



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		上海理工大学
参赛队号		20102520034
队员姓名	1.	彭臣
	2.	孙乾宇
	3.	张汉祖

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：

随着计算机技术的不断发展，近年来脑-机接口一直是被广泛热议的话题之一。将脑-机接口技术与人工智能相结合，使得脑-机接口技术在脑部疾病治愈、娱乐消费、机器控制等领域得到广泛应用，其中关于脑电信号的采集和处理在脑-机接口技术中至关重要。脑电信号按其产生的方式可分为诱发脑电信号和自发脑电信号，其中 P300 电位是一种典型的诱发型脑电信号，这种信号在受外界刺激时诱发大脑产生；而自发脑电信号通常在人睡眠没受外界刺激时产生。对于 P300 信号，其往往出现在刺激后的 300 毫秒左右，此时脑电信号曲线会出现一个正向的波峰。而睡眠脑电信号会在人睡眠的不同阶段呈现出差异化较大的脑电信号波动状态，所以可以根据一个人的睡眠脑电信号来划分出睡眠的状态。本文通过对 5 位被试者的训练数据进行处理分析，通过多种分类算法比较找出 5 个被试测试机中的 10 个待识别目标；根据训练数据和测试数据集运用主成分分析法得到适用于所有被试者的一组最优通道组合；通过选择适量的样本作为有标签样本，其余训练数据作为无标签样本，通过设计基于 SVM 的半监督学习算法在以达到较高的训练效率，并通过问题二的测试数据验证算法的可行性；最后采用支持向量机，K 近邻模型，一维卷积神经网络三类方法进行对比选优，设计了睡眠分期预测模型对特征样本进行预测，结果表明一维卷积能够很好的保留脑电信号的时序特征，因此准确率最高。

（1）数据处理：首先对给的样本数据进行预处理，预处理主要使用带通录波器在滤波范围为 0.2-8Hz 范围内实现滤波，滤波后能够去粗高频噪声，提高数据的信噪比。然后对数据进行单次分割和组块，其目的是将数据中包含 P300 脑电信号的数据部门进行平均划分，以满足分类的准确率，提高脑-机接口系统信息传输效率。同时为了提高模型的训练速度，要对信号片段进行降采样。最后在对数据进行规范化处理，以提高脑-机接口数据的利用效果，降低数据计算成本，使得数据应用于不同算法训练学习速度都有较大提高。

（2）针对问题一，问题一主要目的是找出被试测试集中的 10 个待识别目标，首先对数据预处理，然后设置模型分类评估指标，选用支持向量机、随机森林、一维卷积神经网络对模型进行预测得出相应结果。通过比较三种算法的准确率得出一维卷积神经网络的预测准确率最高。

（3）针对问题二，通过主成分分析法，对数据进行降维处理，去除脑电数据信号中的冗余和无关信息，对测试通道进行筛选，去除对 P300 信号响应不明显的通道，保留对 P300 响应效果好的通道。最后，通过算法自动学习所有通道对应权重大小，由最后各通道权重占比选择最优通道子集。

（4）针对问题三，根据附件 1 中所给数据，选择一定样本数据作为有标签样本，然后将剩下的数据作为无标签训练样本。根据问题二中所选择的最优通道组合，设计了基于自训练半监督方法的脑电信号分类算法，通过部分测试数据检验了该算法的有效性，并利

用该学习方法找出了剩余测试集中的待识别目标。

（5）针对问题四，本文首先采用带通滤波对睡眠脑电信号进行处理，然后对样本数据进行划分，选择部分数据作为训练数据，采用支持向量机、K 近邻模型、一维卷积神经网络三种算法对数据进行训练，最后对剩余样本进行预测，通过各种算法准确率对比发现采用一维卷积神经网络算法准确率最好，能够很好地保留脑电信号的时序特征。

关键词：脑电信号；采集通道；睡眠分期；信号预处理；支持向量机；随机森林；自训练半监督方法；K 近邻模型；一维卷积神经网络；

目录

1. 问题重述

1.1 问题背景

脑机接口(brain-computer interface,BCI)是通过脑电信号(electroencephalogram，EEG）建立起大脑和外部设备信息交互的技术，它不依赖于正常由外围神经和肌肉组成的输出通路[1]。对于一些患有肌萎性侧索硬化、脊髓伤害等疾病的患者，他们丧失了通过肌肉向外界传达信息的能力，利用脑机接口设备就能使患者与外部世界进行简单的交流。P300 事件相关电位是一种典型的诱发型脑电信号，是指在小概率刺激事件发生后约 300ms 出现的正向波峰脑电信号,表征的是知觉、注意、思维、理解和记忆等高级神经心理过程的电位变化[2]。快速、准确的提取和识别 P300 脑电信号对于脑机接口应用在医疗、康复等领域具有十分重要的意义。

睡眠是一项非常重要的生命过程，睡眠质量与人体的健康息息相关，睡眠分期能分析和监控睡眠状态，有利于预防和治疗睡眠相关的疾病。2007 年，美国睡眠医学学会（American Academy of Sleep Medicine，AASM）将睡眠时相划分为五期[3]，包括清醒期（wakefulness）、非快速眼动期 I 期（non-rapid eye movement I）、非快速眼动期 II 期（non-rapid eye movement II）、非快速眼动期 III 期（non-rapid eye movement III）和快速眼动期（rapid eye movement）。睡眠过程中采集到的脑电信号是一种自发型脑电信号，基于脑电信号进行睡眠的自动分期，能有效解决人工分期的繁琐耗时等问题和提高睡眠分期效果，对于改善睡眠质量，诊治睡眠相关疾病，有重要的实际意义。

1.2 问题的提出

问题一：根据 5 名被试者进行 P300 打字机实验的实验数据作为训练样本，找出测试集中的 10 个待识别目标。同时要考虑 P300 脑电信号的分类准确率和计算速度。

问题二：在基于 P300 的 BCI 系统中，如何选出与 P300 特征相关的最优通道子集是决定系统性能的一个关键步骤，一个最优的 P300 相关脑电通道子集不仅可以自动适应受试者个体化差异，且可以提高 BCI 系统的识别效率和识别准确率。本题的 20 个脑电信号采集通道里，存在无关或多余的通道，要找出最优的一组通道。

问题三：选取部分字符数据作为训练，部分字符数据作为测试进行半监督学习。

问题四：依据睡眠状态和脑电信号频率的相关性，将睡眠脑电信号分为以下四种频域无重叠的节律波，分别为 8-13Hz 的α波，14-25hz 的β波，4-7hz 的θ波，0.5-4hz 的δ波。根据已知标签的训练样本集，设计出一个睡眠分期模型。

2. 基本假设

（1）假设采集到的数据真实可靠。

（2）假设采集设备在工作过程中没出问题。

3. 符号说明

符号	符号说明
C	SVM 惩罚函数
ξ	SVM 核参数
Gini(p)	随机森林基尼系数
mg(X,Y)	随机森林边缘函数
Var(a)	方差
Cov(a,b)	协方差
mean_1、mean_2	样本平均值
$d_1(x_i)$、$d_2(x_i)$	欧氏距离
D_L	有标签训练集
D_U	无标签训练集

4. 数据处理

4.1 数据预处理

由于脑电信号是一种微弱低频的生物信号，在实验环境下通过脑-机接口采集诱发性脑电信号易受外界环境因素和生物体自身因素影响。例如周围环境和测量仪器引起的不同频段的噪声干扰信号，以及人体产生的诸如肌电、心电等伪迹生物体信号。为了方便后面问题分析处理，降低干扰信号对实验可靠性的影响，在实际处理问题的时候需要对采集的信号数据进行预处理，去除高频噪声，提高数据的信噪比。

通常 P300 电位一般在小刺激概率发生后 300 毫秒范围左右出现，由于个体间的差异 P300 电位的发生时间也有所不同，但是 P300 电位通常会在特定范围的频段内出现。P300 信号属于低频信号，相关参考文献表明，P300 信号一般最高频率达 8Hz[4]。所以本文利用带通滤波器对采集的信号数据进行滤波降噪，滤波器的滤波范围为 0.2-8Hz。如下图所示选取的为 S1_train_data 中观察字符矩阵中 B 时 FZ 通道的电位信号变化曲线，其中图 1 为滤波前，图 2 为滤波后：

