


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	武汉工程大学
参赛队号	20104900023
队员姓名	1.范玮
	2.罗思吟
	3.邓轶赫

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：

脑电信号(Electroencephalogram, EEG)是由大脑细胞器进行生命活动所产生的一种自发性的电活动，按其产生的方式可分为诱发脑电信号（P300 脑-机接口）和自发脑电信号。诱发脑电信号是通过某种外界刺激使大脑产生电位变化从而形成的脑电活动；自发脑电信号是指在没有外界特殊刺激下，大脑自发产生的脑电活动。

本文旨在对 P300 信号和睡眠脑电进行分类识别研究。首先对题目所给的数据进行预处理，然后对信号进行特征提取，选择支持向量机、K 最近邻算法和在线序列 ELM 算法作为信号识别的算法，对测试集中的待识别目标进行识别分类，并对比各种不同参数条件下的分类性能，得出一种较为准确的测量方法。

针对问题一，需要对附件一中的训练数据以及训练数据的事件标签进行分析，研究其信号变化趋势。考虑到 P300 信号是在刺激发生后 300 毫秒左右出现一个正向波峰，我们选取刺激开始时到开始后 800 毫秒的数据，以确保检测到 P300 信号的发生，再对提取到的数据进行预处理等操作，训练支持向量机模型，对测试数据进行预测，得到测试数据中 char13-char22 对应字符的预测结果为：M, F, 5, 2, I, 6, K, X, A, 0

针对问题二，需要对脑电信号的 20 个通道进行筛选，保留最佳的 10 到 20 个通道。我们采用基于准确率排序的递归通道剔除算法，从原有 20 个通道逐个进行剔除，观察预测准确率的变化，最终针对每个受试者保留 10 个通道，并综合五个受试者的结果，给出了对于所有受试者都能保证较好结果的最优通道：Cz, T7, C4, Fz, CP4, C3, P7, F3, T8, Pz, Oz, CP5, CP3, P3, F4, CP6。

针对问题三，为减少训练时间，我们提出了基于超限学习机 ELM（extreme learning

machine）的在线半监督学习方法，首先利用少量有标签样本训练出一个初始分类器，在线测试的过程中，不断对新获取的无标签数据进行分类，并加入训练集以更新分类器。该半监督的学习算法计算量低，随着不断获取无标签数据，分类性能也逐步得到提高。

针对问题四，我们使用主成分分析（PCA）对数据进行降维处理，再利用 K 最近邻算法（KNN）的分类方法将原数据划分为训练集和测试集，进而根据各个睡眠期的特点预测出睡眠期。在测试多种 K 值和训练集比例后，我们给出了尽可能最少的训练集比例：50%，并且对于所训练出的 K 近邻模型特点进行了评估和分析。

关键词：脑电信号；特征提取；K-最近邻算法；支持向量机；超限学习机

目　录

一、 问题背景与重述

1.1 问题背景

脑电信号(EEG)是由人体皮质内的大量神经元突触后电位同步总和而形成，是很多神经元共同活动的结果。脑电信号中含有大量与人体生理与疾病有关的信息，为康复治疗提供了有效的帮助。在临床医学领域中，脑电信号不仅作为某些脑疾病的临床诊断依据，而且还为某些脑疾病的康复治疗提供了有效的辅助治疗手段。在工程应用中，利用脑电信号实现的脑机接口(BCI)，为瘫痪病人的某些功能重建提供了一种有效的方法。因此在康复医疗领域的实际运用中，对脑电信号的深入研究具有重要意义。

基于事件相关电位的脑机接口通常使用的是 P300 信号，P300 信号是由人脑在受到小概率刺激后 300 毫秒范围左右出现的一个正向波峰，如图 1 是在刺激发生后 450 毫秒左右的 P300 波形。

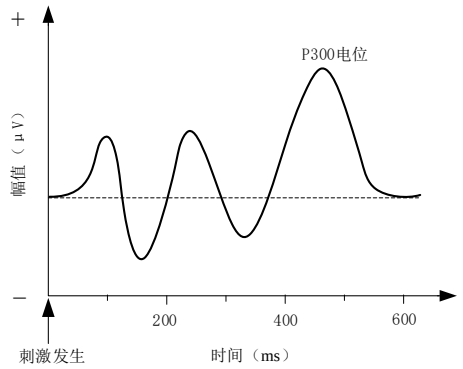

图 1 P300 波形示意图

P300 信号不需要受试者进行预先的训练，只需通过视觉刺激即可自发产生。本文给出了实验设计流程：如图 2 所示，字符矩阵以行或列为单位（共 6 行 6 列）。首先，提示被试者注视“目标字符”，假设出现的“目标字符”为“A” ，在被试注视“目标字符”的过程中，当目标字符所在行或列闪烁时，脑电信号中会出现 P300 电位；而当其他行和列闪烁时，则不会出现 P300 电位。其次，进入字符矩阵的闪烁模式，每次以随机的顺序闪烁字符矩阵的一行或一列，闪烁时长为 80 毫秒，间隔为 80 毫秒，最后，当所有行和列均闪烁一次后，则结束一轮实验，共重复 5 轮。脑电信号在采集过程中使用了 20 个通道，采集频率为 250Hz。

图 2 行/列的标识符

1.2 问题重述

现要求设计有效的模型与方法依次解决脑电信号的几个问题：

1.2.1 问题一

问题需求：根据附件 1 给出的 5 位被测试者数据，使用尽可能少轮次的测试数据，找到 5 个被试测试集中的 10 个待识别目标。

问题目标：用尽可能少的轮次找出附件 1 中 5 个被试测试集中的 10 个待识别目标，并给出具体的分类识别过程。

1.2.2 问题二

问题需求：根据附件 1 所给数据，在 20 个脑电信号采集通道中，设计一个通道选择算法以及通道名称组合，以此来减少无关以及冗余的通道数据。

问题目标：设计出针对每个被试者更有利于分类的通道名称组合。并基于每位被试者的通道选择方案，进一步分析对于所有被试者都较适用的一组最优通道名称组合，并给出具体分析过程。

1.2.3 问题三

问题需求：根据附件 1 所给数据，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，并在问题二选择出的最优通道基础上，设计一种半监督学习的方法。

问题目标：选择恰当的训练集比例，对剩余训练集进行在线学习，优化模型的性能。

1.2.4 问题四

问题需求：根据附件 2 中所给出的特征样本，设计一个睡眠分期预测模型。

问题目标：在尽可能少的训练样本的基础上，得到相对较高的预测准确率，并结合分类性能指标对预测的效果进行分析。

二、 符号及缩写说明

2.1 缩写说明

BCI	脑机接口(Brain Computer Interface)
ERP	事件相关电位(Event Related Potential)
EP	诱发电位(Evoked Potential)
P300	P300 脑电信号
KNN	K 最近邻法(k-nearest neighbors)
SVM	支持向量机(Support Vector Machine)
PCA	主成分分析（Principal Component Analysis）
CS	通道识别准确率
Score	通道评分结果
ELM	超限学习机（extreme learning machine）
SLFN	单隐层前馈神经网络(single-hidden layer feedforward neural network)

2.2 符号说明

符号	说明
α_i	拉格朗日乘子
H	单隐层前馈神经网络输出矩阵
$\boldsymbol{\beta}$	单隐层前馈神经网络输出权向量
W	加权矩阵
T_M	M 个数据的标签向量
Pi	第 i 个受试者通道剔除过程中的次序
CS	通道识别准确率
Score	通道评分结果

三、 问题一的模型建立及求解

3.1 问题分析

问题一所述的字符识别任务，本质上是对于接受刺激后产生的脑电波波形进行分类，区分含有 P300 信号和不含 P300 信号的两类波形。设计的模型通过训练数据集的训练后，将能够对测试数据集的波形数据进行二分类，从中区分出含有 P300 信号的波形数据，从而确定出要寻找的字符所在的行和列，然后根据行列序号即可定位到相应字符，得到字符预测结果。

3.2 数据预处理

数据质量直接影响建模效果，因此在正式构建模型之前往往要对数据进行恰当的处理，通过数据预处理对原始数据进行滤波并提取可能含信号 P300 的单次试验数据，具体的预处理过程如下：

1.滤波：研究指出 P300 事件相关电位的频率主要分布在 10Hz 以下的低频区[1]，因此采用 8 阶的切比雪夫低通滤波器对每个脑电信号采集通道的信号都进行阈值为 10Hz 的低通滤波处理，仅保留频率低于 10Hz 的部分。

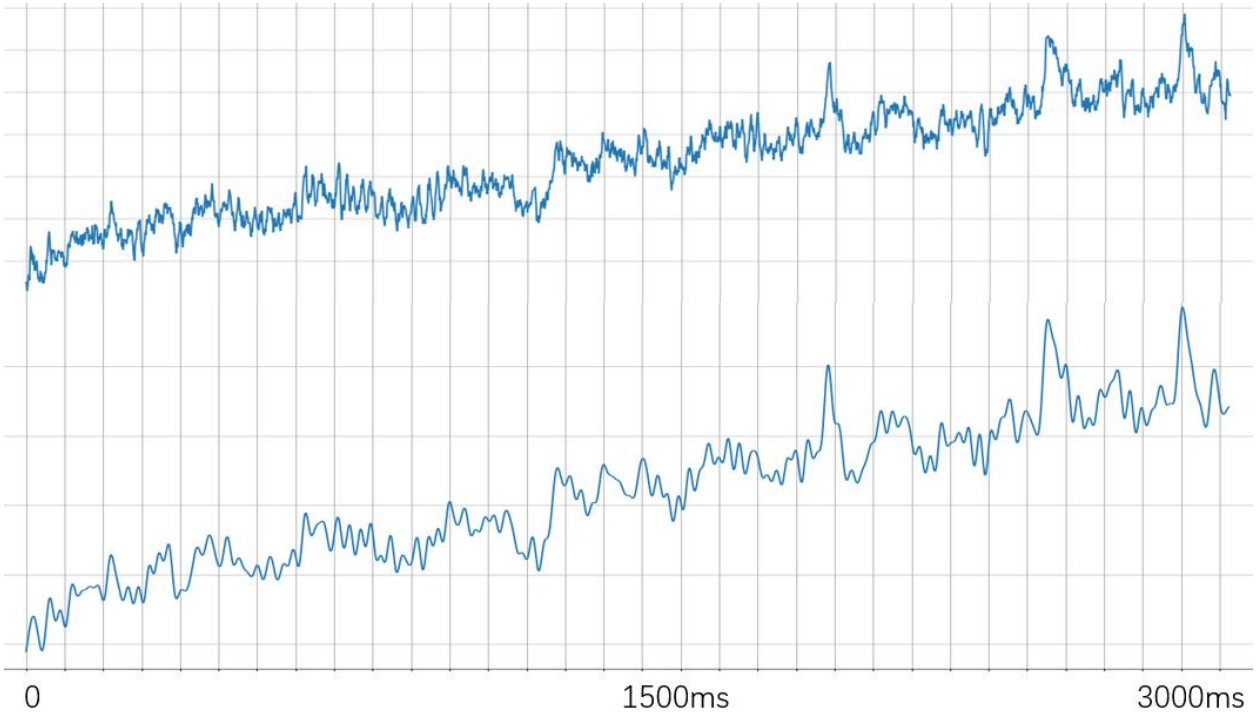

图 3 滤波处理前(上)和滤波处理后(下)对比图

图 3 展示了受试者 S1 的 Fz 通道信号，在进行第一次训练（char01）的过程中经过切比雪夫低通滤波器处理前后的变化，可以看出经过低通滤波处理去掉了一些干扰噪声和其

