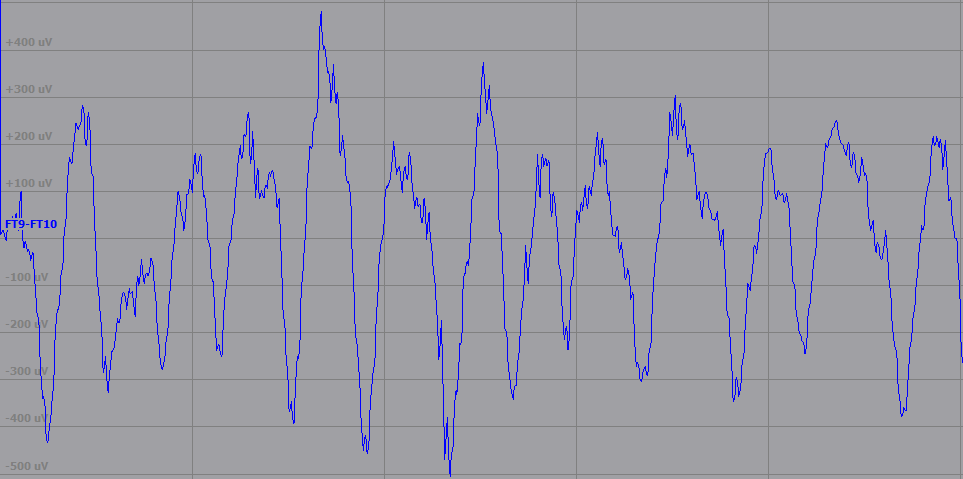


图3·4 chb01\_\_03发作期脑电信号时序图

由于matlab或者python中处理数据时通常用到的数据格式是txt格式或者是csv格式而并非本文上文中提及的edf格式，因此还需要对数据文件的数据格式进行转换。本研究中通过使用EDFbrowser软件，实现了将数据由edf格式到txt格式的转换，这样就可以直接和波恩大学癫痫脑电数据库的数据集进行相同的操作，十分方便。

在将这些数据文件的格式由edf转换为txt之后，对于病例chb01(病例chb03同理，也进行相应的操作)，本研究在其发作间期和发作期均各随机挑选了240s的时间，共计480s，不同于将波恩大学癫痫脑电数据集以1024个数据点分为一个小组，此处以256个数据点即1s的时间为单位进行分组，因此病例chb01的脑电数据共可以分为480个小组，在接下来的研究分析中，均以小组为单位进行操作。

第3.3节 小波变换理论介绍

小波变换与傅里叶变换的关系

小波变换(wavelet transform)是由傅里叶变换(Fourier Transform)发展改良而来的。根据以往学习过的知识， 我们可以知道傅里叶变换是以一个无限长的三角函数作为它的基函数，进而把一个在时域上不属于周期信号的一个连续信号转换为一个在频域上非周期的一个连续的信号，其中傅里叶变换的公式如所示。但是对于像生物医学信号(例如本文中主要讨论的脑电信号)这种自然界大量存在着的有着非平稳、非线性特性的信号，傅里叶变换处理的时候就会出现无法忽视的缺陷，例如对于一段未知的信号，傅里叶变换只可以获取到该信号大体上是由哪一些频率成分构成的，然而傅里叶变换对于这一段未知信号中的各个组成成分出现的时间并不能确定，故而可能会出现这样一个现象，两个信号的频谱图可能十分相似，但是这两个信号在时域上却可能相差甚远。



因此之后就出现了对傅里叶变换进行改进的方法-短时傅里叶变换(Short-time Fourier Transform，STFT)。短时傅里叶变换对于傅里叶变换改进之处在于“加窗”，对于一个时域信号，短时傅里叶变换将其分解为大量的小过程，而这些小过程同时又是等长度的，由于对该时域信号进行了大量的分解，因此可以近似认为每一个小过程都是平稳的，接下来再将该时域信号通过加的窗分为一段一段的子部分，针对这些分好段的时域信号再进行傅里叶变换，通过这种方法就可以明确频率成分随时间变换的情况，其中短时傅里叶变换的公式如所示，其中即为窗口函数。但是对于窗口函数而言，一旦其选定之后，窗口的形状便被固定，切断了窗口框的宽度与频率之间的联系；此外施加的小窗在太宽或者是太窄的情况下都会出现不同的问题，例如当所加窗的框太窄时，自然而然便会导致信号过少，致使得到的频域信号频率分辨率差；而当所加窗的框太宽的时候便会导致时域信号的时间分辨率很低。



而小波变换则是克服了傅里叶变换和短时傅里叶变换的缺点，将傅里叶变换中的基函数(一个无限长的三角函数)替换为了一个有限长度会衰减的小波基函数，该小波基函数可以对窗口进行调变，例如当频率增高的时候，时间窗口的宽度就会相应地变窄，进而达到提高频率分辨率的目的；而当频率降低的时候，时间窗口的宽度就会相应地便宽，已达到提高时间分辨率的目的。因此可以看出小波变换对于非稳定信号的时频分析处理有着极大的优势。

小波变换包括连续小波变换(Continuous Wavelet Transform，CWT)和离散小波变换(Discrete Wavelet Transform，DWT)

CWT公式，其中称之为小波母函数



DWT

框图 离散小波的多层分解过程



第3.4节 由离散小波变换进行脑电特征提取

小波基的选择

上文中已经对离散小波变换进行了详细的介绍，小波变换就是将一个信号分解成为一系列的小波，而其中非常重要的一个步骤就是对小波基的选取。

但是小波基的种类十分繁多，比如常见的Symlet(symN)小波、Daubechies(dbN)小波和Coiflet(coifN)小波等，本研究对这些小波中若干个典型的小波基进行了选取，其图像分别如图3·5、图3·6和图3·7所示。当对小波基进行不同的选择之后，在对一段相同的信号进行分析处理的时候，最终得到的处理效果也可能迥然不同。