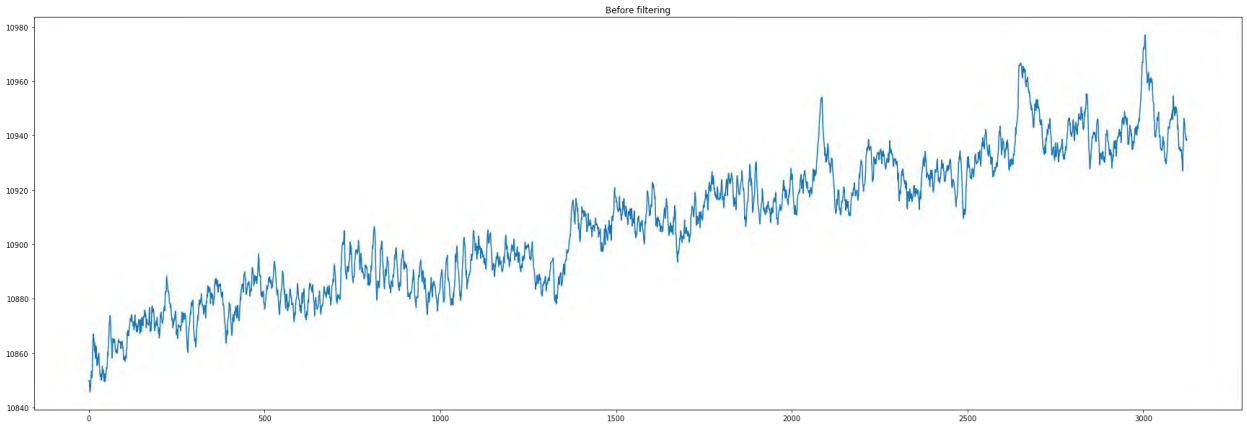


图 1 滤波前

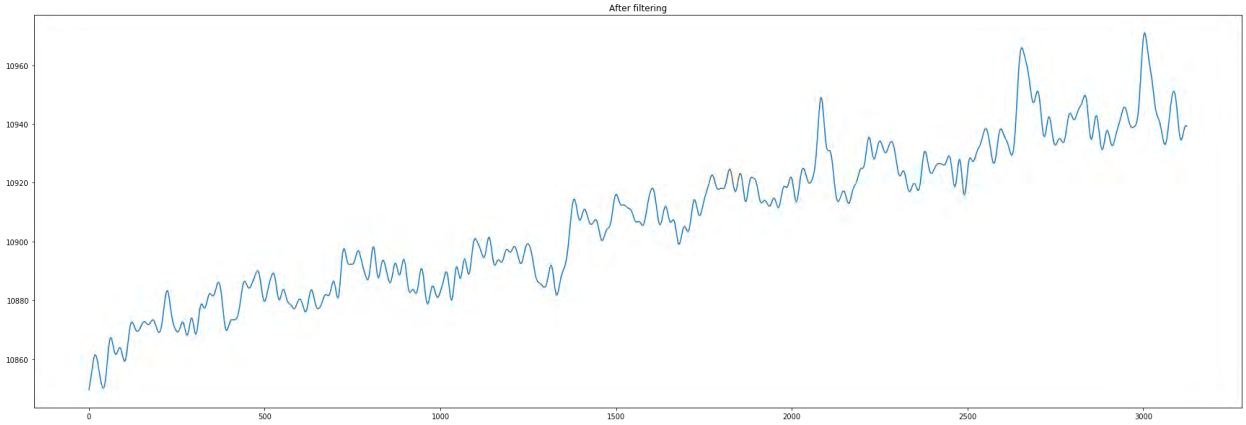
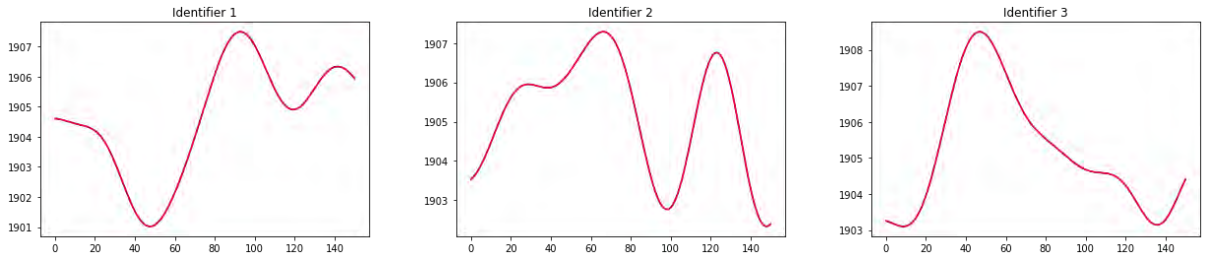


图 2 滤波后

通过对比可以看出滤波后的电位信号变化曲线较滤波前的电位信号曲线更光滑，说明滤波器对原始信号滤波降噪效果好，能够很大程度上消除噪声和伪迹，进而降低对解决后续问题的影响。

4.2 单次分割和组片

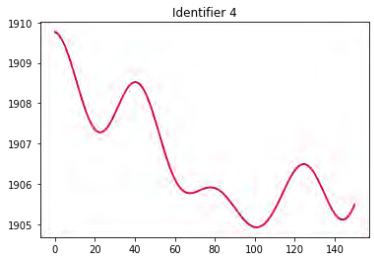
根据附件中的数据，对训练中的数据进行单次切割和组块平均。首先对单个通道从每次闪烁开始每 600ms 长度截取为一个单体样本。由于每轮实验重复 5 次，字符矩阵为 6 行 6 列，每轮闪烁 12 次，重复 5 次后共闪烁 60 次，因此一个目标字符的单个通道可分割为 60 组。题目要求脑-机接口系统中既要满足分类的准确率，同时又要保证一定的信息传输效率。为了实现脑-机接口系统信息传输效率，可以将每轮次行/列标识符相同的组进行平均，这样就可以将单个通道分割成 12 组，与字符矩阵的 12 个行/列相对应。经过平均，其中只有两个平均后的行和列包含被识别字符，另外 10 个不包含。为了进一步提升 P300 峰值电位识别范围的清晰度，对被测字符按在 20 通道上再进行平均，平均后对数据进行可视化，可视化结果是按照字符矩阵的行列标识符进行编号的，所以图的序号对应了字符矩阵的行和列，结果如图 3 所示，



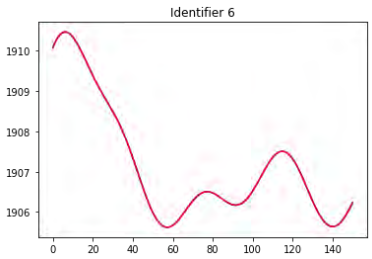
（1）

（2）

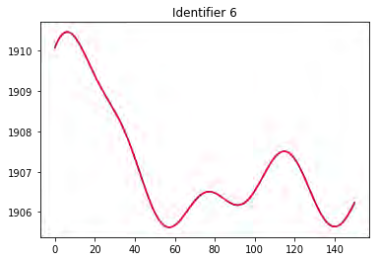
（3）



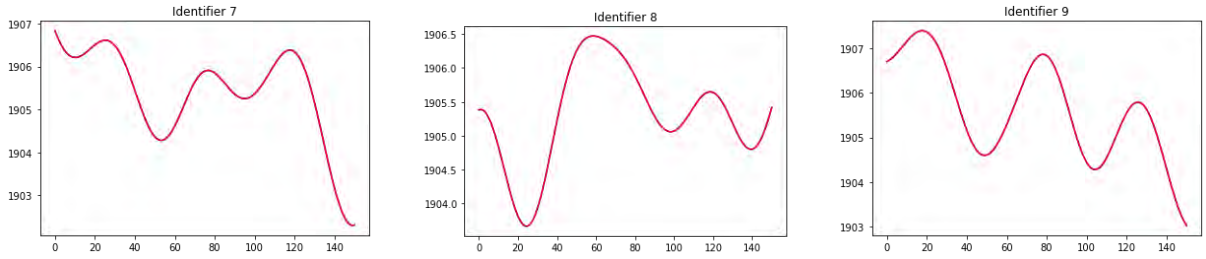
（4）



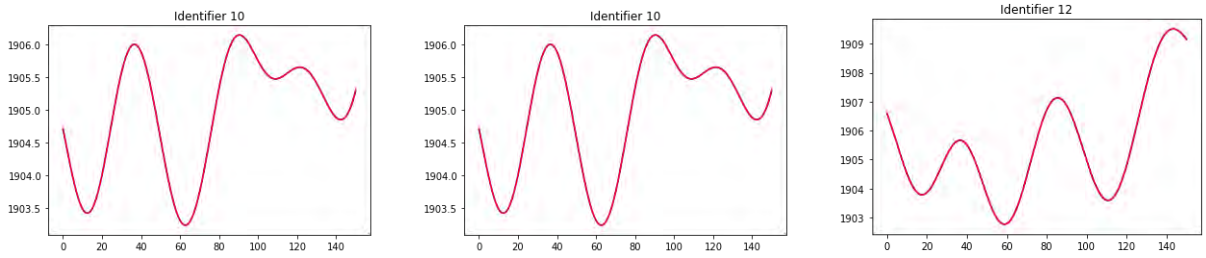
（5）



（6）



（7） （8） （9）



（10） （11） （12）

图 3 按通道平均后各行列信号变化情况

由于实验中采样频率为 250Hz，可知采样一次为 4ms,所以对应于图 3 上的曲线图，大概再 80 次左右，也就是 300ms 左右 P300 信号才会出现明显上升的波峰。通过对图三的 12 个曲线进行对比发现第（1）和（8）曲线有明显的正向波峰。由于数据使用的是测试字符 B 的电位信号，所以根据行列标识符的分布，可以看出 B 对应的标识符为 1 和 8。为了更清晰显示平均后包含目标信号的行列，将包含目标信号的曲线进行加粗处理，最终显示效果如图 4 所示。