他干扰信号后，原信号变得更加平滑，便于后续提取 P300 信号特征。

2.单次试验数据段提取与标注：从数据中提取出持续时间 800ms 的单次试验数据段。提取从刺激开始到刺激开始后的 800ms 结束的这一段试验数据，由于采样频率 250Hz，即每个采样数据点间隔 4ms，所以每段数据需要选取 200 个采样点。由于刺激间隔只有 80ms，相邻试验段数据会有重叠。根据每次试验对应的字符，将该字符对应的行列的试验数据段标记为含有 P300 信号的数据（类别标记为 1），其他实验数据段标记为不含 P300 信号的数据（类别标记为 0）。

3.降采样：将数据切比雪夫滤波后需要对数据进行降采样，对于每条实验数据段进行间隔为 5 的降采样，这样单次试验采样点数降为了 40 个。

4.数据扩充：由于一轮测试中，含有对应字符和行列数仅占总行列数的 1/5，即含有 P300 信号的数据与不含 P300 信号的数据样本数量比值为 1：5，类别数量的不平衡会极大地影响到分类模型的训练和误差计算，因此要对含有 P300 信号的数据样本进行数量上扩充，具体扩充方法利用了步骤 3 的降采样过程：对于不含 P300 信号的样本，不需要进行扩充，直接进行间隔为 5 的降采样；但针对含有 P300 信号的每个样本进行起始索引的不同的五次降采样，这样就得到五条含有 P300 数据的原样本的新降采样数据。上述过程能够获得类别比例为 1：1 的两类数据，避免了类别不平衡的问题。

4.数据归一化：针对每个试验数据段所包含的采样数据，按最大最小归一化处理，即将原始数据线性化的方法转换到[0, 1]的范围内。

5.特征向量构建：在获得上述特征向量之后，我们进行了可视化展示，希望能从视觉上观察到 P300 信号的波形，如图 4 所示。

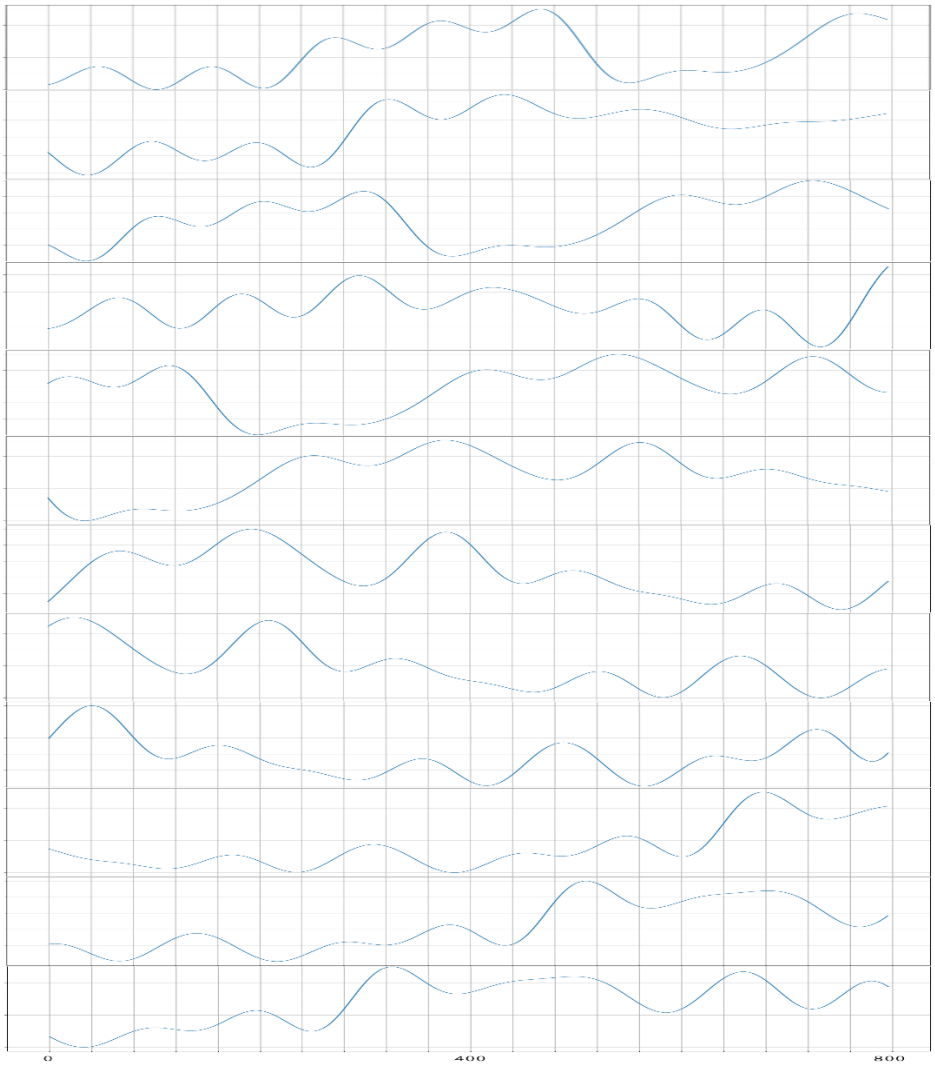

图 4 受试者 S1 对 char01 字符（B）的第一轮测试波形图（Fz 通道）

上图 4 中第 1 行和第 8 列为 char01 字符（B）对应的行列数据段，从图中可以发现，仅凭借单次试验的数据很难从中提取到关于 P300 波形的有价值特征。于是我们针对每个待确定字符的五轮测试数据，按照行列序号对每个行列对应的五轮采样数据分别进行了叠加求平均的操作，获得每个待确定字符测试数据中十二个不同行列对应脑电波平均波形，并且也进行了可视化展示：

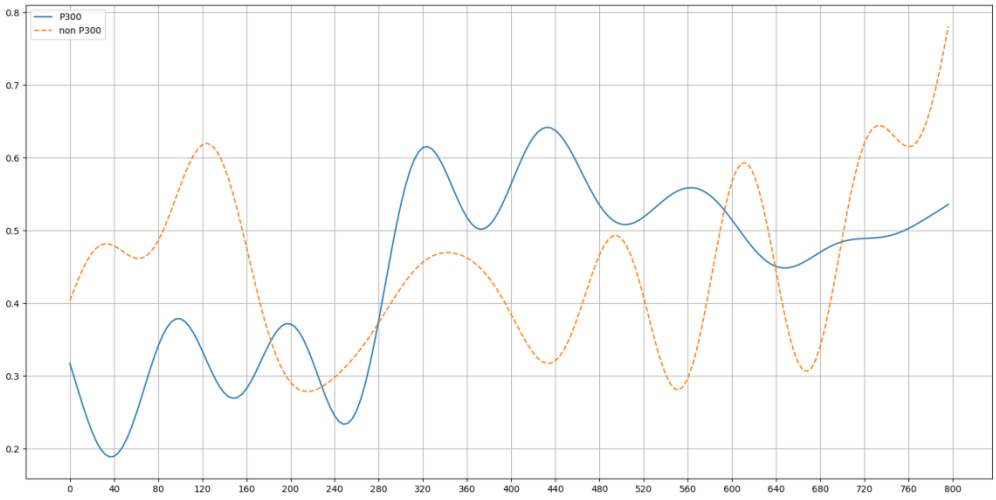

图 5 有 P300 信号（实线）和无 P300 信号（虚线）的平均波形对比

从图 5 中可以发现有 P300 信号的平均波形在 300ms-500ms 之间有明显的正向波峰，符合 P300 信号的特征，因此在单次字符测试中，对每个行列进行叠加求平均的操作可以有效地突出 P300 信号，能够有效提高分类性能。基于上述结果，我们最终对每次字符试验的每个行列进行了叠加求平均操作，即对每个行列对应的五轮采样数据按序对应求和取均值。经过上述操作后，单个字符测试试验中的 60 条有效数据合并为了 12 条，每条数据的特征向量为一维向量，长度为 800，对应类别标注为 0 或 1，并基于类别进行数据扩充。

3.3 模型建立与求解

分类识别就是将未知的事物或现象与各种已知的可以用来比较和模仿的典型的模式进行比较，看它与哪一类最接近。目前常用于脑电信号处理分类识别的算法有多种，可分为线性与非线性两类。线性方法有线性判别分析，感知器等。非线性方法有 K-最近邻，支持向量机等。本文分别采用 K-最近邻算法和支持向量机来研究 P300 信号的特征提取和分类识别。

3.3.1 支持向量机原理

支持向量机（Supprot Vector Machine，简称 SVM），最早由 Vapnik 等人在 20 世纪 90 年代提出，发展至今已有近 30 年的历史。它是一种有监督的机器学习算法，在二分类问题中应用广泛，并且适合于处理小样本数据，符合本题的数据规模。相比于其他单一的分类算法，支持向量机可以得到更高的预测准确率。这主要得益于支持向量机能够将低维的线性不可分空间转换成高维的线性可分空间。

支持向量机的思想是利用支持向量所构成的“超平面”，将不同类别的样本点进行划分。不管样本点是线性可分，近似线性可分还是线性不可分的，都可以利用“超平面”将样本点以较高的准确度分割开来。

