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	武汉大学
参赛队号	20204860084
队员姓名	1.周富
	2.姜祥威
	3.李宗成

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：

脑机接口试图在人脑和外部设备（例如假肢）之间建立一种联系，使得使用者能够直接通过大脑实现对外部设备的控制，而这种方法正好可以弥补残疾人士由于身体缺陷不能完成某些事件的不足，这种研究对推动社会发展有着极其重要的实际意义。P300 事件相关电位是诱发脑电信号的一种，通过对 P300 电位的监听和分析可以很好的获取人脑中的信息，研究者也对 P300 电位进行了深入的研究。睡眠脑电信号是自发性脑电信号的一种，从睡眠脑电信号中可以推断出与人体健康情况有关的一些信息。通过对睡眠中采集的脑信号可以帮助人们更好地做出合理的判断。

问题一：本问题在借鉴其它现有研究的基础上进行试验，通过对人脑 20 个特定位置的脑电波测试数据分析和加工，提取出 P300 相关电位的波形图。在后续的处理过程中以此作为基本的数据集进行分析和判别，本论文中分别采取了两种方法进行处理：1）提取刺激后产生的波形作为训练集，采用神经网络和费希尔线性判别分析作为分类算法，2）引入皮尔森相关系数，衡量波形之间的相关程度，以此作为依据进行判别。实验结果表明第二种方法的精确度高于第一种，主要原因在于数据集中类别不平衡问题。

问题二:本问题需在问题一的基础上进一步处理，论文中选择了问题一中性能最好的一种方法作为本题的实验方法，在这种方法的基础上，分别评估了不同通道对识别结果所产生的影响大小，逐步地剔除掉不相关的通道，从而达到选取最优通道组合的目的，论文中分别对 5 个被试选取了各自最优的通道，并展现了比问题一更优异的实验结果。

问题三：本问题属于一个半监督学习问题，必须在前面两题的基础上进行。半监督学习只需选取适量的样本作为有标签的样本知道，同时使用大量训练样本作为无标签样本对模型的性能进行改进与优化，不仅突破传统方法只用一类样本类型的局限，又能够挖掘出大量无标签数据集中隐藏的信息，在本题中我们选取半监督向量机作为主要模型，重新选取训练集进行处理，并给出了其余部分的识别结果。

问题四：本问题与前三题并无直接联系，是一个分类预测问题，目的在于对一段波形图的不同时间段进行分类。为了实现这一目的，论文中选用了支持向量机、决策树、BP 神经网路等进行分类预测，最后的结果表明实验取得了理想的结果，并获取的几种实验结果这一基础上比较这几种方法在同一问题上的表现差异，进而分析了造成这一结果的原因。最后，还提出了一种对模型进行优化的方法并加以测试。

关键词：脑机接口；脑电信号；半监督学习；分类预测；

1.问题重述

1.1、背景和意义

大脑是人体中高级神经活动的中枢，拥有着数以亿计的神经元，并通过相互连接来传递和处理人体信息。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。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。

（1）诱发脑电信号（P300 脑-机接口）

在日常生活中，人的大脑控制着感知、思维、运动及语言等功能，且以外围神经为媒介向身体各部分发出指令。因此，当外围神经受损或肌肉受损时，大脑发出指令的传输通路便会受阻，人体将无法正常完成大脑指令的输出，也就失去了与外界交流和控制的能力。研究发现，在外围神经失去作用的情况下，人的大脑依旧可以正常运行，而且其发出指令的部分信息可以通过一些路径表征出来。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。

P300 事件相关电位是诱发脑电信号的一种，在小概率刺激发生后 300 毫秒范围左右出现的一个正向波峰（相对基线来说呈现向上趋势的波）。由于个体间的差异性，P300 的发生时间也有所不同，图 1 表示的是在刺激发生后 450 毫秒左右的 P300 波形。P300 电位作为一种内源性成分，它不受刺激物理特性影响，与知觉或认知心理活动有关，与注意、记忆、智能等加工过程密切相关。基于 P300 的脑-机接口优点是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。

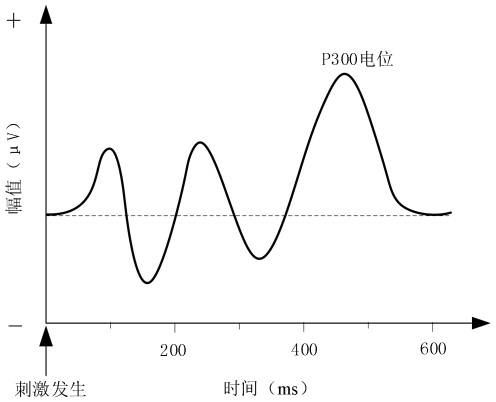
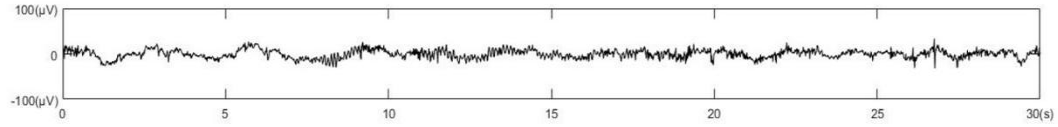


