中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生数学建模竞赛

学　　校	南京邮电大学
参赛队号	**20102930029**
队员姓名	**1.林焜达**
	2.丁海杰
	3.姜宇

中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生

数学建模竞赛

题 目 面向康复工程的脑电信号分析和判别模型

摘 要：

本文旨在利用采集的脑电信号，通过相关的特征分析和数据处理准确反映大脑对外部设备的直接控制信息，以及从复杂的睡眠过程中划分出准确的睡眠阶段。为了准确获得大脑对外界的反应信号，特别是增强识别目标的准确性，需要设计恰当的分类识别方法。考虑到采集的原始脑电数据的冗余性，对原始的采集通道进行组合，得出一组最优的通道组合，再利用这个最优通道组合对样本进行识别。另外，基于脑电信号反映睡眠各个时期的条件，构建一个睡眠分期预测模型，准确划分睡眠数据的各个阶段。通过对脑电信号的充分处理和分析，为康复工程中残疾人的恢复治疗起到辅助作用，以及为专家和学者提供对应睡眠问题治疗的合理建议。

对于问题一，利用小波分解，取出了信号中的低频基线分量以及噪声分量，并设计堆叠双向循环神经网络模型对 P300 信号的进行识别；为了解决数据不平衡导致模型难以训练的问题，使用了一系列策略优化模型：加权损失（Weighted Loss）解决模型极易对负样本过拟合的问题，加权准确率（Weighted Accuracy）替代准确率用于模型选择，利用数据增广构建随机数据集减轻过拟合现象，最后使用交叉验证（Cross Validation）作为评价模型的策略，更好地利用了训练集的数据。

对于问题二，使用显著性图（Saliency Map）来分析 20 个通道的重要性，并分析通道之间的相关性，验证基于显著性图的通道选择方法的合理性。本文选用了 12 个通道进行实验，实验结果表明，减少部分通道并不会明显降低模型表现。

对于问题三，基于问题一的基本模型构建和问题二的通道选择，选用了 30%的标签数据，减低了数据的维度，使得单个轮次训练时间从 3.01 s 下降到 0.12s。训练中，本文基于半监督学习，应用自训练（Self-training）的方法对模型进行训练，对问题一和问题二的结果进行优化。最终在仅仅使用 30%带标签数据的情况下，加权准确率从 54.4%提升到 60.8%，推理准确率（Inference Accuracy）从 9.1%提升到 10.4%。

对于问题四，使用了小波分析处理了脑电数据，并对比分析了 SVM，XGBoosting，BP 神经网络，半监督学习（自训练），在仅仅使用 30%的数据的情况下准确率达到了 99.1%。值得一提的是，即使使用 10%的数据，依然可以达到 86%的准确率。

关键词：脑电波 半监督学习 小波变换 数据不平衡性 自训练 Self-training

目录

1. 问题重述

1.1 引言

肢体运动功能障碍患者常常因为运动功能丧失而导致失去劳动能力，这为许多家庭带来了沉重的负担。脑-机接口技术的出现为肢体运动障碍患者带来了希望。脑-机技术可以让大脑不经外围神经或者肌肉组织直接控制外部设备，而基于 P300 的脑-机接口可以帮助使用者无需通过复杂训练就可以获得较高的识别准确率，具有稳定的锁时性和高时间精度特性，这为康复工程中恢复患者的运动能力，改善治疗结果提供了辅助工具。睡眠是身体休整积蓄能量的重要环节，睡眠质量对人的身心状态有着重大影响。睡眠时产生的脑电信号能够反映身体的变化，可以为医学上诊断和治疗相关睡眠疾病提供依据。因此，如何利用采集的脑电信号，通过相关的特征分析和数据处理准确反映大脑对外部设备的直接控制信息成为了发挥脑电信号作用的关键部分。另外，如何从复杂的睡眠过程中划分出准确的睡眠阶段也成为提高睡眠质量亟待解决的问题。

为了准确获得大脑对外界的反应信号，特别是增强识别目标的准确性，需要设计合理的分类识别方法，考虑到采集的原始脑电数据的冗余性，对原始的采集通道进行组合，得出一组最优的通道组合，再利用这个最优通道组合对样本进行识别。另外，基于脑电信号反映睡眠各个时期的条件，构建一个睡眠分期预测模型，准确划分睡眠数据的各个阶段。

1.2 需要解决的问题

根据所采集的脑电信号以及脑电信号在两种产生方式下表现的不同特征，提出了以下问题：

1) 使用尽可能少的测试数据，找出 5 个测试集中的待识别目标，并与几种方法进行对比。

2) 对 20 个采集通道进行处理，过滤掉多余的通道数据，对通道进行组合，并得出一组最优的通道名称组合。

3) 从训练集中选择适量的样本作为有标签样本，将其余样本作为无标签样本，基于问题二得到的一组最优通道组合，设计出一种学习方法，用测试数据（char13-char17）检验方法的有效性，并利用该方法找出测试集中其余待识别目标。

4) 利用尽可能少的训练样本进行训练数据和测试数据的分配划分，并设计出一个睡眠分期预测模型能够准确地对样本进行预测。

2. 符号说明和模型假设

2.1 符号说明

符号	意义
τ	小波变化位移量
a	小波变换尺度因子
c_0	负样本的预测结果
c_1	正样本的预测结果
(x_i, y_i)	第 i 个训练数据点
X_i	输入随机变量的第 i 分量
n	小波变换分解水平
N	小波阶数，样本容量
ρ_{ij}	随机变量X_i和X_j的相关系数
R	实数集，随机向量的相关矩阵
$\|\cdot\|$	*L2* 范数
L	损失函数，拉格朗日函数

2.2 模型假设

- P300 事件相关电位在小概率刺激发生后 300 毫秒范围左右出现一个正向波峰。
- 在 20 个脑电信号采集通道中，存在无关或冗余的通道数据。
- 脑电数据会出现基线漂移现象，总体的数据呈逐渐向上漂移的趋势。
- 采集到的脑电信号满足狄利克雷条件。

3. 问题分析

肢体运动功能障碍患者常常因为运动功能丧失而导致失去劳动能力，这为许多家庭带来了沉重的负担。脑-机接口技术的出现为肢体运动障碍患者带来了希望。脑-机技术可以让大脑不经外围神经或者肌肉组织直接控制外部设备，而基于 P300 的脑-机接口可以帮助使用者无需通过复杂训练就可以获得较高的识别准确率，且具有稳定的锁时性和高时间精度特性，这为康复工程中恢复患者的运动能力，改善治疗结果提供了辅助工具。睡眠是身体休整积蓄能量的重要环节，睡眠质量对人的身心状态有着重大影响。睡眠时产生的脑电信号能够反映身体的变化，可以为医学上诊断和治疗相关睡眠疾病提供依据。因此，如何利用采集的脑电信号，通过相关的特征分析和数据处理准确反映大脑对外部设备的直接控制信息成为了发挥脑电信号作用的关键部分。另外，如何从复杂的睡眠过程中划分出准确的睡眠阶段也成为提高睡眠质量亟待解决的问题。

3.1 问题一分析

问题一要求尽可能使用较少轮次的测试数据来识别出 5 个测试集中的 10 个目标，并给出具体的分类识别过程。首先需要分析所给的数据集，了解训练集和测试集中各样本代表的含义以及数据的采集策略，根据脑电数据的采样频率，分析标签文件中相邻两行的采样频率。然后利用去噪方法将脑电信号数据中存在的噪声去除，另外还要剔除数据集中一些无用的数据，如每轮实验的起始标签行和一轮实验的结束标签“100”。最后将脑电数据在时序的条件下对脑电数据进行分析，得出最终的识别结果，并与脑电数据在瞬时条件下的结果进行对比。

3.2 问题二分析

问题二要求对脑电信号的 20 个采集通道进行处理，删除一些无关冗余的通道数据，选取对被试者都较为适用的一组最优通道名称组合。P300 脑-机接口同时采集多个通道的数据，由于传感器之间位置、硬件、环境等因素，最终采集到大量包含噪声、冗余信息的数据。大量冗余数据的存在会引入大量无关的噪声，使模型过拟合，妨碍训练，影响模型识别 P300 信号。为了去除无用通道，需要将各个通道数据对整个脑电信号分析的重要性具体化。通过使用问题一中构建的深度学习模型，分析各个通道对模型的贡献程度。在数学上，每个位置的梯度可以表示在该点上，求导对象对被求导对象的影响，即每个位置对目标的重要性，通过使用问题一中训练得到的深度学习模型的目标函数对输入数据进行求导，得到输入数据的显著性图（Saliency Map）。首先应该使用 Saliency Map 对各个通道的数据计算其重要性，用 Saliency Map 生成的像素点亮度表现通道的重要性。然后需要分析 20 个通道之间的相关性，筛选出相关性高的通道，并按照通道的重要性在相关性高的通道中进行选取，从而得出一组最优通道名称组合。为了验证选取的通道组合的最优性，将选取的最优通道组合放在测试集中实验，并用已给的测试数据（char13-char17）的结果进行验证。

3.3 问题三分析

问题三要求选择适量的样本作为有标签样本，其余训练样本作为无标签样本，在选取

的最优通道组合的基础上设计一种学习的方法，并利用已知标签的测试数据检验方法的有效性，最终识别出所有测试集的目标。为了满足题目对训练时间的要求，在选择样本数量时，需要尽可能地减少有标签样本的数量，减少训练时间。同时对于最优通道组合，经过问题二的处理，通道数量相较于原始通道数量减少一些，则原始脑电数据中一些冗余的通道数据将不参与模型的训练，训练时间将进一步减少。考虑到题目要求尽可能利用少量的样本数据，结合需要减少训练时间的条件，这里采用半监督学习作为学习方法，最终找出测试集中的其余待识别目标。

3.4 问题四分析

问题四要求在尽可能少的训练样本基础上，设计一个睡眠分期预测模型，并给出具体的分类识别过程，检验分类性能指标对预测的效果。睡眠脑电数据主要分为五个子表，分别是清醒期、快速眼动期、睡眠 I 期、睡眠 II 期以及深睡眠期的睡眠脑电数据。由于睡眠时采集的脑电数据中也存在大量干扰信号，因此需要对睡眠脑电数据去除干扰信号。然后应该设计合适的分类器，根据提取到的各项特征将睡眠脑电信号的各个阶段进行有效划分。最后需要对所给的睡眠脑电数据进行合理划分，利用一小部分充当训练集，另一部分充当测试集，以尽可能少的样本训练并用测试集验证设计的分类器的预测分类效果。

4. 问题一的建模与求解

4.1 问题分析

问题一要求尽可能使用较少轮次的测试数据来识别出 5 个测试集中的 10 个目标，并给出具体的分类识别过程。首先需要分析所给的数据集，了解训练集和测试集中各样本代表的含义以及数据的采集策略，根据所给的脑电数据，分析数据集是否存在噪声，需要对脑电数据进行预处理。根据脑电数据的采样频率，分析标签文件中相邻两行的采样频率，设置合理的时序窗口。然后利用去噪方法将脑电信号数据中存在的噪声去除，另外还要剔除数据集中一些无用冗余的数据，如每轮实验的起始标签行和一轮实验的结束标签“100”。最后将脑电数据在时序的条件下对脑电数据进行分析，得出最后的识别结果，再用瞬时条件的结果与时序条件下的结果进行对比，阐明在时序条件下所设计方法的合理性。

4.2 脑电数据及事件标签分析

在所给的 P300 脑电数据中，数据文件主要分为两大类：脑电数据以及事件标签。脑电数据是在 P300 脑机接口实验中对每位被试按照 250Hz 的采样频率采集的样本数据，也就是按照 4ms 的间隔对被试的脑电数据进行采集。每条脑电数据有 20 列，分别代表 20 个记录通道采集的电极数据，每个电极数据是作用电极与参考电极之间的差值。事件标签文件中的数据共有两列，第一列表示标签，第二列表示采样点序号。对于标签训练集，文件起始行的标签列为目标字符对应的标识符。接下来的标签列为闪烁的行或列的标识符，每轮实验的结束标签为“100”。而在测试集标签文件中，首行的标签列用“666”表示，以此代表尚未识别的标识符。每轮实验均有 5 次，由于闪烁的字符矩阵每次以随机的顺序闪烁矩阵的一行或一列，闪烁时长与间隔均为 80ms，因此对标签文件中的数据进行分析，发现每行数据的间隔采样事件约为 160ms。

考虑到 P300 事件相关电位是在小概率刺激发生后 300ms 范围左右出现的一个正向波峰，相对于基线来说是一种向上趋势的波，如图 1 所示，尽管个体间存在差异性，导致 P300 的发生时间不同，但 P300 峰值信号大概仍是在接收刺激后的 300ms 左右出现。由于电位信号具有连续性的特点，因此涉及到时间序列相关问题，在处理脑电数据中，采用深度学习中时间序列的方案对脑电数据进行处理。鉴于 P300 信号是在接收外界刺激 300ms 范围左右出现的一个正向波峰，考虑到个体之间的差异会导致反应时间上的迟缓，这里将闪烁矩阵一行或一列对应的采集样本点前面 200ms 范围内的样本数据以及后面 500ms 内范围的样本数据作为一个时间序列，并对这个时间序列进行分析处理。

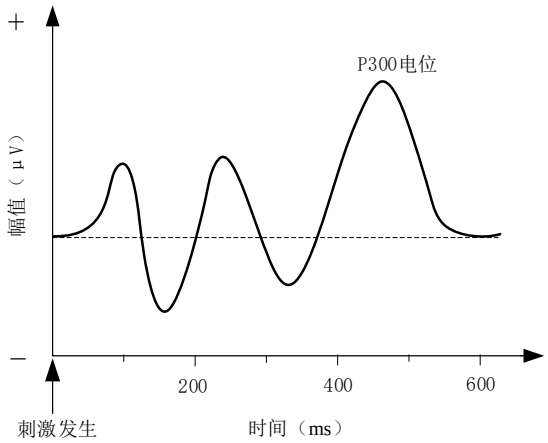

图 1 P300 波形示意图

4.2.1 基线漂移现象

脑-机接口实验数据主要分为两类数据：脑电数据和标签文件。脑电数据的训练集数据中包含着 12 个子表（char01~char12），这 12 个子表中的脑电数据样本分别有 3000 多条，均对应这字符矩阵上显示的灰色字符的闪烁实验，每个字符均为已知字符，来自 5 轮实验，每轮实验都是 12 个字符的随机轮流闪烁。