运用支持向量机模型具有几个显著的优点：(1)支持向量机模型在增加或删除非支持向量的样本点时，不会改变分类器的效果，具有较好的“鲁棒”性。(2) 支持向量机模型因其良好的泛化能力，在一定程度上避免模型的假设变得过度严格。(3)支持向量机模型可以有效的避免在计算过程中出现局部最优的情况。

常见的支持向量机模型：

1. 线性可分的支持向量机

线性可分支持向量机处理的是严格的线性可分的数据集。如图 6 所示：

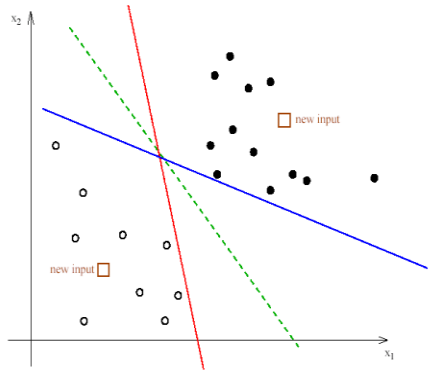

图 6 线性可分支持向量机示意图

对于线性可分支持向量机，需要给定一个严格线性可分的训练数据集。通过间隔最大化或等价地求解相应的凸二次规划问题学习，可以得到一个分离“超平面”：

$$w^T \cdot x + b = 0 \tag{1}$$

其中 w^T 表示分离超平面的法向量, b 表示截距，位于分离“超平面”之上的样本为正样本，之下的为负样本。对于线性可分的支持向量机的目标函数可以表示为如下表达式：

$$\begin{cases} \min\limits_{\alpha} \left(\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n} \alpha_i \right) \\ \text{s.t.} \sum_{i=1}^{n} \alpha_i y_i = 0 \alpha_i \geq 0 \end{cases} \tag{2}$$

其中，$(x_i \cdot x_j)$ 表示两个样本点的内积。利用拉格朗日乘子的 α_i 值计算“超平面”的参数 w 与 b 的值为：

$$\begin{cases}\widehat{w}=\sum_{i=1}^{n}\hat{\alpha}_l y_i x_i \\ \hat{b}=y_j-\sum_{i=1}^{n}\hat{\alpha}_l y_i\left(x_i\cdot x_j\right)\end{cases} \tag{3}$$

2.近似线性可分支持向量机

近似线性可分支持向量机通常被称为线性支持向量机。如图 7 所示，在近似线性可分支持向量机上有少量样本点不满足函数间隔大于 1 的情况。

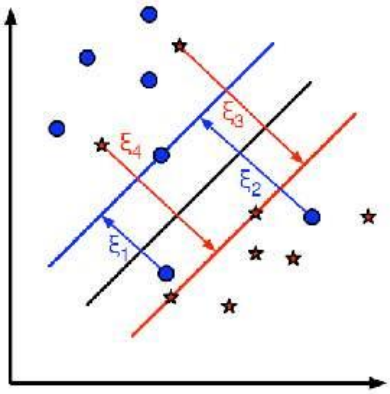

图 7 近似线性可分支持向量机示意图

近似线性可分支持向量机的“超平面”与线性可分的支持向量机相同。对于此类问题，通常的解题思路是在所有样本点的函数间隔上均加上一个松弛因子 ξ_i，并且满足 $\xi_i \geq 0$。因此，根据线性可分支持向量机的目标函数我们可以得到近似线性可分支持向量机的目标函数为：

$$\begin{cases}\min\limits_{\alpha}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)-\sum_{i=1}^{n}\alpha_i\right) \\ \text{s.t. } \sum_{i=0}^{n}\alpha_i y_i=00\leq\alpha_i\leq C\end{cases} \tag{4}$$

利用拉格朗日乘子 α_i 的值计算“超平面”的参数 w 与 b 的值为：

$$\begin{cases}\widehat{w}=\sum_{i=1}^{n}\hat{\alpha}_l y_i x_i \\ \hat{b}=y_j-\sum_{i=1}^{n}\hat{\alpha}_l y_i\left(x_i\cdot x_j\right)\end{cases} \tag{5}$$

3.非线性支持向量机

非线性支持向量机用于无法使用某个线性的“超平面”对样本点进行分割的情况。如图 8 所示：

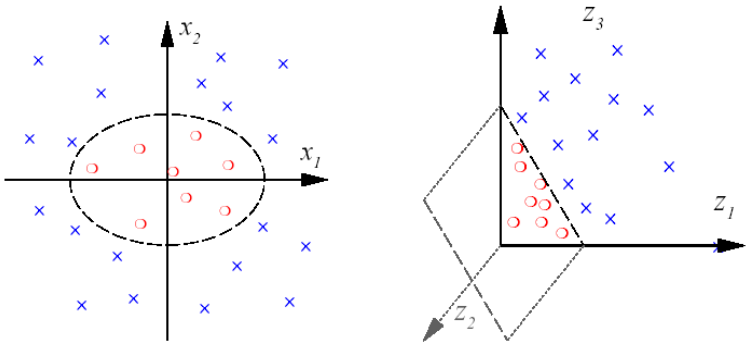

图 8 非线性可分支持向量机示意图

非线性支持向量机的核心思想是把原始的数据扩展到更高维度的空间中，然后在高维空间中实现样本点的线性可分。非线性向量机模型的构建一般可以分为如下两个步骤：

（1）将原始空间中的样本点进行映射使其变换到高维度的特征空间

（2）在特征空间中寻找一个可以用于识别各类别样本点的线性“超平面”。

假设在原始空间中的样本点为 x，如若将所有样本通过某种转换 $\Phi(x)$ 映射到高维空间中，则其对应的非线性支持向量机的目标函数可以表示为：

$$\begin{cases} \min\limits_{\alpha} \left(\frac{1}{2}\sum_{i=1}^{n}\ \sum_{j=1}^{n}\ \alpha_i\alpha_j y_i y_j \left(\Phi(x_i)\cdot\Phi\left(x_j\right)\right) - \sum_{i=1}^{n}\ \alpha_i\right) \\ \text{s.t. } \sum_{i=0}^{n}\ \alpha_i y_i = 00 \le \alpha_i \le C \end{cases} \tag{6}$$

其中，内积 $\Phi(x_i)\cdot\Phi(x_i)$ 可以利用核函数进行替换，即 $K(x_i,x_i)=\Phi(x_i)\cdot\Phi(x_i)$。计算利用拉格朗日乘子 α_i，同样可以获得线性“超平面”的参数 w 与 b 的值为：

$$\begin{cases} \widehat{w} = \sum_{i=1}^{n}\ \hat{\alpha}_l y_i x_i \\ \hat{b} = y_j - \sum_{i=1}^{n}\ \hat{\alpha}_l y_i K\left(x_i,x_j\right) \end{cases} \tag{7}$$

支持向量机引入核函数的主要目的是通过核函数将原本在高维特征空间中的计算放到输入空间中完成。从本质上来说，核函数只是低维的运算结果，并没有采用低维到高维的映射。使用核函数可以帮助研究者省去高维空间中的复杂运算过程，核函数主要包括以下几种：

a) 线性核函数

线性核函数的表达式为 $K\left(x_i,x_j\right)=x_i\cdot x_i$，故其所对应的分割“超平面”为

$$f(x)=\sum_{i=1}^{n}\ \hat{\alpha}_l y_i x_i x + \left(y_i - \sum_{i=1}^{n}\ \hat{\alpha}_l y_i x_i\cdot x_j\right) \tag{8}$$

线性核函数实际上就是线性可分的支持向量机模型。

b) 多项式核函数

多项式核函数的表达式为 $K(x_i, x_i) = (\gamma(x_i \cdot x_i) + r)^p$，故其所对应的分割“超平面”为

$$f(x) = \sum_{i=1}^{n} \hat{\alpha}_l y_i \left(\gamma\left(x_i \cdot x_j\right) + r\right)^p + \left(y_i - \sum_{i=1}^{n} \hat{\alpha}_l y_i \left(\gamma\left(x_i \cdot x_j\right) + r\right)^p\right) \quad (9)$$

其中 γ 和 p 均为多项式的参数。

c) 高斯核函数

高斯核函数的表达式为 $K\left(x_i, x_j\right) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$，故其所对应的分割“超平面”为

$$f(x) = \sum_{i=1}^{n} \hat{\alpha}_l y_i \exp\left(-\gamma \mid x_i - x_j \parallel^2\right) + \left(y_i - \sum_{i=1}^{n} \hat{\alpha}_l y_i \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)\right) \quad (10)$$

其中 γ 为高斯核函数的参数，高斯核函数也被称为径向基核函数（Radial Basis Function, RBF），是一种沿径向对称的标量函数。

d) Sigmoid 核函数

Sigmoid 核函数的表达式为 $K(x_i, x_i) = \tanh(\gamma(x_i \cdot x_i) + r)$， 故其所对应的分割“超平面”为

$$f(x) = \sum_{i=1}^{n} \hat{\alpha}_l y_i \tanh\left(\gamma\left(x_i \cdot x_j\right) + r\right) + \left(y_i - \sum_{i=1}^{n} \hat{\alpha}_i y_i \tanh\left(\gamma\left(x_i \cdot x_j\right) + r\right)\right) \quad (11)$$

3.3.2 模型的建立

针对问题一，我们采用非线性支持向量机进行训练和分类的步骤如表 2 所示。

表 2 支持向量机的训练和分类步骤

支持向量机的训练和分类步骤
1.输入训练数据$(x_i, y_i), i = 1,2,\cdots,n, x \in R^n, y \in \{0,1\}$，如果 $y = 0$ 则表示没有出现 P300 电位；如果 $y = 1$ 表示出现了 P300 电位。
2.选择核函数及其他参数 核函数可以选择线性核、rbf 核、多项式核、sigmoid 核等；其他参数诸如惩罚系数、gamma 值、多项式最高次数等也可以选择不同的值进行分别训练和评估。
3.模型训练和评估 使用预处理后的数据输入模型进行训练，样本数为 240 个。使用交叉验证法，对模型进行训练中的评估，从而对核函数和其他参数进行择优。
4.预测数据 使用前一步骤择优后的模型，对测试数据进行预测，得出结果。

我们选取了多种不同的核函数，分别进行训练和 10 折交叉验证，比较他们的交叉验证

准确率根据交叉验证的结果，我们最终选择线性核函数。而对于其他参数，也使用类似的方法进行分别训练和验证，然后择优选取。最终使用择优后的各种参数训练得到的模型对测试数据进行预测，核函数的交叉验证结果与最优参数如表 3 所示：