图 1 P300 波形示意图

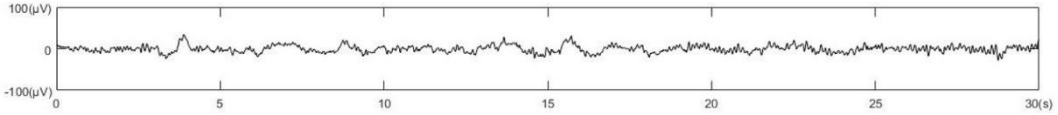
（2）自发脑电信号（睡眠脑电）

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也有着重大影响。如何提高睡眠质量，减少睡眠相关疾病对健康的影响，日益受到广泛关注。睡眠过程中采集的脑电信号，属于自发型的脑电信号。自发型的睡眠脑电信号能够反映身体状态的自身变化，也是用来诊断和治疗相关疾病的重要依据。

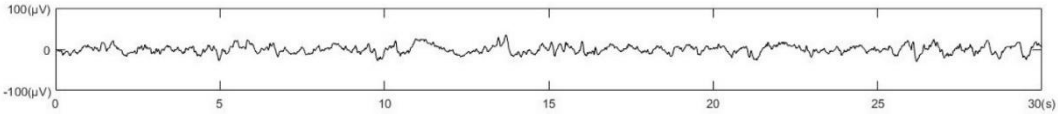
睡眠过程是一个动态变化的复杂过程。在国际睡眠分期的判读标准 R&K 中，对睡眠过程中的不同状态给出了划分：除去清醒期以外，睡眠周期是由两种睡眠状态交替循环，分别是非快速眼动期和快速眼动期；在非快速眼动期中，根据睡眠状态由浅入深的逐步变化，又进一步分为睡眠 I 期，睡眠 II 期，睡眠 III 期和睡眠 IV 期；睡眠 III 期和睡眠 IV 期又可合并为深睡眠期。图 2 给出了不同睡眠分期对应的脑电信号时序列，自上而下依次为清醒期、睡眠 I 期、睡眠 II 期、深睡眠和快速眼动期。从图 2 中可以观察到，脑电信号在不同睡眠分期所呈现的特点有所不同。基于脑电信号进行自动分期，能够减轻专家医师的人工负担，也是评估睡眠质量、诊断和治疗睡眠相关疾病的重要辅助工具。



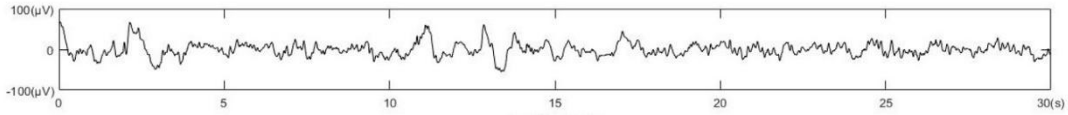
(a) 清醒期



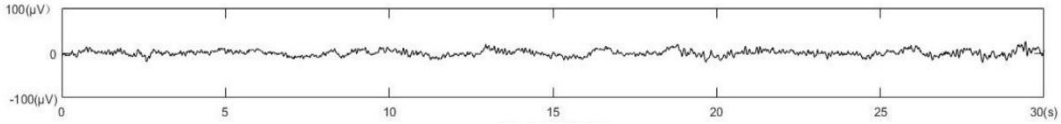
(b) 睡眠 I 期



(c) 睡眠 II 期



(d) 深睡眠期



(e) 快速眼动期

图 2　各睡眠分期的睡眠脑电信号时序列

1.2、问题提出

问题一：目标识别

在脑-机接口系统中既要考虑目标的分类准确率，同时又要保证一定的信息传输速率。请根据题目所给数据，设计或采用一个方法，在尽可能使用较少轮次（要求轮次数小于等于 5）的测试数据的情况下，找出其中 5 个被试测试集中的待识别目标。其中 S1、S4、S5 需给出 10 个识别结果，S2、S3 只要求并给出 9 个识别结果。具体的分类识别过程，可与几种方法进行对比，来说明设计方法的合理性。

问题二：通道选择

由于采集的原始脑电数据量较大，这样的信号势必包含较多的冗余信息。根据题目信息显示，在 20 个脑电信号采集通道中，无关或冗余的通道数据不仅会增加系统的复杂度，且影响分类识别的准确率和性能。请分析题目所给数据，并设计一个通道选择算法，给出针对每个被试的、更有利于分类的通道名称组合（要求通道组合的数量小于 20 大于等于 10，每个被试所选的通道可以不相同）。基于通道选择的结果，进一步分析对于所有被试都较适用的一组最优通道名称组合，并给出具体分析过程。