随机从 12 个子表中摘取几个子表，进行粗略分析，再从每个子表中随机取几列脑电数据，由于各列脑电数据来自于不同的采集通道，互不影响，则将这几列随机选择的数据分别生成图标，观察大概的信号趋势，以其中一列数据为例，如图 2 所示。

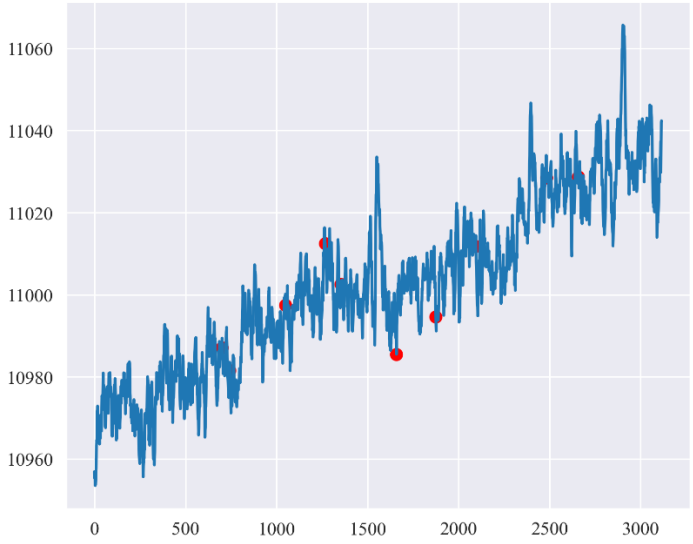

图 2 脑电数据原始信号图

如上图所示，原始的脑电数据出现了基线漂移现象，总体的数据是逐渐向上漂移的趋势，由于漂移的影响，导致无论怎样划分，每一段的数据的起点都不在同一个地方。对于采集的脑电数据，由于字符闪烁对被试者是随机间隔性的出现，只有当被试者看到闪烁行

或列出现“目标字符”时，脑电信号中才会出现 P300 电位，当其它行或列闪烁时，脑电信号中不会出现 P300 信号。在理想状态下，总体的数据趋势应该是不定间隔的 P300 电位出现，其余时间内的电位信号处于相对平静的状态。

4.2.2 电信号关系研究

随机找两个被测试者，将从这两个被试者相同通道采集的脑电数据进行概率分布，并用图表展示，由图 3 可以明显看到同一通道采集的数据电位信号差别很大，同一通道采集的数据平均差别高达 13000，结果表明从不同被试者相同识别目标字符的同一通道采集的电位信号存在很大的差别，需要区别处理。

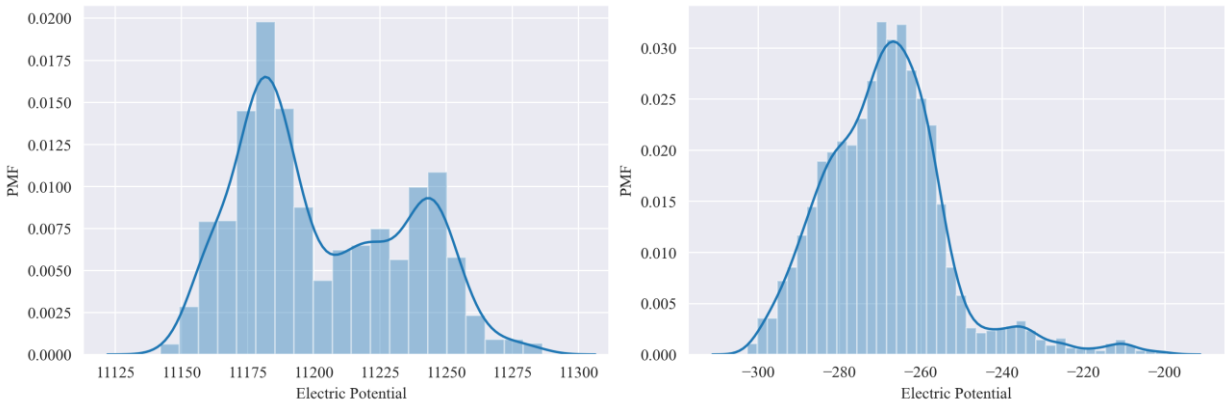

图 3 不同受试者相同通道采集的脑电信号分布图

为了确定同一通道采集的电位数据存在差异的原因，这里对同一被试者的不同识别目标字符的脑电信号进行分布对比，如图 4 所示，结果表明，对于同一被试者相同通道采集的电位信号分布结果相差不大。

基于以上两种脑电信号分布的对比，可以知道：不同受试者相同通道采集的数据存在差异，这就要求在数据标准化时，对于不同受试者应该存在差异化处理，不能按照统一标准，而对于同一受试者不同目标识别字符的脑电信号数据，可以进行相同标准的数据标准化处理。

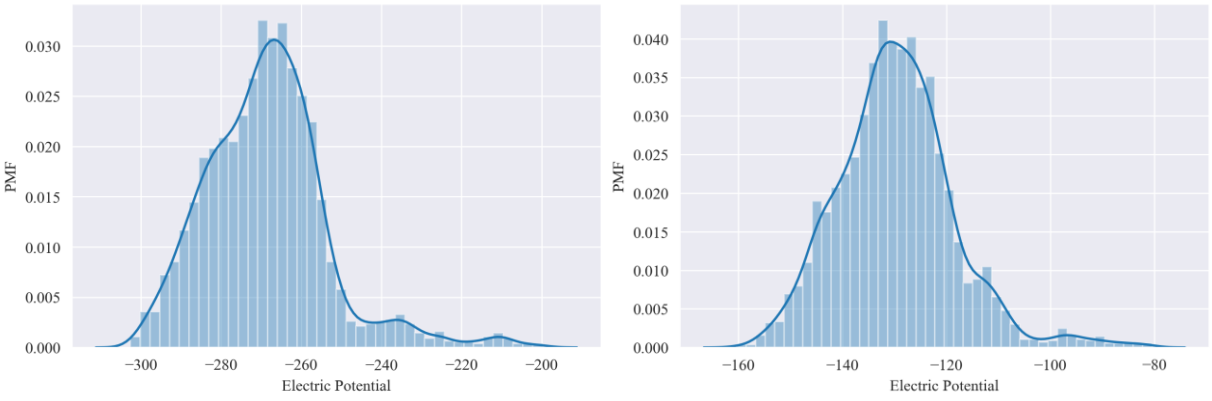

图 4 同一受试者相同通道采集的脑电信号分布图

从脑电数据训练集中随机找一组已知识别字符的通道数据，并用图表展示其趋势，其中已知 P300 电位信号用红点标记，如图 5 所示。发现 P300 电位信号隐藏于很多高频信号中，与图 1 展示的 P300 电位信号的波形有很大出入，显然采集的脑电信号中包含着许多噪声，这些噪声影响整体脑电信号的波形，给数据分析和处理增加了困难，需要在数据预处理部分处理噪声，将数据校正到正常的位置。

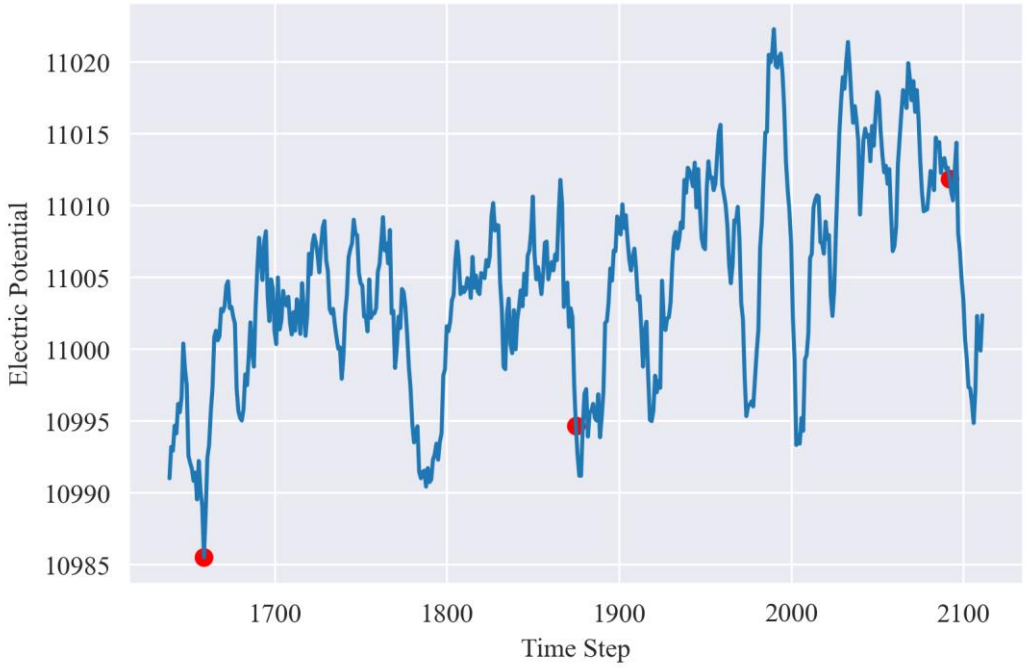

图 5 脑电数据某通道 P300 电位图

4.3 P300 脑-机接口实验数据预处理

4.3.1 小波变换

一般信号数据发生基线漂移的原因分为生理原因和非生理原因，这些原因会影响脑电信号的采集和分析，例如人正常的眨眼、咀嚼、吞咽以及眨眼，这些正常的生理动作会影响脑电信号的采集，还有电极错位、电缆移动以及家用供电的噪音都会给采集的脑电数据增加噪声，导致采集的脑电数据不能直接用于分析。当基线漂移严重时，往往会导致波形识别和参数测量成为不可能，甚至无法记录[1]。基线漂移是一种缓慢变化的低频信号，对数字信号分析会产生不利的影响，需要通过预处理消除信号基线。一般可以使用高通滤波的方法加以消除，另一种思路就是利用曲线拟合对基线漂移进行分段纠正，但当脑电信号质量不高时，插值点难以提取[2]。

为了消除脑电信号的基线漂移，将数据校正到正常的位置，这里采用了小波变换对脑波信号进行去噪，解决信号的基线漂移现象。

小波变换[3]是一种时间-尺度分析方法，具有多分辨分析的特点，在低频部分具有较低的时间分辨率和较高的频率分辨率，在高频部分具有较高的时间分辨率和较低的频率分辨率，很适合分析非平稳的信号和提取信号的局部特征，它是一种窗口大小不变但形状可变，属于一种时域和频域都可以改变的时频局部化分析方法。

小波变换的原理是把基本小波（Mother Wavelet）的函数作位移τ在不同尺度a下，与

待分析信号 $x(t)$ 左内积，即

$$WT(a,\tau)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}x(t)\cdot\psi(\frac{t-\tau}{a})dt \tag{1}$$

其中，$a>0$，称为尺度因子，其作用是对基本小波 $\psi(t)$ 函数作伸缩，τ 反映位移，可正可负。小波变换是将信号分解为一系列小波函数，用一族小波函数表示或逼近信号。而小波变换去噪的原理可以简单阐述为：将信号经小波分解后小波的系数较大，噪声的小波系数较小，并且噪声的小波函数小于信号的小波系数，通过利用多尺度分解后所得到的低频逼近信号充分逼近脑电信号中基线漂移噪声的特性，对其进行均一化处理。

对于小波变换，首先应该选择合适的小波函数，在小波分解中，小波函数将成为每个水平上构成其分解对象的基本单元[4]。因此，小波函数的选择直接决定了小波变换去除基线漂移的效果。

这里选择 db5 小波函数，db 小波函数全称为 Daubechies 小波，一般简写为 dbN，N 是小波的阶数。dbN 小波具有良好的正则性，即该小波作为稀疏基所引入的光滑误差不容易被察觉，使得信号重构过程比较光滑。而 dbN 小波的特点是随着阶次的增大消失矩阶数越大，其中消失矩越高光滑性就越好，频域的局部化能力就越强。小波函数的有效支撑长度为 2N-1，消失矩为 N。db5 小波波形如图 6 所示。

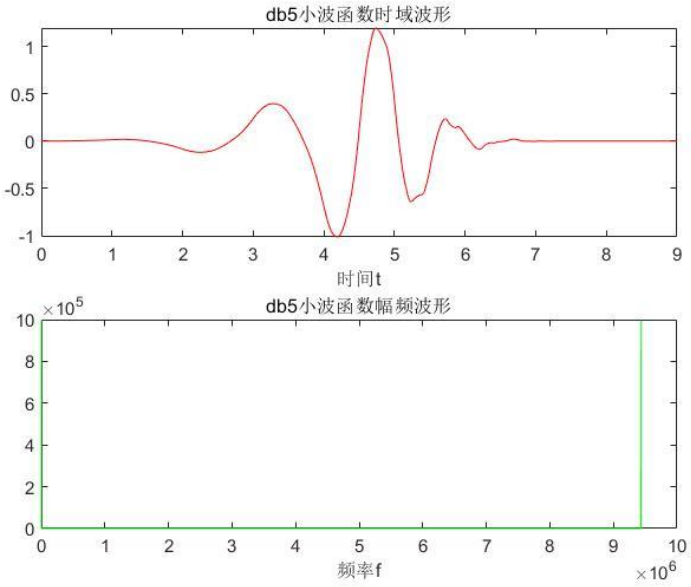

图 6 db5 小波函数波形图

假设 $P(y)=\sum_{k=0}^{N-1}C_k^{N-1+k}y^k$，其中 C_k^{N-1+k} 为二项式的系数，则有

$$|m_0(\omega)|^2=\left(\cos^2\frac{\omega}{2}\right)^N P(\sin^2\frac{\omega}{2}) \tag{2}$$