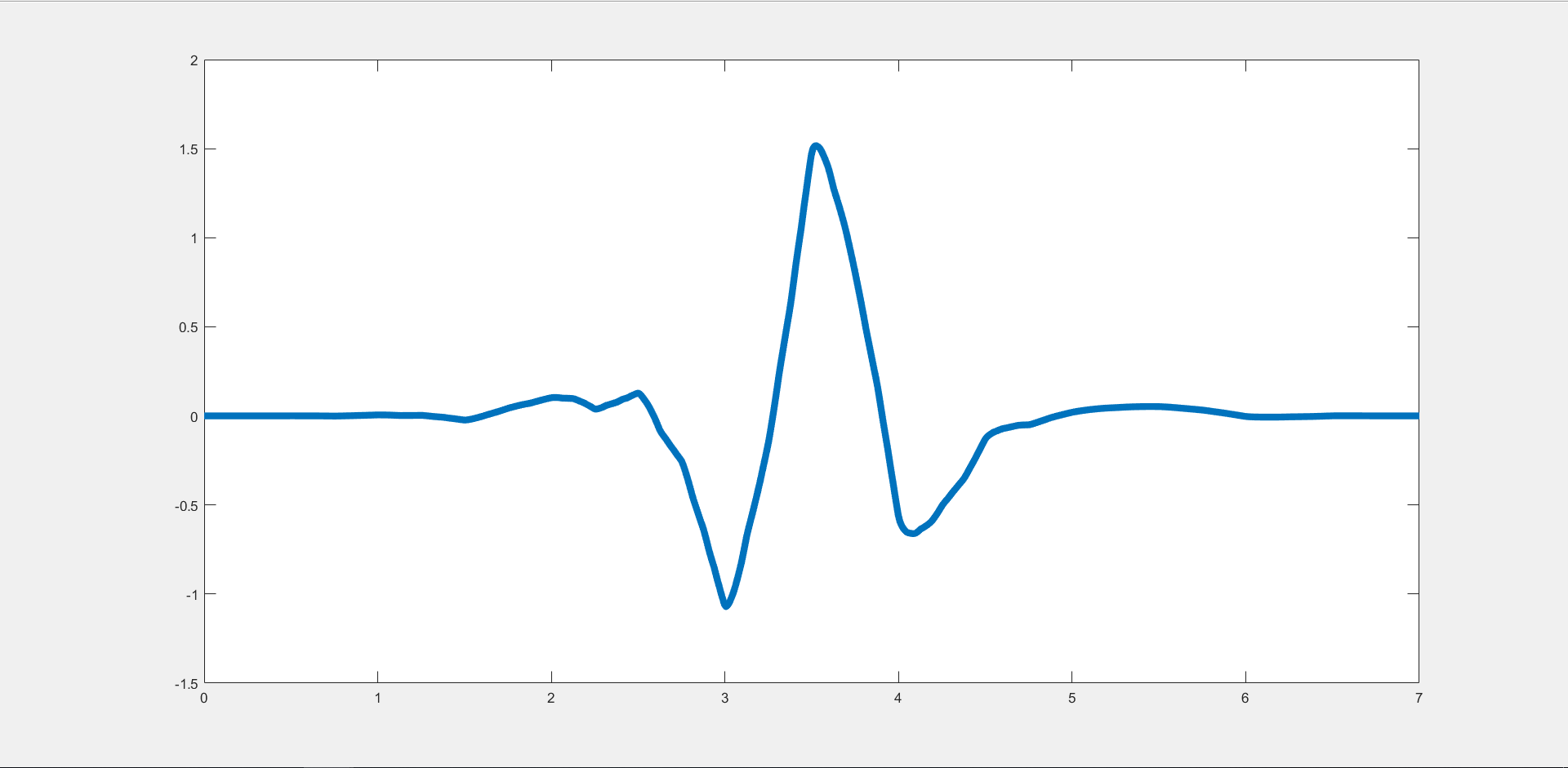


图3·5 Sym4小波基函数图像

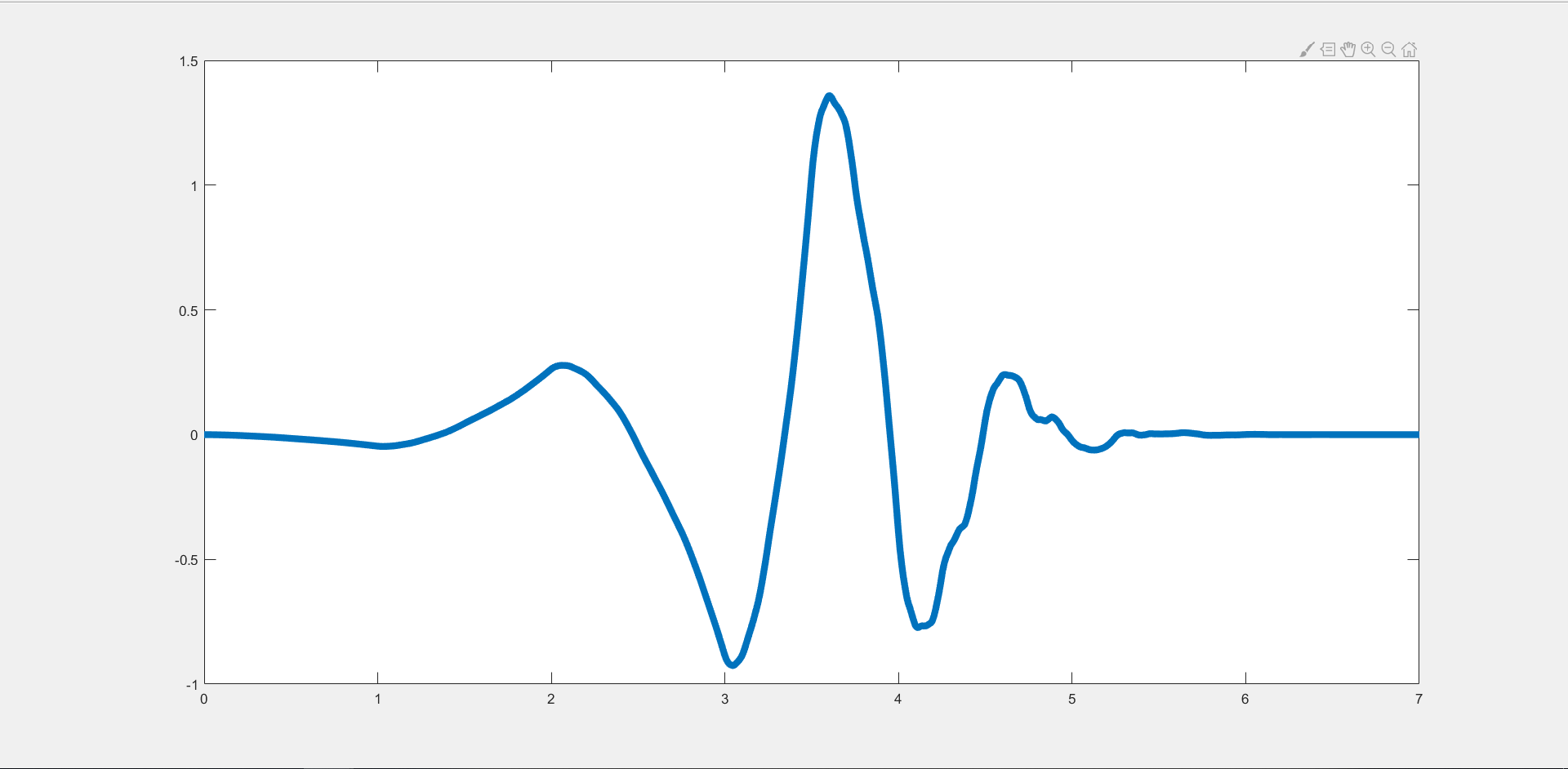


图3·6 Db4小波基函数图像

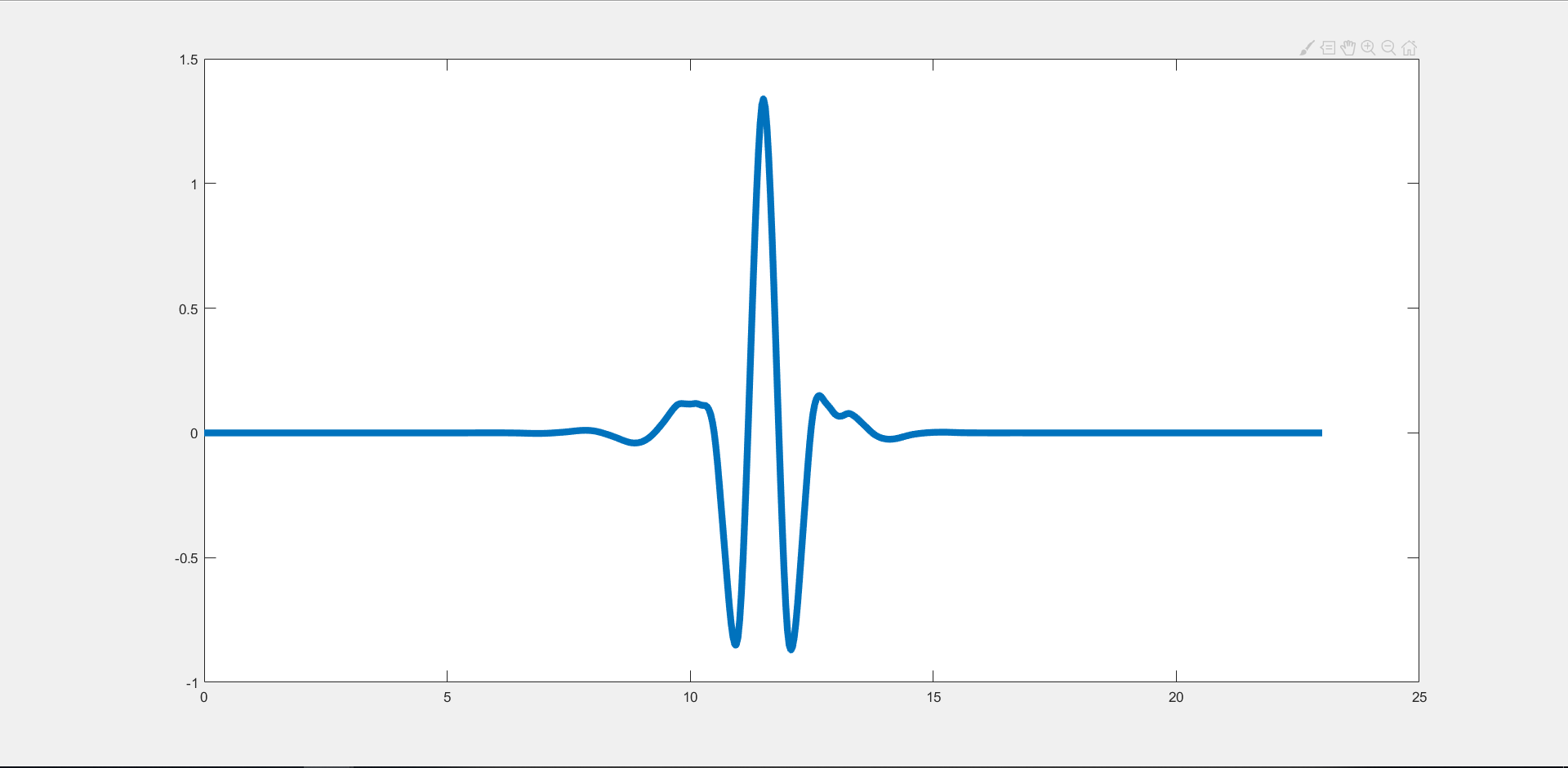


图3·7 Coif4小波基函数图像

小波基的选取要从正交性、支撑长度、对称性和正则性等几个方面的因素进行考虑，同时还要考虑所要处理的信号和小波基函数波形的相似性，如果两者的相似性越高，那么经过离散小波变换之后得到的小波系数就会越大，而如果相似性越低，那么小波系数就会越小。本研究应该尽量选取与癫痫脑电信号最为相似的小波基，综合以上所有的因素考虑之后，在图3·5、图3·6和图3·7这三张图像之中选择了’db4’小波基函数进行后续的离散小波变换。

癫痫脑电信号的特征提取

本研究已经确定要使用’db4’为小波基来进行5层的离散小波变换(DWT)，进而对脑电信号进行处理提取特征值，经过离散小波变换我们便可以得到待处理信号在不同频率范围上的近似系数和细节系数。

根据对波恩大学癫痫脑电数据库的数据集描述可知，采集该数据集时的采样频率为173.61Hz，由香农采样定理(Shannon Sampling Theorem)的相关描述可知，波恩大学癫痫脑电数据库的数据集的最大有用频率应该是采样频率的一半。那么可进一步得出由离散小波变换进行分解的频率带应该是0~86.81Hz。具体的分解过程如图3·8所示，对图3·8中提及的符号的详细描述如表3·5所示，cD1和cA1等与表3·5中的cA5和cD5类似，均为相应层数分解后得到的小波系数。

离散小波变换分解的过程如下：首先，在第一层分解中，输入的整个一维EEG脑电信号同时经过一个低通滤波器和一个高通滤波器，这两个滤波器将输入的EEG信号的频率分为了相等的两部分得到了详细小波系数cD1(频带为0~43.4Hz)和cA1(43.4~86.8Hz)，经过滤波之后，信号的时间分辨率减半，而信号的频率分辨率变为原来的两倍；然后cA1继续进行第二层的分解，同样也是经过一个低通滤波器和一个高通滤波器，进而得到cD2(频带为0~21.7Hz)和cA2(频带为21.7~43.4Hz)。该过程重复进行五次之后，可以得到cA5、cD1、cD2、 cD3 、cD4和 cD5六个小波系数，而这六个小波系数覆盖了0~86.8Hz的整个频带。按照这一流程，本研究将上文中经过分组的2000小组脑电信号记录均分别进行了五层离散小波变换，这样便可得到各为2000组的cA5、cD1、cD2、 cD3 、cD4和 cD5小波系数。

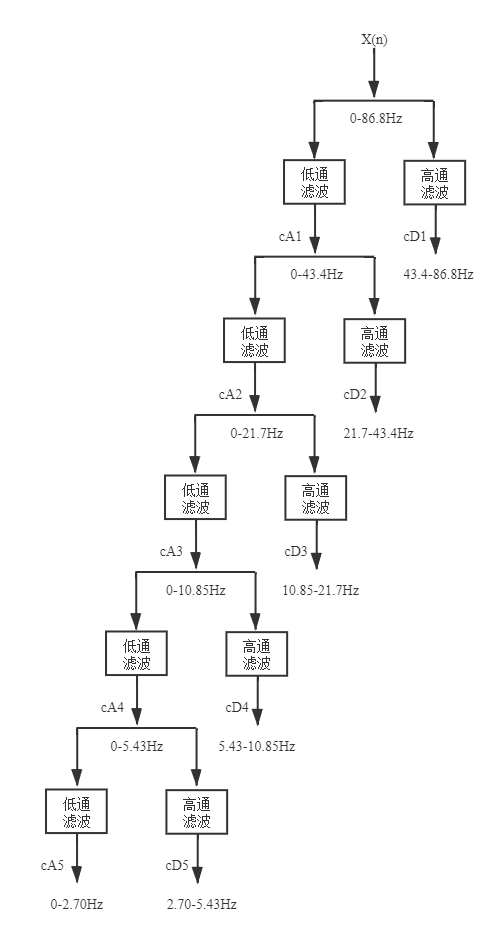


图3·8 基于离散小波变换的EEG信号分解过程

表3·5 离散小波变换分解过程符号解释

|  |  |
| --- | --- |
| 名称 | 描述 |
| X(n) | 一维离散输入信号 |
| cA5 | 第五层分解的近似小波系数 |
| cD5 | 第五层分解的近似小波系数 |

本研究以波恩大学癫痫脑电数据集中A组脑电数据中的第一条脑电记录的第一个小组(上文提及的依时间顺序以1024个数据点为一组)为例，将其脑电时序图与经过离散小波变换之后得到的六个小波系数进行了绘制，如图3·9中所示。

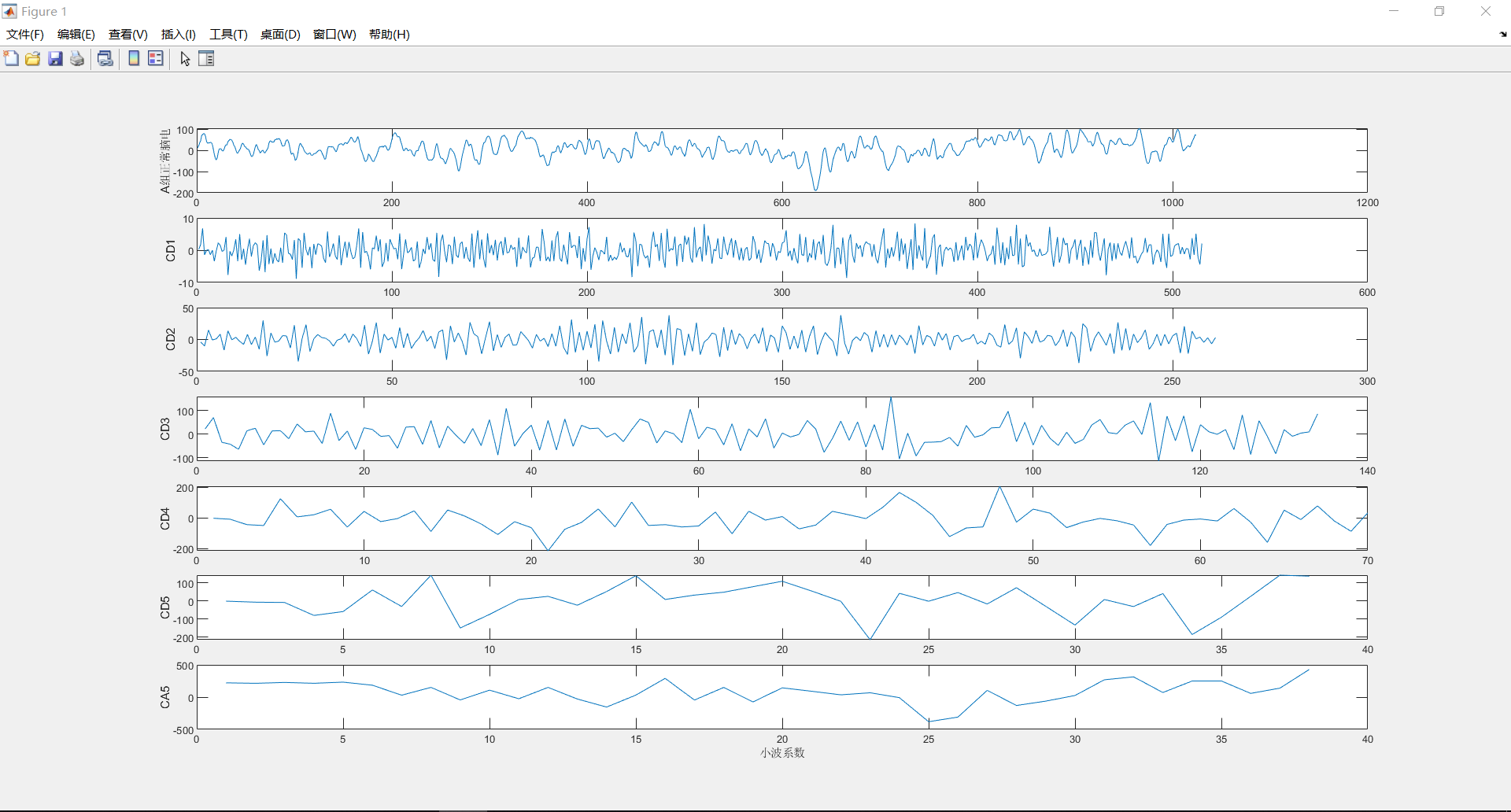


图3·9 A组脑电信号小波系数图像

特征选择与提取

在肖文卿等人的研究中，（改为引用文献，文献18）由于当小老鼠在癫痫发作时脑电信号的标准差是会低于小老鼠在正常生理情况下的脑电信号标准差，因此他们对小老鼠在两种不同状态之下的脑电信号自身的标准差进行了提取，并将其作为分类器的输入来对正常小鼠和癫痫小鼠进行了分类。