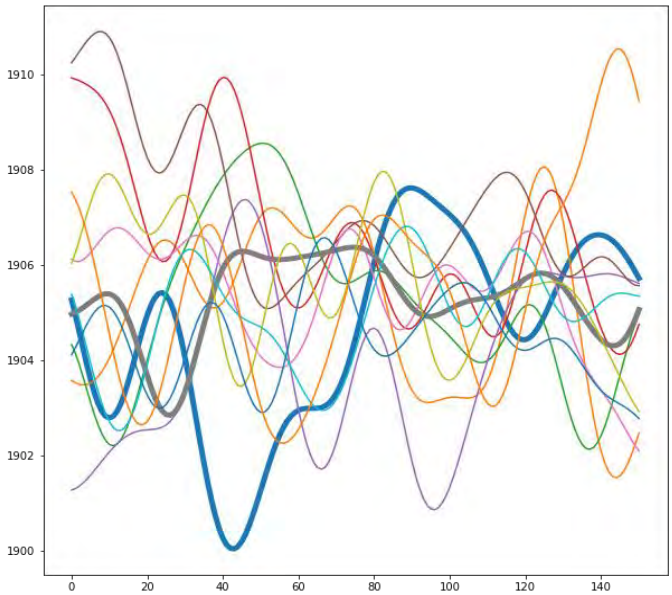


图 4 某字符按行列和通道组块平均后的信号

4.3 信号片段降采样

模型训练的速度与数据量有很大关系，本实验样本点信号采样频率为 250Hz,所以可得出每 4ms 采样一次，本文是在第一次闪烁后 0 到 600ms 截取一段，所以在每段内共有 151 个数据点。为了提高模型训练速度，在 0 到 600ms 内每隔 5 个点选取一个点作为最后特征，

这样就会得到 31 个信号段的特征采样点，每个字符采样通道为 20 个，所以对每轮信号平均后可得到每段包含 31×20 个特征。

4.4 数据规范化

通常我们在处理一批数据的时候，都希望让数据满足某种规律，达到某种规范化，以方便对数据进行挖掘。特别对于本文中的问题解决过程用到多种算法，在建模前必须对数据进行规范化处理，以降低数据计算成本，提高训练数据学习的速度。本文采用的规范化方法是采用向量单位化方法将每段的特征属性数据按比例投射到[0.0, 1.0]这种小范围内，用以消除因数据大小属性不统一带来的特提取结果偏差，对后面用到的随机森林算法、支持向量机（SVM）、循环神经网络（CNN）等算法的应用有明显的作用效果。所以，为了使脑机接口数据的利用效果更好，有必要对数据进行规范化处理。数据预处理流程框图如图 5 所示。

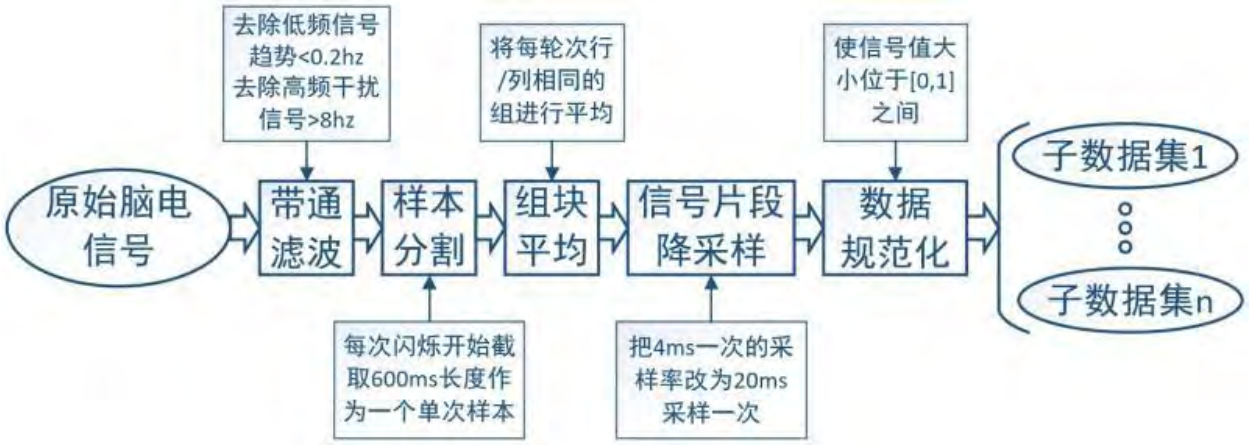


图 5 数据预处理流程图

5. 问题一求解

5.1 问题分析

题目要求使用训练集数据通过一定算法对测试集数据进行预测以得到测试集中的 10 个待测目标。根据问题要求在分析附件 1 脑机接口数据的基础上，通过对五个人的训练集数据进行预处理以得到包含 P300 信号的子数据集用于相应算法训练，最后通过分类算法对测试集数据进行分类训练，然后再测试集分类训练的基础上，对测试机数据进行分类处理，根据分类结果中字符矩阵行列出现的概率来确定测试集中的待识别目标。具体处理流程框图如图 6 所示，主要包括预处理、特征提取和分类 3 大步骤。

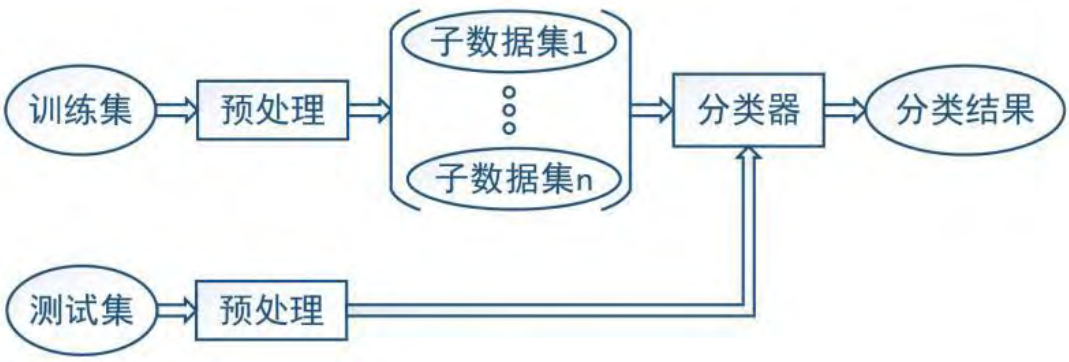


图 6 脑电信号处理流程图

5.2 训练数据预处理

原始数据经第四章切片、滤波、下采样等处理方法之后，单个被试单目标字符原始数据变成了[12,20x31]的维度。其中 12 表示 12 个行/列的闪烁，20 个通道以及每个通道 31 个采样信号。这样单个被试单目标字符一共有 12 个样本，每个样本有 20*31 个特征，其中有两个样本具有 P300 电位信号。

由于每个被试的 P300 点位信号出现的时间和幅度相差不多，为了增加训练样本量，提高模型精度.本文将五个被试以及每个目标字符的样本全部合并在一起，这样就得到了 720 个训练样本。720 个训练样本中定义有 P300 电位信号的样本标签为 1，没有 P300 电位信号的样本标签为 0 如下表 1。

表 1 样本标签与数量

样本	标签	数量
有 P300 电位信号	1	120
无 P300 电位信号	0	600

由于标签为 1 的样本只占样本数的 17%，这就造成了数据不均衡问题，直接放入模型训练必然为降低模型的精度。为了解决数据不均衡问题，本文将标签为 1 的样本随机复制，同时将标签为 0 的样本随机删除最终一共得到 600 个训练样本，其中 300 个正样本，300 个负样本。

5.3 模型评估指标

本文以准确率作为模型分类的评估指标，准确率定义如下：

准确率 = 分类正确的样本数量 / 样本总数量

5.4 分类算法比较

5.4.1 支持向量机

支持向量机是一种有监督的机器学习算法，SVM 可以应用于数据分类、模式识别、回归分析等问题。SVM 是基于结构风险最小化原则的统计理论方法，通过将数据向高维空间映射，进而构建最优超平面进行数据分类。SVM 具有如下优点[5]：（1）能有效解决小样本问题，SVM 引入了置信区间，同时要求经验风险和置信区间最小化，限制了模型复杂度的增加，有效地增强了 SVM 处理小样本数据的能力；（2）在非线性问题处理上具有较好的适应性，SVM 引入了核函数的思想，通过将低维度的原始样本映射到高维空间中，使样本在高维空间中变得线性可分，这样就将线性不可分问题转换为了线性可分问题，从而可以在高维度空间找到最优划分超平面，同时，SVM 只需要在核函数中求得内积，然后向高维空间进行映射，避免了维度过高带来的计算困难及无法计算的问题，具有较好的算法适应性。