在式(2)中，有

$$m_0(\omega)=\frac{1}{\sqrt{2}}\sum_{k=0}^{2N-1}h_k e^{-ik\omega} \tag{3}$$

选择适当的小波函数之后，需要选择合适的分解水平 n，这个分解水平的选择关系着脑电信号的分解层次。随着小波分解尺度的增加，时间分辨率的降低，逼近信号所含有的高频信息就会随之减少。当分解到下一个层次时，就会有一些更高的一些频率信息被过滤掉，剩下的部分就逐步逼近基线漂移分量。但当分解水平超过某一程度时，就会同时过滤掉基线漂移中一些频率较高的成分，从而丢失大量基线漂移的信息。因此，选择适当的分解水平，也是去除基线漂移现象的关键。

为了确保选择适合的分解水平能对脑电数据进行分解，这里分别对 n 取 4，5，6，7 四个值，在相同的通道数据下，比较分别应用的处理结果。由下列重构信号图可见，当 n=5 时，通道数据可以获得最好的基线漂移去除结果。

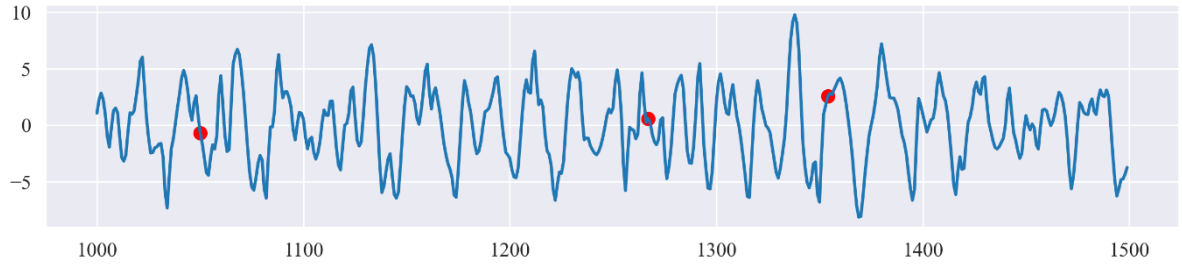

图 7 *n*=4 的重构脑电信号

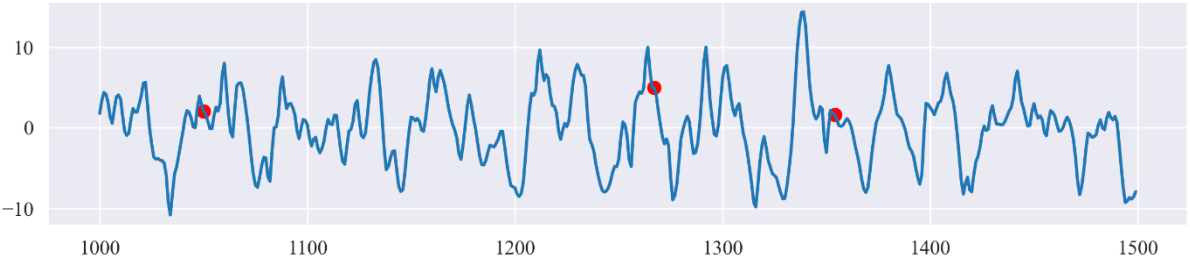

图 8 *n*=5 的重构脑电信号

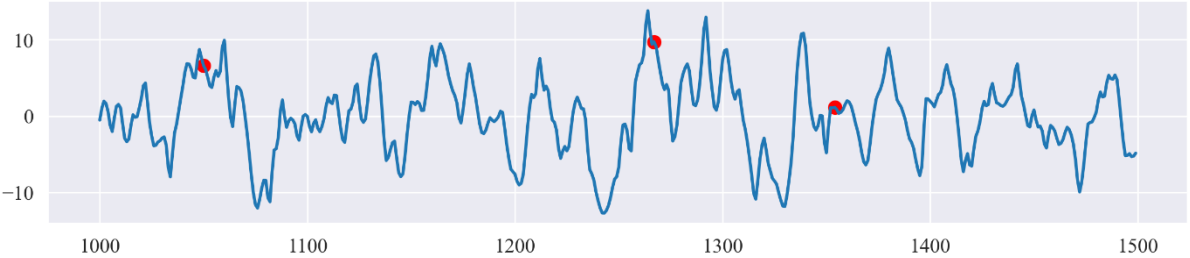

图 9 *n*=6 的重构脑电信号

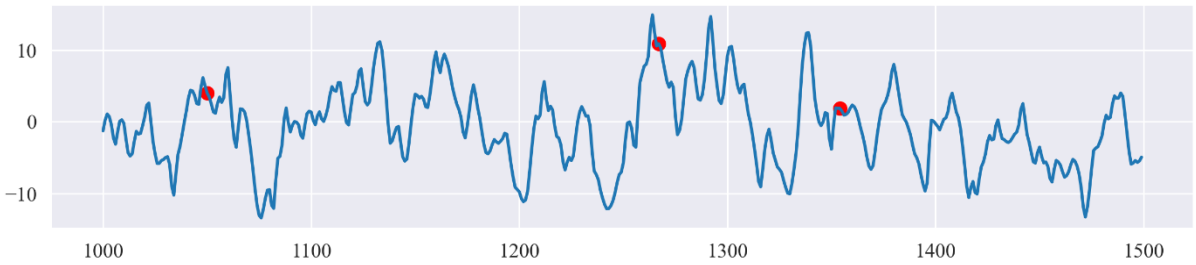

图 10 *n*=7 的重构脑电信号

依据以上对比，在本文中对于分解水平 n 取 5，进行脑电数据的分解，依据以上对比，在本文中对于分解水平 n 取 5，进行脑电信号的分解。在去掉基线低频信号和部分高频的毛刺信号，合成了最终用于分类识别的数据，处理后的信号如图 11 所示。

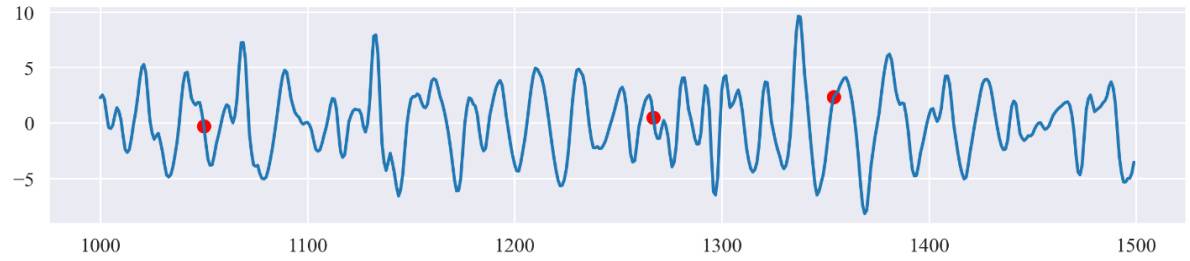

图 11 最终分类识别的信号

图 12 是 db5 小波选取 n=5 的情况下进行小波分解的结果，将原信号分解为 1 个低频分量和 5 个高频分量。 基线漂移主要是信号中的低频分量对信号造成的影响，从 Approximated Component（低频分量）可以明显看出信号有整体呈现上升的趋势。信号中的干扰分为生理信号伪迹干扰，包括 EOG（眨眼等），EMG（咀嚼等），ECG（心电脉冲）；非生理伪迹干扰，包括环境干扰，电极移位，运动伪影等。

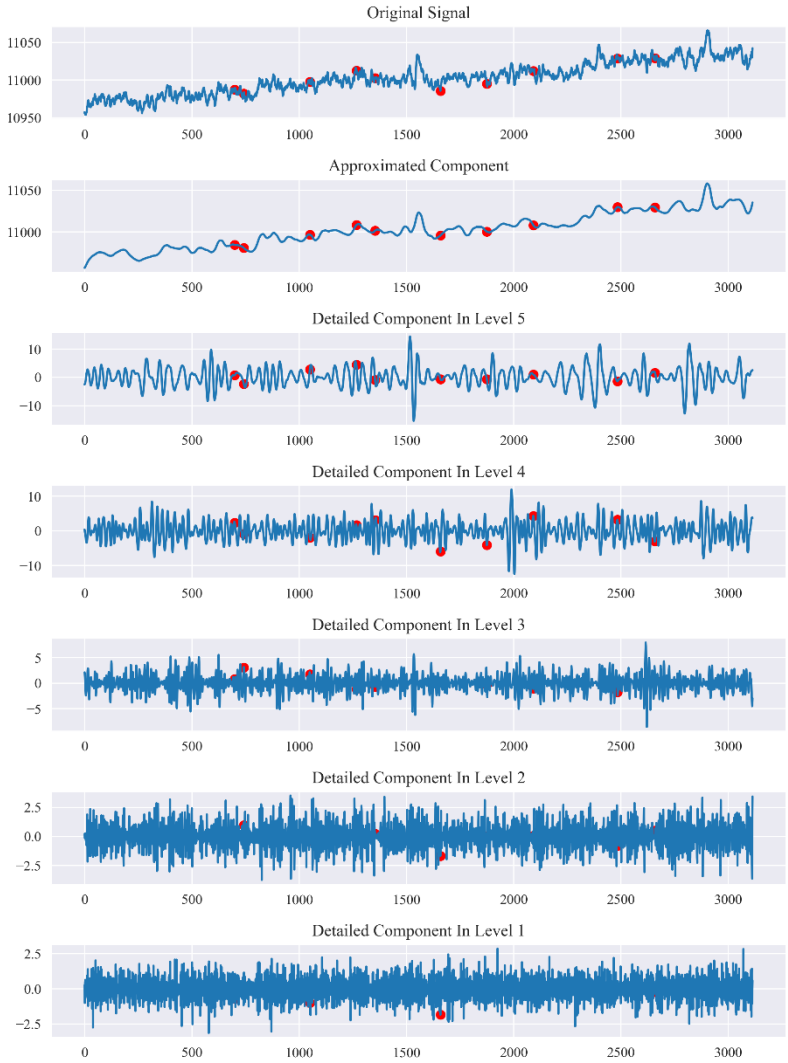

图 12 db5 在 n=5 的分解层次图

4.3.2 数据标准化

由于不同受试者在相同通道识别相同目标字符的电位信号差异很大，同时同一受试者在识别不同字符时电位信号强度比较接近，考虑到数据需要进行标准化，但数据标准化策略的目标是在不改变数据分布的情况下，尽可能将数据缩放到同一尺度，如果所有数据同时做标准化处理就容易改变单人的数据分布，因而这里分别对不同受试者使用相同的传感器采集到的数据进行标准化处理。

最常见的标准化方法是标准差标准化（Zero-meannormalization），其原理如下：

对序列 $x_1, x_2, \cdots, x_n$ 进行变换：

$$y_i = \frac{x_i - \bar{x}}{s} \tag{4}$$

其中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$，$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$，则新序列 $y_1, y_2, \cdots y_n$ 的均值为 0，而方差为 1，

且无量纲。

1. 瞬时数据

本实验中先考虑将该问题视为一个典型的机器学习问题，将采集的 20 个通道的信号视为 20 个不同的特征，通过不同特征来判别该点信号是否为 P300 电位信号。在这种情况下，处理方法比较直观，需要从 train_event 以及 test_event 找出数据记录点与 train_data 以及 test_data 原始采样数据的关系，从中找到相应的行，提取出 20 个分量的特征，分别组成训练集和测试集。

2. 时间序列

考虑到电信号为时间上连续的数据，P300 电位出现在一个时间窗口的信号变化中，这里需要引入时序特征，以保证数据的完备性。同样地，也可以在 train_event 和 test_event 中找到记录点对应信号出现的位置，设定一个时间窗口捕获 P300 信号上下文是信息。在这种方式下，核心之处在于时间窗口的选取。据相关研究，P300 电位信号是在施加特定刺激后经过大约 300ms 的潜伏期之后出现的正向波。由于采集频率为 250Hz，即每 4ms 采集数据，P300 信号时序窗口应该大致在刺激出现后 75~112 时间步内，由于信号出现时间有所区别，这里利用采集刺激点之后 50~125 时间步内的数据，组成该刺激点时间序列。此时每个样本的大小为（75，20），表示刺激点之后 200ms 到 500ms 的电位变化。

根据 train_event，test_event 的映射关系，一共提取出训练样本 3960 个，其中负样本（非 P300 刺激点）3360 个，正样本（P300 刺激点）600 个，测试样本 2880 个。

4.4 模型构建

在本文中，基于数据预处理结果构建了多个模型，按照预处理方式不同，可以分成两类，以下对模型的构建做简单的阐释。

4.4.1 时序模型

基于双向 LSTM（Long short-term Memory）构建了堆叠双向循环神经网络模型（Recurrent Neural Networks, RNN），如图 13 所示，这主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。RNN 跟传统神经网络最大的区别在于每次都会将前一次的输出结果，带到下一次的隐藏层中一起训练，因此很适合处理序列数据。另外此模型中考虑了 P300 的尖峰信号，其前后的电信号上下文能给模型带来更多的信息，因此可以使用双向的循环神经网络以更好地捕获时序上下文的信息。

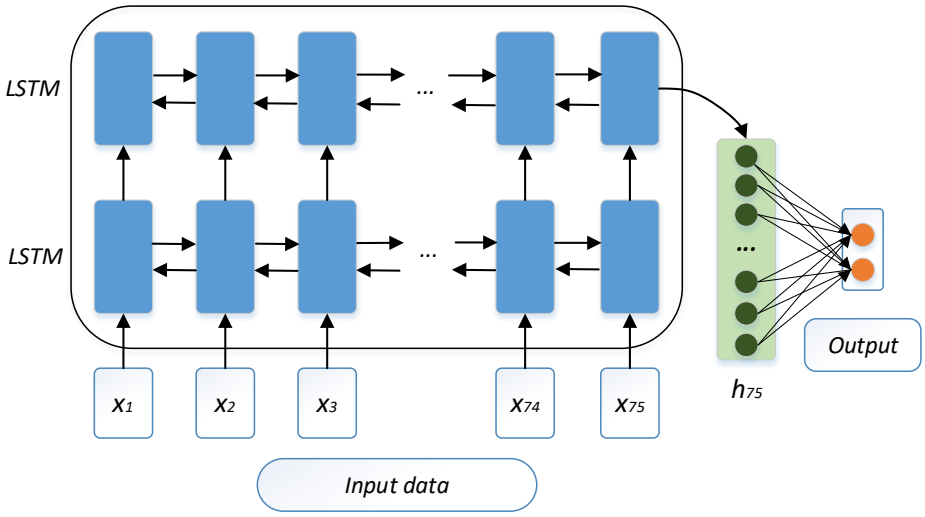

图 13 模型结构图

对于本模型的参数设置如表 1 所示。

表 1 模型参数设置

Layer Name	Input	Output	Parameters
Bidirectional LSTM	(batch, 75, 20)	(batch, 75, 128)	143,360
Feedforward network	(batch, 128)	(batch, 2)	258