本研究参考了这一思想，对波恩大学癫痫脑电数据集中经过分组得到的2000组原始脑电信号记录和各为2000组的cD1、cD2、 cD3 、cD4和 cD5小波系数计算其标准差，这样便可得到共计6组(原始脑电信号、cD1、cD2、 cD3 、cD4和 cD5)标准差，每一组中则有2000个标准差。

此处以表3·2中的第一种情况即A vs E两组进行二分类为例，根据计算得到的原始脑电信号、cD1、cD2、 cD3 、cD4和 cD5的标准差绘制了图3·10所示的箱型图。经过对比可以发现，正常志愿者和正处于癫痫发作期的志愿者的自身脑电信号标准差值和cD1~cD5这五组小波系数的标准差值均有着十分明显的差别，而且处于癫痫发作期志愿者自身的EEG与cD1~cD5这五组小波系数的标准差值均大于正常志愿者的标准差值。所以根据图3·10，本研究选取志愿者自身的脑电信号标准差值与cD1~cD5这五组小波系数的标准差值作为接下来研究中对癫痫脑电信号进行分类的特征。

对于表3·2中的其他三种情况，同样进行上述类似的操作，并且分别经过与上述类似的六种方差的对比，也进行了相同的选择，并将其分别作为各自进行分类时的特征。

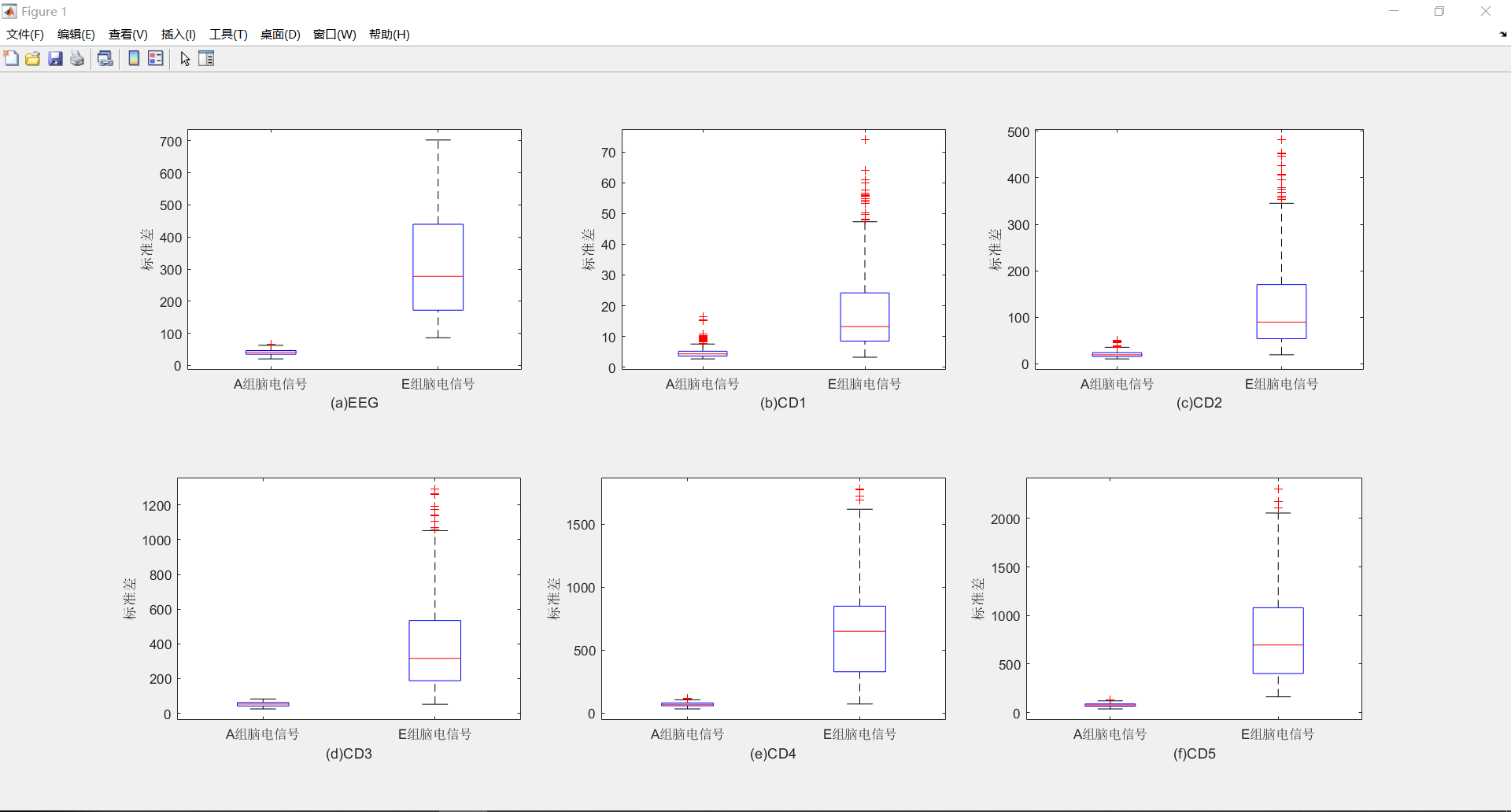


图3·10 AE两组标准差箱型图对比

对于经过预处理和时间选择的波士顿儿童医院的CHB-MIT头皮脑电图数据库的EEG脑电数据集，与上述对波恩大学癫痫脑电数据集进行的操作一样，也分别对病例chb01和chb03的各480组的脑电数据中的每一组也别进行了五层的离散小波变换，并且相应得到了cA5、cD1、cD2、 cD3 、cD4和 cD5这六组小波系数，也计算得到了原始脑电信号与这六组小波系数的标准差。在此处，本研究以病例chb01为例，同上述内容进行相同的操作，绘制了病例chb01在发作间期和发作期时的原始脑电、cD1、cD2、 cD3 、cD4和 cD5这五组小波系数的标准差相对比的箱型图，具体如图3·11所示。

观察图3·11可以发现，病例chb01当分别处于发作间期和发作期时，这两个阶段的自身脑电信号标准差值和cD1~cD5这五组小波系数的标准差值均有着十分明显的差别，而且其处于发作期的自身的EEG与cD1~cD5这五组小波系数的标准差值均大于其处于发作间期时的标准差值，而这一结论与对图3·10的分析一致。因此本研究选取病例chb01自身的脑电信号标准差值与cD1~cD5这五组小波系数的标准差值作为接下来研究中对癫痫脑电信号进行分类的特征。对于病例chb03的脑电信号数据也进行与病例chb01类似的操作，此处不再赘述。

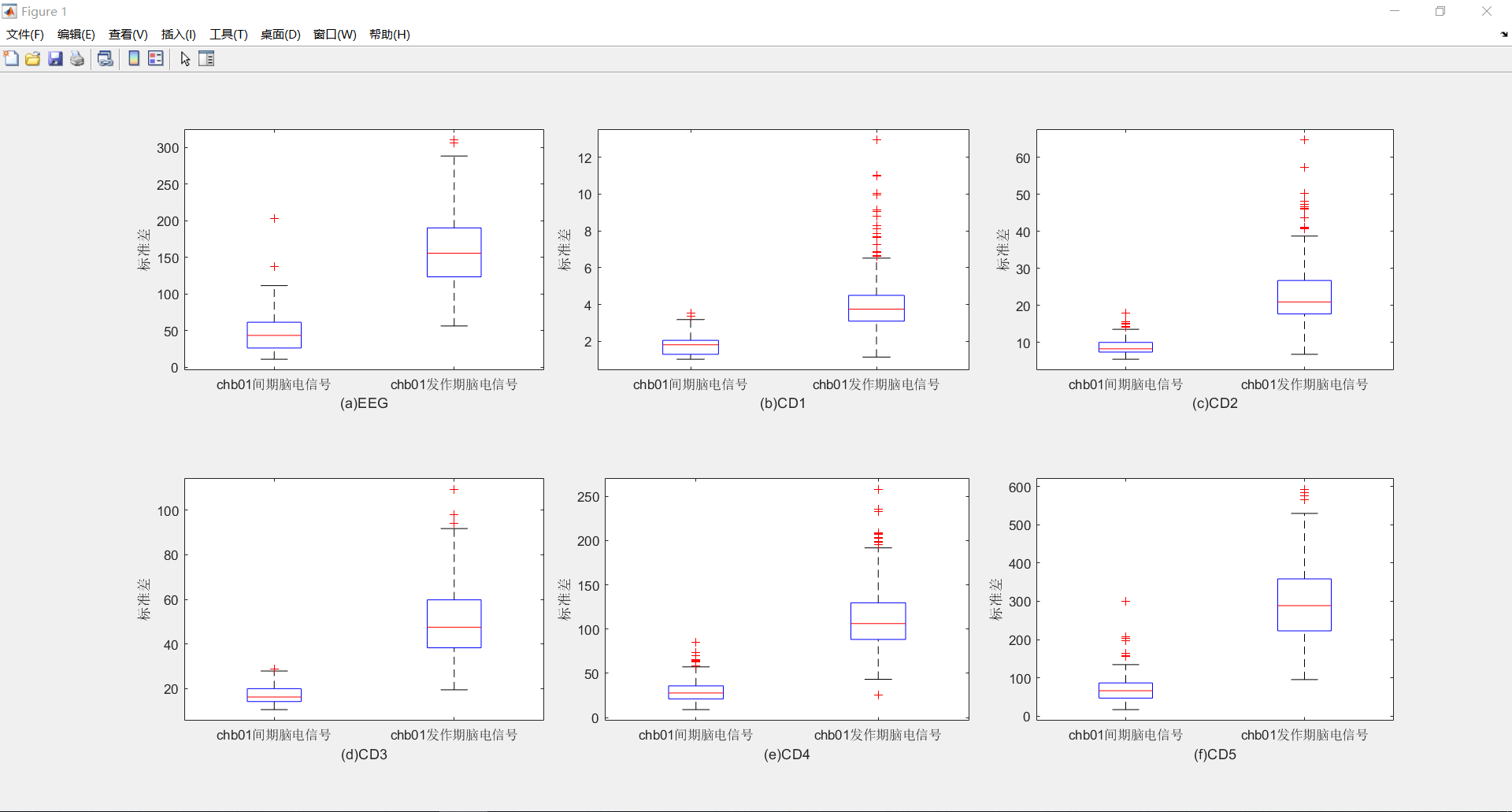


图3·11 病例chb01发作间期和发作期脑电信号方差箱型图对比

样本熵(Sample Entropy)最早由Richman等人提出，它从数据长度独立性和相对无故障的实现两个方面对近似熵(Approximate Entropy)进行了改进，它经常用来评估生理时间序列信号的复杂性，诊断疾病状态。根据Cheol Seung Yoo等人的研究，他们某某某发现在人体在癫痫发作过程中，脑电信号的样本熵值会下降，在癫痫发作终止时达到最小值，癫痫发作终止后，样本熵迅速恢复到癫痫发作前的数值。根据这一结论，本研究将脑电信号的样本熵这一非线性特征选择作为除去标准差之外的对脑电信号分类的另一个特征。