问题三：半监督学习

在 P300 脑-机接口系统中，往往需要花费很长时间获取有标签样本来训练模型。为了减少训练时间，请根据题目所给数据，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，在问题二所得一组最优通道组合的基础上，设计一种学习的方法，并利用问题二的测试数据检验方法的有效性，同时利用所设计的学习方法找出测试集中的其余待识别目标。

问题四：分类模型

根据题目中所给出的特征样本，请设计一个睡眠分期预测模型，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，给出训练数据和测试数据的选取方式和分配比例，说明具体的分类识别过程，并结合分类性能指标对预测的效果进行分析。

2.符号说明

在本题的解答过程中，鉴于模型构造和问题求解方面会涉及到许多专业术语，为确保论文的可读性及可理解性，本节将给出所涉及到的符号及其含义，并将其汇总于一个表中。如下所示，表 1 中包含了论文中使用到的所有符号以及它代表的实际意义。

表 1 论文中出现的符号及意义

符号	实际意义
T	P300 事件相关电位的延迟时间/t
f	采样频率/Hz
R	皮尔森相关系数
EEG	脑电图（Electroencephalogram）
ERP	事件相关电位（Event-Related-Potentials）
SVM	支持向量机（Support Vector Machine）
KNN	K 近邻算法（K-Nearest-Neighbor）
LDA	线性判别法（Linear Discriminant Analysis）
MLP	多层感知器（Multi-Layer-Perceptron）

3.模型假设

考虑到模型的复杂度会影响到后续求解过程的速度及准确率，有必要依据题目信息作出一些理想的假设。根据题目信息可知，针对问题一、二、三而言，P300 事件相关信号是一种内源性信号，它的产生和性质不受刺激物理特性影响，而与注意、记忆、智能等加工过程密切相关，每个被试之间由于个体的差异，会导致产生的 P300 事件相关信号会有个体差异，现做出如下假设：

1、单个个体对相同刺激所产生的 P300 事件相关信号导致的 EEG 波形保持不变或在小范围内变化，即刺激发生延时 T 基本保持不变。

2、单个个体对不同字符刺激所产生的的 P300 事件相关信号导致的 EEG 波形同样保持不变或在小范围内变化，即刺激发生延时 T 基本保持不变。

3、由于不同个体间的内在差异，对于 P300 事件相关信号产生的 EEG 波形有较大差异，即存在个体差异，刺激发生延时 T 因个体不同而不同。

4、P300 事件相关信号产生于刺激发生后 300ms 左右，故将其实际区间限制在 800ms 以内，以便于对信号进行处理。

5、不同通道对最终的波形图有同等的贡献，也就意味着最终生成的波形图是影响 P300 信号的所有通道的一个平均。

6、人体脑电波的电位会受到当前状况、环境的影响，同种刺激下电位会有一定差异。这种差异既包含噪音，也包含着人体自发的脑电信号变化。

7、人体对刺激有一个逐步适应的过程，会导致最开始接受刺激时，脑电图电位变化较高。随着刺激次数的增加，人体适应了这种偶然刺激，脑电图电位变化较低。

8、P300 信号有关的通道并非分散分布，它们之间存在直接或间接的联系，这也符合神经信号传导的理论。

4.问题分析及求解

4.1、问题一

4.1.1 数据分析

在 P300 相关电位实验中，被试需从 36 个字符中选定一个目标字符（A~Z 以及 0~9），实验过程中若出现选定字符时就相当于大脑接受一次刺激，且依据 Sutton 采取的一个典型的 ERP 实验范式——Oddball 实验范式的实验结果：就单个通道而言，当其接受两种刺激时，由于其中一种刺激出现的概率远高于另一种刺激，当实验过程中出现小概率刺激时大脑中则会立刻做出反应并产生事件相关电位 ERP，而 P300 信号就属于其中一种。在本实验中 36 个字符被模拟为一个 6×6 的矩阵，如图 1 所示。因而当特定字符所在的行列被确定以后就可以知道在何时产生了相应的 P300 信号来获取 P300 信号的波形图；反之，当确