4.4.2 瞬时数据模型

随机森林（Random Forest，RF）是 Bagging 的一个扩展变体。Bagging 是并行式集成学习方法最著名的代表，其主要思想是：给定包含 m 个样本的数据集，先随机取出一个样本放入采样集中，再把该样本放回初始数据集，使得下次采样时该样本仍有可能被选中，这样，经过 m 次随机采样操作，可以得到含 m 个样本的采样集，初始训练集中有的样本在采样集中多次出现，有的则从未出现，因此可采样出 T 个含 m 个训练样本的采样集，然后基于每个采样集训练出一个基学习器，再将这些基学习器进行结合，这就是 Bagging 的基本流程。

RF 在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入了随机属性选择。具体来说，传统决策树在选择划分属性时是在当前结点的属性集合（假定有 d 个属性）中选择一个最优属性，而在随机森林中，对基决策树的每个结点，先从该结点的属性集合中随机选择一个包含 k 个属性的子集，然后再从这个子集中选择一个最优属性用于划分。

随机森林简单，容易实现，计算开销小，且往往能展现出强大的性能。并且，与 Bagging 相比，随机森林通常会收敛到更低的泛化误差，其训练效率也往往优于 Bagging。

梯度提升决策树（Gradient Boosting Decision Tree，GBDT）是一种强大的机器学习模型，其主要思想是利用决策树迭代训练以得到最优模型，具有训练效果好、不易过拟合等优点。XGBoost 就是一种典型的 GBDT 工具，这是一种基于预排序方法的决策树算法。首先，对所有特征都按照特征的数值进行预排序。其次，在遍历分割点的时候用 O(N)的代价找到一个特征上的最好分割点。最后，在找到一个特征的最好分割点后，将数据分裂成左

右子节点。但也有明显缺点：首先，空间消耗大。XGBoost 算法需要保存数据的特征值，还保存了特征排序的结果，这就需要消耗训练数据两倍的内存。其次，时间上也有较大的开销，在遍历每一个分割点的时候，都需要进行分裂增益的计算，消耗的代价大。

LightGBM 在保证了准确率的条件下加快了 GBDT 模型的训练速度，进行了如下优化：

1. 采用了直方图算法将遍历样本转变为遍历直方图，极大的降低了时间复杂度；
2. 在训练过程中采用单边梯度算法过滤掉梯度小的样本，减少了大量的计算；
3. 采用了基于 Leaf-wise 算法的增长策略构建树，减少了很多不必要的计算量；
4. 采用优化后的特征并行、数据并行方法加速计算，当数据量非常大的时候还可以采用投票并行的策略；
5. 对缓存也进行了优化，增加了缓存命中率；

但是 LightGBM 同样存在一些缺点：

1. 可能会长出比较深的决策树，产生过拟合；
2. 在寻找最优解时，依据的是最优切分变量，没有将最优解是全部特征的综合这一理念考虑进去。

4.5 分类识别过程

该部分的模型为基本的二分类架构，输出每个样本二分类的概率，即判断每个样本是否为 P300 电位。由于测试数据比较特殊，并不是简单判断 P300 电位。受试者在判断一个目标字符需要经过 5 轮实验，每一轮实验需要进行 12 次随机的屏幕行或列的闪烁，命中目标行时，脑电信号中会出现 P300 电位。因此需要在 12 个测试样本中分别找出行和列概率最大的结果，再依据行列信息得到目标字符，以上推理过程如图 14 所示。

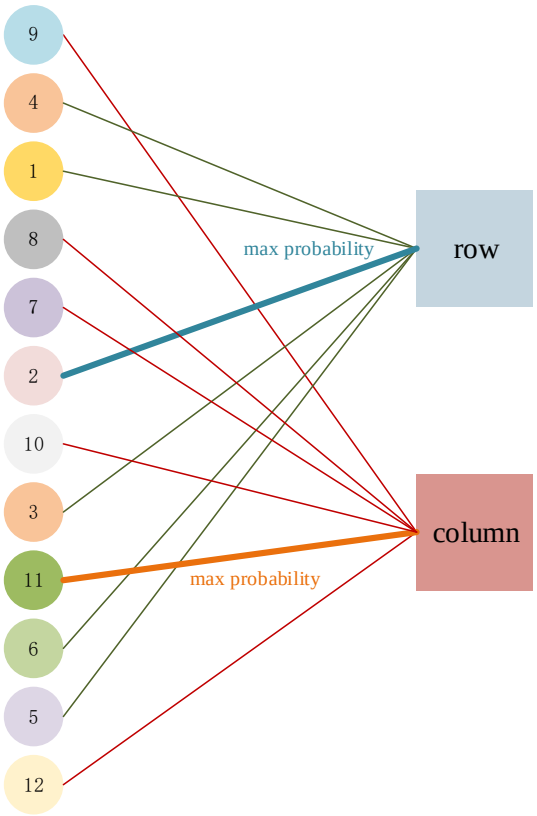

图 14 推理过程图

由于对于每个目标字符的识别，都需要经过完整的 5 轮，也就是在测试数据中每 5 轮的样本，P300 信号存在的行列信息是相同的。为了提高准确率，这里采用投票的方式得到每一轮的结果。

4.6 模型优化

4.6.1 加权损失

经实验发现，给定的数据即使经过小波分解处理仍难以训练，模型容易对标签为 0（无 P300 的信号）的样本产生严重过拟合，极易出现预测全 0 的现象。经推测，这是因为当前数据本身质量一般或者对脑电波信号的分解需要改进，导致模型高维解空间中存在陡峭坡面，因而导致预测全 0 现象。即使使用了多类模型以及交叉熵损失（Cross Entropy Loss），都很难进行训练。为了缓解这个问题，这里使用加权损失（Weighted Loss）来解决这个问题，在 Cross Entropy Loss 的基础上，对两个分类进行加权。加权损失函数如式（5）所示，根据正负样本数量，设定负样本权重为 0.1，正样本权重 0.9。实验结果表明，使用了 Weighted Loss 之后，模型训练变得容易许多，很少出现预测全 0 或全 1 的现象。

$$loss = \sum_{class}(-w_{class} \cdot \log(\frac{e^{y_{class}}}{\sum_{j} e^{y_j}})) \tag{5}$$

其中，w_{class} 表示某个分类，y_{class} 表示对分类的预测分量，y_i 表示对某一分类的预测分量，$\frac{e^{y_{class}}}{\sum_{j} e^{y_j}}$ 表示 *softmax* 函数，用于分类问题。

4.6.2 加权准确率

在分类模型中，通常使用准确率来作为评价指标之一。在神经网络训练中，通常使用验证集的准确率来作为模型选择的策略。但由于数据不平衡以及模型敏感性问题，使用准确率作为模型选择的策略无法正确反映出模型的变动。例如本问题中训练集里的正负样本比例为 1:6，则当模型预测为全 0（全负）时，准确率高达 85%。另一方面，使用准确率作为评估指标时，难以从准确率中得出模型的具体性能。

对于脑电数据中不是 P300 电位的预测值，定义负样本的预测结果为c_0，如下：

$$c_0 = \begin{cases} 1 & ,\hat{y}_0 = y_0 \\ 0 & ,\hat{y}_0 \neq y_0 \end{cases} \tag{6}$$

对于脑电数据中是 P300 电位的预测值，定义正样本的预测结果为c_1，如下：

$$c_1 = \begin{cases} 1 & ,\hat{y}_1 = y_1 \\ 0 & ,\hat{y}_1 \neq y_1 \end{cases} \tag{7}$$

为了解决上述问题，这里决定使用每个样本各自的准确率，在做均值处理作为模型选择的指标，其公式如下：

$$acc_{weighted} = (\frac{\sum c_0}{N_0} + \frac{\sum c_1}{N_1}) / 2 \tag{8}$$

4.6.3 数据增广

脑电数据中一共给出了 180000 个采样样本，但是在 train_event 中，实际记录的只有 3960 个。对于深度学习而言，提高数据量有诸多好处，能减轻过拟合现象，增加模型的通用性，直观上看，这些大量的脑电信号能提供额外的有效信息帮助模型训练。考虑从未记录在 train_event 和 test_event 中的数据中采样，取出更多样本进行训练。然而正如上面提及的不平衡问题，采样的新增数据中并没有正样本，这会进一步加剧过拟合的现象。为了解决这个问题，这里设计了如下随机数据集构建方法，如图 15 所示。

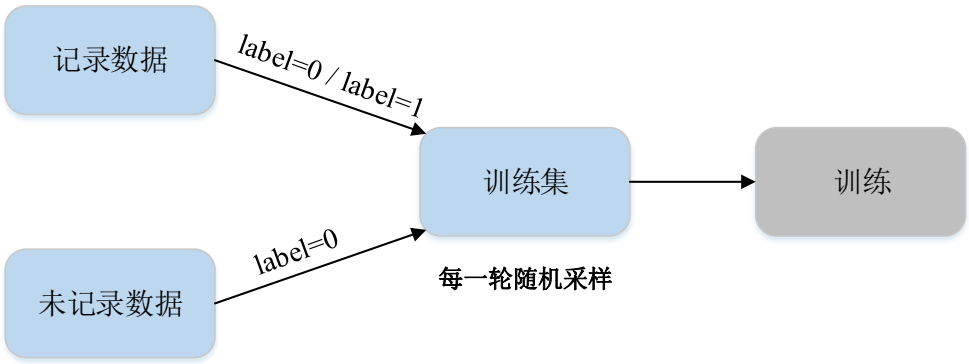

图 15 随机数据集构建方法

在已记录行列的数据中，在对每个 epoch 训练时，从未记录的 180000 个样本点中采集少量样本点，提取各个点上的一段时间窗口的数据作为采样样本，每次随机采样，构建随机数据集。

4.6.4 交叉验证

交叉验证的基本思想是将原始数据进行分组，一部分作为训练集训练模型，另一部分作为测试集评价模型。交叉验证用于评估模型的预测性能，尤其是训练好的模型在新数据上的表现，可以在一定程度上减少过拟合，还可以从有限的数据中获取尽可能多的有效信息。

k 折交叉验证通过对 *k* 个不同分组训练的结果进行平均来减少方差，因此模型的性能对数据的划分就不那么敏感，如图 16 所示。

步骤 1：不重复抽样将原始数据随机分为 *k* 份。

步骤 2：每一次挑选其中 1 份作为测试集，剩余 *k*-1 份作为训练集用于模型训练。

步骤 3：重复第二步 *k* 次，这样每个子集都有一次机会作为测试集，其余机会作为训练集。在每个训练集上训练后得到一个模型，用这个模型在相应的测试集上测试，计算并保存模型的评估指标。

步骤 4：计算 *k* 组测试结果的平均值作为模型精度的估计，并作为当前 *k* 折交叉验证下模型的性能指标。

根据数据集的规模，这里 *k* 取 5，也就是使用 5 折交叉验证。

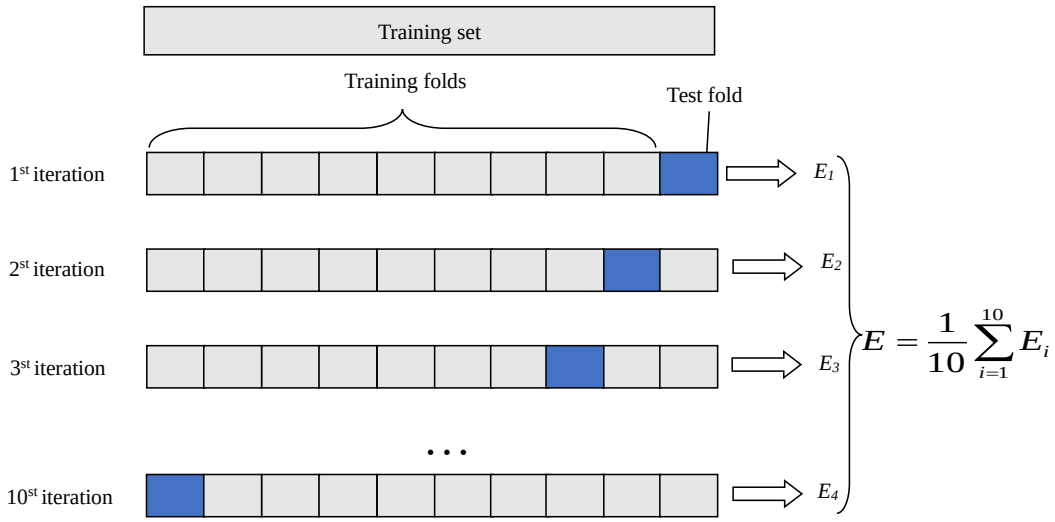

图 16 k 折交叉验证

4.7 模型结果对比

4.7.1 时序数据模型

使用模型构建部分提出的堆叠双向循环神经网络来捕获时间序列特征。在模型选择阶段，主要使用加权准确率（Weighted Accuracy）作为评估指标，由于数据不平衡问题，准确率高低并不代表模型性能。另外，为了防止数据集划分对模型的影响，以及尽可能利用训练数据集（样本数量较少），这里使用交叉验证来评估模型。图 17 中展示了模型优化中多种优化策略的实验结果。