假定一个时间序列为，其时间间隔是一个固定的值，其中N为该序列中包含数据的个数。定义一组向量序列，该序列的维数是m，而，其中。定义距离函数为切比雪夫距离(Chebyshev distance)，即



定义样本熵为：



其中为阈值，，而则是对每一个值，均对的数目进行了计算与统计，并将该数目记作，对于，有：



而和也可以经过和上述类似的操作得到，只是在构成向量序列的时候与有所不同，在此处则是将原来的m维增加到了m+1维，即与，而其他的步骤与上面的步骤一样。

从中对样本熵的定义可以看出，的值始终小于或者等于的值，因此将始终为零或者正值。在研究中通常采用的参数值，其中代表信号的标准偏差。在本研究中，采用的参数。

同样与上文中对波恩大学癫痫脑电数据集中经过分组得到的2000组原始脑电信号记录和五组各为2000组的小波系数计算标准差类似，此处可计算得到共计6组(原始脑电信号、cD1、cD2、 cD3 、cD4和 cD5)样本熵，每一组中则有2000个样本熵。此处同样以表3·2中的第一种情况即A vs E两组进行二分类为例，根据计算得到的原始脑电信号、cD1、cD2、 cD3 、cD4和 cD5的样本熵绘制了图3·10所示的箱型图。

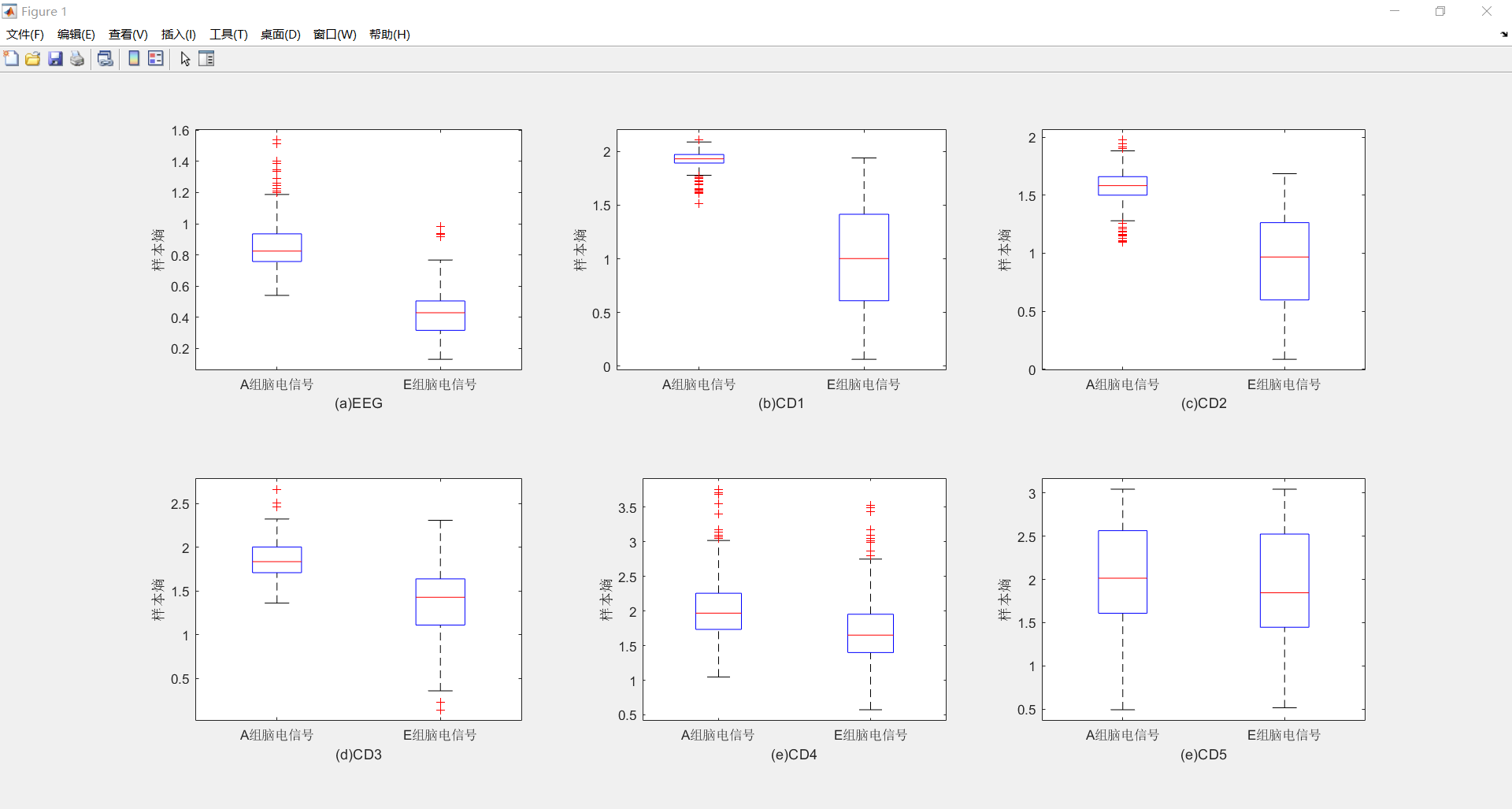


图3·12 AE两组样本熵箱型图对比

观察过图3·12可发现，与上文中对标准差作为特征时的选择不同，此处正常志愿者和正处于癫痫发作期的志愿者只有自身脑电信号样本熵值和cD1~cD3这四组小波系数的样本熵值均有着十分明显的差别，而且处于癫痫发作期志愿者自身的EEG与cD1~cD3这五组小波系数的样本熵值均小于正常志愿者的样本熵值，这与Cheol Seung Yoo等人的研究结论一致。所以本研究选取志愿者自身的脑电信号与cD1~cD3这四组数据的样本熵值作为接下来研究中对癫痫脑电信号进行分类的特征，其余三种情况的二分类也进行了相似的选择。

同样，经过预处理和时间选择的波士顿儿童医院的CHB-MIT头皮脑电图数据库的EEG脑电数据集计算样本熵值时与其计算标准差值时的方法类似，也可分别计算处病例chb01和chb03发作间期和发作期时自身的EEG与cD1~cD5这五组小波系数的样本熵值，并可分别绘制出如图3·13和图3·14所示的箱型图。

通过图3·13和图3·14，对于病例chb01和chb03分别选择小波系数cD3~cD5三组小波系数的样本熵值和自身脑电信号与小波系数cD3的样本熵值作为其分类的特征。