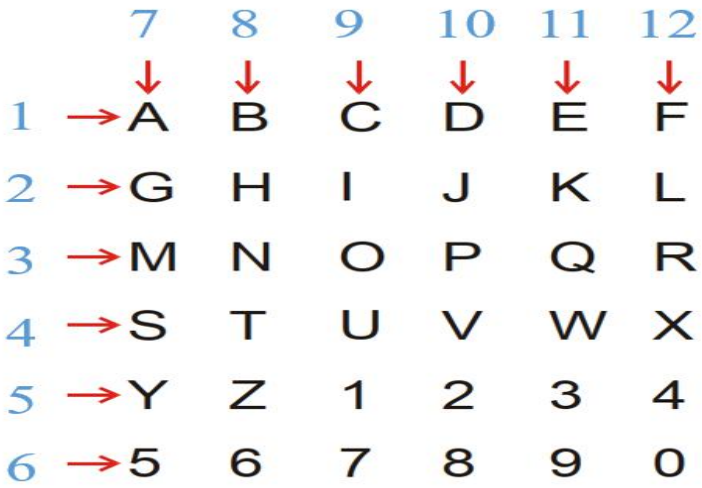


图 3 字符矩阵对应关系

定了在何时产生 P300 信号以后就可以反推其对应的行列数，进而便可以很轻松地确定相应的字符。

附件 1 中所给数据中共有五个被试的实验结果，每个被试的结果中可分为波形数据和事件流[1]数据。其中，波形数据中包含了 20 个通道在非测试时间以及测试时间内的波形图，其中给出了 12 个测试结果对应的字符；事件流数据包含了测试开始的时间节点、不同事件所对应的时间节点以及时间的先后顺序。为了更加明显的看到数据的具体形式，我们也在实验过程中对部分数据进行可视化。

依据模型假设，同一个体产生 P300 时间相关电位的延迟时间约为 300ms 且基本保持不变，即接收相同刺激后会产生一个相似的波形。为了获取个体产生刺激后产生的波形以便于后续对 P300 信号的特征提取和识别，我们采取了和相关论文[2][4]中相似的处理方法，截取相应事件后 600ms 内的脑电图（EEG）波形作为评判标准。进一步地，可以知道单次实验每个通道会产生 12*5=60 个波形图（即 12 个事件以及单次实验中有 5 轮测试），而附件 1 中为单个个体分别进行了 12 个字符的测试，于是这种划分方法总共可以获取 60*12=720 个波形图（每个字符产生 60 个波形图，实验共选取了 12 个字符）作为训练集，其中每个数据的标签对应于是否产生 P300 信号（其中标签 1 等价于是，代表正例；标签 0

[1] 事件流指行列闪烁所对应的数字，变化范围为 1~12。

等价于否，代表负例），于是在所有训练集中共包含 60*10=600 个标签为 1 的数据以及 60*2=120 个标签为 0 的数据，比例为正例：负例=1:5。采取同样地处理方法对 10 组未知数据（每组数据对应与一个未知字符，实验中有不同于训练集中已出现的 10 个未知字符）进行处理，于是可以得到 5*12*10=600 个带有未知标签的波形图，其中正例：负例的比例仍为 1:5。

4.1.2 问题定义

问题一虽然看似是一个如何从 24 个字符中选取 10 个未知字符的问题，但在进行仔细的数据分析以后可以明显发现，我们想要确定一个未知字符可以通过计算其对应的行列来定位字符，而行列在实验中是以事件流的形式出现，如果在一次字符实验中 1~6 中的某个事件出现了 P300 事件相关信号且 7~12 中也存在某个事件引发了 P300 事件相关信号，那么依据这两个事件便可以在字符矩阵中将未知字符唯一确定下来，从而达到本次实验通过采集的 EEG 来预测字符的目的。

那么该问题就转换成单个个体对于不同刺激所产生的的波形图的一个二分类问题：如果是 P300 信号波形就分为第一类，不是 P300 信号波形则分为第二类，本论文中用 1 和 0 来进行区分。当确定了一次测试每个波形图的分类结果，结合数据中给出的事件流就可以确定字符所在的行列，进而确定未知字符。在本实验中，每次测试分为 5 轮进行，每一轮都进行了随机的 12 次刺激，一次测试最终可生成 60 个波形图。于是需要从上述 60 个波形图的二分类结果中选出最符合 P300 信号的波形图来确定本次测试的两个事件，进而即可确定未知字符。举个简单的例子，如果在波形图分类结果正确的假设下，在 60 个波形中会出现 10 个波形为 1 的刺激事件，这刚好对于一个数据组合（2,8），那么在图 3 中就对于字符“H”。但是在分类不准确的情况下则不可能如此顺利的达到实验目的，影响实验的精度，因而本实验的问题在于如何设计一个好的方法来对波形图进行高精度的分类，这也是实验成败的关键所在。

4.1.3 问题求解

附件一提供了 5 个测试者对 22 个字符进行实验时，脑部 20 个测试点的脑电波数据。每个测试者对于每个字符均会进行 5 轮测试，我们观察 5 轮测试中有刺激点时某些通道的脑电波如图 4 所示（数据已经过等比例压缩平移）。我们可以看到图中曲线变化很大，这对于我们估计出 P300 的位置很有利，同时我们也注意到波形图中有很多信息应该属于干扰信息，不利于对类别的准确分类。