表 3 参数选择结果

核函数	十折交叉验证准确率										平均
linear	0.958	0.916	0.875	0.875	1	0.962	0.875	0.916	0.875	0.875	0.912
rbf	0.833	0.791	0.75	0.791	0.833	0.791	0.875	0.916	0.833	0.958	0.837
sigmoid	0.916	0.87	0.916	0.833	0.875	0.791	0.958	0.75	0.75	0.791	0.845
poly	0.791	0.833	0.833	0.916	0.75	0.916	0.75	0.791	0.833	0.916	0.833
最佳参数	核函数 kernel=linear，惩罚系数 C=1，决策函数 decision function=ovr										

3.4 预测结果及分析

使用前述的支持向量机模型对 S1-S5 被试者的测试数据进行预测，获得其行列刺激对应数据样本的分类标签，即含有 P300 电位（类别标记为 1）和不含 P300 电位（类别标记为 0），然后定位出待测字符，总共得出对 char13-char22 共 10 个待测字符的结果。

与训练数据保持一致，我们对单次试验中不同轮次内相同行/列的标识符对应的数据进行叠加求平均的处理，同时为了讨论尽可能使用较少轮次的效果，我们对 5 轮试验的数据进行逐次剔除最后一轮的操作，分别使用 5、4、3、2、1 轮数据，将 5 个被试者得出的行列预测结果分别进行取众数操作，得到最终的行列结果从而得出预测字符，如表 4 到表 8 所示。

表 4 五轮测试数据预测结果

	char13	char14	char15	char16	char17	char18	char19	char20	char21	char22
S1	7	1,2,11,12	6,7,11	6,10	1,4,9	1,6,8	1,2,8,11	2,5,12	1,2,7,8	12
S2	7	4,10,12	7	10	2	7,8,9,12	2,3,11	4,12	5,6,7	缺失
S3	3,7	1	7	4,5,10	2	4,8	2,11	4,12	1	缺失
S4	3,7	1	7	None	2,9	6,8	11	12	1,2,7,9	5,6,8,12
S5	7	1,12	6,7	5,10	9	6	2,9,11	12	1,7	6,12
结果	3,7	1,12	6,7	5,10	2,9	6,8	2,11	4,12	1,7	6,12

字符	M	F	5	2	I	6	K	X	A	0

表 5 四轮测试数据预测结果

	char13	char14	char15	char16	char17	char18	char19	char20	char21	char22
S1	7	1,2,12	6,7,11	6,10	1,4,9	1,6,8	1,2,4,8,11	2,5,12	1,2,4,7,8	6,12
S2	7	4,10,12	5,7	5,10	2,5	8,9,12	2,3	4,12	6,7	缺失
S3	3,7	1	7	4,5,10	2	4,8	2,5,11	4,8,12	1	缺失
S4	3,7,10	1	6	4	2	4,5,6,7,8	8,10,11	12	1,2,7,9	6,8,12
S5	7	1,12	6,7	5,10	2,9	6	2,11	12	1,7	6,12
结果	3,7	1,12	6,7	5,10	2,9	6,8	2,11	4,12	1,7	6,12
字符	M	F	5	2	I	6	K	X	A	0

表 6 三轮测试数据预测结果

	char13	char14	char15	char16	char17	char18	char19	char20	char21	char22
S1	7	1,2,12	6,7,11	5	1,4,9	1,5,8	1,3,4,8,11	2,12	1,2,4,8	6,12
S2	1,7	4,10,12	3,7	10	2	7,8,12	3	4,12	6,7,11	缺失
S3	3,7	None	7	4,5	2	4,8	2,5,11	4,8,12	1	缺失
S4	3,7,8,10	1,12	6	None	None	5,6,7,8	10,11	2,10	1,2,3,7,9	3,6,8,12
S5	7	1,12	1,6,7,8	5,10	2,9	6,8	11	8,11,12	1,5,7	6,12
结果	3,7	1,12	6,7	5,10	2,9	5 或 6,8	3,11	2 或 4,12	1,7	6,12
字符	M	F	5	2	I	Z 或 6	Q	L 或 X	A	0

表 7 二轮测试数据预测结果

	char13	char14	char15	char16	char17	char18	char19	char20	char21	char22
S1	3,7	1,2,12	6,7,11	5	1,9	1,4,5,8	1,2,8,11	1,2,8,12	1,4,5,6,7,8	3,6,12
S2	1,4,5,7,12	1,4,10,11,12	1,7	5,10	2	7,8,9,12	2,3,5	12	2,8,11,12	缺失
S3	3,7	1,3	3,7	4,5,10	2	4,8	2,5,11	4,5,8	1,8	缺失

S4	3,7,8,10	1,12	6	11	None	5,7,8,11	8,9,10	2,10	1,2,3,7,8,9	3,5,6,8,12
S5	5,7,12	1,2,12	6,7	2,5	9	1,6,8	2,9,11	8,12	1,5,7	6,12
结果	3,7	1,12	6,7	5,10	2,9	1/4/5,8	2,11	2,12	1,8	6,12
字符	M	F	5	2	I	B/T/Z	K	L	B	0

表 8 一轮测试数据预测结果

	char13	char14	char15	char16	char17	char18	char19	char20	char21	char22
S1	3,4,7	6,9,10,12	2,6,7,8,9,11	5	1,8,9,12	1,5,8	1,2,5,8,11	1,2,3,7,8,11,12	1,2,4,5,6,7,8	3,6,12
S2	1,7,10,12	1,4,10,11,12	1,5	4,5,9,10	2,3,7	6,7,8,9,1,12	2,7,11	12	1,8,11,12	缺失
S3	3,7	1,3,6,9	3,4,5,6,7,11	3,4,5,9	2,3,7	4,8,10	11,12	4,5,8,9,10,11	1,5,8,9	缺失
S4	3,7,8,10	2,12	1,6,10	None	10	1,7,8,11	9,10	2,9,10	1,2,7	1,6,8,10,11,12
S5	1,6,7,8,10	2,12	1,5,6,7,9	5,7	1,4,9,12	6,8	5,11	1,7,8,12	1,3,5,7	1,6,7,10
结果	3,7	1/2,12	6,7	5,9	1/2/3,9/12	1/6,8	2/5,11	1/2,8/12	1,7/8	6,10/12
字符	M	F 或 L	5	1	C,F,I,L,O,R	B/6	K/3	B/F/H/L	A/B	8/0

从表 4 到表 8 中可以发现，在逐步减少测试数据轮次的过程中，综合预测结果逐渐变得不唯一，即模型对测试数据进行预测的结果随着数据轮次的减少而变得不稳定。针对题目所涉及到的数据，在仅使用四轮测试数据时，能够得出唯一的结果，当测试数据的轮次继续减少时，无法再得出确切的结果。

最终我们模型给出的 char13-char22 对应字符的预测结果分别为：M, F, 5, 2, I, 6, K, X, A, 0，使用了 4 轮测试数据。

四、 问题二的模型建立及求解

4.1 问题分析

针对问题二，由于采集的脑电数据量较大，可能存在大量冗余信息。问题二的实质就是设计一个通道选择算法，给出针对每个被试者更有利于分类的通道名称组合。每个通道对于模型分类的贡献评估是该问题的关键，我们采用了一种基于识别准确率指标的递归剔除无效导联的方法，结合在问题一中提出的 SVM 算法模型，对题目所给定的前五个测试字符进行测试和准确率评估后计算识别准确率，对 20 个通道进行逐步筛选最终得到最优通道组合。

4.2 模型建立与求解

通道选择在脑电信号采集中是十分重要的一步，选择合适的通道能够有效地提升脑部活动的检测效率，也能够减小数据的处理难度和耗时。图 9 显示了各个通道的名称和位置，标红的是题目所给的 20 个通道，表 8 为记录通道的标识符。

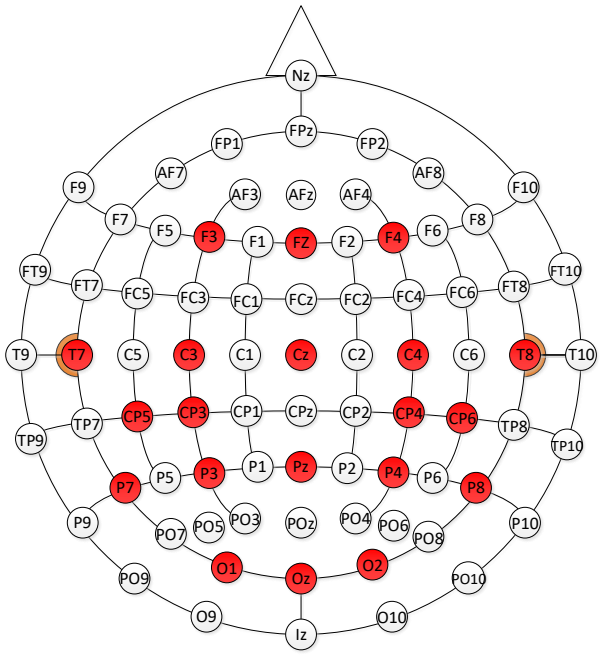

图 9 脑电信号采集通道图

表 9 采集通道的标识符

标识符	通道名称	标识符	通道名称
1	Fz	**11**	CP5
2	F3	**12**	CP6
3	F4	**13**	Pz
4	Cz	**14**	P3
5	C3	**15**	P4
6	C4	**16**	P7
7	T7	**17**	P8
8	T8	**18**	Oz
9	CP3	**19**	O1
10	CP4	**20**	O2

通道选择方法就是选择 P300 电位最为显著的通道，同时也是对分类识别最有利的通道，这也是减少数据量最为直接的办法。通常可以针对每一个被试者进行详细的通道选择

计算，剔除无效通道，得出最优通道组合。

递归通道剔除是依据对通道对分类结果准确率的贡献对通道进行分析选择的方法，该方法使用了一种递归的思想来剔除无效通道获取较大贡献通道的。识别准确率指标的计算公式如下：

$$CS = \frac{TP}{TP+FP+FN} \tag{12}$$

其中 TP 为正确识别的正值数（即预测结果和正确结果均为含有 P300 信号），FP 为错误识别的正值数（即预测结果为含有 P300 信号，正确结果为不含 P300 信号），FN 为错误识别的负值数（即预测结果为不含 P300 信号，正确结果为含有 P300 信号）。可以注意到该公式中并没有考虑 TN（正确识别的负值），这是由于数据中正负样本不均衡，去掉 TN 项可以使得该准确率公式更注重于正样本。