支持向量机找到最优分类平面的过程如下[6]。首先，假设线性可分的样本集为：

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$$

其中，$x_i \in R^d$，d 为训练样本向量的维数，R^d 为所有样本的向量空间。$y_i \in \{+1, -1\}$，$i = 1, 2, \cdots, n$，y_i 为样本所属类别。最优分类超平面可以表示为：

$$\omega \cdot x + b = 0$$

其中 ω 是方向向量，b 是偏置项。对于输入的样本 x_i，可通过决策函数得到其所属类

别：

$$f(x_i) = sign(\omega \cdot x_i + b) = \begin{cases} \omega \cdot x_i + b > 0, \ y_i = 1 \\ \omega \cdot x_i + b < 0, \ y_i = -1 \end{cases}$$

寻找最优分类超平面就是通过样本数据不断调整参数ω和b的过程，从上式中可以看出，在保证分类正确的情况下，公式$\omega \cdot x_i + b = 0$的计算结果与实例x_i所对应的标记y_i在符号上是保持一致的，并且满足以下公式：

$$y_i(\omega \cdot x_i + b) - 1 \geq 0, \ i = 1, 2, \cdots, n$$

可以计算得出，此时分类间隔$m \arg in = 2/\|\omega\|$，要想最大化分类间隔，只需要将$\|\omega\|^2$最小化即可，为了后续求导计算方便，在前面添加系数1/2，同时，针对样本数据中可能会带有噪声，个别样本出现线性不可分，而剩余大部分样本仍可以进行线性划分的情况，可以适当放宽上述不等式的约束条件，引入松弛变量，形成软间隔分类超平面，以适应那些异常数据。于是可以将求解最优分类超平面转换为最优化问题：

$$\min_{\omega, b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$s.t \quad y_i(\omega \cdot x_i + b) \geq 1 - \xi_i, \ i = 1, 2, \cdots, n$$

其中，C为惩罚函数，用来控制模型对错误分类的惩罚程度，为了方便求解，引入拉格朗日乘子，可以表述为如下形式：

$$L(\omega, b, \alpha) = \frac{1}{2}(\omega \cdot \omega) - \sum_{i=1}^{n} \alpha_i \{y_i[\omega \cdot x_i + b] - 1\}, \ \alpha_i \geq 0, \ i = 1, 2, \cdots, n$$

其中，$\alpha = (\alpha_1, \alpha_1, \cdots, \alpha_n)$为样本所对应的拉格朗日乘子，将拉格朗日函数$L(\omega, b, \alpha)$分别对$\omega$，$b$求偏导数，并令结果等于0，可以得到

$$\frac{\partial L(\omega, b, \alpha)}{\partial \omega} = \omega - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$

$$\frac{\partial L(\omega, b, \alpha)}{\partial \omega} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

求解上式可以得到：

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

通过将原优化问题转换为对应的对偶问题，可得如下公式：

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(x_i \cdot x_j \right) - \sum_{i=1}^{n} \alpha_i$$

$$s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \ \alpha_i \geq 0, \ i = 1, 2, \cdots n$$

进而可求得偏置项b的值：

$$b = y_i - \sum_{i=1}^{n} \alpha_i y_i \left(x_i \cdot x_j \right)$$

将求解得到的参数ω，b，α的值带入最优超平面公式中，可得最优分类超平面为：

$$\sum_{i=1}^{n} \alpha_i y_i \left(x \cdot x_i \right) + b = 0$$

支持向量机算法的核心问题就是根据给出的样本数据点找出最优分类超平面，超平面可以将不同类别的样本进行区分，同时与两侧数据样本的间隔达到最大，增强了算法的泛化能力。在超平面公式的基础上可以得到如下的决策函数：

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i \left(x \cdot x_i \right) + b \right) = \begin{cases} \sum_{i=1}^{n} \alpha_i y_i \left(x \cdot x_i \right) + b > 0, \ y = 1 \\ \sum_{i=1}^{n} \alpha_i y_i \left(x \cdot x_i \right) + b < 0, \ y = -1 \end{cases}$$

通过将样本 x 输入到决策函数中，根据其结果的符号即可判断出输入样本所属类别。模型训练参数如下表 2：

表 2 模型训练参数

SVM 参数名称	参数设定
核函数	线性核(linear)
正则系数 C	0.9

经 10 折交叉验证评估，SVM 分类的准确率为 80.2%。

5.4.2 随机森林算法

随机森林是集成算法中最具有代表性的一种，随机森林的基本思想是通过随机抽取样本及特征构建多个弱分类器决策树，最后由这些弱分类器通过投票的方式选择得出预测标签[7]。随机森林算法的优点有：（1）它可以判断特征的重要程度，可以判断出不同特征之间的相互影响；（2）不容易过拟合，且训练速度比较快；（3）能够有效地运行在大数据集上，能够处理具有高维特征地输入样本，而且不需要降维。

随机森林算法的示意图如图 7 所示。随机森林以决策树为基学习器，其算法可以总为：

（1）首先通过随机有放回的抽样，从原始数据集中生成 k 个训练集；

（2）然后利用抽样产生的训练集构建决策树。在构建每棵树时，会先从所有的特征 N 中随机选择出 n 个特征，在这 n 个特征中，根据 Gini 系数最小的准则选取最优特征生成子节点：

$$Gini(p) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

（3）每棵树都有得出一个分类结果，根据多数投票原则，随机森林选择得票最多的结果作为输出。

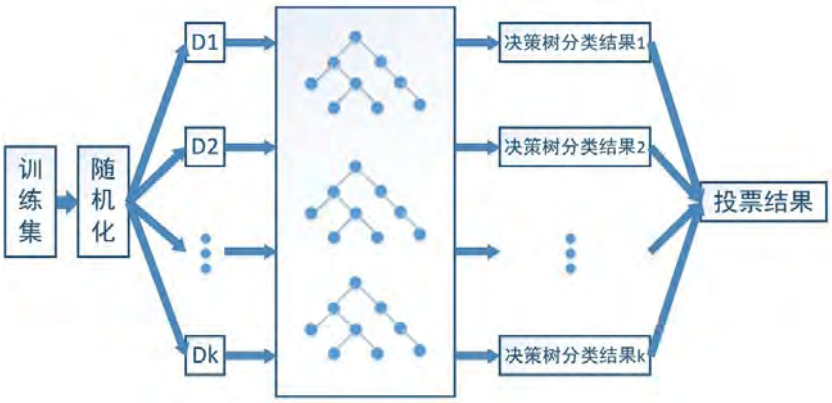


图 7 随机森林算法的示意图

对于一组基分类器$\{h(x, \beta_k), k = 1, 2, \cdots, n\}$构成的随机森林模型，$x$ 是输入变量，X 为原始数据集，Y 为分类标签。

定义边缘函数为：

$$mg(X, Y) = av_k I[h_k(X) = Y] - \max_{Z \neq Y} av_k I[h_k(X) = Z]$$

其中，I 为示性函数，Y 为正确分类，Z 为错误分类，av 表示对分类器取平均。通过边缘函数来评估模型的预测能力。

随机森林模型训练参数如下表 3：

表 3 随机森林模型训练参数

随机森林参数名称	参数设定
n_estimators	500
max_depth	5
Random_state	3

经 10 折交叉验证随机森林的训练准确率为 85.3%

5.4.3 基于卷积神经网络的分类器

脑电波信号是一种多通道时域信号，通常具有较低的信噪比，且不同的受试者所产生的 EEG 信号差异较大。对于这种可变信号而言，基于局部核函数的模型在表示它时效率低下，因为它并不是全局光滑的。由于采集的信号的时候就是按时间顺序采集的，因此脑电波信号经过数字化后，可以看成时序信号。下图 8 为频率 250HZ 单一通道采集的电信号。

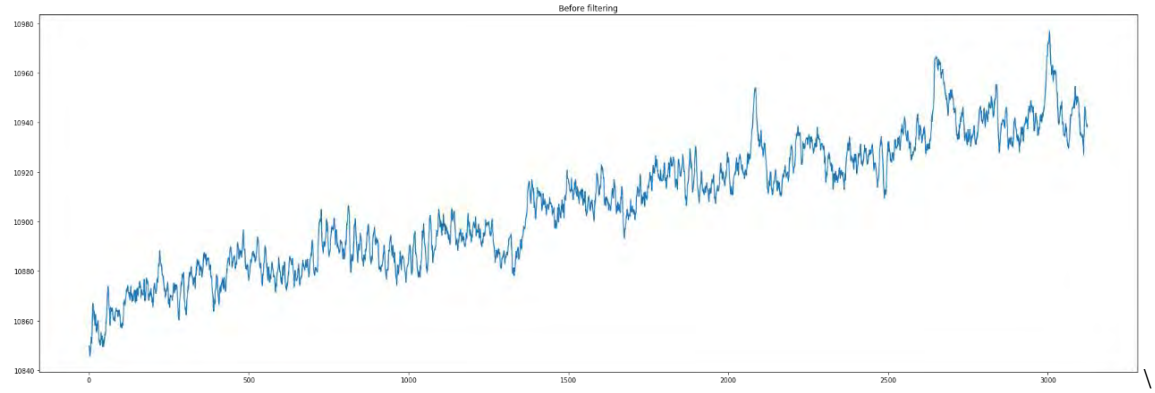


图 8 频率 250HZ 单一通道采集的电信号

对时序信号做分类，传统的机器学习分类模型在分类和提取信号特征的过程中并不能够提取到时序信号的时序特征，如上 SVM 和随机森林的分类准确率并不是很好。由于脑电波信号是一维的，将多通道的脑电波信号合并成二维数据类比为图像数据，然后用卷积神经网络去分类是不合理的。因为每个信号采集通道之间并没有很强的空间相关度。