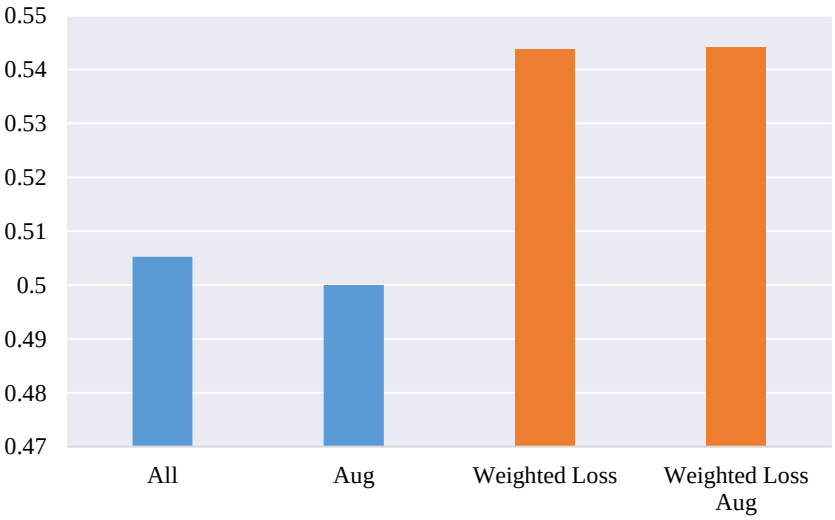

图 17 模型优化中多种优化策略的实验结果

图中使用加权准确率来衡量实验结果的好坏。其中 All 使用所有数据，Aug 使用数据增广部分我们提到的随机数据构建方案，Weighted Loss 是使用所有数据加上加权损失来解

决数据不平衡问题，Weighted Loss Aug 表示在 Aug 的基础上使用 Weighted Loss，意在解决数据不平衡问题的同时尽可能防止过拟合现象。

值得注意的是，All 和 Aug 中都发现了对负样本严重的过拟合现象，Weighted Loss 和 Weighted Loss and Aug 两者则修复了该问题，其预测结果中正负样本的数量如图 18 所示。加权准确率只有当模型在所有分类的准确率都很高的情况下才能得到比较好的性能。由图 18 中，可以看出模型已经不再对负样本过拟合，在问题二的求解中将进一步优化实验结果。

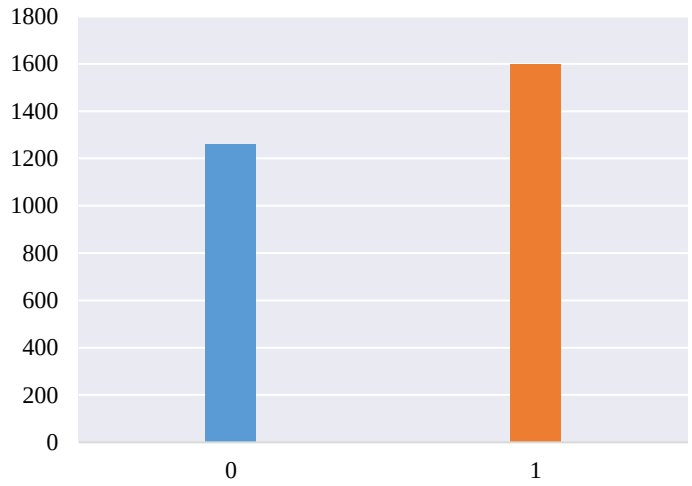

图 18 预测结果的正负样本数量

4.7.2 瞬时数据模型

在瞬时数据中，将该问题看作一个典型的机器学习问题，将 20 个传感器收集到的信号作为 20 个特征，并选用了一些机器模型算法，得到了 86%的准确率。但经分析，模型只是简单地将结果预测为负样本，如图 19 所示。

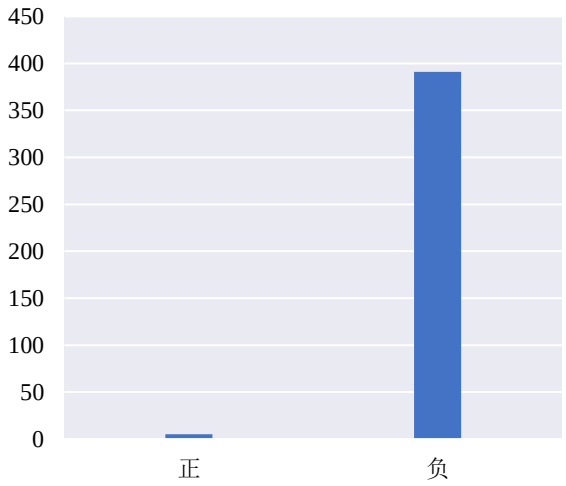

图 19 过拟合预测中正负样本数量

经分析，P300 信号探测不适合用瞬时数据来研究，因为 P300 信号指的是一个时间窗

口内信号变化的一段信号波，而且该信号并不是出现在刺激点处，而是出现在刺激点后若干毫秒内，仅仅使用瞬时的方法并无法捕获到什么特征。从数据可以发现，横向对比同一时间点内的数据，发现数据之间变化很小，所以很难通过瞬时数据之间的相关大小来判别是否为目标点，显然该问题不适合用瞬时数据的模型解决。

表 2 为推理预测的目标字母。

表 2 预测目标字符

Name	Detection Letters
S1	I, 2, 1, X, V, M, H, 4, P, I
S2	0, Y, R, Y, 0, Q, 8, U, F
S3	1, L, 1, U, 4, 4, M, 9, R
S4	O, P, N, Q, Y, 7, W, L, 3, 0
S5	8, S, 3, 9, M, R, P, 4, R, 5

5. 问题二的建模与求解

5.1 问题分析

P300 脑-机接口同时采集多个通道的数据，由于传感器之间位置、硬件、环境等因素，最终采集到大量包含噪声、冗余信息的是数据。大量冗余数据存在会引入大量的无关噪声，使模型过拟合，妨碍训练，影响模型识别 P300 信号。为了去除无用通道，需要将各个通道数据对整个脑电信号分析的重要性具体化。这里使用问题一中构建的深度学习模型，分析各个通道对模型的贡献程度。在数学上，每个位置的梯度可以表示在该点上，即表示求导对象对被求导对象的影响，也就是每个位置对目标的重要性。可以使用问题中训练得到的深度学习模型的目标函数对输入数据进行求导，得到输入数据的显著性图（Saliency Map）。首先应该使用 Saliency Map 对各个通道的数据计算其重要性，用 Saliency Map 生成的像素点亮度表现通道的重要性。另外需要分析 20 个通道之间的相关性，筛选出相关性高的通道，并按照通道的重要性在相关性高的通道中进行选取，从而得出一组最优通道名称组合。为了验证选取的通道组合的最优性，将选取的最优通道组合放在测试集中实验，并用已给的测试数据（char13-char17）的结果进行验证。

5.2 通道数据分析

当利用脑-机接口采集脑电信号时，由于利用了多通道对脑电数据进行采集，在采集的过程中会存在通道间采集信号时互相影响，导致通道采集的数据受到干扰，而这种情况还受到通道之间相互接触，造成信号相互串扰。另外，20 个通道中也会存在一些通道的数据冗余无用，比如双侧乳凸点，又比如眼电通道的数据，显然脑电数据中不需要眼电通道数据的参与。需要将一些通道数据进行剔除，能够对测试数据上获得更好的识别效果。

5.3 通道重要性分析及处理

为了确定每个通道的重要性，进行通道数据的冗余去除，这里使用了显著性映射（Saliency Map）来得出每个通道的重要性。

显著性映射[5]源于一个视觉注意力系统，指的是在视觉处理的背景下图像的独特特征（像素，分辨率等）。这些独特的特征描绘了图像中视觉上引人入胜的位置，而 Saliency Map 可以看作是它们的重要性表示。当在分析一幅图像时，可以重点关注或只关注这些显著的特征和位置。根据 Saliency Map，像素点的亮度表示所在像素点的显著程度，即像素点的亮度与其显著性成正比。可以通过以下例子得到：

$$\begin{aligned} &\{x_1, x_2, \cdots, x_k, \cdots, x_n\} \to \{x_1, x_2, \cdots, x_k + \Delta x, \cdots, x_n\} \\ &y_k \to y_k + \Delta y \end{aligned} \tag{9}$$

对于输入的数据，通过模型的预测得到其输出，利用对应数据正确的标签计算每一个样本的损失值，然后通过损失值对输入数据中不同通道的信号计算其梯度，这个梯度即表示不同通道处的数据变动会对模型的影响，表示不同通道的显著程度。

当一个向量 X 中某个分量的变化，会对 y_k 的改变，等于下面的导数计算：

$$|\frac{\Delta y}{\Delta x}|\rightarrow|\frac{\partial y_k}{\partial x_k}| \tag{10}$$

根据诱发脑电信号的特征：在小概率刺激发生后 300 毫秒范围左右才会出现一个正向波峰，所以将每次采样得到的瞬时数据之后的 200ms~500ms 的数据融合到一起作为有效数据。并且考虑到信号的采集频率为 250Hz，将每 4ms 看作一个时间步，则这些数据可以被分成 75 个时间步。这里使用问题一构建的堆叠双向循环神经网络模型求取输入的梯度，每个点上的梯度的大小代表该位置对模型的贡献程度，也代表模型对输入的每个位置的感兴趣程度。为了求取 20 个通道的重要性，对所有数据的梯度进行累积取均值，利用每个通道上累计梯度大小反映该通道的重要性。

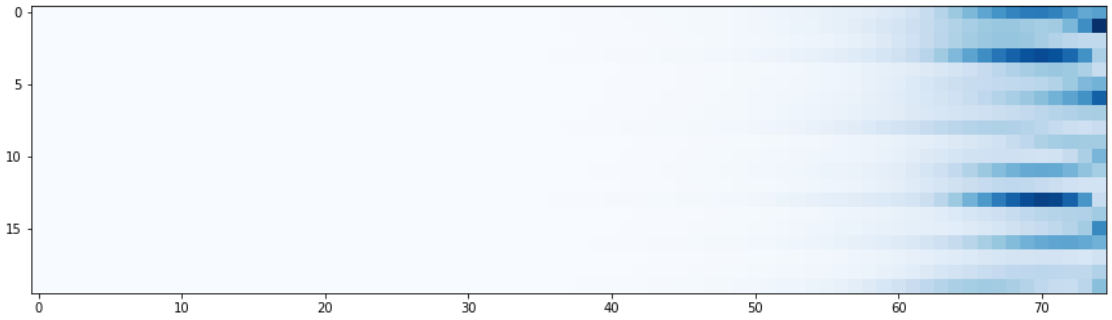

图 20 75 个时间步内不同通道求出的 Saliency Map

图 20 展示了一次测量的 75 个时间步内不同通道求出的 Saliency Map，X 轴输入时序序列 75 个时间步，Y 轴为 20 个通道，图中颜色越深代表该通道对神经网络模型的贡献程度越大，也代表了该通道的重要性。可以看出，前 40 个时间步中的所有 20 个通道的显著性都几乎为 0，而经过 50 个时间步之后，不同通道间的显著性开始出现较大的偏差，这正符合了诱发脑电信号在 300ms 左右出现波峰的特性，这说明了在问题一中构建的模型符合直观观察，模型对于 P300 信号出现的前后时间窗口更感兴趣。

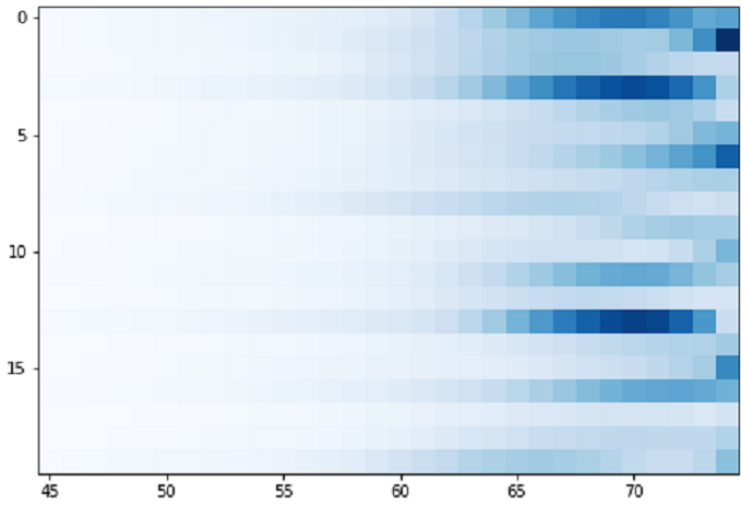

图 21 45-75 时间步的 Saliency Map

为了方便观察，这里截取了图 19 中 45 到 75 时间步的 Saliency Map，从图中可以观察到各个通道的重要性。图 21 中显示出各个通道的显著值。

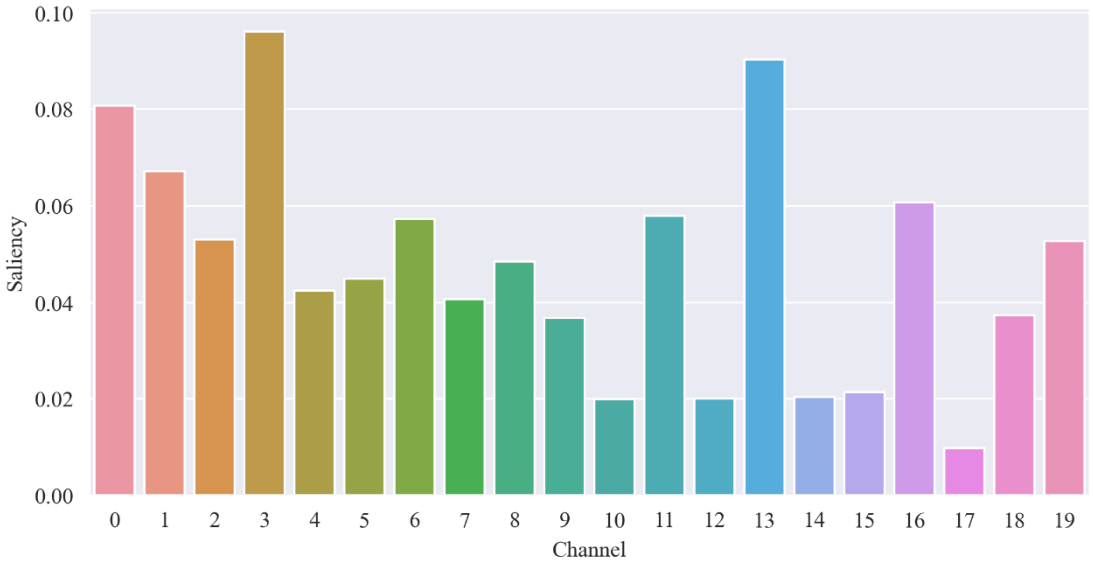

图 22 20 个通道显著值图

最终选取了显著性最高的十二个通道，包含[0, 1, 2, 3, 4, 5, 6, 8, 11, 13, 16, 19]十二个通道，如图 22 所示。

5.4 通道的相关性分析及处理

考虑到 20 个通道中可能存在通道数据之间相互影响，彼此相关的情况，对 20 组通道数据进行相关性分析，找出彼此相关性高的通道，再基于通道的重要性对相关性高的通道进行筛选，从而得出一组最优通道组合。