针对每个受试者，为尽可能保证准确性，模型使用测试数据的全部五轮试验结果进行预测，算法的具体步骤如下：首先将 20 路通道都参与进算法，计算得到一个 CS 值，然后剔除其中一个通道，余下通道继续参与算法得到一个的新的 CS 值，所计算出的 20 个值中最大的对应通道可以认为其 P300 电位最小，将其剔除，剩余的 19 个通道继续进行递归运算，最后计算出的通道组合就是对识别贡献最大的。

我们使用问题一训练得到的 SVM 模型，使用上述递归通道剔除法对每个受试者进行逐个的通道剔除，具体步骤如表：

表 10 基于 SVM 的递归通道剔除法步骤

基于 SVM 的递归通道剔除
1.输入 20 个通道的数据，对测试数据进行预测，将预测结果带入 CS 计算公式计算得到一个 CS
2.逐个剔除通道 i，计算得到新的 CS 值，从中选择贡献最小的通道剔除
3.剩余通道进行步骤 2 的递归运算
4.最终得出最优通道组合

4.3 问题求解与结果分析

使用上述的递归剔除算法，使用问题一中训练的 SVM 算法模型对前五个待测字符（M, F, 5, 2, I）进行预测，根据准确率指标对每个受试者分别进行逐个通道剔除，每次剔除后的分类准确率如下图所示：

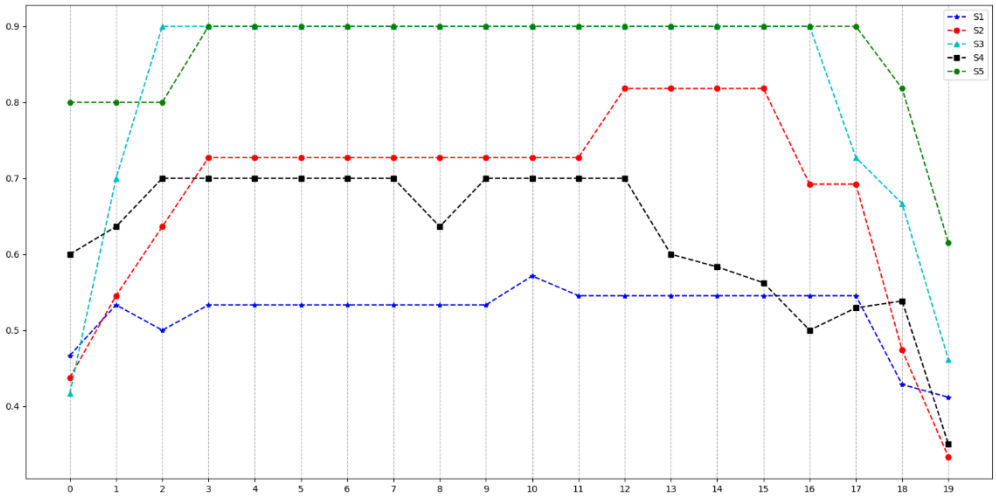

图 10 通道剔除过程中的准确率指标变化

各个受试者的通道剔除顺序如表 11 所示：

表 11 各受试者的通道剔除次序

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	最后剩余
S1	O1	O2	P3	CP6	CP5	T7	CP3	P4	Pz	Oz	C4	Cz	C3	CP4	T8	Fz	P7	F4	F3	P8
S2	F3	F4	Fz	O2	O1	CP6	P8	P4	Oz	P7	P3	C3	Pz	Cz	T8	CP3	T7	CP4	CP5	C4
S3	Pz	T8	O2	P8	O1	P4	CP5	P7	CP3	Oz	F3	F4	C4	Fz	Cz	CP4	CP6	C3	T7	P3
S4	F4	O2	P8	P7	P4	P3	Oz	CP6	T8	CP4	CP5	CP3	C3	F3	C4	O1	Pz	Cz	T7	Fz
S5	O2	O1	CP3	P8	P4	CP6	P3	CP5	C3	F3	F4	CP4	C4	Oz	Pz	T7	Fz	Cz	Oz	P7

图 10 中的横坐标为通道剔除数目，从原始 20 通道（0）逐个剔除到仅剩一个通道（19）；纵坐标为识别准确率指标；五种不同颜色的数据点对应五个受试者。可以结果中发现，在逐个剔除通道的过程中，准确率指标都出现了先上升进入平台期，然后在通道非常少的时候发生大幅下降的现象。针对此现象进行分析，我们认为前面 4-5 个左右被剔除的通道对于分类性能有负面效果，在通道筛选中进行优先剔除，最后剩余的 5 个左右的通道对模型的分类具有关键性作用，在通道筛选中优先进行保留。

根据上表中的通道剔除次序，我们提出了一种对通道进行评分的算法，对每个通道依据以下公式计算评分：

$$\text{Score} = \sum_{i=1}^{5} p_i \quad (12)$$

其中 Pi 为该通道在第 i 个受试者通道剔除过程中的次序，该公式对优先剔除的通道的评分会较低，对最后剔除的通道评分较高，根据剔除次序体现了不同通道在筛选过程中的重要性。

对每个通道计算上述通道评分，并将五个受试者的通道评分进行对应求和，对所有 20 个通道进行总体排序，从而得到出总体的最优通道，排序结果如表 12 所示

表 12：各通道评分

通道	Cz	T7	C4	Fz	CP4	C3	P7	F3	T8	Pz	Oz	CP5	CP3	P3	F4	CP6	P8	P4	O1	O2
评分	77	77	72	70	70	65	59	55	55	55	55	50	47	47	44	41	38	32	29	12

针对题目所要求的筛选出 10 到 20 个最优通道，对上述排序结果再进行一次通道剔除，从全部 20 个通道开始，逐步从后向前剔除通道，直至剩余 10 个为止，使用五个受试者的平均准确率指标，得出通道剔除过程中的变化率如下图所示：

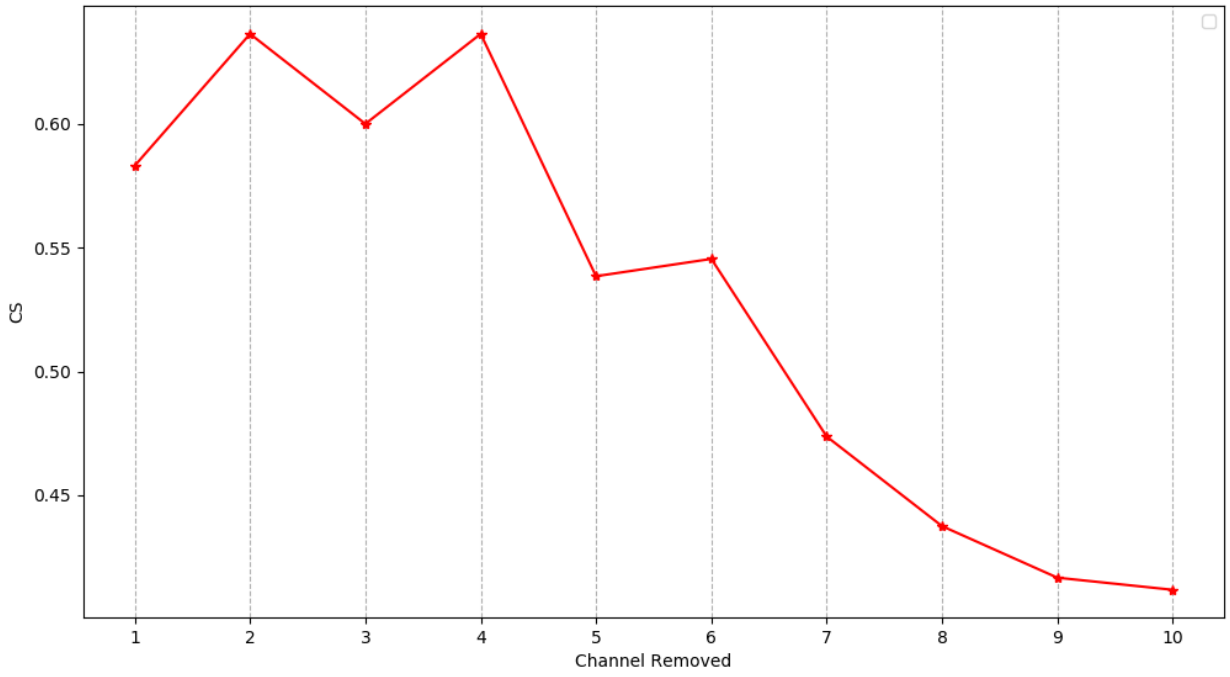

图 11 通道剔除过程的准确率变化

可以得出当通道数下降到 16 个时，出现了准确率指标的大幅下跌，因此以保留十六个通道最为恰当。因此我们对五个受试者分别选取的最优通道为：

S1：CP5, T7, CP, P4, Pz, Oz, C4, Cz, C3, CP4, T8, Fz, P7, F4, F3, P8

S2：O1, CP6, P8, P4, Oz, P7, P3, C3, Pz, Cz, T8, CP3, T7, CP4, CP5, C4

S3：O1, P4, CP5, P7, CP3, Oz, F3, F4, C4, Fz, Cz, CP4, CP6, C3, T7, P3

S4：P4, P3, Oz, CP6, T8, CP4, CP5, CP3, C3, F3, C4, O1, Pz, Cz, T7, Fz

S5：P4, P6, P3, CP5, C3, F3, F4, CP4, C4, Oz, Pz, T7, Fz, Cz, Oz, P7

综合选择适合于所有受试者的十六个最优通道为：Cz, T7, C4, Fz, CP4, C3, P7, F3, T8, Pz, Oz, CP5, CP3, P3, F4, CP6。

五、 问题三的模型建立及求解

5.1 问题分析

鉴于监督学习训练过程长，往往需要花费很长时间获取有标签样本来训练模型，容易造成被试者疲劳，为减少训练时间，针对问题二我们提出了基于超限学习机 ELM（extreme learning machine）的在线半监督学习方法，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。首先在有标签的数据样本上训练出一个初始分类器，在线测试的过程中，该分类器对在线获取的无标签数据进行分类，并加入训练集以更新分类器。本文对 ELM 算法进行了改进，针对在线学习的形式，我们引入了序列更新的形式，而且引入了正则化项以获得更好的泛化能力。该方法能在较短的训练时间里满足在线计算要求，从实验结果来看，基于 ELM 的在线半监督学习算法在时间和性能上有较好的平衡。