基于这个问题，本文设计了一维卷积神经网络对脑电波信号进行分类。一维卷积过程如下图 9 所示

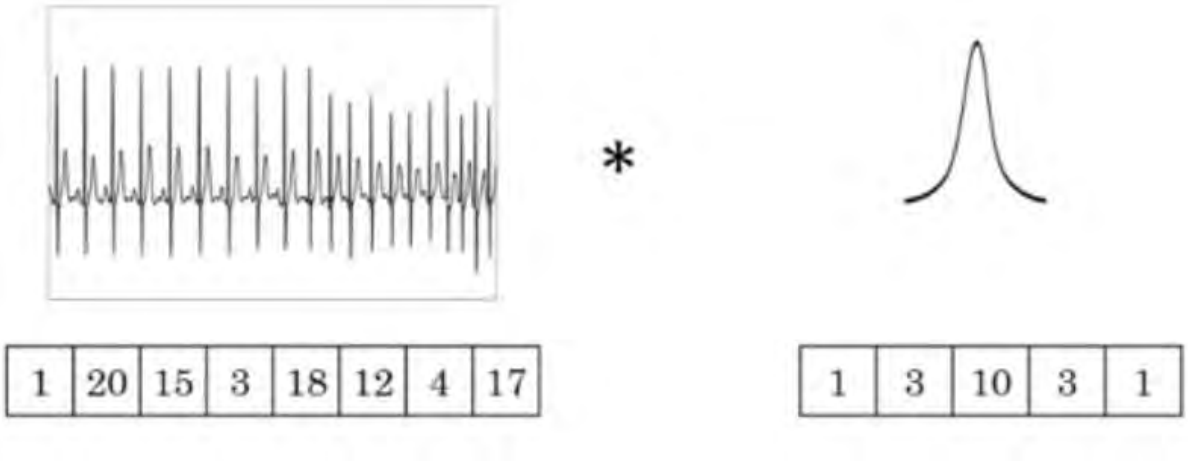


图 9 一维卷积过程

如上图，和二维卷积类似，一维卷积核在一维时序信号上平移相乘求和。一维卷积核能够在不改变时序的情况下很好的提取出脑电波信号的特征。本文所设计的一维卷积神经网络主要包括三层卷积层、两层最大池化层、dropout 层和两层全连接层。网络结构如下图 10

```
Model: "MyCNN"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
Conv1 (Conv1D)               (None, 248, 16)           144
_________________________________________________________________
max_pooling1d_34 (MaxPooling (None, 62, 16)            0
_________________________________________________________________
Conv2 (Conv1D)               (None, 62, 10)            1290
_________________________________________________________________
max_pooling1d_35 (MaxPooling (None, 15, 10)            0
_________________________________________________________________
Conv3 (Conv1D)               (None, 15, 4)             164
_________________________________________________________________
dropout_23 (Dropout)         (None, 15, 4)             0
_________________________________________________________________
flatten_20 (Flatten)         (None, 60)                0
_________________________________________________________________
dense_36 (Dense)             (None, 16)                976
_________________________________________________________________
dropout_24 (Dropout)         (None, 16)                0
_________________________________________________________________
dense_37 (Dense)             (None, 2)                 34
=================================================================
Total params: 2,608
Trainable params: 2,608
Non-trainable params: 0
_________________________________________________________________
```

图 10 一维卷积神经网络结构

模型训练参数如下表 4：

表 4 模型训练参数

一维卷积神经网络参数名称	参数设定
训练轮数	15
优化器	Adam
学习率	0.01
损失函数	交叉熵损失

以 20%的数据作为测试数据，经过卷积神经网络的 15 轮训练，最终精度能达到 90.6%，达到预期效果。

5.5 结果分析比较

以上三种分类模型评估精度如下表 5：

表 5 三种分类模型评估精度

分类模型	评估准确率
支持向量机	80.2%
随机森林	85.3%
一维卷积神经网络	90.6%

由于一维卷积神经网络在提取特征的过程中保留了脑电信号原有的时序特征，因此其准确率最高，最终选择一维卷积神经网络作为问题一的分类模型。

将 S1 到 S5 五个被试的测试数据按训练数据相同的方法经行预处理，经一维卷积神经网络预测，最终结果如下表 6：

表 6 测试目标字符结果

被试	测试目标字符
S1	M F 5 2 I T K X A 0
S2	M F 5 2 I T K X A
S3	M F 5 2 I T K X A
S4	M F 5 2 I T K X A 0
S5	M F 5 2 I T K X A 0
最终结果	M F 5 2 I T K X A 0

6. 问题二求解

6.1 问题分析

由于实验中采集的原始脑电数据较大，所以这些信号必然包含冗余的信息，由于冗余和无关信息的存在使得系统复杂度进一步加深，而且对分类识别的准确率和性能也有较大影响。为了解决这个问题，往往要对实验中使用的通道进行筛选，去除对 P300 信号响应不明显的通道，保留对 P300 信号影响较大的通道，以选择最优的通道组合。因此，将问题转化为主成分分析问题。所以，对于问题二的解答主要采用主成分分析法，完成最优通道组合的筛选。先对 5 个测试对象分别选择一套适用于自己的最优通道名称组合，最后，根据 5 个测试对象的筛选的最优通道组合对比分析选择出一套对于所有被试都适用的一组最优通道组合名称。具体实施过程中，通过算法自动学习所有通道对应的权重大小，由权重占比来选择最优通道子集。

6.2 主成分分析法

主成分分析法能够简化变量间互相关联的复杂关系，它是一种矩阵降阶算法，能够在尽可能保留原始矩阵信息的前提下实现矩阵维数的降低[9]。通过降维能够去除数据中的噪声，即去除数据集中的无用信号。例如本文研究的脑-机接口问题，在没有进行最优通道筛选的情况下，对每轮信号平均后可得到每段包含 31×20 个特征，但是如果实现最优通道筛选，那么通道数会低于 20，进而实现矩阵维数的降低。这样能够在保证计算精度的前提下减少数据量，节省运算空间，提高运算效率。

在运用主成分分析法的过程中要将数据转化为向量表示并进行基变换，其数学推导主要包括两个方面：其一是在划分方差最大的优化条件下从最大可分性方面入手，另外一个就是以点到划分平面距离最小的优化条件为依据的最近重构性方面进行。如下框图 11 是主成分分析法运用过程中可能会涉及的方面。

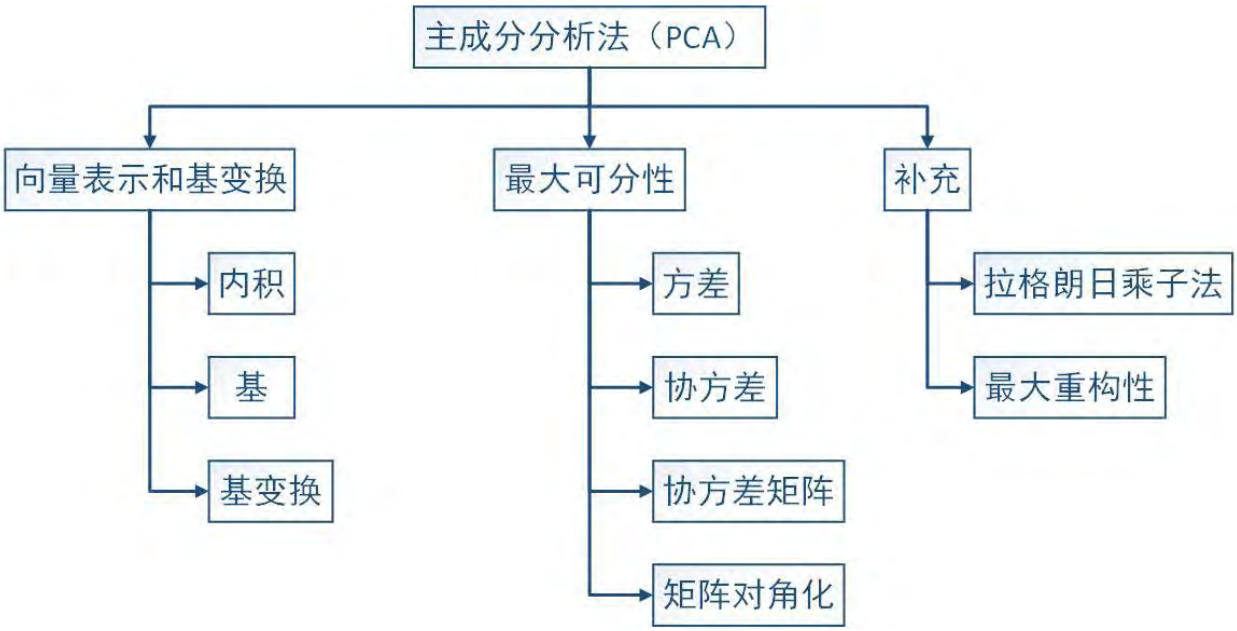


图 11 主成分分析基本方面

关于降维过程可以这样理解，假设有一组二维数据，可以将其表示在二维平面坐标系中，用一个点表示如下图 12（a）所示，当数据点较多时，我们可以假设在该二位空间中存在一个一维子空间。我们选取基向量，将其投影到该基向量上面，则二维向量就能用基向量表示成一维向量，实现降维效果如图 12（b）所示。