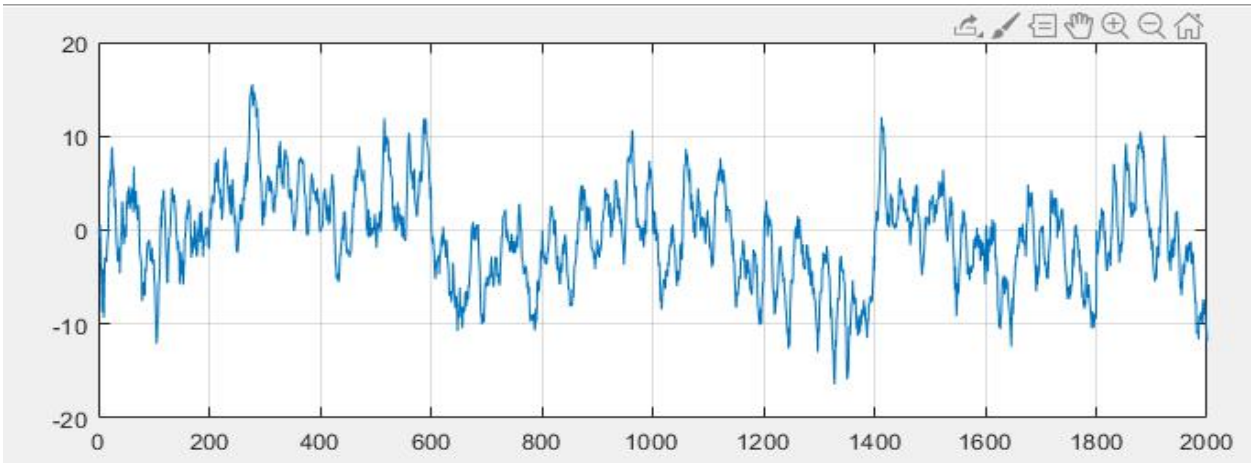


图 4　原始数据波形

数据预处理　由于我们要检测的对象是 P300 信号，它大概会在刺激点后 300ms 出现，于是我们将刺激点之后 800ms 的波段作为采样片段，当然采取其余长度的时间间隔同样可以达到实验目的，并且将五轮实验同一事件的波段求和平均用来消除环境白噪声，并对波段进行滤波处理，这也是传统方法中的一种常用技巧[4]，最终可以生成我们的数据集。一般情况下，对原始数据的处理可以大致分为以下几个步骤：滤波、单次刺激数据提取、降采样、数据调整等多个步骤，本论文中从中选取了一部分必要步骤进行试验，并将片段提取结果及滤波结果进行展示，结果如下图：

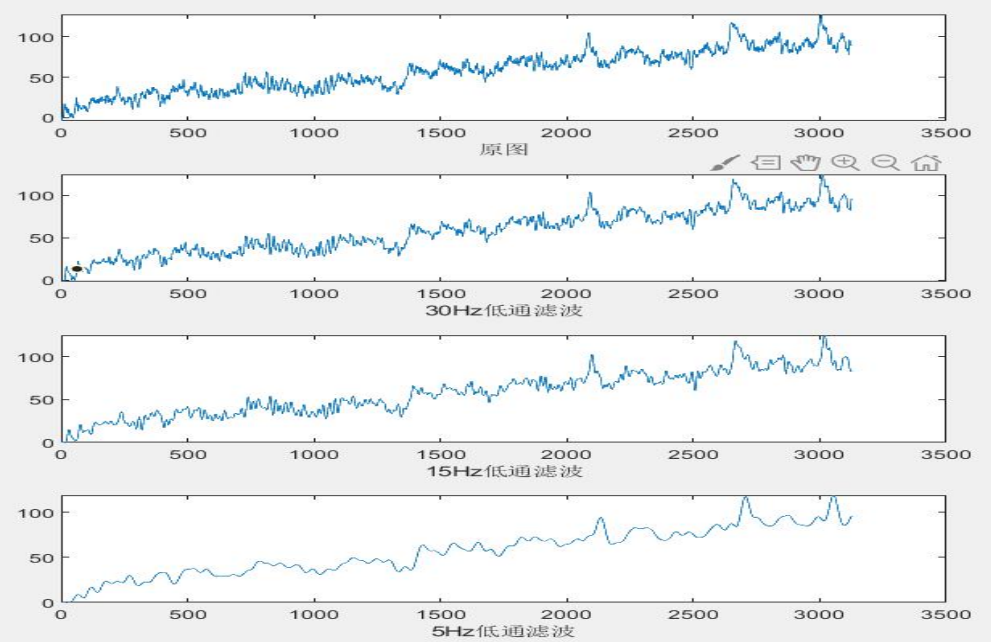


图 5　滤波处理（8 阶巴特沃斯滤波器）

我们采用了巴特沃斯滤波器滤除图像的高频部分，图中有 30Hz，15Hz，5Hz 低通滤波处理的结果，我们可以看到 5Hz 的滤波器只能保留图像大致形状，对于之后对每个点的刺激片段截取分析是非常不利的，于是我们采用 10Hz 滤波器处理所有数据。