为了计算出通道间存在的相关性，这里使用相关矩阵来表现通道间的相关性。

相关矩阵也叫相关系数矩阵，是由矩阵各列间的相关系数构成的，有如下定义：

设 $(X_1, X_2, \cdots, X_n)$ 是一个 n 维随机变量，任意的 X_i 和 X_j 的相关系数 $\rho_{ij}, (i, j = 1, 2, 3, \cdots, n)$ 存在，则以 ρ_{ij} 为元素的 n 阶矩阵称为该维随机向量的相关矩阵，并记作 R，如式（11）所示，

$$R = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{bmatrix} \tag{11}$$

相关系数 ρ_{ij} 的计算公式如下：

$$\rho_{ij} = \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{DX_i}\sqrt{DX_j}} \tag{12}$$

其中，协方差的计算公式如下：

$$\mathrm{cov}(X_i, X_j) = E((X_i - E(X_i)) \cdot (X_j - E(X_j))) \tag{13}$$

20 个通道各自与其他的通道之间存在相关性，图 23 展示了 20 个通道间的相关性矩阵。颜色越深代表相关性越弱，由图 22 可以得知，通道 5/6/7 以及通道 16/17/18 之间存在极高的相关性，因此这些通道都可以被各自另外的两个通道代替，这正和上文中选出的[0, 1, 2, 3, 4, 5, 6, 8, 11, 13, 16, 19] 相符。

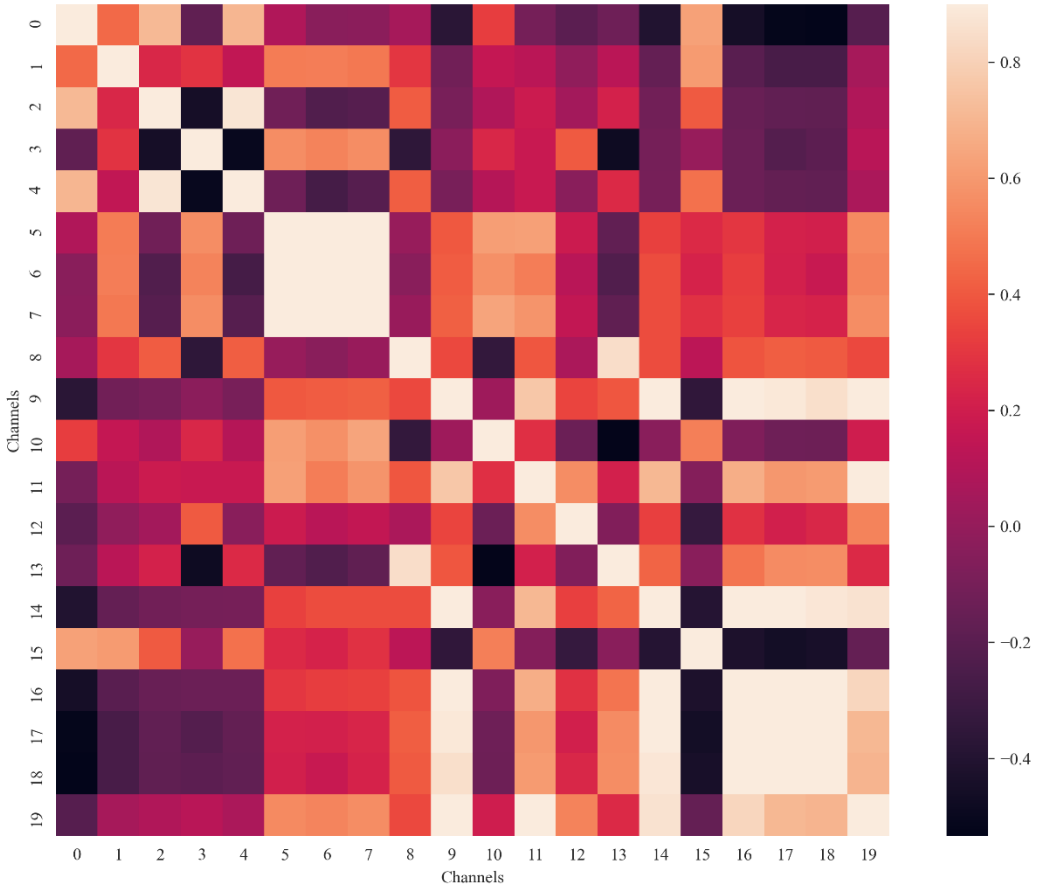

图 23 20 个通道间的相关性矩阵

5.5 推理准确性定义

由于模型目标和任务目标不同，前者为 P300 二分类检测，后者为行列到目标字符的推理。在正常情况下应该以字符命中的准确率作为衡量的标准，但经过实验发现，推理字符需要在行列同时命中的情况下才能显示反映出模型的提升，但这种条件对于模型前期训练过程来说比较苛刻，无法灵敏地反映出模型的变化。所以最终决定执行宽松的准确率评估策略，定义了推理准确率。

这里定义行列预测的结果c，如式（14）所示：

$$c=\begin{cases}1 & ,\hat{y}_{row}=y_{row}\,\|\,y_{col}=\hat{y}_{col}\\ 0 & ,\text{其它}\end{cases} \tag{14}$$

推理准确率的公式如下：

$$InferAcc=(\sum c)/2N \tag{15}$$

5.6 最优通道组合及验证

依据上文中提到的通道选择方法，这里选择了 12 个通道并进行相应的实验，与完整通道的实验数据相比，选择 12 个通道在性能上只有轻微损失，其加权准确率对比如图 24 所示，在问题三部分将会继续优化实验结果。

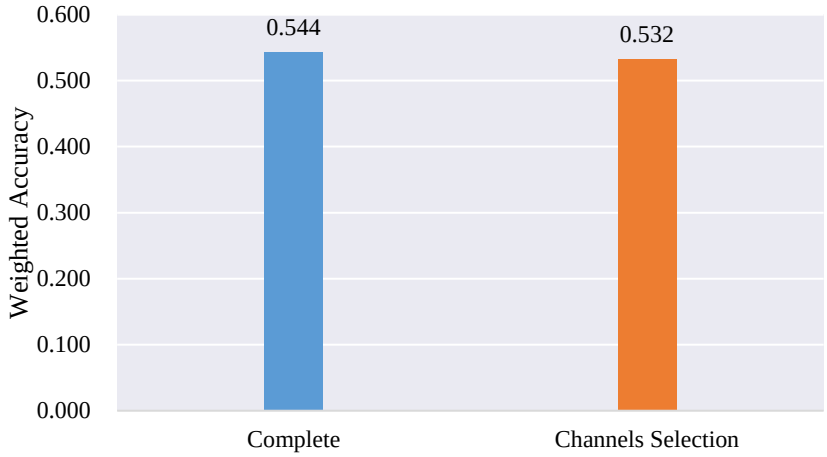

图 24 完整通道选择与部分通道选择的加权准确率对比

依据测试数据（char13-char17）的结果，它们的字符分别是：M、F、5、2、I，根据推理准确率的定义，这里计算出了 12 个通道情况下的推理结果，如图 25 所示。其中展示了完整数据和 12 个通道数据的推理准确率。实验结果将在问题三中进一步优化。

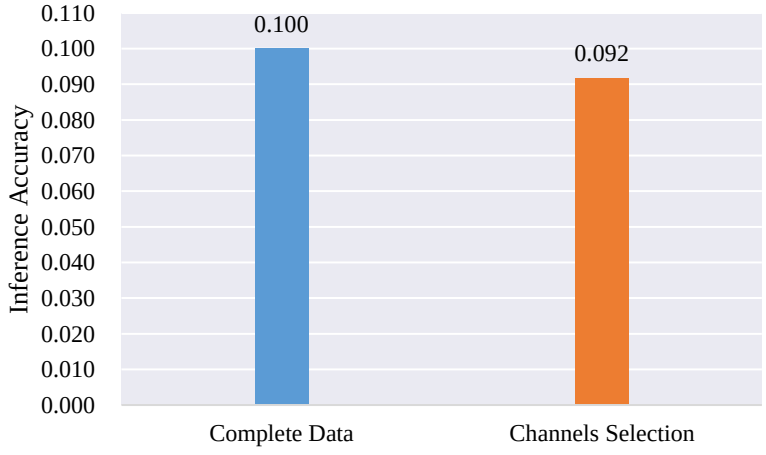

图 25 完整通道选择与部分通道选择的推理准确率对比

表 3 是问题二预测目标字符。

表 3 预测目标字符

Name	Detection Letters
S1	6, 3, 4, V, E, 5, A, 9, 0, 8
S2	D, Q, P, 6, U, 2, L, 8, 6
S3	O,1, B, 7, 3, W, 3, A, Q
S4	4, N, 9, O, 7, A, V, R, P, X
S5	X, U, 8, 7, Y, 5, C, 4, 7, M

6. 问题三的建模与求解

6.1 问题分析

为了满足题目对训练时间的要求，在选择样本数量时，需要尽可能地减少有标签样本的数量，减少训练时间，但鉴于样本数据本身存在样本不平衡的特点，在选择有标签样本时需要考虑出现 P300 电位的样本数量和没有出现 P300 电位的样本数量不能存在明显的不平衡。同时对于最优通道组合，经过问题二的处理，通道数量相较于原始通道数量减少一些，则原始脑电数据中一些冗余的通道数据将不参与模型的训练，训练时间将进一步减少。考虑到题目要求尽可能利用少量的样本数据，结合需要减少训练时间的条件，这里采用半监督学习作为学习方法，最终找出测试其中的其余待识别目标。

6.2 样本数据的选择

在解决问题一时发现，所给出的数据比较难训练，得到的模型非常敏感（Sensitivity），数据样本不平衡和模型的变动都极容易导致预测全 0 或全 1 的现象。这可能也是由于问题本身引起的，长序列中 P300 电位点的捕获与识别相对困难，而且我们还发现，P300 电位的位置可能会有所变动。推测存在某种信号分解处理方法，可以得到鲁棒的数据集。

经过问题二的最优通道选择之后，去除了一些包含冗余信息的通道数据，并且由于每次测量的 75 个时间步内只有 45~75 时间步内的信息有效，因此去除了 0~45 时间步内的数据，这样一来，通过减少了数据维度使模型训练的速度再一次加快。

同时，为了解决样本不平衡的问题，我们采用了欠采样的方法。选择和所有标签为 1 的数据同样规模的标签为 0 的数据，两者相加共占到了大约 30%的数据，这样虽然抛弃了部分样本，可能造成较大的偏差，但我们通过半监督学习训练方法弥补了这一缺陷，同时大幅缩减了模型训练的时间。

6.3 学习方法的设计

考虑到对训练样本的划分，其中 30%的训练样本作为有标签样本，其余的训练样本作为无标签样本，这里使用自训练的半监督学习方法。

自训练的工作原理在于：

步骤 1：将标记的数据实例拆分为训练集和测试集，然后对标记的训练数据训练一个分类算法。

步骤 2：使用经过训练的分类器来预测所有未标记数据实例的类标签，在这些预测的类标签中，正确率最高的被认为是“伪标签”。（其中，所有预测的标签可以同时作为“伪标签”使用，而不考虑概率，“伪标签”数据可以通过预测的置信度进行加权。）

步骤 3：将“伪标记”数据与正确标记的训练数据连接起来，在组合的“伪标记”和正确标记训练数据上重新训练分类器。

步骤 4：使用经过训练的分类器来预测已标记的测试数据实例的类标签，使用选择的度量来评估分类器性能。

重复步骤 1 到 4，直到步骤 2 中的预测类标签不再满足特定的概率阈值，或者直到没

有更多未标记的数据保留。

通过在训练的迭代过程中持续加入新的包含伪标签的数据，减小了数据量不足对模型的影响。对于脑电数据自学习训练的流程图如下：

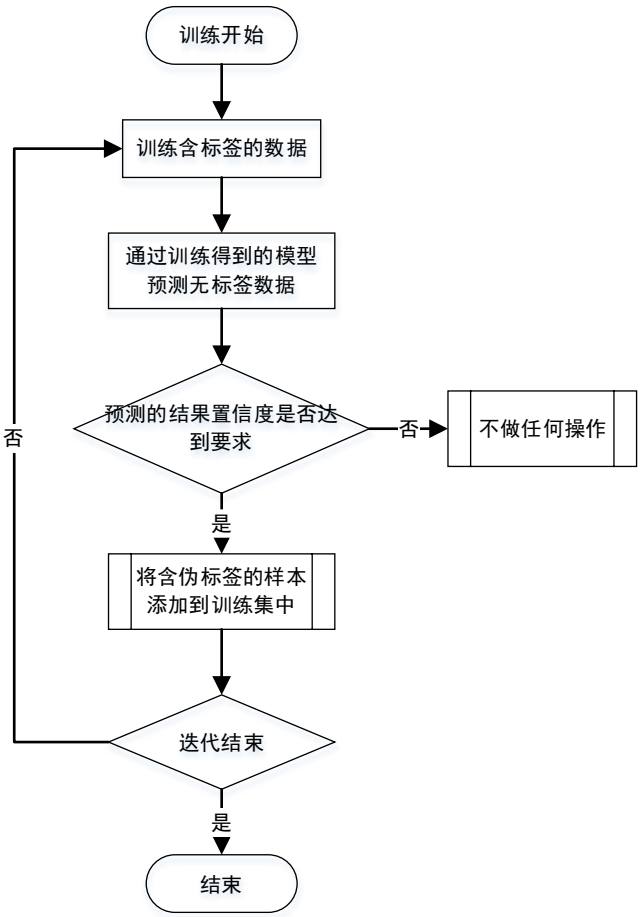

图 26 脑电数据自训练流程图

由图 26 可知，在自训练的过程中，首先训练含标签的脑电数据，将 30%的有标签数据放到模型中训练，通过训练得到的模型预测其余的无标签脑电数据，再从预测的结果中找到置信度达标的伪标签样本，将这些伪标签样本添加到训练集中继续训练无标签的样本数据，依次迭代下去，直到迭代次数已经达到设置的迭代阈值或者已经没有更多未标记的数据保留。