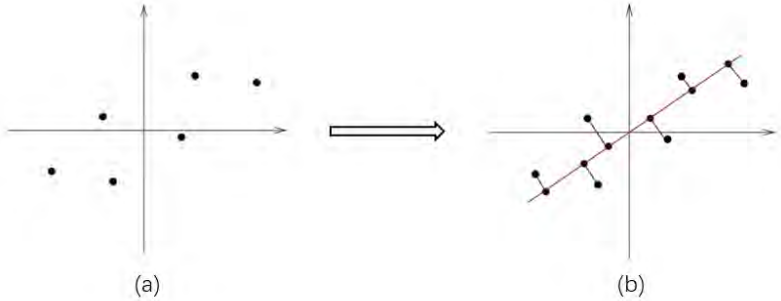


图 12 降维过程

相关参数计算公式：

（1）方差，按照方差定义，可将一个变量的方差看成变量中所有元素与变量均值差的平方和的均值，即：

$$Var(a)=\frac{1}{m}\sum_{i=1}^{m}(a_i-\mu)^2$$

通常为了计算方便，可以将变量的均值设置为 0，进而将方差的计算公式简化为如下式所示，

$$Var(a)=\frac{1}{m}\sum_{i=1}^{m}a_i^2$$

对于上面的降维问题则可转化为求数据坐标转化为基坐标后，使得方差值最大。

（2）协方差，协方差是用来表示两个变量的相关性的。通常我们希望让变量能够尽

可能多地表示原始信息，所以就要式变量间的相关性为 0，然而实验过程中各通道采集的数据必然存在一定的相关性，所以要通过求解协方差为 0 来去除各变量间的重复信息，去掉数据的冗余度，节省运算空间，提高运算效率。
协方差计算公式为：

$$Cov(a,b)=\frac{1}{m-1}\sum_{i=1}^{m}(a_i-\mu_a)(b_i-\mu_b)$$

在样本量很大的情况下，分母可用 m 代替 $m-1$，能够方便计算。

当两个变量是相互独立时其协方差为零，所以若要满足协方差为 0 的要求就要保证选择的基向量间是相互正交的。在这种情况下，降维过程的优化目标就转化为求向量转化中的单位正交基问题，同时在保证协方差为 0 的前提下，使得变量的方差最大。

6.3 求解过程

实际情况下，不同的受测者个体间也存在较大差异，每个受试者对不同通道的敏感程度也就不同，所以测得的脑-机接口信号也就存在差异化。因此对于测试数据集和训练数据集进行数据处理，首先对原始数据进行滤波处理，然后对数据进行分段截取完成数据切片，具体处理过程按照前有关文数据处理章节所述方法进行，直到生成训练及验证样本集。

关于本问题运用算法的求解过程如下图 13 所示：

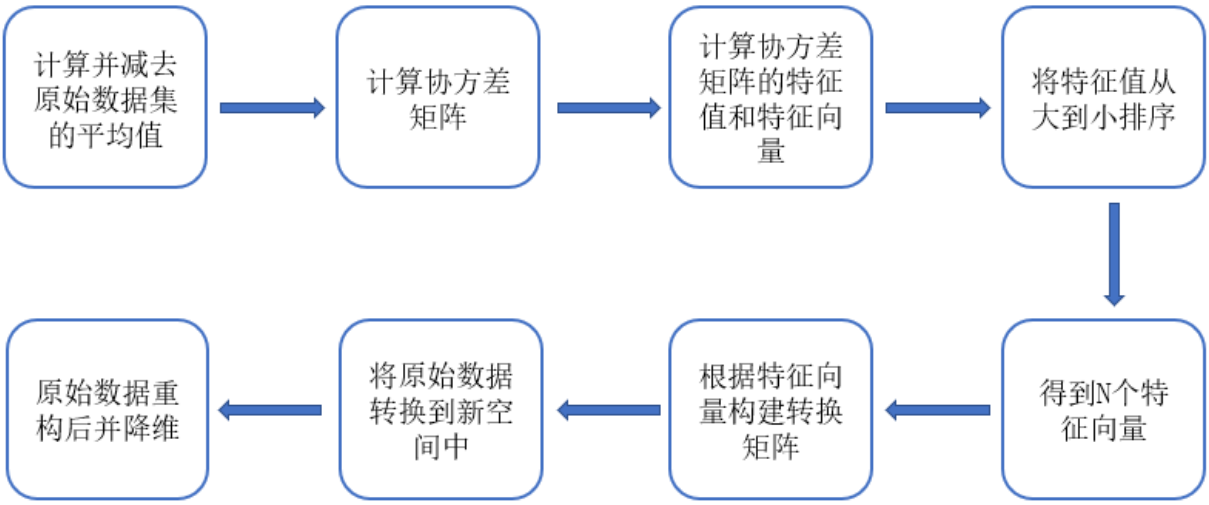
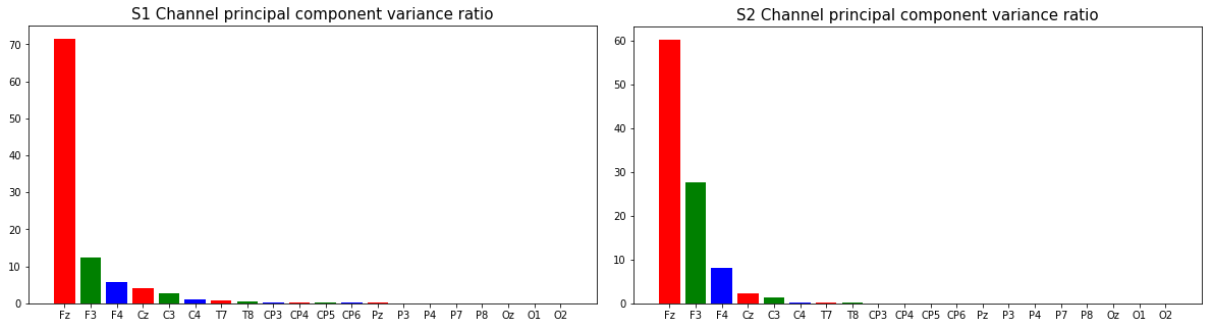