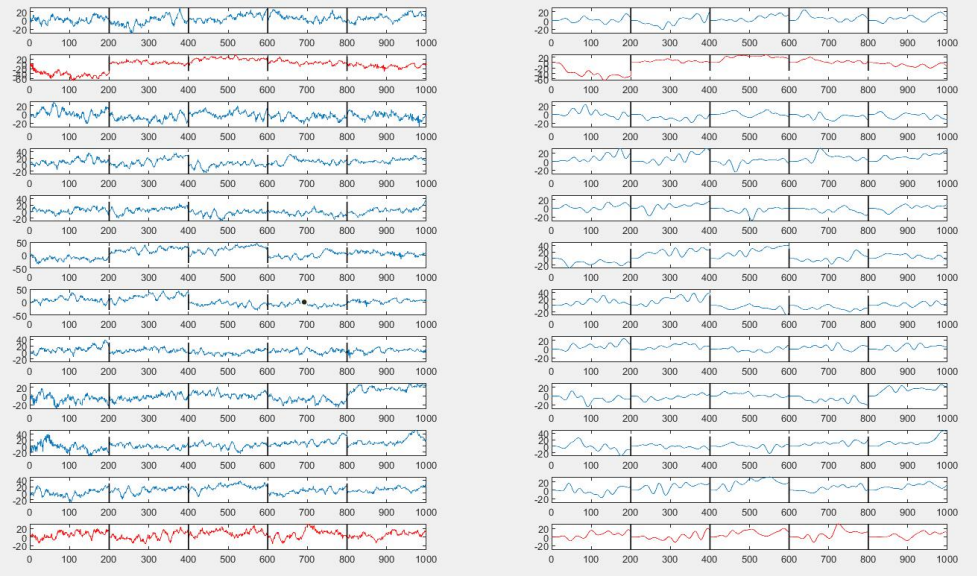


图 6　滤波前后 12 事件的形状（红色代表正样本）

观察图 6，我们发现，即使经过了滤波处理，也不是所有的正样本（红色部分）都与负样本有明显的区别，为了显示出正样本的特性，我们将 5 轮试验结果取平均之后得到了图 7。最后一列代表前五次的平均，我们可以看到正样本的特征的确更明显，同时负样本也因为平均之后随机误差变小而变得平缓。

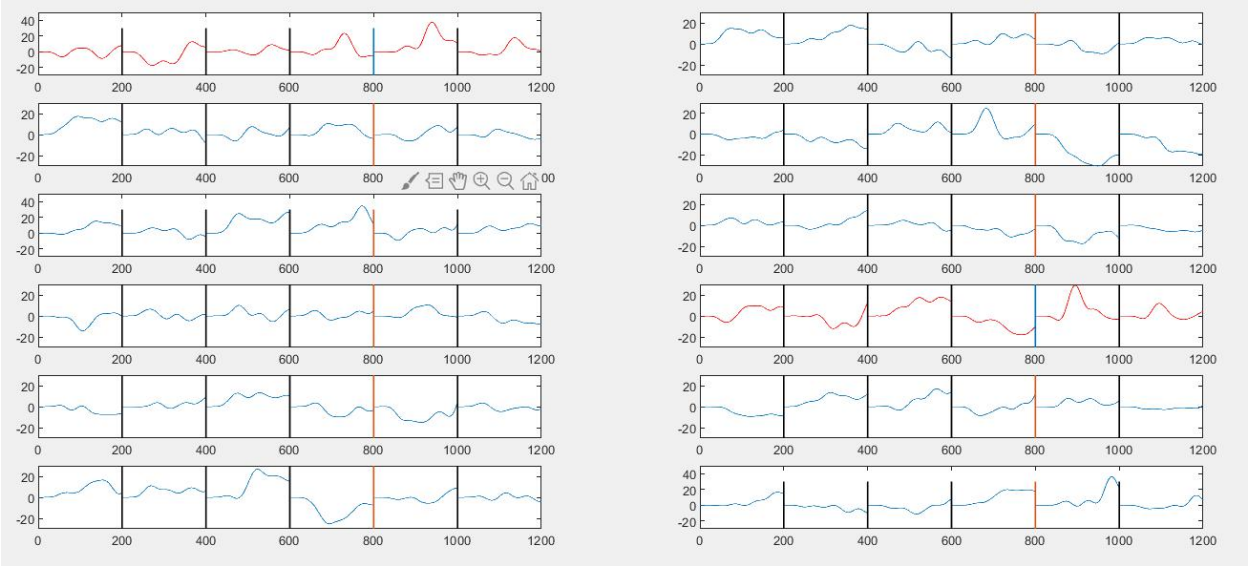


图 7 平均化之后的波形图

每一个测试者的训练集里面都有 2880 条数据，每个数据包含一个事件标签和一个分类标签，并且包含 200 个采样点，训练集中分类标签为 0、1，测试集则不包含分类标签。这样得到的数据集就可以用一些常见的二分类算法进行处理。