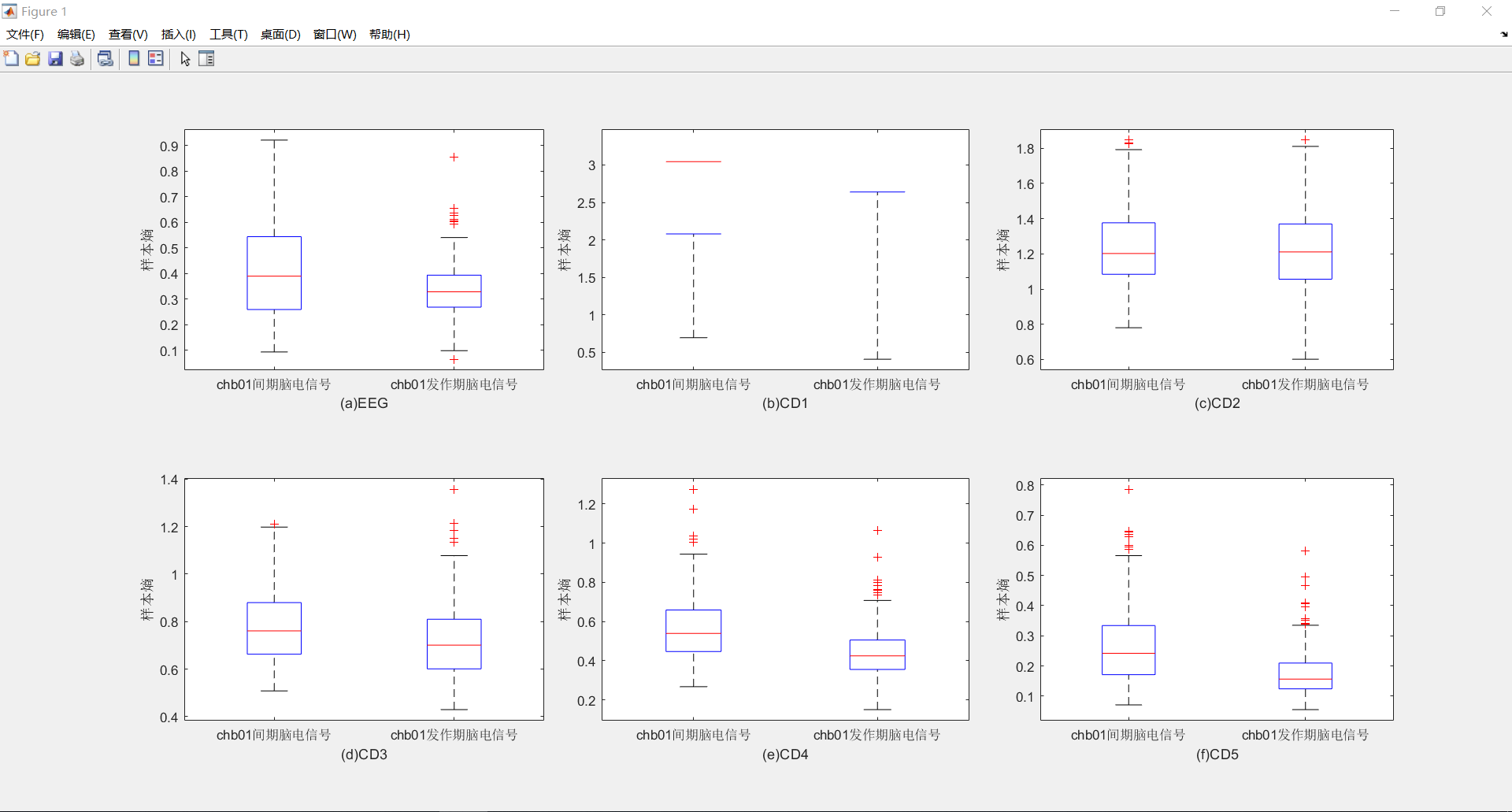


图3·13 病例chb01发作间期和发作期脑电信号方差箱型图对比

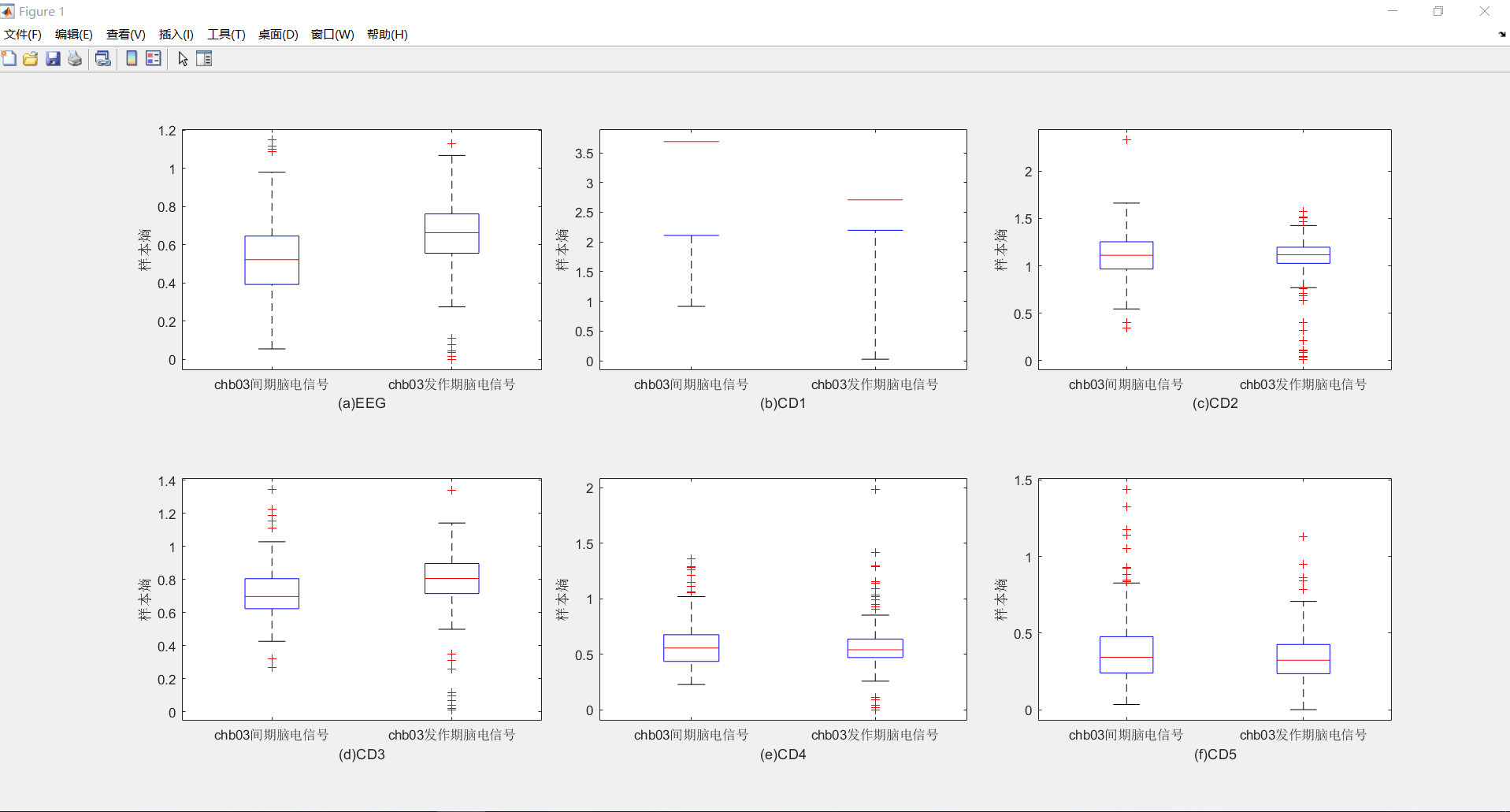


图3·14 病例chb03发作间期和发作期脑电信号方差箱型图对比

第3.5节 结果与分析

本章的主要内容和目的是基于离散小波变换(DWT)对脑电数据集进行特征的提取。首先本文对要用到的两个数据集进行了详细的介绍并进行了相应的预处理操作，之后得到了可用于分析研究的数据并确定了本研究要进行的脑电信号分类任务。然后对傅里叶变换、短时傅里叶变换和小波变换均进行了介绍，并对连续小波变换(CWT)和离散小波变换(DWT)的原理进行了阐述，之后本文采取离散小波变换进行后面的研究分析。对于小波基的选取，本文在对比了sym4小波、db4小波和coif4小波之后，最终选择了db4这一与脑电信号更加相似的小波。然后本研究对上述两个数据集均以db4为小波基进行了五层的离散小波变换并得到了相应的小波系数。通过前人论文中的研究结论，本研究选取标准差和样本熵作为癫痫脑电信号分类的特征。通过计算原始脑电信号与所选取的小波系数的标准差和样本熵与绘制箱型图进行比较，对这两个数据集要进行的分类任务，针对标准差和样本熵分别确定了分类所要用到的特征，为接下来癫痫脑电信号分类奠定了基础。

第4章 基于多种机器学习算法的癫痫脑电信号分类

通过前文的论述我们已经了解到脑电信号具有非平稳和非线性的特性，而癫痫脑电信号则更是具有突发性和长期反复发作的特征。经由上文中论述的工作，本研究已经将脑电信号的特征成功提取出来，而本研究最终要达成的目标则是要将人体正常生理状况下的脑电信号与癫痫发作时期的脑电信号进行分类。

现如今，应用十分广泛的脑电信号分类方法包括支持向量机(support vector machine，SVM)、决策树(Decision Tree，DT)、线性判别分析(Linear Discriminant Analysis，LDA)、朴素贝叶斯(Naïve Bayes，NB)、K最近邻算法(K-NearestNeighbor，KNN)、深度学习(Deep Learning，DL)和人工神经网络(Artificial Neural Network，ANN)等。本研究采用了三种机器学习算法(支持向量机、K最近邻算法和决策树)根据上文中得到的脑电信号特征来对脑电信号分类。

第4.1节 支持向量机

支持向量机理论

支持向量机(SVM)在1964年被提出，经过不断的发展，该算法已经在多个领域得到了广泛的应用。它是一种可被用于回归与分类分析中的监督式学习算法，它不仅可以应用于线性分类之中，而且在使用核函数(Kernel Function)将数据映射到高维度的空间之后进行非线性分类。支持向量机最重要的思想就是寻找一个超平面使得两类样本的样本点有最大间隔。

SVM线性分类

对于一个包含有n个点的测试点集S，，其中表示一个n维的向量，而的值则表明了的种类，其值可以是1或者-1(在本文中，1代表癫痫发作时期的脑电信号，而0代表其他时间段的脑电信号)。要将这两类样本数据进行分类，我们可以寻求一个超平面，将这两类样本数据分割为两部分，这样便达到了分类的目的。这样的超平面可能有很多个，如图4·1所示。而支持向量机的目的是找到一个最大间隔超平面，可以让超平面与样本点数据之间距离最大化，具体如图4·2所示，其中实线表示决策边界，虚线表示间隔边界，虚线上经过的两个点称之为支持向量。

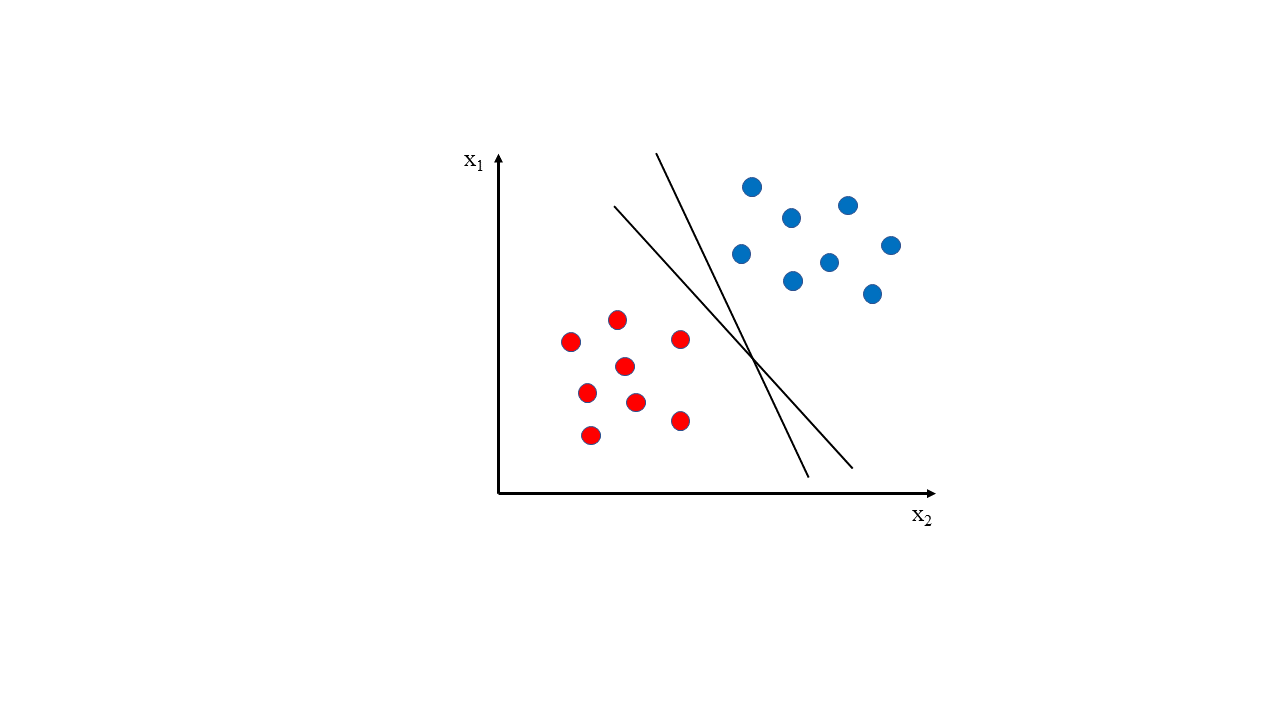


图4·1 超平面对样本数据的分隔

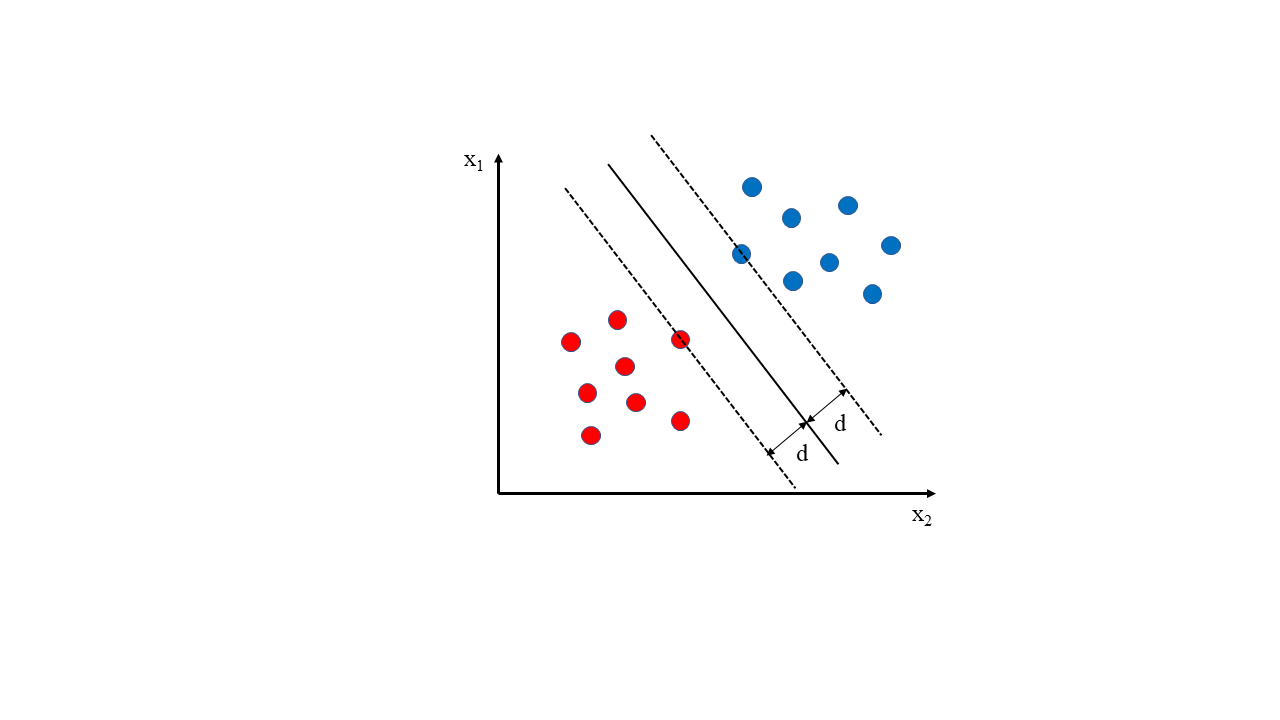


图4·2 最大间隔超平面

任意超平面均可用进行表示，其中。根据二维平面上的点到直线的公式，可扩展得出对于一个n维空间上的点，其到超平面的距离为，而。如图4·2所示，对于支持向量来说，其到超平面的距离d等于，而除去支持向量外的其他点到超平面的距离则大于d，那么对于两类样本数据则有