5.2 模型建立与求解

5.2.1 ELM 简介

ELM 最初是作为训练单隐层前馈神经网络(single-hidden layer feedforward neural network,SLFN)的一种方法，该网络可以随机产生输入层权值，只需确定隐含层神经元个数及隐含层神经元的激活函数，并利用最小二乘法即可直接获得输出层的权值。由于该方法具备封闭解形式，计算复杂度低，且泛化性能较好，因此本文采用基于 ELM 的分类算法进行分类识别。

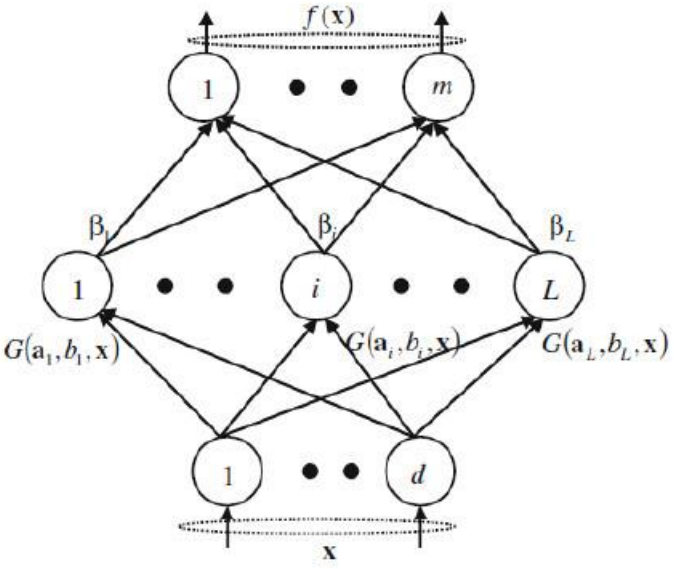

图 12 单隐层前馈神经网络示意图

含 L 个隐层节点的单隐层前馈神经网络如图 5.1 所示。其中输入层为 d 维向量，隐含层包括 L 个隐含神经元，输出层的输出为 m 维的向量，对于二分类问题，显然该向量是一维的。

对于含有 L 独立样本的数据集 $(x_i, t_i), i = 1,2,\cdots,L$ 对应的单隐层前馈神经网络输出可以由一个 $L \times L$ 的矩阵表示：

$$\mathrm{H} = \begin{pmatrix} G(v_1, \mathrm{x}_1, b_1) & \cdots & G(v_L, \mathrm{x}_1, b_L) \\ \vdots & \ddots & \vdots \\ G(v_1, \mathrm{x}_L, b_1) & \cdots & G(v_L, \mathrm{x}_L, b_L) \end{pmatrix} \tag{13}$$

其中 $v_i, i = 1,\cdots,L_3$ 表示输入向量与第 i 个隐层节点的权向量， $b_i, i = 1,\cdots,L$ 是第 个隐层节点的偏置项。由于在 ELM 算法中， 和 都是随机确定的，且是固定的。 是激活函数，可以是线性函数，也可以是 sigmoid 函数。于是，该单隐层前馈神经网络输出函数可以表示为：

$$H\beta = T \tag{14}$$

对于包含正则项的 ELM 二分类问题，其优化目标为：

$$\begin{aligned} &\min_{\beta} \frac{1}{2} \| \beta \|^2 + \frac{C}{2} \sum_{i=1}^{N} \| \xi \|^2 \\ &\text{s.t. } \mathrm{h}(\mathrm{x}_i)\boldsymbol{\beta} = t_i - \xi_i, i = 1,\cdots,N \end{aligned} \tag{15}$$

使用最小二乘可以找到 $\boldsymbol{\beta}$ 的封闭解形式：

当 $N > L$ 时，$\boldsymbol{\beta} = \left(\mathrm{H}^{\mathrm{T}}\mathrm{H} + \frac{\mathrm{I}}{C}\right)^{-1} \mathrm{H}^{\mathrm{T}}\mathrm{T}$ (16)

当 $N < L$ 时，$\boldsymbol{\beta} = \left(\mathrm{H}\mathrm{H}^{\mathrm{T}} + \frac{\mathrm{I}}{\mathrm{C}}\right)^{-1} \mathrm{T}$ (17)

5.2.2 加权序列化 ELM

在二分类问题中，往往会出现样本不平衡问题，本文直接采取加权的方式，对样本较少的数据加上较大的惩罚系数，这样就极大降低了样本较少类别的分错概率。常用的加权处理方法有对样本的误差平方和进行加权和对样本的误差绝对值进行加权。经过加权处理后，两种类别的样本基本达到平衡，产生的误差也相当。本文使用的是对对误差平方和进行加权。

对误差平方和进行加权的 ELS 优化目标可以表示为：

$$\min_{\beta} \frac{1}{2} \| \boldsymbol{\beta} \|^2 + \frac{C}{2} \sum_{i=1}^{N} \omega_i \|\xi_i\|^2$$
$$\text{s.t. } \mathrm{h}(\mathrm{x}_i)\boldsymbol{\beta} = t_i - \xi_t, i = 1, \cdots, N$$

(18)

当 $N > L$ 时，$\boldsymbol{\beta} = \left(\mathrm{H}^{\mathrm{T}}\mathrm{WH} + \frac{\mathrm{I}}{\mathrm{C}}\right)^{-1} \mathrm{H}^{\mathrm{T}}\mathrm{WT}$ (19)

当 $N < L$ 时，$\boldsymbol{\beta} = \left(\mathrm{HWH}^{\mathrm{T}} + \frac{\mathrm{I}}{C}\right)^{-1} \mathrm{WT}$ (20)

其中 $\mathrm{W} = \mathrm{diag}\{w_i\}_x \mathrm{i} = 1, \cdots, N$ 为加权矩阵。对于权值 w_i 的选择，可以按照如下公式计算：

$$\omega_i = \begin{cases} \frac{1}{*(\text{ positive })}, \text{ ith sample } \in \text{ positive class} \\ \frac{1}{*(\text{ negtive })}, \text{ ith sample } \in \text{ negtive class} \end{cases} \quad (21)$$

其中 *(positive) 表示正样本个数， * (negtive) 表示负样本个数；本题中一轮试验中包含正样本个数为 2，负样本个数为 10，即 $*(\text{positive}) = 10, *(\text{negtive}) = 2$ 。

结合上面的方法，针对在线更新的形式，我们可以得到正则化的加权序列 ELM 方法如下：

假设在线获得了包含 N 个样本的数据块，我们可以将输出矩阵更新为：

$$\mathrm{H}_{(n+1)} = \begin{bmatrix} \mathrm{H}_{(n)} \\ \mathrm{H}_M \end{bmatrix} \quad (22)$$

其中，$(\cdot)_{(n)}$ 表示第 n 次的风险，H_M 表示新到达的 M 个数据样本形式的隐层输出矩阵。

当 $(N + M)$ 大于 L 时，更新后的输出权向量为：

$$\boldsymbol{\beta}_{(n+1)} = \left(H_{(n+1)}^T W_{(n+1)} H_{(n+1)} + \frac{1}{C}\right)^{-1} H_{(n+1)}^T W_{(n+1)} \begin{bmatrix} T_{(n)} \\ T_M \end{bmatrix} \quad (23)$$

其中 $W_{(n+1)}=\begin{bmatrix} W_{(n)} & \\ & W_M \end{bmatrix}$，$T_M$ 为新到达的 M 个数据样本的标签向量。令

$$K_{(n)} = H_{(n)}^T W_{(n)} H_{(n)} \tag{24}$$

于是有：

$$\begin{aligned} K_{(n+1)} &= H_{(n+1)}^T W_{(n+1)} H_{(n+1)} \\ &= K_{(n)} + H_M^T W_M H_M \end{aligned} \tag{25}$$

计算可得输出权值为：

$$\boldsymbol{\beta}_{(n+1)} = \boldsymbol{\beta}_{(n)} + \left(\frac{\boldsymbol{I}}{C} + \boldsymbol{K}_{(n+1)}\right)^{-1} \boldsymbol{H}_M^T \boldsymbol{W}_M \left[\boldsymbol{T}_M - \boldsymbol{H}_M \boldsymbol{\beta}_{(n)}\right] \tag{26}$$

表 13 基于 ELM 的在线半监督算法步骤

基于 ELM 的在线半监督算法
1.根据（13）式获取初始隐层输出矩阵 $H_{(0)}$，根据（20）式计算输出权向量 $\boldsymbol{\beta}_{(0)}$
2.当含 M 个数据的新的待分类的数据块到来时，使用训练好的分类器进行分类，得到其标签 $t_i, i = 1, \cdots, M$。根据（20）式构建新的输出矩阵 $H_{(n+1)}$ 根据 $H_{(n+1)} = \begin{bmatrix} H_{(n)} \\ H_M \end{bmatrix}$ 创建新的标签向量，使用（20）式计算输出权向量 $\boldsymbol{\beta}_{(n+1)}$
3.重复步骤 2 直到 $N > L$
4.根据（25）式计算 $K_{(0)}$，使用（20）式获取更新后的权值向量 $\boldsymbol{\beta}_{(n)}$
5.当含 M 个数据的新的待分类的数据块到来时，使用训练好的分类器进行分类，得到其标签 $t_i, i = 1, \cdots, M$。根据（20）式构建新的输出权向量 $\boldsymbol{\beta}_{(n+1)}$.并按照（25）式更新 $K_{(n+1)}$
6.重复步骤 5 直到没有新的数据