求解方法　这里我们采用两种方法来进行实验，借鉴于论文[1]中的思想，实验中我们考虑引入神经网络的方法来进行处理。第一种分类就是线性判别法（LDA），它的核心思想就是将特征空间中的样本投影到该空间的一条直线上以实现高维数据的特征提取和划分。第二种就是多层感知器（MLP），它的主要实现原理就是模仿神经系统的信号处理模式，构造一个个神经元来处理信号，因而比较适合复杂信息的分类。

实验过程　我们采用同一组数据用两种方法进行处理，便于对实验结果进行比较。具体而言，我们将训练集的一部分作为验证集来测试模型有效性，分别在两个模型上进行试验。分类方法一：线性判别法：将测试集导入模型做二分类，实验结果如表 2 所示：

表 2　线性判别法的分类结果

	S1	S2	S3	S4	S5
char13	(3, 7)	(2, 9)	(3, 7)	(1, 8)	(2, 7)
char14	(2, 12)	(1, 8)	(1, 7)	(3, 12)	(3, 12)
char15	(4, 11)	(5, 7)	(3, 11)	(5, 10)	(6, 7)
char16	(6, 10)	(2, 7)	(5, 11)	(6, 10)	(5, 10)
char17	(1, 9)	(5, 9)	(2, 7)	(4, 9)	(1, 9)
char18	(6, 8)	(4, 8)	(4, 11)	(6, 8)	(1, 9)
char19	(1, 7)	(2, 11)	(1, 11)	(1, 10)	(5, 9)
char20	(4, 12)	(5, 9)	(4, 12)	(2, 7)	(1, 12)
char21	(4, 7)	(4, 7)	(1, 7)	(2, 10)	(4, 8)
char22	(4, 11)			(2, 7)	(4, 10)
二分类正确率	0.78219697	0.861742424	0.721590909	0.865530303	0.920454545

分类方法二：我们采用一个三层的感知器模型进行模拟，其中每一层的参数为（200，13，24，1），激活函数选取较为常用的线性整流函数。十二种有刺激的事件加上无刺激产生的情况一共十三种情况，24 代表 12 类正负情况，结果如表 3 所示。

表 3　感知器模型的分类结果

	S1	S2	S3	S4	S5
char13	(3, 7)	(2, 9)	(6, 12)	(2, 10)	(5, 11)
char14	(2, 12)	(1, 8)	(3, 11)	(2, 12)	(3, 12)
char15	(6, 7)	(5, 7)	(3, 11)	(6, 7)	(6, 7)
char16	(1, 10)	(5, 11)	(5, 10)	(6, 10)	(5, 10)
char17	(4, 9)	(2, 9)	(2, 7)	(5, 10)	(4, 9)
char18	(5, 8)	(1, 8)	(1, 8)	(2, 8)	(1, 12)
char19	(3, 7)	(2, 11)	(5, 7)	(6, 10)	(1, 11)
char20	(4, 8)	(4, 12)	(2, 9)	(5, 11)	(2, 12)
char21	(5, 12)	(1, 7)	(1, 8)	(1, 7)	(6, 8)
char22	(1, 8)			(1, 7)	(6, 12)
二分类正确率	0.962121212	0.981060606	0.948863636	0.926136364	0.982954545

从上述两个表的对比中我们可以看到方法一和方法二的部分分类结果有较大的差异，但对于一部分字符产生了相同识别的结果，可以说明两者在理论上都可以实现本问题的求解，但是不可否认的两种方法都有一定的精度缺陷。除此之外，通过对比两种方法在训练集上的表现可以得出这样的结论：方法二（感知器模型）的二分类正确率要明显高于方法一。因此权衡多种因素后，实验中选取了 3 层感知器模型作为最终分类的模型并给出未知数据点字符。实验中首先使用前一节划分出来的训练集对模型进行训练，然后将未知数据放入模型中进行分类识别，识别结果以一个二元数组给出，通过此二元数组可以对字符进行定位，识别结果如下表 4 所示：

表 4 识别结果图

	S1	S2	S3	S4	S5
char13	M	I	0	J	3
char14	L	B	Q	L	R
char15	5	Y	Q	5	5
char16	D	3	2	8	2
char17	U	I	G	2	U
char18	Z	B	B	H	F
char19	M	K	Y	8	E
char20	T	X	I	3	L
char21	4	A	B	A	6
char22	B			A	0

针对减少数据量的问题，实验中依次采用 2,3,4 轮的平均数进行实验，实验平均准确率说明将多轮实验的结果进行叠加可以很有效的消除由于随机噪音给波形带来的影响。在突出 P300 信号波形的同时，优化相关的数据集，间接提升了模型分类的准确率。

不同于上述方式，实验中还从 P300 信号的组成成分这一方面入手，考虑到采集到的 EEG 中包括一部分自发电位、噪音影响、P300 信号波动。从可视化以后的数据来看，数据波动较大，其中自发电位造成大范围的上下波动，而噪音电位会导致微小的起伏。考虑噪音的浮动实际上是一个正态分布的形式，这种噪音会导致波形轻微的上下浮动。如果在多次刺激下，将含有噪音的 P300 信号进行叠加[6]，其中的噪音由于会相互抵消而减少对波形的影响，相反地 P300 信号是一种突出的正向波峰，在经过多次叠加后波形会更加突出，从而完成对 EEG 的特征提取，进而实现对 P300 波形的识别。除了这种方法以外，现有特征提取的方法还包括小波变换[5]、自回归模型、带通滤波器等其他方法[7]。