令，那么对于支持向量而言则有，最大化图4·2中两超平面之间的距离，则有：



让最大，即让最大，换言之，即令最小，因此便转换为了求解，其中对平方是为了去除其中的根号。



对于的不等式约束优化问题，引入拉格朗日乘子来进行解决，将不等式约束转换为了等式约束，并且得到了拉格朗日函数：



然后分别对求偏导并使之等于0，可解出：





将求解得到的结果代入中，可将其转换为二次规划寻优问题，即求下式的最大值，其约束条件为式子：



通过使用SMO(Sequential Minimal Optimization)算法可求解出最优解，求故可得，因此也可求解得到的值。将求得的和代入中便可得最佳超平面，其模型为：



之后可将新样本点带入分类决策函数中便可得到其类别(sign(x)表示符号函数)。

线性不可分

上述介绍的内容都是针对样本数据是线性可分得，然而当样本不是线性可分时，就需要用到核函数，核函数将原始低维空间中的点映射到了高维空间然后实现了线性可分。

其中原始空间的特征向量映射到了高维空间后得到了新的特征向量，故可得新的超平面：



因此中的同便变成了，但是当低维空间中的数据映射到高维空间之后数据的维度会非常大进而使得计算量巨大。

而核函数就是将把和代入并计算得到的结果来代替直接计算和计的内积，这样可以极大地降低计算量，节省成本。常用的核函数包括线性核函数、多项式核函数、RBF核函数和Sigmoid核函数等。在实际应用过程中，选择核函数也十分重要，如果核函数选择不合适，那么样本会因为映射不当而导致分类效果差。

第4.2节 K最近邻算法

K最近邻算法(KNN)是监督学习中最常用也是最简单的分类算法之一。该算法分类的依据十分简单，就是对一个待分类数据进行判断预测时，通过与该值距离最近的K个点的类别来判断该值的类别。如果这K个点中的大多数属于某一类，那便将该数据归为这一类。

而对空间中点的距离进行度量时，有很多种方式，例如曼哈顿距离(Manhattan Distance)、欧几里得距离(Euclidean Distance)和切比雪夫距离，而这些度量方式都是由明可夫斯基距离在不同取值下的特例。

对于两个n维实数向量和，，，其中明可夫斯基距离的计算公式如下：



其中为常量，且。当时，可得到计算曼哈顿距离的公式：



当时，可得到计算欧几里得距离的公式：



当时，可得到计算切比雪夫距离的公式：



本研究选取欧几里得距离作为度量特征向量之间距离的方式，对于KNN算法中K值的选取，具体如图2·4所示。当K值选取为4时，对于测试数据而言，其周围有三个B类数据和一个A类数据，由于B类数据占了多数，故将测试数据归为B类。

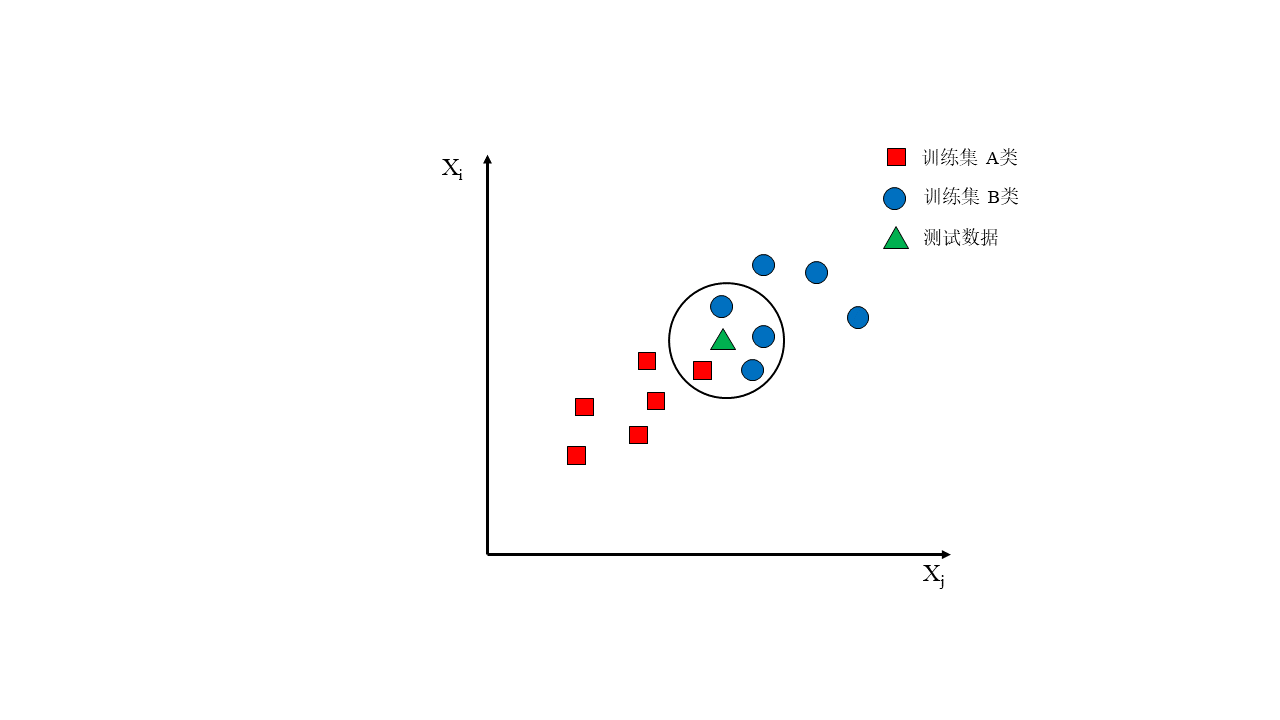


图2·4 KNN算法示例

第4.3节 决策树

决策树(Decision Tree)是机器学习中的一种非线性的监督学习算法，顾名思义，该算法是一个树状结构，该结构由节点和边两种元素组成，而节点则包括根节点、叶子节点、父节点和子节点。父节点可按照一定的规则分裂出子节点(若不是二叉树，会分裂出多个子节点)，子节点也可以作为新的父节点继续分裂，直至不能分裂为止。根节点即为无父节点的节点，它代表了整个样本；而叶子节点则为无子节点的节点，它不能进一步继续分裂。

在本研究中，仅对分类树进行介绍(回归树与之类似)。该算法的决策过程是这样子的：从根节点开始，对待测试分类项中对应的特征属性进行测试，按照其值选择输出分支，直到到达叶子节点出，然后将叶子节点中存放的类别作为最终的决策结果。

其中决策树的构建流程分为三个步骤：

1.选取对训练数据具有分类能力的特征；

2.在决策树各个点上按照一定方法选择特征，递归构建决策树；

3.对于已经生成的树，裁剪一些节点，进而简化分类树模型；

对于步骤1，可选择包括信息增益、信息增益比和基尼指数在内的三种特征，对着三类特征的简单介绍如表4·1所示。

表4·1 三类特征的介绍

|  |  |  |  |
| --- | --- | --- | --- |
| 特征 | 对应算法 | 树结构 | 支持模型 |
| 信息增益 | ID3 | 多叉树 | 分类 |
| 信息增益比 | C4.5 | 多叉树 | 分类 |
| 基尼指数 | CART | 二叉树 | 分类、回归 |

对于信息增益和信息增益比的计算，首先要计算熵，其计算公式如下：



熵值越大代表数据的混乱程度也就越高，其中代表种类i在样本中占的百分比，若数据全部为某一类时，熵值此时为0。

通过公式计算得出熵值之后，便可计算信息增益的值，父节点的熵值减去对应分裂得到的若干子节点的熵的和便可而得到信息增益，其计算公式如下：



当信息增益越大，表示父节点分裂后复杂度减少的越多，分类效果越明显。

针对ID3算法中存在的问题，改良得到了C4.5算法，信息增益比的计算公式如下：



其中即为计算得到的信息增益，而而则表示分裂得到的子节点数据量的信息增益，其计算公式为:



其中m代表子节点的数量，代表第i个节点的量，N代表父节点的数据量。其中信息增益率越高，分类效果越好。

基尼指数的计算公式如下:



其中代表分裂得到的子节点中种类i在样本中占的百分比。基尼值越大，代表数据越不纯，分类效果越差。

对于步骤2，在对步骤1中提及的三种具有分类能力的特征进行选择之后，需要按照训练数据中不同的属性分别进行分裂，然后能令选定的分类特征的值更优的属性来进行分类，构建决策树。

对于步骤3，为了让模型的泛化能力更强，需要对步骤2中生成的决策树进行简直，剪枝的方法分为预剪枝和后剪枝。预剪枝是预先对决策树的的高度或者阈值(实例个数等)进行定义，当达到时候，便停止决策树的生长；后剪枝则是先构造出完整的决策树，然后对置信度不够的子节点用叶节点来代替。

第4.4节 癫痫脑电信号的分类

评价指标的选取

在医院临床上经常要对病人进行及疾病的诊断，对于一项诊断结果，通常需要对其真实性、可靠性和实用性等方面进行考虑，进而方便医护人员进行下一步的治疗。

本研究便选取了三个在机器学习和医学等领域上常用的评价指标来对本研究中癫痫脑电信号的分类效果进行评价。这三个评价指标分别为准确度(Accuracy，acc)、灵敏度(Sensitivity，sen)和特异度(Specificity，spe)。

为了便于理解，此处引入混淆矩阵，对其的详细介绍如表4·2所示。在这张表中，True Positive表示将实际患病的病人成功预测为患病的状态，而False Negative表示将实际患病的病人错误地预测为健康的状态，剩下的False Positive和True Negative也与之相似。

表4·2 混淆矩阵详细介绍

|  |  |  |
| --- | --- | --- |
|  | 患病(预测状态) | 健康(预测状态) |
| 患病(实际状态) | True Positive | False Negative |
| 健康(实际状态) | False Positive | True Negative |

其中灵敏度(真阳性率)表明了分类算法对于实际患病的病例正确判断的能力大小，该值越大，将实际患病的患者进行正确判断的能力也就越强，其公式为：



而特异度(真阴性率)则表明了对实际状态为健康的人正确判断出其未患病的能力，该指标通灵敏度一样，也是越大越好，其公式为：



在综合了灵敏度(sen)和特异度(spe)这两个评价指标之后，便可以得到准确度(acc)这一评价指标，它表示对实际状态分别为患病和健康的人进行正确判断的能力，同样，该评价指标也是越大越好，其计算公式为：



癫痫脑电信号分类安排

上文已经对所要用到的两个数据集分别选定了分类要使用到的特征并对要使用的三种分类算法进行了理论上的介绍，接下来便进行癫痫脑电信号的分类。

对于波恩大学癫痫脑电数据集，本研究将进行如表3·2所示的四种情况二分类；

对于CHB-MIT头皮脑电图数据库的EEG脑电数据集，本研究将针对病例chb01和病例chb03分别进行发作间期和发作期的脑电信号的二分类。

对分类时所需要的特征，本研究将其分为了三类，分别是标准差特征、样本熵特征以及融合特征，对于这些特征的对应的变量定义及其含义如表4·3所示。

对于标准差特征，本文将以ori\_std、cd1\_std~cd5\_std共计六组的标准差值，每一组均单独作为特征，进行上述的分类。

对于样本熵特征，虽然本研究针波恩大学癫痫脑电数据集、病例chb01和病例chb03所选取的样本熵特征各不相同，但类似于上文中的标准差特征，本文将以每一组脑电数据的样本熵特征分别作为分类器的特征向量来进行分类。

对于融合特征，本研究将分别线性组合标准差特征和样本熵特征，将其重新构成多维的特征矩阵，并作为分类器的输入进行分类。在分类结果中，本文将融合后的标准差特征和样本熵特征分别与融合前的标准差特征和样本熵特征放在了一起进行对比。

表4·3 变量定义及其含义

|  |  |  |  |
| --- | --- | --- | --- |
| 特征名 | 含义 | 特征名 | 含义 |
| ori\_std | 原始脑电信号标准差 | ori\_sampen | 原始脑电信号样本熵 |
| cd1\_std | 小波系数cd1标准差 | cd1\_sampen | 小波系数cd1样本熵 |
| cd2\_std | 小波系数cd2标准差 | cd2\_sampen | 小波系数cd2样本熵 |
| cd3\_std | 小波系数cd3标准差 | cd3\_ sampen | 小波系数cd3样本熵 |
| cd4\_std | 小波系数cd4标准差 | cd4\_ sampen | 小波系数cd4样本熵 |
| cd5\_std | 小波系数cd5标准差 | cd5\_ sampen | 小波系数cd5样本熵 |
| multi\_std | 融合后的标准差特征 | multi\_sampen | 融合后的样本熵特征 |