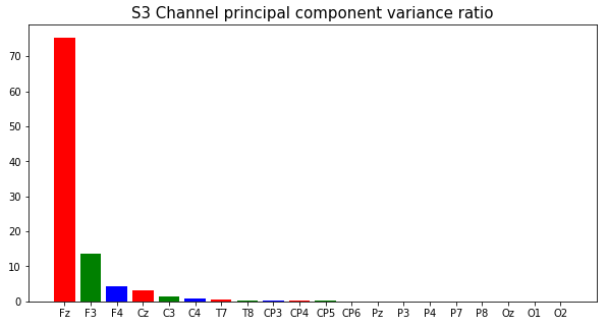
图 13 PCA 算法求解过程

6.4 结果讨论

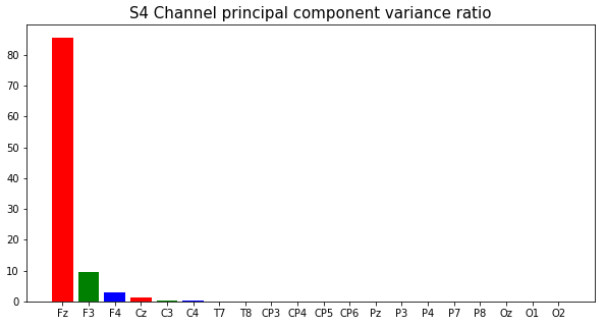
运用算法求解后可以得到各主成分方（20 个通道）的百分比，如图 14，



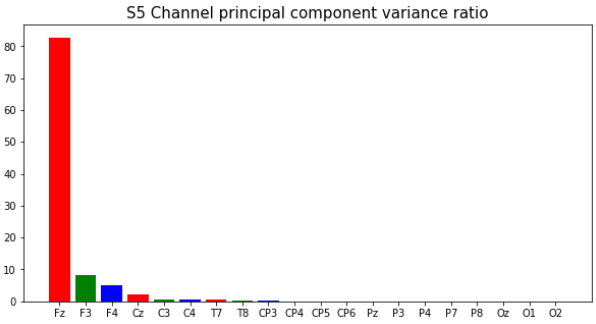
S1　　S2



S3



S4



S5

图 14 各主成分方（20 个通道）的百分比

根据上图 14 统计百分比分析选择出每个被试者所选的通道见下表 7：

表 7 被试者通道选择

被试者	通道名称	通道编号
S1	Fz、F3、F4、Cz、C3、C4、T7、CP3、Pz、P3、P4、Oz	1、2、3、4、5、6、7、9、13、14、15、18
S2	Fz、F3、F4、Cz、C3、C4、T7、T8、CP5、Pz、P3、P4、Oz	1、2、3、4、5、6、7、8、11、13、14、15、18
S3	Fz、F3、F4、Cz、C3、C4、T8、CP6、Pz、P3、P4、Oz	1、2、3、4、5、6、7、8、12、13、14、15、18
S4	Fz、F3、F4、Cz、C3、C4、T7、CP3、Pz、P3、P4、Oz	1、2、3、4、5、6、7、9、13、14、15、18
S5	Fz、F3、F4、Cz、C3、C4、CP4、Pz、P3、P4、01	1、2、3、4、5、6、7、10、13、14、15、18

通过对表格中被测者对比，选着出对所有被试者都适用的通道名称组合为 Fz、F3、F4、Cz、C3、C4、Pz、P3、P4、Oz，各通道对应的标识符为 1、2、3、4、5、6、13、14、15、18。

7. 问题三求解

7.1 问题分析

在 P300 脑机接口系统中，有标签数据的获取往往需要花费更多的时间和精力。现在打算通过半监督学习的方式，根据问题二所得的一组最优通道组合：Fz、F3、F4、Cz、C3、C4、Pz、P3、P4、Oz，在五名受试者的数据中，选择适量的数据作为有标签样本，其余数据作为无标签样本，以少量有标签的样本和大量无标签的样本作为训练集来训练分类模型。半监督学习示意图如图 15 所示。

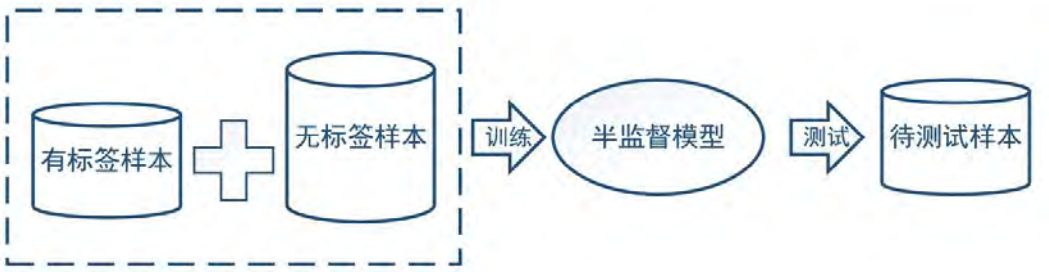


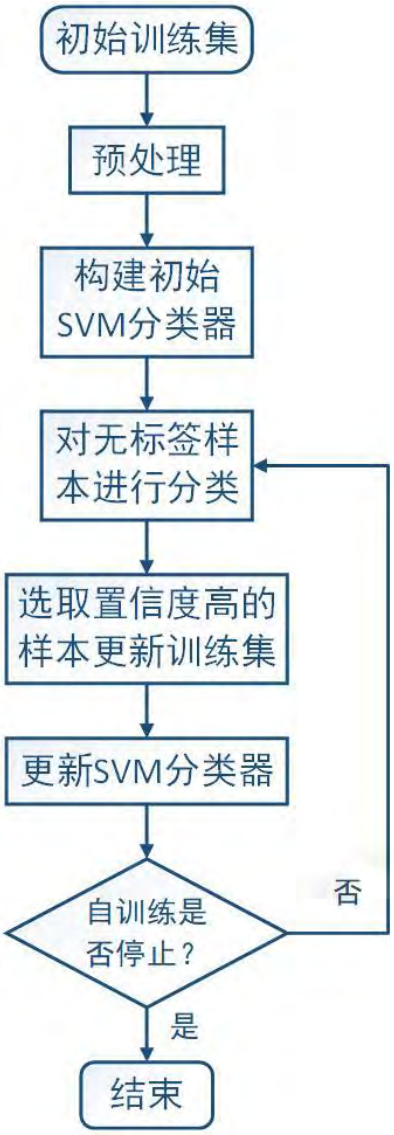
图 15 半监督学习示意图

7.2 基于 SVM 的半监督自训练算法

传统机器学习分为监督学习和无监督学习。监督学习使用的是有标签样本，通过大量有标签的样本训练出来的模型效果才好，但是获取有标签样本需要大量的时间和精力，同时也忽略了无标签样本的潜在价值。无监督学习通过无标签样本训练学习。半监督学习就是介于这二者之间的一种学习器。半监督学习的方法一方面通过少量有标签的样本训练初始分类器模型，另一方面可以通过大量的无标签样本进一步提升模型的分类性能。

目前，半监督学习方法[10]主要包含：自训练方法（self-training）、协同训练方法（co-training）、基于分歧的训练方法（tri-training）等。自训练（Self-training）因其运算简单、易于实现等优点，逐渐成为半监督分类中较为热门的结构类型。自训练半监督学习的具体步骤为：对原始数据首先进行预处理和特征提取，将少量有标签样本作为初始训练集训练分类器；再把训练集中其余无标签样本分组，用先前训练好的分类器完成第一组无标签样本的预测，通过相关置信度准则判断其预测结果置信度高低，选取适量高置信度样本更新训练集，再去训练分类器，对剩余几组无标签样本重复上述操作，如此循环，得到越来越精确的分类模型。

本文选择基于 SVM 的无监督自训练算法[11]，首先利用少量有标签样本构建初始训练集，再利用构建好的 SVM 完成无标签样本的预测任务，选取置信度高的样本来更新训练集，进而更新 SVM 分类器。算法结构流程图如图所示[12]。



在计算扩展训练集内无标签样本结果对应样本的置信度高低时，我们采用欧氏距离[13]作为评估标准。在训练 SVM 分类器之后可以得到对应样本的判别分数 $f(x_i)$，通过判别分数可以计算有标签样本集 D_L 中两类样本的平均值 mean_1 和 mean_2，且与 mean_1 和 mean_2 相距越近的无标签样本具有更高的置信度。

$$\text{mean\_1} = mean(f(x_i)) \qquad (x_i \in D_{L1})$$

$$\text{mean\_2} = mean(f(x_i)) \qquad (x_i \in D_{L2})$$

其中 D_{L1} 和 D_{L2} 分别为类 1 和类 2 对应的初始训练集，$D_L = D_{L1} \cup D_{L2}$。

通过无标签样本的判别分数计算置信度值，即求出各类无标签样本与 mean_1、mean_2 的欧氏距离 $d_1(x_i)$、$d_2(x_i)$。

$$d_1(x_i) = |f(x_i) - mean\_1| \qquad (x_i \in D_{U1})$$

$$d_2(x_i) = |f(x_i) - mean\_2| \qquad (x_i \in D_{U2})$$

将$d_1(x_i)$、$d_2(x_i)$按从小到大的顺序排序，选取置信度高的样本，用于扩充有标签样本集D_L。

7.3 实验结果分析

实验中分别随机选择总样本数据中的 10%、20%、30%、40%的数据来构建初始有标签训练集D_L，剩下的数据作为对应D_U中的样本。D_U中的样本以总样本数量的 10%为一组，进行对无标签样本的分类，则对应的自训练迭代组数分别为 9、8、7、6。已知 char13-char17 字符是 M、F、5、2、I，得到几种情况下的平均分类正确率如下表 8 所示。

表 8 几种情况下的平均分类正确率

有标签样本占比	10%	20%	30%	40%
平均分类正确率	74.13%	76.65%	79.47%	82.02%

对五名受试 char18-char22 的分类结果如下表。

被试	char18-char22
S1	T K X A 0
S2	T K X A
S3	T K X A
S4	T K X A 0
S5	T K X A 0
最终结果	T K X A 0