6.4 待识别目标的识别结果

图 27 是不同数据选择下模型训练时间的对比图，其中，Complete Data 是使用完整数据的时间消耗，每个 epoch 耗时 3 s，平均 200 个 epoch 整体训练时间为 600 s。Selection A，是选择 30%样本下的训练时间，每个 epoch 耗时 0.96 s，正好是 30%的时间，平均 200 个 epoch 整体训练时间为 192s。Selection C 使用了 30%样本，结合通道选择和时间步选择，每个 epoch 耗时 0.12 s，平均 200 个 epoch 耗时 24 s。

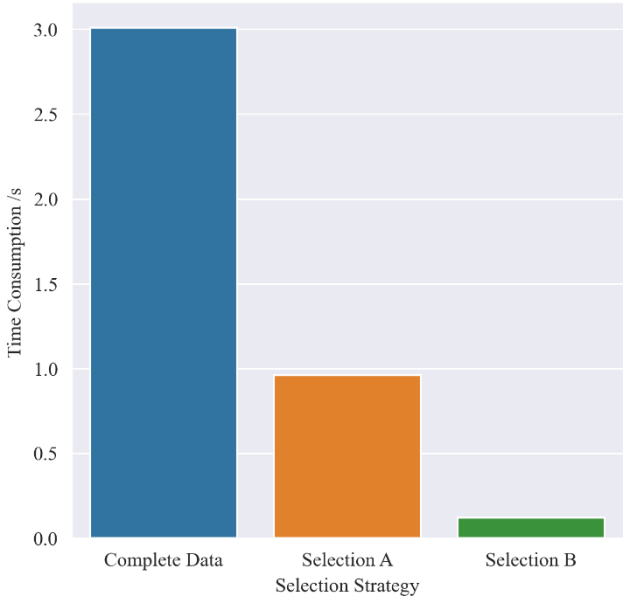

图 27 不同数据选择下模型训练时间的对比图

自训练的过程中，验证集抖动比较剧烈，这是因为训练集数据分布一直在变化，图 28 是自训练过程中训练集和测试集加权准确率变化情况。可以发现，采用自训练，验证集比较不容易发生过拟合的现象，同时，在验证集上的表现更好。

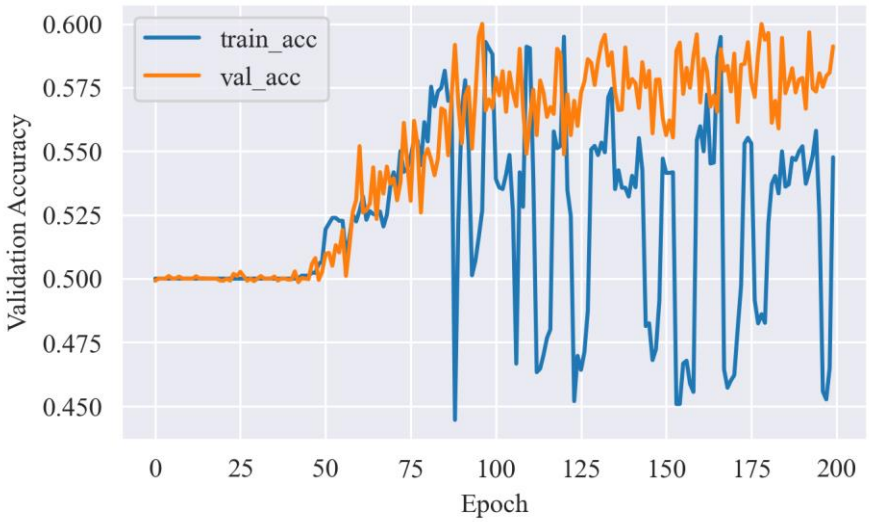

图 28 训练集和测试集加权准确率变化情况

问题一中构建了模型解决分类问题，问题二中对输入通道进行了选择，问题三中基于上述解决方法使用自训练对训练结果进行优化，并且仅使用了 30%的数据。图 29 中综合了问题一到问题三的实验结果：使用自训练方法，以 30%的数据，训练得到加权准确率达到 60.8%。

图 29 问题 1~问题 3 的实验对比

图 30 展示了某次预测的结果，黄色表示行预测正确，绿色表示列预测正确，蓝色表示行和列都预测正确。在此次预测结果中，模型输出的结果分别为 (3, 7)，(3, 8)，(4, 7)，(6, 8)，(6, 11)，分别表示预测的行和列，并且通过对结果的加权平均操作之后真实输出为(3, 7)。而此次真实的目标预测值正是(3, 7)，目标字符为（M），因此判定预测结果正确。

图 30 推理结果示例

由图 31 可见，通过自学习的方法得到的模型的准确率达到 0.104，比使用完整数据得到的结果 0.1 更高。并且，自学习方法仅使用了 30% 的数据，由此分析得出自学习可以在尽可能少的带标签训练样本的基础上得到较高的预测准确率。

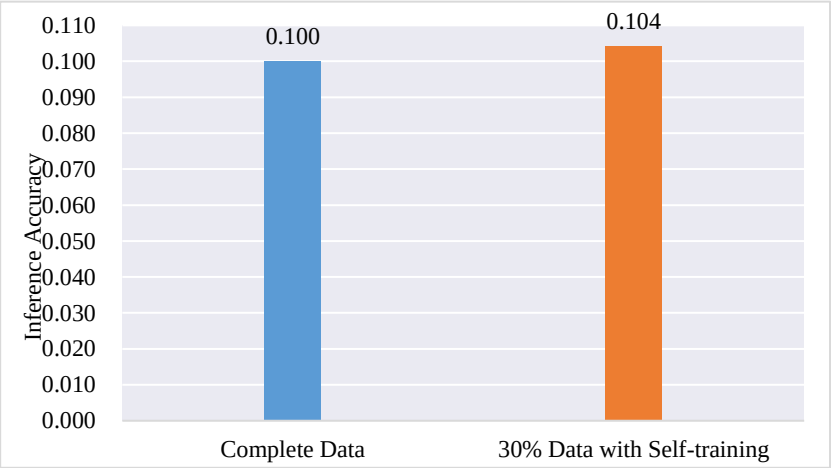

图 31 自训练方法的推理准确率对比

表 4 是问题三预测目标字符。

表 4 预测目标字符

Name	Detection Letters
S1	9, L, Z, R, D, W, R, I, 3, R
S2	G, 9, 2, 0, 7, 5, 4, M, L
S3	1, L, 1, U, 4, 4, M, 9, R
S4	O, F, H, Y, 7, E, A, 0, Z, L
S5	J, 2, Z, R, D, V, M, P, 6, 4

7. 问题四的建模与求解

7.1 问题分析

脑电信号是大脑生理活动的记录，随着睡眠深度的不同，脑电信号会呈现出不同的特点，对睡眠分期具有重要的研究意义和实用价值。睡眠脑电数据集主要分为五个子表，分别是清醒期、快速眼动期、睡眠 I 期、睡眠 II 期以及深睡眠期的睡眠脑电数据。由于睡眠时采集的脑电数据中存在大量干扰信号，因此需要对睡眠脑电数据去除干扰信号。然后应该设计合适的分类器，根据提取到的各项特征将睡眠脑电信号的各个阶段进行有效划分。最后需要对所给的睡眠脑电数据进行合理划分，按照合理比例，利用一小部分充当训练集，另一部分充当测试集，以尽可能少的样本训练并用测试集验证设计的分类器的预测分类效果。

7.2 睡眠脑电数据分析

睡眠脑电数据集主要分为五个子表，分别是清醒期、快速眼动期、睡眠 I 期、睡眠 II 期以及深睡眠期的睡眠脑电数据。每个子表中包含着两列数据，第一列为“已知标签”，其它四列为从原始序列中计算得到的特征参数，依次包括“Alpha”，“Beta”，“Theta”，“Delta”，分别对应了脑电信号在“8-13Hz”，“14-25Hz”，“4-7Hz”和“0.5-4Hz”频率范围内的能量占比。由于睡眠时采集的脑电数据存在大量干扰信号，因此需要对睡眠脑电数据去除干扰信号。

7.3 睡眠脑电数据预处理

从数学角度看，人的大脑可以算是一个非常复杂的非线性系统，并且采集到的脑电信号具有以下三个特点：信号强度微弱，易被噪声干扰；波形不固定，随机特性显著；非平稳，非线性。为得到较为纯净的脑电信号，本部分将对选取的数据进行去噪预处理。

小波变换因具有良好的时频局部化特性和多分辨特性，被广泛的应用到脑电信号的去噪过程中，相关介绍在问题一中已详细阐述。本文选择的小波基为常用于脑电信号的 db4 小波函数，分解水平 n 置为 7。图 32 展示了 alpha 频率的脑电信号在处理前后的对比，经过小波变换之后可以得到其近似分量。

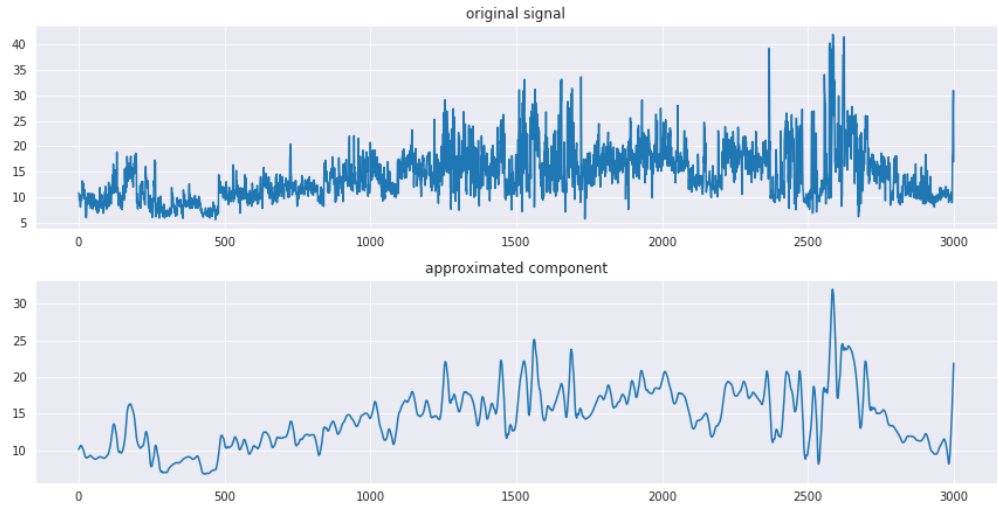

图 32 alpha 频率的脑电信号在小波变换处理前后的波形

从上图的处理结果可知，alpha 频率信号的尖峰和噪声被显著消除，得到了较为纯净的信号，这为接下来的实验提供了特征更明显的脑电信号数据，同时也从侧面说明小波变换在对脑电信号的处理中是十分有效的。

睡眠脑电数据去噪后，需要对数据进行一定形式的处理，首先将睡眠脑电数据提取成训练所需的形式：（x, y），其中，x 表示每个样本的 alpha、beta、theta、delta 的能量占比，y 表示该样本对应的睡眠分期类别。通过离散小波变换方法对数据进行降噪处理，取分解之后的低频近似分量作为下一步的数据输入。然后对数据作标准化处理（normalization）。对 alpha、beta、theta、delta 四个频率的信号数据分别进行均值归一化操作，将数据按比例缩放，并映射到均值为 0，标准差为 1 的空间中，提升模型的收敛速度和预测精度。以 alpha 频率的信号为例，图 33 展示了 alpha 频率的信号经过标准化处理后，数据的分布保持不变，但数据被映射到了均值为 0，标准差为 1 的变化范围较小的分布空间中最后按比例切分不同规模的训练集和测试集，训练集用于学习模型中的超参数，测试集用于测试模型的性能。

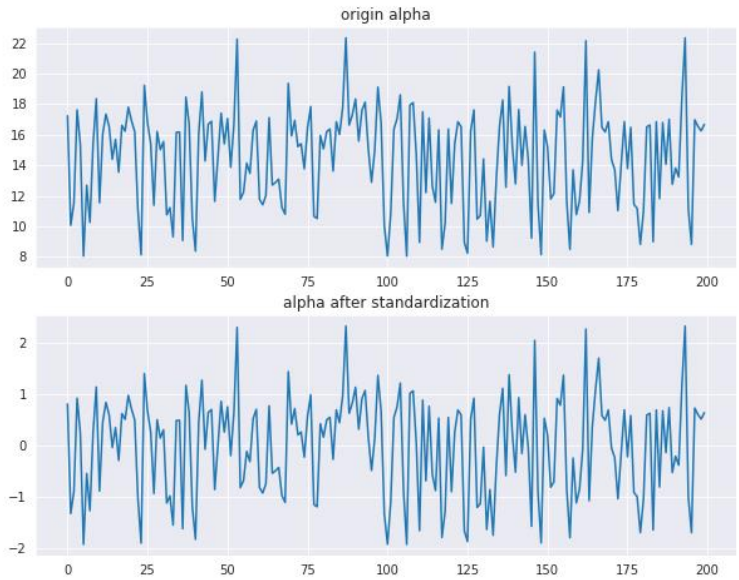

图 33 alpha 频率的信号在标准化前后的分布

7.4 自动睡眠分期模型的建立与实现

在睡眠分期领域中可供分类的方法众多，主要以 BP 神经网络和支持向量机（Support Vector Machine, SVM）为主，可以通过对这些经典的方法进行改进来提高睡眠分期准确率。