在使用支持向量机训练与分类时，本文在matlab环境下使用了林智仁教授开发的libsvm工具箱。对于标准差特征或样本熵特征，本研究采用线性核函数；对于融合特征后得到的特征矩阵，则采用RBF径向基核函数，对于其参数的选择则是在对特征矩进行归一化后，使用网格搜索算法得出最优的cost和gamma参数。

在使用KNN训练与分类时，本文在matlab环境下使用fitcknn等函数进行了实现。其中最近邻参数k经过比较选为9，距离的度量方式选择了欧几里得距离。

在使用决策树训练与分类时，同样在matlab软件中使用统计工具箱进行了实现，其中使用fitctree函数实现了决策树的创建，并打开了决策树的自动剪枝，然后使用predict函数实现了对数据的分类。

在使用上述三种方法进行癫痫脑电信号的分类时，每次分类均进行了五折交叉验证，每次分类均进行了三次，在对得到的三个评价指标取平均值之后作为最终的结果，

分类结果与分析

将上述三类特征分别输入上文中介绍的三种分类算法中均可得出准确度、灵敏度和特异度三个评价指标，对其性能进行评价。

表4· 4 AE两组分类结果(标准差特征)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC (%) | | | SEN(%) | | | SPE(%) | | |
| SVM | KNN | DT | SVM | KNN | DT | SVM | KNN | DT | |
| ori\_std | 99.54 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 99.10 | 100.0 | 100.00 | |
| cd1\_std | 86.33 | 84.75 | 82.63 | 86.49 | 84.69 | 83.12 | 86.23 | 84.79 | 82.32 | |
| cd2\_std | 94.50 | 93.92 | 91.92 | 94.50 | 92.48 | 91.56 | 94.56 | 95.27 | 92.38 | |
| cd3\_std | 98.50 | 99.21 | 98.50 | 98.59 | 98.41 | 98.47 | 98.41 | 100.00 | 98.52 | |
| cd4\_std | 99.25 | 99.38 | 99.17 | 99.25 | 98.72 | 98.71 | 99.27 | 100.00 | 99.56 | |
| cd5\_std | 98.50 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 97.03 | 100.00 | 100.00 | |
| multi\_std | 99.67 | 99.88 | 100.00 | 99.36 | 99.75 | 100.00 | 100.00 | 100.00 | 100.00 | |

表4·5 BE两组分类结果(标准差特征)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC (%) | | | SEN(%) | | | SPE(%) | | | |
| SVM | KNN | DT | SVM | KNN | DT | | SVM | KNN | DT | |
| ori\_std | 94.46 | 96.08 | 95.21 | 95.82 | 94.95 | 94.55 | | 93.09 | 97.33 | 95.87 | |
| cd1\_std | 81.83 | 81.04 | 76.96 | 81.59 | 82.48 | 76.45 | | 82.12 | 79.65 | 77.42 | |
| cd2\_std | 90.63 | 90.33 | 87.04 | 84.29 | 84.27 | 84.70 | | 96.90 | 96.42 | 89.20 | |
| cd3\_std | 82.29 | 82.79 | 81.17 | 83.81 | 78.44 | 79.93 | | 81.11 | 87.03 | 82.27 | |
| cd4\_std | 88.58 | 91.00 | 89.21 | 91.95 | 89.40 | 88.54 | | 85.08 | 92.66 | 90.06 | |
| cd5\_std | 97.42 | 99.54 | 99.38 | 99.85 | 99.50 | 99.32 | | 95.03 | 99.59 | 99.41 | |
| multi\_std | 99.79 | 99.25 | 99.42 | 99.58 | 98.49 | 99.37 | | 100.00 | 100.00 | 99.50 | |

表4·6 CE两组分类结果(标准差特征)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | SPE(%) | | |
| SVM | KNN | DT | SVM | KNN | DT | SVM | KNN | DT | |
| ori\_std | 97.79 | 98.25 | 97.75 | 98.89 | 98.75 | 97.76 | 96.65 | 97.79 | 97.77 | |
| cd1\_std | 88.25 | 86.92 | 84.38 | 93.70 | 90.27 | 85.80 | 82.87 | 83.61 | 83.17 | |
| cd2\_std | 96.92 | 96.33 | 95.04 | 97.09 | 95.70 | 94.76 | 96.71 | 96.98 | 95.41 | |
| cd3\_std | 97.40 | 98.29 | 97.54 | 98.50 | 98.49 | 97.37 | 95.60 | 98.08 | 97.73 | |
| cd4\_std | 96.63 | 97.88 | 96.63 | 97.77 | 97.90 | 96.93 | 95.58 | 97.81 | 96.36 | |
| cd5\_std | 93.71 | 96.38 | 94.96 | 96.24 | 95.06 | 94.97 | 91.24 | 97.75 | 95.09 | |
| multi\_std | 98.88 | 98.29 | 97.96 | 98.96 | 97.80 | 98.14 | 98.73 | 98.73 | 97.76 | |

表4·7 DE两组分类结果(标准差特征)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | | SPE(%) | | |
| SVM | KNN | DT | | SVM | KNN | DT | SVM | KNN | DT |
| ori\_std | 90.88 | 92.46 | 89.38 | | 95.34 | 97.39 | 90.49 | 86.51 | 87.28 | 88.23 |
| cd1\_std | 90.08 | 88.79 | 84.96 | | 90.39 | 90.13 | 84.81 | 89.73 | 87.53 | 85.24 |
| cd2\_std | 93.46 | 93.63 | 92.67 | | 96.06 | 95.32 | 93.81 | 90.73 | 91.87 | 91.53 |
| cd3\_std | 94.67 | 96.50 | 94.17 | | 97.55 | 97.73 | 93.74 | 91.88 | 95.23 | 94.70 |
| cd4\_std | 91.50 | 94.38 | 92.00 | | 94.45 | 94.26 | 92.50 | 88.60 | 94.48 | 91.39 |
| cd5\_std | 87.29 | 92.04 | 88.13 | | 92.39 | 93.61 | 87.53 | 82.18 | 90.48 | 88.38 |
| multi\_std | 97.04 | 96.63 | 96.21 | | 96.91 | 97.27 | 96.60 | 97.12 | 95.97 | 95.74 |

样本熵特征

表4·8 AE两组分类结果(样本熵特征)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | SPE(%) | | |
| SVM | KNN | DT | SVM | KNN | DT | SVM | KNN | DT |
| ori\_sampen | 95.92 | 96.00 | 94.63 | 94.05 | 95.16 | 94.0 | 97.84 | 96.83 | 95.14 |
| cd1\_ sampen | 95.67 | 95.63 | 94.38 | 93.82 | 95.45 | 94.21 | 97.51 | 95.79 | 94.57 |
| cd2\_ sampen | 89.58 | 89.33 | 85.00 | 86.41 | 88.67 | 83.60 | 92.92 | 90.04 | 86.39 |
| cd3\_ sampen | 79.17 | 76.83 | 75.83 | 70.42 | 71.1 | 74.87 | 87.83 | 82.65 | 76.50 |
| multi\_sampen | 99.13 | 98.75 | 98.58 | 98.24 | 97.45 | 99.21 | 100.00 | 100.00 | 98.21 |

表4·9 BE两组分类结果(样本熵特征)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | SPE(%) | | |
| SVM | KNN | DT | SVM | KNN | DT | SVM | KNN | DT |
| ori\_sampen | 95.17 | 95.63 | 93.75 | 90.98 | 92.19 | 93.20 | 99.51 | 98.99 | 94.32 |
| cd1\_ sampen | 91.54 | 91.33 | 88.25 | 87.56 | 87.95 | 86.96 | 95.68 | 94.72 | 89.51 |
| cd2\_ sampen | 90.38 | 88.79 | 87.54 | 86.14 | 85.95 | 85.76 | 94.57 | 91.59 | 89.31 |
| cd3\_ sampen | 66.13 | 61.63 | 60.50 | 46.68 | 58.01 | 60.06 | 85.62 | 65.51 | 61.04 |
| multi\_sampen | 98.54 | 97.71 | 98.08 | 97.62 | 96.81 | 98.19 | 99.41 | 98.57 | 98.06 |

表4·10 CE两组分类结果(样本熵特征)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | SPE(%) | | |
| SVM | KNN | DT | SVM | KNN | DT | SVM | KNN | DT |
| ori\_sampen | 69.13 | 66.63 | 61.83 | 63.09 | 65.61 | 60.68 | 75.72 | 67.95 | 63.15 |
| cd1\_ sampen | 97.00 | 96.25 | 95.46 | 95.07 | 95.61 | 95.85 | 99.03 | 96.94 | 95.21 |
| cd2\_ sampen | 94.96 | 95.00 | 92.75 | 93.70 | 95.33 | 92.33 | 96.04 | 94.65 | 93.09 |
| cd3\_ sampen | 76.67 | 74.38 | 70.38 | 66.43 | 70.23 | 70.16 | 87.13 | 78.64 | 70.70 |
| multi\_sampen | 97.63 | 97.88 | 97.71 | 97.00 | 97.57 | 97.16 | 98.20 | 98.17 | 98.26 |

表4·11 DE两组分类结果(样本熵特征)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | SPE(%) | | |
| SVM | KNN | DT | SVM | KNN | DT | SVM | KNN | DT |
| ori\_sampen | 64.75 | 64.75 | 63.46 | 79.33 | 68.33 | 62.37 | 50.52 | 61.42 | 64.66 |
| cd1\_ sampen | 88.13 | 87.54 | 84.58 | 91.37% | 91.33 | 85.93 | 84.92 | 83.69 | 83.35 |
| cd2\_ sampen | 79.79 | 78.63 | 74.00 | 89.2 | 87.97 | 76.65 | 70.50 | 69.19 | 71.30 |
| cd3\_ sampen | 63.96 | 62.67 | 58.58 | 64.09 | 63.97 | 60.40 | 64.02 | 61.69 | 56.66 |
| multi\_sampen | 89.46 | 89.54 | 87.00 | 88.54 | 92.54 | 87.35 | 90.43 | 86.51 | 86.80 |

表4·12 病例chb01发作间期与发作期分类结果(标准差特征)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | | SPE(%) | | |
| SVM | KNN | DT | | SVM | KNN | DT | SVM | KNN | DT |
| ori\_std | 90.62501 | 92.98609 | 91.45833 | | 91.78 | 91.24 | 91.19 | 89.40 | 94.71 | 91.49 |
| cd1\_std | 91.18055 | 93.61111 | 91.94444 | | 86.05 | 93.71 | 92.46 | 96.27 | 93.59 | 91.26 |
| cd2\_std | 91.38889 | 90.97221 | 87.3611 | | 85.19 | 85.85 | 86.83 | 97.52 | 96.01 | 87.67 |
| cd3\_std | 97.84722 | 97.98611 | 97.15277 | | 96.78 | 96.91 | 96.51 | 98.87 | 99.07 | 97.80 |
| cd4\_std | 94.93055 | 94.99999 | 95.20833 | | 96.05 | 94.34 | 95.31 | 94.08 | 95.87 | 95.11 |
| cd5\_std | 93.19445 | 95.97222 | 94.86111 | | 96.55 | 95.99 | 94.74 | 89.92 | 95.96 | 95.26 |
| multi\_std | 99.23611 | 98.05555 | 98.54165 | | 99.53 | 98.26 | 98.12 | 98.94 | 98.02 | 98.88 |

表4·13病例chb03发作间期与发作期分类结果(标准差特征)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | | SPE(%) | | |
| SVM | KNN | DT | | SVM | KNN | DT | SVM | KNN | DT |
| ori\_std | 90.56 | 92.08 | 88.47 | | 95.20 | 95.85 | 88.72 | 86.27 | 88.16 | 88.23 |
| cd1\_std | 87.92 | 86.60 | 82.92 | | 96.86 | 93.57 | 84.48 | 79.04 | 79.75 | 82.09 |
| cd2\_std | 87.43 | 87.64 | 82.01 | | 95.87 | 94.87 | 82.11 | 79.09 | 80.45 | 82.11 |
| cd3\_std | 89.65 | 90.14 | 86.11 | | 95.79 | 94.99 | 87.10 | 83.62 | 85.08 | 85.28 |
| cd4\_std | 94.17 | 95.63 | 94.38 | | 95.74 | 95.26 | 93.81 | 92.76 | 96.04 | 94.95 |
| cd5\_std | 93.89 | 94.51 | 92.15 | | 95.67 | 94.24 | 91.96 | 92.03 | 94.74 | 92.26 |
| multi\_std | 96.67 | 96.46 | 95.90 | | 95.77 | 95.54 | 95.43 | 97.54 | 97.38 | 96.25 |