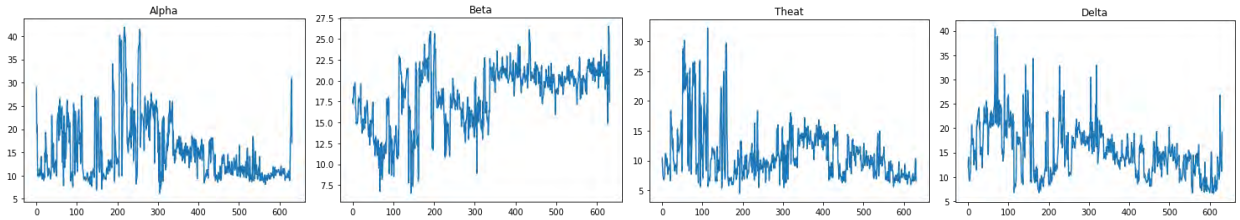
8. 问题四求解

8.1 问题重述

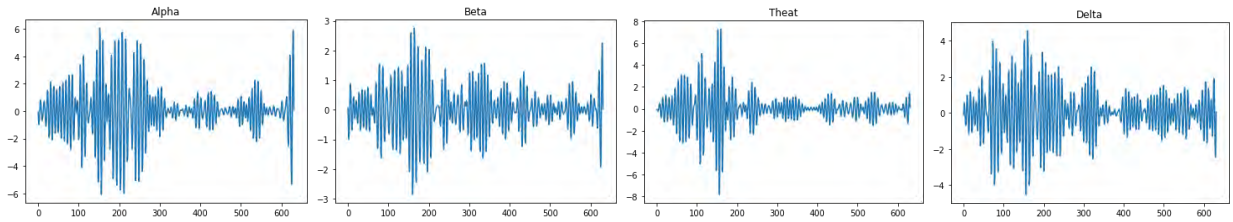
问题四要求设计睡眠分期预测模型，针对清醒期，快速动眼期，睡眠期，睡眠 2 期，深睡眠期，以 Alpha，Beta，Theat，Delta 为信号特征，此问主要解决如何能够放大识别不同睡眠期的这四类特征，通过对输入测试特征进行判别计算从而确定各组不同百分比的四类特征所对应的睡眠周期。针对此问题，我们采用支持向量机，K 近邻模型，一维卷积神经网络三类方法进行对比选优。

8.2 数据预处理

首先我们进行睡眠脑电信号的处理。采用带通滤波进行数据的处理，根据采样定理，采样频率要大于两倍的信号本身最大的频率，才能还原信号截止频率一定小于信号本身最大的频率，经过计算选取 25Hz-50Hz 的范围之内，下图 16 为清醒期四类特征的原始波形图与滤波之后的波形图。



(a) 原始波形图



(b) 滤波后波形图

图 16 清醒期四类特征的原始波形图与滤波之后的波形图

对于输入到训练器中的数据集，将五个周期的滤波之后的数据按照四类特征进行按行叠加，并且将每个周期作为一类结果值，将周期名与结果值进行如下匹配对应：清醒期-6，快速动眼期-5，睡眠期-4，睡眠 2 期-3，深睡眠期-2。

8.3 算法预测

进行睡眠脑电信号数据处理完成之后，将处理完成的数据送到各个模型中进行训练预测。

首先采用支持向量机进行训练预测。将数据集划分为训练集以及测试集，选取数据的 30%作为测试集。对于 SVM 中的参数选择如下表 9：

表 8 SVM 中的参数选择

SVM 参数 shed	参数值
核函数	高斯核 Rbf
惩罚系数 C	4

对于各个周期的预测准确性如下表 9 所示:

表 9 各个周期的预测准确性

周期名	Precision	recall	F1-score
2	0.91	0.97	0.94
3	0.73	0.83	0.78
4	0.66	0.44	0.53
5	0.82	0.92	0.87
6	0.98	0.92	0.98

SVM 模型进行的预测整体准确率为 83.9%。结果较优，在于其合理范围之内。

其次，采用 K 近邻模型对模型进行预测。模型的参数选择如下表 10。

表 10 K 近邻模型参数选择

K 近邻模型参数名	参数值
n_neighbors	15

对于各个周期的预测准确性如下表 11 所示。

表 11 各个周期的预测准确性

周期名	Precision	recall	F1-score
2	0.89	0.97	0.93
3	0.74	0.80	0.77
4	0.66	0.51	0.57
5	0.82	0.89	0.85
6	0.99	0.97	0.98

K 近邻模型进行的预测整体准确率为 80.6%。

以上两种方式都是通过机器学习模型进行预测分类。最后使用自主搭建一维卷积神经网络进行数据的训练以及预测。

在进行建模预测之前，首先要对结果值进行 one-hot 编码。之后进行训练集测试集的划分，测试集为总数据的 30%

一维卷积神经网络模型参数如下表 12。

表 12 一维卷积神经网络模型参数

Layer	Output Shape	Param
conv1d_33 (Conv1D)	(None, 2, 4)	16
conv1d_34 (Conv1D)	(None, 2, 16)	80
max_pooling1d_22 (MaxPooling)	(None, 1, 16)	0
conv1d_35 (Conv1D)	(None, 1, 64)	1088
max_pooling1d_23 (MaxPooling)	(None, 1, 64)	0
conv1d_36 (Conv1D)	(None, 1, 128)	8320
max_pooling1d_24 (MaxPooling)	(None, 1, 128)	0
flatten_9 (Flatten)	(None, 128)	0
dense_9 (Dense)	(None, 5)	645

经 40 轮迭代准确率如下图 17。

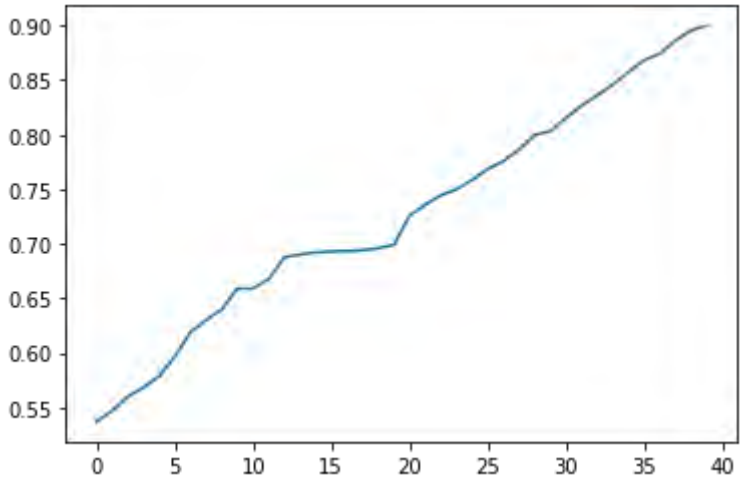


图 17 经 40 轮迭代准确率

一维卷积进行的特征提取精确性为 89.3%。

对于具体的分类识别过程，由函数 predict，进行分类预测，将特征值识别对应输出周期所对应的数字。

8.4 结果讨论

问题对于脑电信号的分类主要选取了传统的机器学习算法支持向量机和 K 近邻算法，以及自行设计的一维卷积神经网络模型。各个算法的评估准确率如下表 13。

表 13 各个算法的评估准确率

分类模型	准确率
SVM	83.9%
K 近邻	80.5%
一维卷积神经网络	89.3%

由此可知，一维卷积能够很好的保留脑电信号的时序特征，因此准确率最高。

9 参考文献

[1] Wolpaw J R, Birbaumer N, Heetderks W J, et al. Brain-computer interface technology: a review of the First International Meetings. IEEE Transactions on Rehabilitation Engineering, 8(2):164-173, 2000.

[2] Postelnicu C C, Talaba D. P300-based brain-neuronal computer interaction for spelling application. IEEE Transactions on Biomedical Engineering, 60(2):534-543. 2013.

[3] 王菡侨，有关美国睡眠医学学会睡眠分期的最新判读标准指南解析，诊断学理论与实践，8（6）：575-578，2009。

[4] S. Andrews, R. Palaniappan, A. Teoh, and L. C. Kiong, "Enhancing P300 Component by Spectral Power Ratio Principal Components for a single Trial Brain-Computer Interface," Amer. J. Appl. Sci. 5, No. 6, 639-644, 2008.

[5] 董宝玉，支持向量技术及其应用研究[D]. 大连海事大学，2016。

[6] Cortes C, Vapnik V. Support-vector networks, Machine learning, 20(3):273-297, 1995.

[7] 方匡南，吴见彬，朱建平，随机森林方法研究综述[J]，统计与信息论坛，26（3）：32-38，2011。

[8] Liaw, A, Wiener, M, Liaw, A. Classification and Regression with Random Forest[J]. R News, 2002, 23(23).

[9] Peter Harrington，机器学习实战[M],北京：人民邮电出版社，242-251，2013。

[10] 薛贞霞，支持向量机及半监督学习中若干问题的研究[D]，西安电子科技大学，2009。

[11] 周志华，机器学习，北京：清华大学出版社，298-300，2016。

[12] 刘静，基于半监督算法的在线运动想象 BCI 研究[D],重庆大学，2019。

[13] Y. Li,C.Guan,H. Li, et al. A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system[J]. Pattern Recognition Letters, 29(9):1285-1294。