5.4 模型验证与结果分析

为了验证上述 ELM 模型的效果，我们实验了不同比例初始训练集下模型的在线半监督学习过程的准确率变化情况，如图所示。

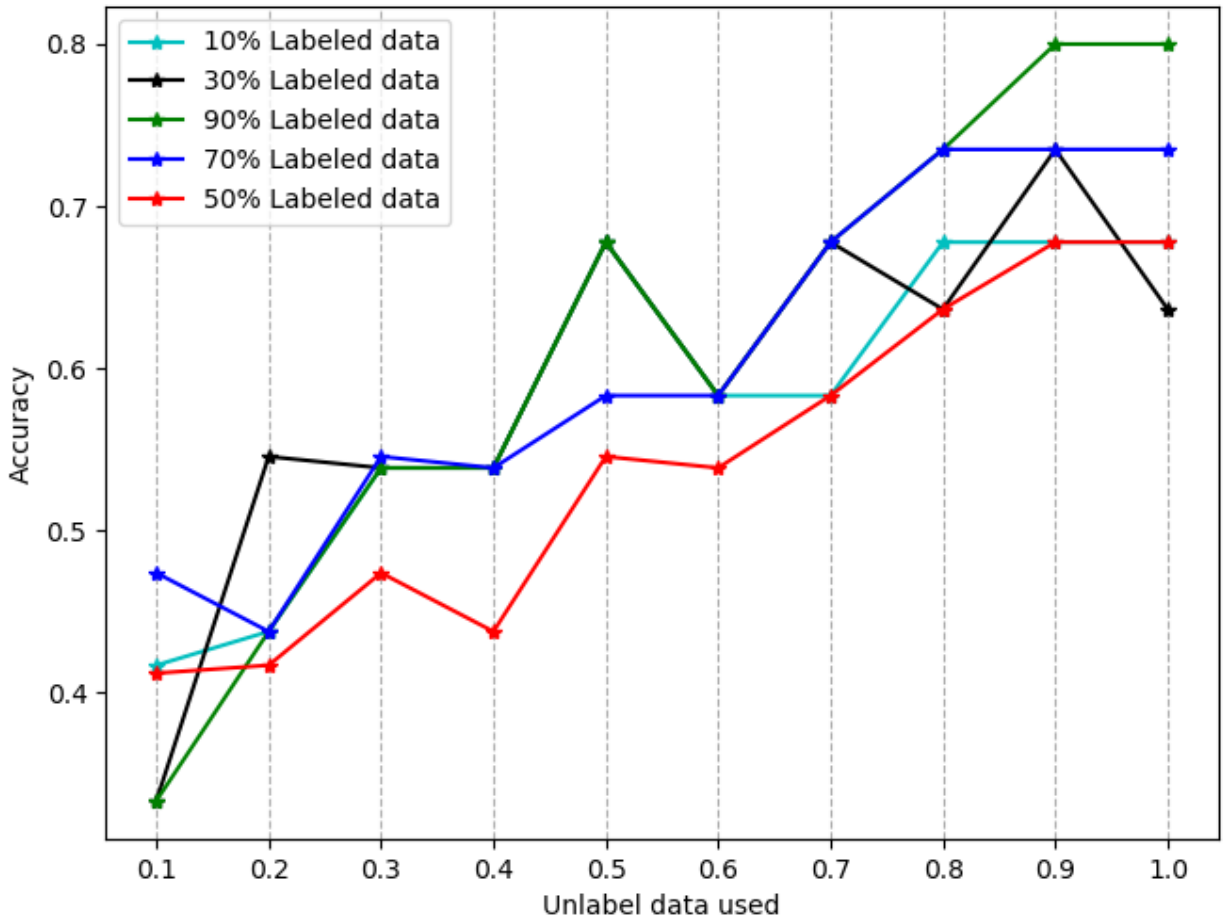

图 ELM 模型在线半监督学习的准确率变化

图中不同颜色的数据点分别代表使用不同比例有标签样本的模型情况，横坐标为加入剩余无标签样本的比例，纵坐标即为准确率。从数据的趋势可以看出，当逐步将剩余无标签样本加入学习过程时，模型的预测准确率有着比较大的提升，在仅使用部分有标签的样本的情况下，ELM 模型取得了较好的结果。

六、 问题四的模型建立及求解

6.1 问题分析

问题四所述的睡眠分期预测，本质上是根据各个睡眠期的四种波(Alpha，Beta，Theta，Delta)在各个睡眠期的特点对样本点进行分类，进而预测出睡眠期。问题要求对原数据集进行恰当的划分，分为训练集和测试集，使用训练集训练模型对五种不同的睡眠期数据进行分类，并使用测试集来评估模型的分类性能。解决问题的关键在于选取恰当的数据集划分方法和模型，并使用恰当的评价指标对模型进行评估。

原数据集中包含 3000 个样本，每个样本可以看作一个长度为 4 的一维向量，样本的维度并不高，我们尝试将数据集的样本进行三维空间的可视化展示。采用了主成分分析

（PCA）的方法，将样本数据的维度从 4 维降至 2 维后，在二维坐标系中进行散点图展示，结果如下图所示：

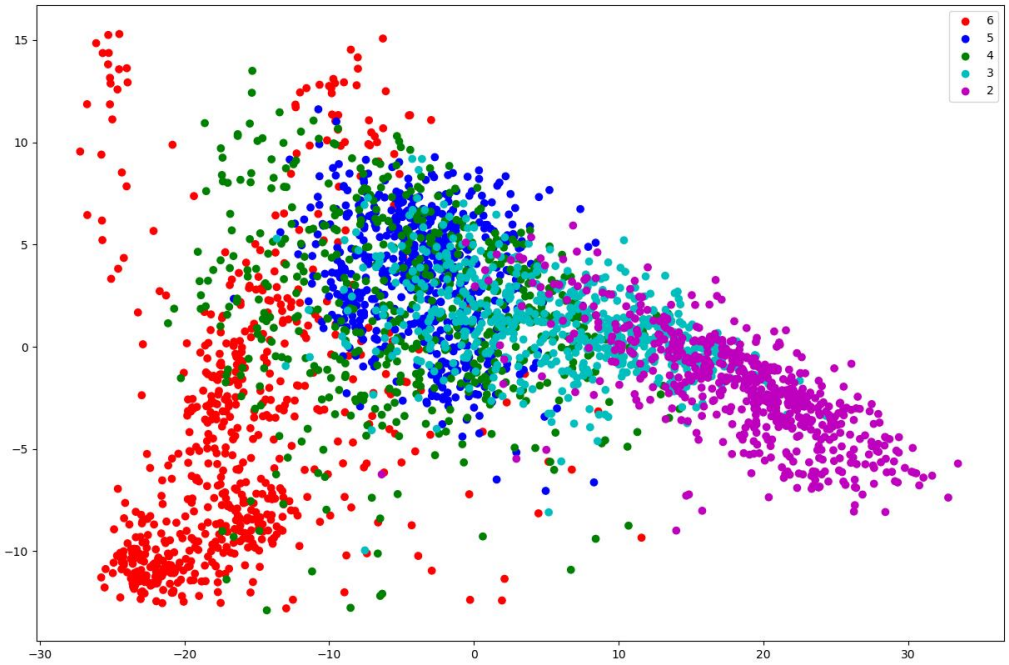

图 13 睡眠脑电数据的二维可视化

图中 2（紫色）、3（浅绿）、4（绿色）、5（蓝色）、6（红色）分别为深睡眠期、睡眠 II 期、睡眠 I 期、快速动眼期、清醒期的数据，数据也反映出了深睡眠期和清醒期的脑电数据是有着比较明显的差异，而中间三个时期的数据处于深睡眠期和清醒期的过渡状态，彼此差异并不明显。

另外，数据集的各个类别数量比较平衡（类均 600 个样本），我们采用随机洗牌的方式将原数据集打乱，然后根据一定比例划分训练集和测试集，在训练集上使用监督学习的分类方法进行分类，然后在测试集上进行效果的评估。

6.2 模型建立与评估

针对问题四使用了 K 最近邻算法（KNN）的分类方法，主要包括 KNN 模型的训练和 KNN 模型的分类两部分。

KNN 模型也被称为 K 最近邻算法，它的基本思想是搜寻最近的 k 个已知类别样本的位置以此来对未知类别样本进行预测。在该算法中，k 的取值对该模型的预测准确率至关重要。如果 k 值偏小，可能会导致 KNN 模型出现假设变得过度严格即进入到过拟合状态；而如果 K 值偏大，则会导致模型进入到欠拟合状态。通常 K 最近邻算法的训练和分类分为如下步骤：

表 14　K 最近邻算法的训练和分类步骤

K 最近邻算法的训练和分类步骤
1. 计算测试数据与各个训练数据之间的距离
2. 按照距离的递增关系进行排序
3. 选取距离最小的 K 个点
4. 确定前 K 个点所在类别的出现频率
5. 返回前 K 个点中出现频率最高的类别作为测试数据的预测分类

为了评价 KNN 模型中选择不同 K 值，以及测试集占总数据集的比值给模型预测准确率带来的影响，对 K 值和测试集比率都进行了各种取值的测试，K 值从 1 到 50、测试集比例从 0.1 到 0.9 的结果如下图所示：

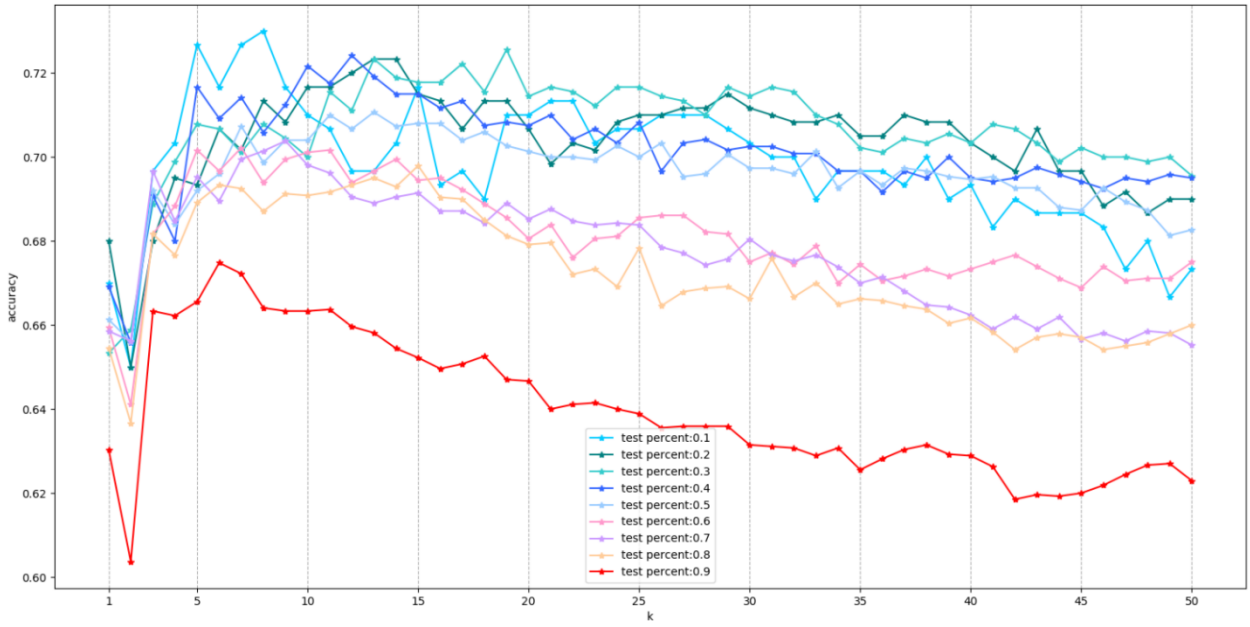

图 14 KNN 模型预测准确率对比 1

其中横坐标为 K 值，纵坐标为模型在测试集上的预测准确率，不同颜色的数据点为不同的测试集比例。从结果中可以得出，当 K 值取 5 到 15 之间时，能够在不同测试集比例的情况下都能得到较高的预测准确率。