表4·14病例chb01发作间期与发作期分类结果(样本熵特征)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | SPE(%) | | |
| SVM | KNN | DT | SVM | KNN | DT | SVM | KNN | DT |
| cd3\_ sampen | 57.08 | 55.69 | 53.26 | 52.42 | 53.59 | 53.08 | 62.50 | 58.06 | 53.43 |
| cd4\_ sampen | 65.76 | 63.82 | 58.82 | 74.45 | 63.69 | 58.97 | 57.19 | 63.89 | 59.48 |
| cd5\_ sampen | 66.04 | 61.18 | 58.06 | 85.12 | 63.34 | 58.70 | 47.43 | 59.01 | 57.52 |
| multi\_sampen | 71.25 | 68.75 | 63.47 | 80.37 | 71.93 | 62.07 | 62.47 | 66.22 | 64.95 |

表4·15病例chb03发作间期与发作期分类结果(样本熵特征)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 特征 | ACC(%) | | | SEN(%) | | | SPE(%) | | |
| SVM | KNN | DT | SVM | KNN | DT | SVM | KNN | DT |
| ori\_ sampen | 65.35 | 61.74 | 60.28 | 72.35 | 64.06 | 62.29 | 59.05 | 59.94 | 57.70 |
| cd3\_ sampen | 64.72 | 63.47 | 58.89 | 56.30 | 63.27 | 58.65 | 73.69 | 64.10 | 59.45 |
| multi\_sampen | 67.64 | 69.58 | 64.65 | 64.40 | 70.96 | 64.48 | 71.34 | 68.33 | 65.20 |

第5章 集成学习

第5.1节 集成学习理论介绍

集成学习(Ensemble Learning，EL)的主要思想是将多种学习算法通过某种策略将它们进行组合从而得到一个预测模型的元算法。该元算法可提升性能，得到更优的预测结果。通常集成学习分为三类：用于减少方差的bagging、用于减少偏差的boosting和用于提升预测结果的stacking。

Bagging方法

Bagging(Bootstrap aggregating)方法又叫装袋法，它通过减少结果的方差以避免过拟合的发生，其常用的集成算法模型是随机森林。该算法进行的步骤如下：

1.抽取训练集：由原始样本中采用有放回的抽取方式抽取得到p个训练样本(有些样本可能一次也没被抽中而有些样本可能被抽中多次)。这样的抽取共进行m轮，可得到m个训练集(这些训练集相互独立)。

2.训练模型：将步骤1中抽取得到的m个训练集分别训练同一个模型，这样可得到m个相同的模型。对于训练时所使用的回归或分类算法可根据具体问题的不同采用不同的算法，例如感知器、决策树等。

3.对于回归问题可通过计算上述m个模型的均值作为最终的结果；而对于分类问题可将上述m个模型中分类结果采取投票的方式得到最终的分类结果。

Boosting方法

Boosting方法和Bagging方法都是将某一种分类或回归算法组合得到一个性能更优的强分类器，均属于同质集成。然而不同的是boosting方法每一轮训练集不变，只是每个数据在分类器中的权重发生改变；而bagging方法从数据集中有放回地抽取，训练集之间相互独立，且每个数据的权重是相等的，不变的。Boosting方法的主要思想是使用加权的数据来训练一系列弱学习器，然后通过一定的组合策略将这一系列弱学习器转换为强学习器，其中常用的算法有AdaBoost算法和提升树系列算法(最常用的是梯度提升树)。该算法的执行步骤如下：

1.根据初始权重使用训练集训练得到一个弱学习器1；

2.由弱学习器的学习误差率更新训练集权重(提高弱学习器1学习错误率高的样本点)；

3.根据调整后的权重训练弱学习器2；

4.重复上述步骤，直至弱学习器数目达到这指定数量T；

5.将这T个弱学习器结合(回归问题采用加权结合；分类问题采用投票法结合)

Stacking方法

Stacking方法可以看作是一种结合策略，不同于bagging和boosting方法，它是一种异质集成方法。它是一种分层模型的集成框架，即它可以由多层模型构成，一般由两层模型构成，其中第一层模型由若干个初级学习器构成，这些初级学习器根据训练集进行训练得到学习结果；第二层模型由一个次级学习器构成，由第一层模型输出的学习结果对其进行训练得到最终的预测结果。其中stacking方法过程的示意图如图5·1所示，图5·1采用了两层模型和五折交叉验证。

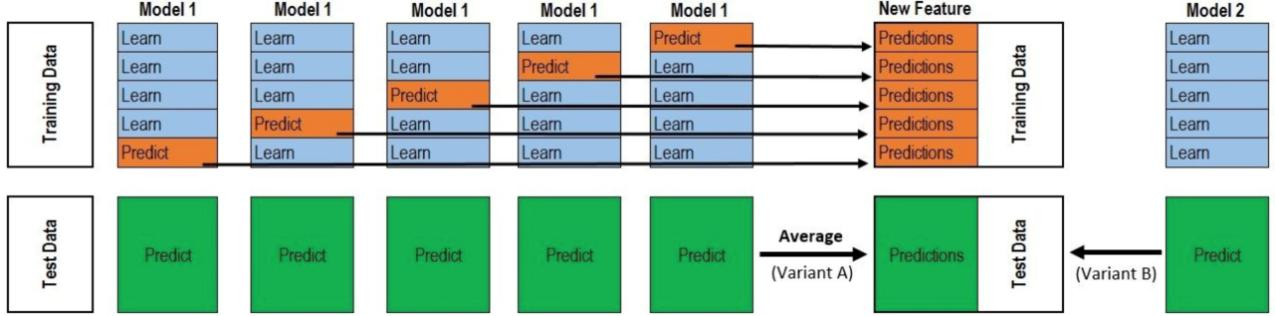


图5·1 stacking方法过程示意

首先将数据集以一定的比例分为训练集和测试集两部分，然后将训练集分为五份，使用五折交叉验证将训练集中的每一部分单独作为测试数据，其余四部分用来训练第一层模型中的model1，五折交叉验证后便可得到五组预测结果，并将该预测结果作为第二层模型中次级学习器的一组输入特征；另外每一折验证之后均对测试集的数据进行预测，这样便可得到五组预测结果，对这五组预测结果取平均后的结果便作为第二层模型的一组测试集。对于第一层模型中的其他分类算法，也采用相同的方法，便可相应得到输入特征和测试集，将第一层模型中得到的所有输入特征和测试集分别线性组合便可得到第二层次级分类器的训练集和测试集，之后便可得到分类结果。

第5.2节 集成学习1

Stacking 结果

波恩大学癫痫脑电数据集

标准差

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACC | SEN | SPE |
| A vs E | 100.00 | 100.00 | 100.00 |
| B vs E | 100.00 | 100.00 | 100.00 |
| C vs E | 100.00 | 100.00 | 100.00 |
| D vs E | 98.75 | 98.61 | 98.86 |

样本熵

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACC | SEN | SPE |
| A vs E | 99.38 | 98.63 | 100.00 |
| B vs E | 98.75 | 97.30 | 100.00 |
| C vs E | 96.88 | 95.89 | 97.70 |
| D vs E | 89.38 | 93.65 | 86.60 |

CHB-MIT数据集

标准差　％

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACC | SEN | SPE |
| 病例chb01 | 100.00 | 100.00 | 100.00 |
| 病例chb03 | 98.96 | 100.00 | 98.08 |

样本熵　％

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACC | SEN | SPE |
| 病例chb01 | 64.58 | 62.22 | 66.67 |
| 病例chb03 | 61.46 | 57.41 | 66.67 |

结 论

本课题采用XXXXXXXXXXXXXXXXXXXXXXXXXX得到以下结论：

1、XXXXXXXXXXXXXXXXXXXXXX

2、XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX（宋体，小四，固定值22磅，首行缩进2字符）

参考文献

1. 唐颖莹,陆璐,周东.中国癫痫诊断治疗现状[J].癫痫杂志,2019,5(03):161-164.

[2] 刘光知．螺杆元件排列组合的混炼效果研究[J]．中国塑料，1993，7(3)：58．

[3] 耿孝正，张沛．塑料混合及设备[M]．北京：中国轻工业出版，1992：1．

[4] 李鹏，耿孝正，马秀清．啮合同向双螺杆挤出过程组合流道（捏合块+螺纹元件）三维流场分析[J]．中国塑料，2001，15(7)：1．

[5] Li Shaofan，Liu WK．Mesh free and particle methods and their applications[J]．ApplMech Rev，2002，55(3)：1-34．

[6] 张平亮．双螺杆挤出机的进展及其应用[J]．工程塑料应用，2005，33(5)：2-3．

[7] 李鹏，耿孝正．同向啮合双螺杆挤出机捏合块流道三维流场分析[J]．中国塑料，2000，14(3)：75．

[8] 王文飞．啮合同向双螺杆挤出机捏合块组合的研究[D]．北京：北京化工大学，2011．

[9] 邱国庆．同向旋转双螺杆挤出机螺旋状捏合盘元件混炼机理研究[D]．北京：北京化工大学，2010．

[10] Yang．H．H，Manas-Zloczower．I．Flow field analysis of the kneading disc region in a co-rotating twin screw extruder[J]．Polymer Engineering & Science，1992，32(19)：1411-1417．

[11] Shaffiq A．Jaffer，Victor L．Bravo，Philip E．Wood，Andrew N．Hfwmak．Experimental validation of numerical simulations of the kneading disc section in a twin screw extruder[J]．Polymer Engineering & Science，2000，40 (4)：892-901．

[12] White．J．L，Chen Ziyun．Simulation of non-isothermal flow in modular co-rotating twin screw extruder[J]．Polymer Engineering & Science，1994，34(3)：229-237．

[13] D．J．Van Der Wal，D．Goffart，E．M．Klomp，H．W．Hoogstraten, L．P．B．M．Janssen．Three-dimensional flow modeling of a self-wiping co-rotating twin screw extruder．Part II： the kneading section[J]．Polymer Engineering & Science，1996，36(7)：912-924．

[14] Yasuya Nakayama，EijiTakeda，TakashiShigeishi，HidekiTomiyama，Toshihisa Kajiwara．Melt-mixing by novel pitched-tip kneading disks in a co-rotating twin-screw extruder[J]．Chemical Engineering Science，2011，66 (1)：103-110．

[15] Takeshi Ishikawa，Shin-Ichi Kihara，Kazumori Funatsu．3-D numerical simulations of nonisothermal flow in co-rotating twin screw extruders[J]．Polymer Engineering＆Science．2000，40(2)：357-364．

[16] V．L．Bravo，A．N．Hrymak．Numerical simulation of pressure and velocity profiles in kneading elements of a co-rotating twin screw extruder[J]．Polymer Engineering & Science，2000．40(2)：539．

[17] Yoshinaga，Makoto，Katsuki，Shingo，Miyazaki，Masam．Mixing mechanism of three-tip kneading block in twin screw extruders[J]．Polymer Engineering & Science，2000，40(1)：168-178．

[18] 李鹏，耿孝正．啮合同向双螺杆挤出机捏合块流道流场分析[D]．北京：北京化工大学，2000．

[19] 刘青峰，马秀清．同向双螺杆挤出过程不同螺杆构型的混合性能分析[J]．橡胶技术与装备，2006，32(9)：1-6．

[20] Xian-Ming Zhang，Lian-Fang Feng，Wen-Xing Chen，Guo-Hua Hu．Numerical simulation and experimental validation of mixing performance of kneading discs in a twin screw extruder[J]．Polymer Engineering & Science，2009，49(9)：1772-1783．

[21] 尹燕玲，耿孝正，马秀清．啮合同向双螺杆挤出过程新型混合元件——六棱柱元件实验研究[J]．中国塑料，2002，16(6)：76-80．

（宋体，小四，固定值22磅，首行缩进2字符，至少15篇中文文献，5篇外文文献）

致 谢

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXxXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

（宋体，小四，固定值22磅，首行缩进2字符）