本文中对分类器模型的建立如下：

首先将处理完成的训练集作为模型的输入喂入模型中，模型将不断学习并尝试拟合训练集的非线性特征，训练完成后对测试集进行预测得出结果。

为了实现睡眠分期，这里采用了支持向量机实现分类。支持向量机是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可以形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题，支持向量机的学习算法是求解凸二次规划的最优化算法[6]。

支持向量机的学习算法具体如下：

对于训练样本 $\{x_i, y_i\}, i=1,2,3,\cdots,l, x\in R^d, y_i\in\{-1,1\}$，有超平面 $\omega x+b=0,(\omega\in R, b\in R)$，引入拉格朗日函数：

$$L=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{l}\alpha_i(y_i(\omega\cdot x_i+b)-1) \tag{16}$$

经求解之后，分类决策函数可以表示为：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}\alpha_i\cdot y_i(x_i\cdot x)+b^*)b^* \tag{17}$$

而对于非线性决策函数可以表示为：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}\alpha_i\cdot y_i K(x_i,y_i)+b^*)b^* \tag{18}$$

而当前构造 SVM 多分类器的方法主要有两类：一类是直接法，直接在目标函数上进行修改，将多个分类面的参数求解合并到一个最优化问题中，通过求解该最优化问题“一次性”实现多类分类。另一类是间接法，主要是通过组合多个二分类器来实现多分类器的构造，常见的方法有一对一（one-versus-one）和一对多（one-versus-rest）两种。

本文采用一对多的方法，训练时依次把某个类别的样本归为一类,其他剩余的样本归为另一类，这样 5 个类别的样本就构造出了 5 个 SVM。分类时将未知样本分类为具有最大分类函数值的那一类即可。

为了提高分类的性能，又对模型进行了修改，使用基于 GBDT 的 XGBoost 算法。XGBoost 是一种高效的决策树算法[7]，其中 Boosting 是一种常用的统计学习方法，应用广泛且有效，在分类问题中通过改变训练样本的权重来学习多个分类器，并将这些分类器进行线性组合，提高分类的性能。它的主要思想是将多个专家对同一个复杂任务的判断进行适当地综合，综合出的判断往往要比其中任一个专家单独的判断好。

这里通过将处理好的 numpy 格式的数据转换为 DMatrix 格式的数据集以供训练，并设置 XGBoost 的主要参数：分类类别为 5，树的最大深度为 12，剪枝概率为 0.1，学习率为 0.001，随机种子为 1，迭代次数为 10。

这里设计了一种基于 BP 神经网络的深度学习模型，模型的主体包括三层全连接层神

经网络，正则化方法采用 dropout，激活函数选择 relu 函数。通过 pytorch 深度学习框架提供的一些高阶 API 快速处理数据和搭建模型，并设置模型的主要参数：最大迭代次数为 100，隐藏层维度为 128，dropout 比率为 0.5，学习率 0.0003，随机种子为 1。

最后还实验了自训练在分类器模型上的效果。模型主体采用上面提到的三层全连接层神经网络，在训练集上训练完成之后，将测试集的数据（不含标签）作为输入进行预测，通过 softmax 函数得到 5 个睡眠分期类别各自的概率，筛选不小于预先设定的阈值的概率选项，并选择概率最大值所在的分期类别设置为这个样本的伪标签，在下一轮训练时将这些含伪标签的数据添加到训练集中继续训练，直到迭代结束。经过一些调参工作后设置模型的主要参数：阈值为 0.92，最大迭代次数为 5。

上述模型主体涉及到半监督学习，这里作简要介绍。

半监督学习（Semi-supervised Learning，SSL）是将监督学习（Supervised Learning）与无监督学习（Unsupervised Learning）相结合的一种学习方法。半监督学习使用大量的未标记数据，且同时使用标记数据，来进行模式识别工作。当前大部分 SSL 利用的数据是无噪声干扰的，而且依赖的基本假设没有充分考虑在噪声干扰下无类标签数据分布的不确定性以及复杂性，不过在实际应用中通常难以得到无噪声数据。半监督学习总体结构图 34 所示。

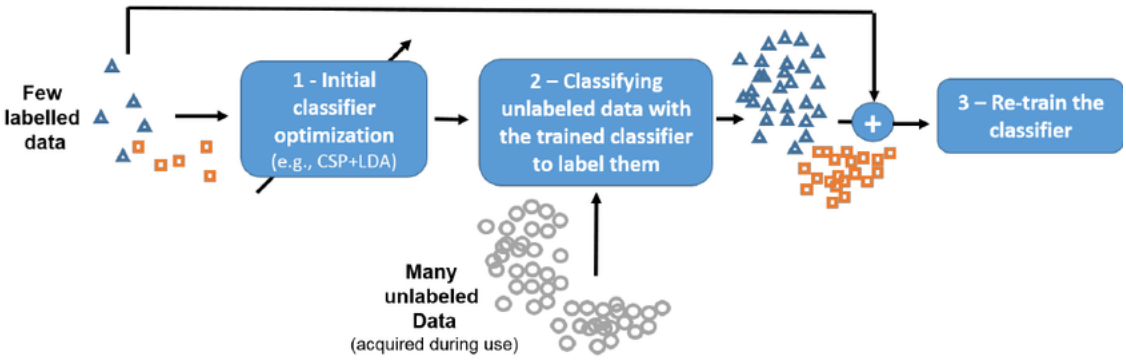

图 34 半监督学习

7.5 睡眠分期模型的测试

在构建了上述分类器之后，对比分析小波变换以及不同的训练集/测试集分配比例对模型性能的影响。

首先分析小波变换对模型性能的影响，在训练集分配比例为 0.3，即取 900 条数据训练的情况下，未经过和经过小波变换的数据训练得到的准确率对比见下表 1。从表 5 中可以看出，小波变换对模型性能有很大的影响，可以将平均预测准确率提升大概 20%~30%。

表 5 小波变换对模型性能的影响

	SVM	XGBoosting	BP 神经网络	半监督学习
未经过小波变换预处理	69.0 %	66.8 %	69.8 %	71.1 %
经过小波变换预处理	90.8 %	98.6 %	92.1 %	99.2 %

为了确定不同的训练集/测试集的分配比例对模型性能的影响，将分配比例设置为从 10%~90%分成 15 个间隔，对于不同的分配比例下的训练集训练得到的准确率对比见下图。从图 35 中可以看出，对于三种模型，即使只分配全部数据的 10% 给训练集，依然能得到

86% 以上的准确率。并且对于半监督学习来说，分配 20%~90% 的数据对模型性能的影响很小，说明半监督学习可以在尽可能少的训练样本的基础上得到较高的预测准确率。

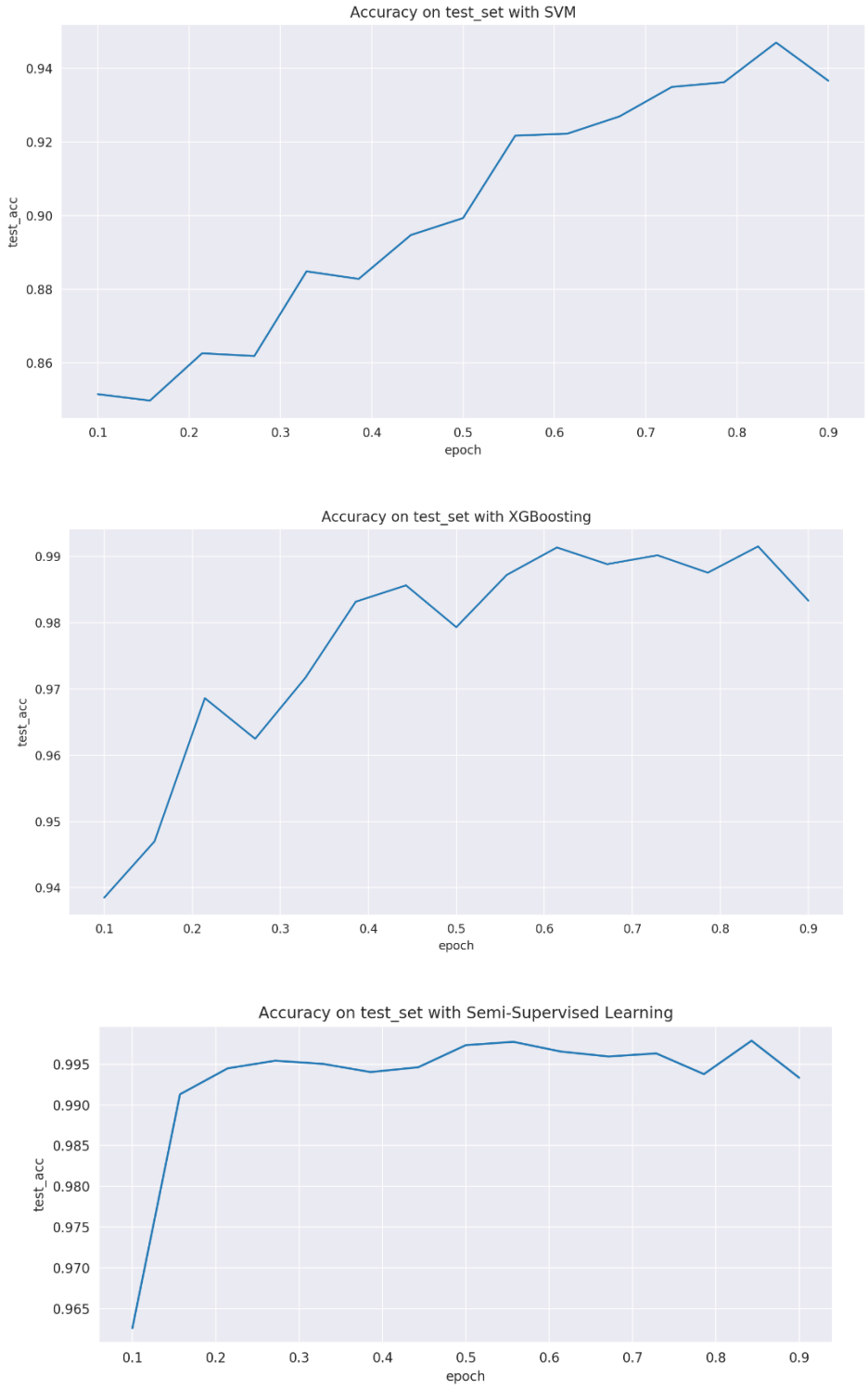

图 35 不同的训练集/测试集分配比例对模型性能的影响

经过实验比较，可以得出 XGBoost 和 BP 神经网络有着近似的分类效果。这是因为在少样本的训练集下，神经网络并不能发挥出能够自动提取特征的能力；相反的，其训练时

间要略大于训练 SVM 和 XGBoost 的时间，而且存在可能出现过拟合的问题。但是，将 SVM、XGBoost 和半监督学习在一起对比时，可以看出，半监督学习弥补了训练集不足的缺陷，同时由于每轮迭代时都会加入不同规模的新的伪标签样本，也在一定程度上抑制了过拟合问题。由图 35 可知，当分配全部数据的 20% 给训练集时，半监督学习训练得到的模型已经能达到 99%以上的准确率，而 SVM 和 XGBoost 则只能达到 89%/95% 左右的准确率，这正说明半监督学习可以在尽可能少的训练样本的基础上得到较高的预测准确率。

8. 模型评价和未来展望

8.1 问题一

在问题一利用小波分解，取出了信号中的低频基线分量以及噪声分量，并设计堆叠双向循环神经网络模型对 P300 信号的进行识别；为了解决数据不平衡导致模型难以训练的问题，使用了一系列策略优化模型：加权损失（Weighted Loss）解决模型极易对负样本过拟合的问题，加权准确率（Weighted Accuracy）替代准确率用于模型选择，利用数据增广构建随机数据集减轻过拟合现象，最后使用交叉验证（Cross Validation）作为评价模型的策略，更好地利用了训练集的数据。

8.2 问题二

在问题二中使用显著性图（Saliency Map）来分析 20 个通道的重要性，并分析通道之间的相关性，验证基于显著性图的通道选择方法的合理性。本文选用了 12 个通道进行实验，实验结果表明，减少部分通道并不会明显降低模型表现。

8.3 问题三

基于问题一的基本模型构建和问题二的通道选择，在问题三中，选用了 30%的标签数据，减低了数据的维度，使得单个轮次训练时间从 3.01 s 下降到 0.12s。训练中，本文基于半监督学习，应用自训练（Self-training）的方法对模型进行训练，对问题一和问题二的结果进行优化。最终在仅仅使用 30%带标签数据的情况下，加权准确率从 54.4%提升到 60.8%，推理准确率（Inference Accuracy）从 9.1%提升到 10.4%。

8.4 问题四

在问题四中使用了小波分析处理了脑电数据，并对比分析了 SVM，XGBoosting，BP 神经网络，半监督学习（自训练），在仅仅使用 30%的数据的情况下准确率达到了 99.1%。值得一提的是，即使使用 10%的数据，依然可以达到 86%的准确率。

8.5 未来展望

针对本赛题中，本文还发现一些值得改进的方向，由于时间原因无法一一实现。首先，脑电波信号处理方面需要更多专业知识对信号分解做进一步的优化，从而更好识别 P300 信号。其次，P300 信号在时域上出现的时间容易变动，可以使用连续小波变换得到时频图像，在时频图像上应用深度学习图像识别技术来识别 P300 信号。最后，本题中关注无标签数据的训练问题，可以使用半监督、无监督学习的方法来处理无标签数据，比如自编码模型（Auto-Encoder），孪生网络等模型进行处理。

参考文献

[1] 朱洪俊. 心电信号深层识别机理的研究与虚拟式心电图仪的研制[D]. 重庆大学.
[2] 张文琼,刘肖琳,吴涛.一种利用小波变换逼近信号滤除心电图基线漂移的方法[J].计算机工程与应用,2005,41(20):222-224.
[3] Yong L, Shengxun Z. Apply wavelet transform to analyse EEG signal[C]//Proceedings of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 1996, 3: 1007-1008.
[4] 胡昌华. 基于 MATLAB 的系统分析与设计:小波分析[M]. 西安电子科技大学出版社, 1999.
[5] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on pattern analysis and machine intelligence, 1998, 20(11): 1254-1259.
[6] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
[7] Chen Tianqi, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA: ACM, 2016: 785-794.