针对从 5 到 15 之间的 K 值，进一步测试结果进行分析，如下图所示：

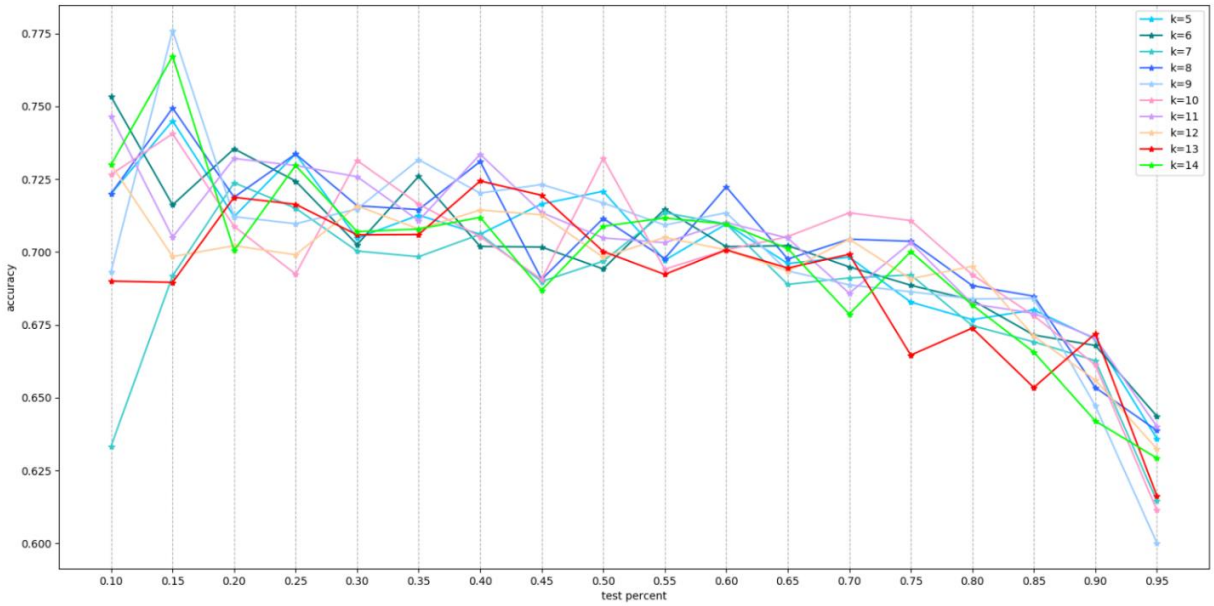

图 15 KNN 模型预测准确率对比 2

上图的横坐标为从 0.1 到 0.95 的测试集比例，纵坐标为预测准确率，不同颜色的数据点为不同的 K 值，可以得出预测准确率在测试集占比为 0.15 的时候达到最大，即训练集与测试集的最佳比值为 0.85：0.15，另外当测试集比例不断增加到 0.5 时，都没有出现准确率的明显下滑。

综上所述，可以认为用于此 KNN 模型的最少训练集比例为 50%，最佳训练集比例为 15%，合适的 K 值在 5 到 15 之间，针对这些条件对应的模型和分类结果，也进行了热图展示，如图 16 所示（测试集数量为 1500 条）：

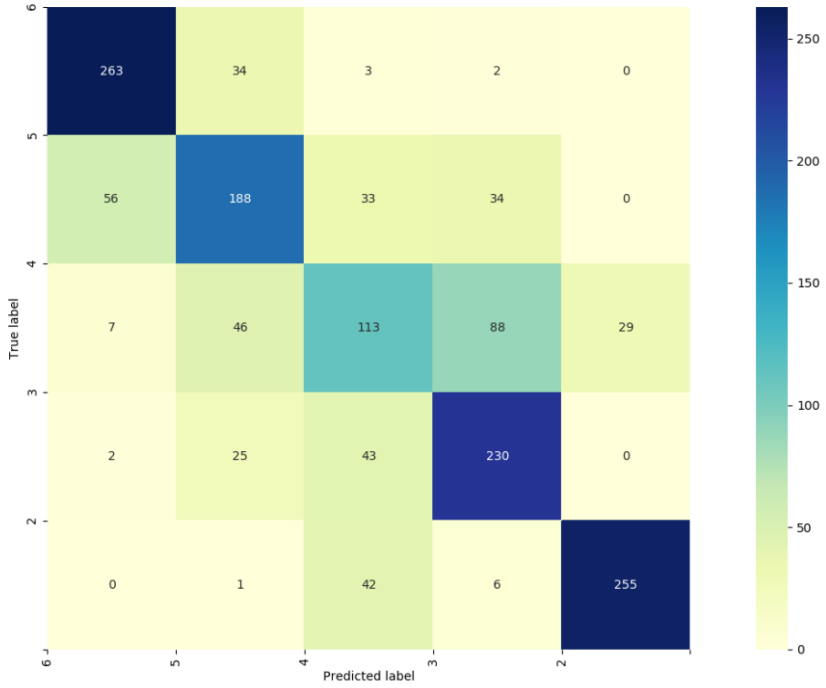

图 16 模型预测标签与正确标签的关系热图

热图的横纵坐标分别为模型对数据的预测标签和数据的真实标签，不同标签的交叉部分即表示模型对不同类数据进行预测时得出的类别情况，颜色越深的部分数据越多。可以看出，本模型对于深睡眠期（2）和清醒期（6）的预测效果最好，主要预测误差出现在中间三种过渡状态的分类结果中，在这一方面需要进一步的改进。

七、 总结

脑-机接口技术旨在不依赖正常的由外围神经或肌肉组织组成的输出通路的通讯系统，实现大脑与外部辅助设备之间的交流沟通，其中 P300 电位和睡眠脑电波的分析都是这一研究领域的热点。

针对问题一和问题二的 P300 电位对应字符的预测，本质上都是分类问题，在考虑多种分类方法之后，我们选择了基于线性核函数的支持向量机模型进行 P300 电位的分类识别，并在此基础上通过递归的通道剔除算法，为每位受试者选择了最优通道选择方案，并给出了适合所有受试者的总体最优方案。我们提出的模型对于对问题一和问题二都提出了较好的解决方案，其优点在于对数据做了充分的预处理，预测字符的准确率很高，也能够进一步推广到对脑电波的其他信号识别任务中。不足之处在于需要综合多个受试者才能得出较为准确的结果，需要对模型的识别算法进行进一步的改进来提高在单个受试者数据上

的预测准确性。

针对问题三涉及的半监督学习问题，我们选择了基于超限学习机 ELM（extreme learning machine）的在线半监督学习方法，仅利用一部分有标签的训练数据，对无标签数据进行在线学习，实现模型的自我更新，在时间和性能上都有着较好的效果。

针对问题四的睡眠脑波分类问题，我们选择建立 K 最近邻模型并对其参数进行优化，通过选择不同比例的训练集和测试集，择优得到较好的训练集和测试集的划分比例。最后对模型的预测效果进行了细节分析，模型对于深睡眠期和清醒期取得了较好的识别效果，但对中间过渡时期的分类还有待进一步的改进和优化。

本文对于各问题所提出的模型，都有比较大的灵活性和通用性，不仅可以用于本题所涉及的 P300 信号或睡眠脑波的分析，也可以推广到其他电子信号的分类识别任务上，而且还有比较大的改进和提升空间。

参考文献

[1] 吕竟雷. 基于支持向量机的 P300 脑电信号分类研究[D].西北工业大学,2005.

[2] 黄安湖. P300 脑电诱发电位的分类识别及在脑机接口中的应用[D].山东大学,2008.

[3] 薛晓炎. 脑电信号分类研究的方法[D].南京理工大学,2008.

[4] 林亚静. 基于 P300 的脑机接口的数学模型与算法研究[D].长沙理工大学,2014.

[5] 沈之芳. 基于 P300 的脑机接口及其在线半监督学习[D].华南理工大学,2014.

[6] 王俊杰. P300 脑机接口的半监督和无监督学习算法研究[D].华南理工大学,2017.

[7] 辛雨航. 基于半监督与时序模型的脑电信号特征提取方法研究[D].山东大学,2018.

[8] 肖郴杰. 基于深度学习的 P300 脑机接口分类算法研究[D].华南理工大学,2019.

[9] 冯宝,张绍荣.组稀疏贝叶斯逻辑回归的 P300 信号通道自动选择算法[J].东北大学学报(自然科学版),2019,40(09):1245-1251.

[10] 庄玮,段锁林,徐亭婷, 等.基于支持向量机的脑电信号分类方法研究[J].科学技术与工程,2014,14(9):73-77. DOI:10.3969/j.issn.1671-1815.2014.09.015.

[11] 汤伟,耿逸飞.基于余弦相似的改进 CEEMD 脑电信号去噪方法[J].南京邮电大学学报（自然科学版）,2020,40(3):8-14. DOI:10.14132/j.cnki.1673-5439.2020.03.002.

[12] Cecotti Hubert,Gräser Axel. Convolutional neural networks for P300 detection with application to brain-computer interfaces.[J]. IEEE transactions on pattern analysis and machine intelligence,2011,33(3).

附录

本文的代码均由 Python 实现，运行环境为 Anaconda3 Python 3.7.1，所有代码文件均包含在附件内，附件内的文件介绍：

ReadData.py 从 xlsx 中读取数据

TrainDataProcess.py 对训练数据进行预处理 1

TrainDataProcess2.py 对训练数据进行预处理 2

TestDataProcess.py 对测试数据进行预处理 1

TestDataProcess2.py 对测试数据进行预处理 2

TrainModel.py 模型的训练与评价

ModelPredict.py 模型对测试数据的预测

Sleep.py 睡眠脑波数据的分类

Filter.py 滤波函数

Clustering.py 聚类相关算法，包含 KNN

SupportVectorMachine.py 支持向量机相关算法

Metrics.py 分类评价指标

DimensionReduction.py 降维及可视化相关函数

Utils.py 一些工具

Paint.py 绘图代码