在进行试验之前，通过对五轮试验中波形图的分析可以看出，最后两轮的波形图较为相似，而与前三轮的波形图差异较大，因此在本方法中忽略前三轮的测量结果，引入皮尔森相关系数 R 来衡量波形与 P300 波形的相似性，实验也取得了相应的结果。

4.2、问题二

4.2.1 问题分析

在前一节中我们分析了几种不同的方法来完成对字符的检测，等价于对是否存在 P300 信号的检测。从实验结果来看的确可以很好地实现对泛化数据的检测，但就精准度而言尚有缺陷，具体原因可以归咎于以下几点：采集数据的通道过多，导致获取的 EGG 含有过多的无关信息以及大量的噪音；采用的模型和方法本身存在缺陷，无法精确的完成对字符的分类识别。

在不考虑模型的条件下，通道中的信息成了影响分类结果的决定因素，一个合适的通道组合可以最大程度的拟合 P300 信号的波形，这种方法获取到的波形特征最为明显和准确。问题一中在求解过程中使用了 20 通道的全部信息，其中一部分通道携带的无关信息降低了不仅没能帮助判别不同的类别，反而降低了不同波形图之间的差异，降低了实验结果。因此可以很自然地想到我们可以通过选取不同通道来提高分类的准确率，因为这可以视为一种去除噪音以提高信噪比的方法，问题二可以归结为一个通道组合的选择问题，题目要求通道组合数目需大于十个。

4.2.2 问题求解

通过上一节的分析可以很清晰的指导，如何获取一个最优的通道组合是本题的核心任务，在第三部分我们假设相关的通道组合之间存在相应的联系，并非一种分散式的模式，于是考虑到结合人脑的采样通道分布图可以进行经验性的选择[4],这种方法较为简单，但对进行选择的人在实践经验上有较高要求，在缺乏这种实践经验的基础上，论文中并不将这种方法纳入考虑范围，以免由于经验的缺乏而导致实验精度并未得到显著提高，反而浪费过多的时间在经验性的选择上。

依据相关论文[4]中的思想，可以使用递归特征通道移除的方法在为每个通道评分的基础上进行移除，这里的得分实际上可以看做是不同通道对分类结果准确率的一个贡献值，通过评选最后的得分高低来决定是否将该通道从实验中移除。在不断的递归迭代过程中修改每个通道的得分，最终可以获得最准确的通道得分作为评选的标准。除此以外，单个特征移除法给定每个通道一个评价指标，通过对比指标的高低来决定移除某个通道，这种方式的思想和递归通道移除法很相似，但也存在一点的差别。

依据对模型的假设，实验中不仅需要考虑逐步移除通道或给定评价指标的思想，还需要额外的考虑每个通道之间的连接关系，于是在理论可行性的基础上采取块移除的方式进行通道选择，即按照下图中的分布情况进行块分类，从而进行通道移除。移除的思想采取了引入评价指标的方法，为上述采样通道进行分块，每个块赋予一个评价指标，即是这一

块对模型精度的影响大小，多次试验以逐步去除掉与结果无关或负相关的部分通道，达到实验目的。

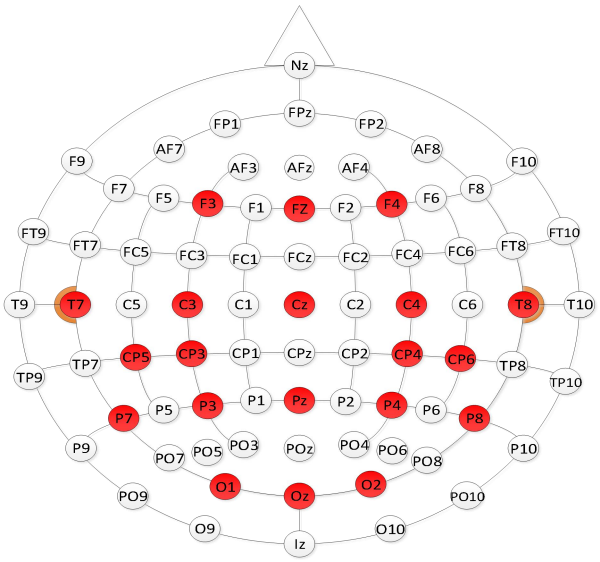


图 8　人脑通道采集点位图

为了较为直观的鉴别不同通道对 P300 信号的影响，实验对比了 20 个通道在实验中接受刺激后产生的波形差异，进一步说明通道选择的必要性，实验结果如图 9 所示。

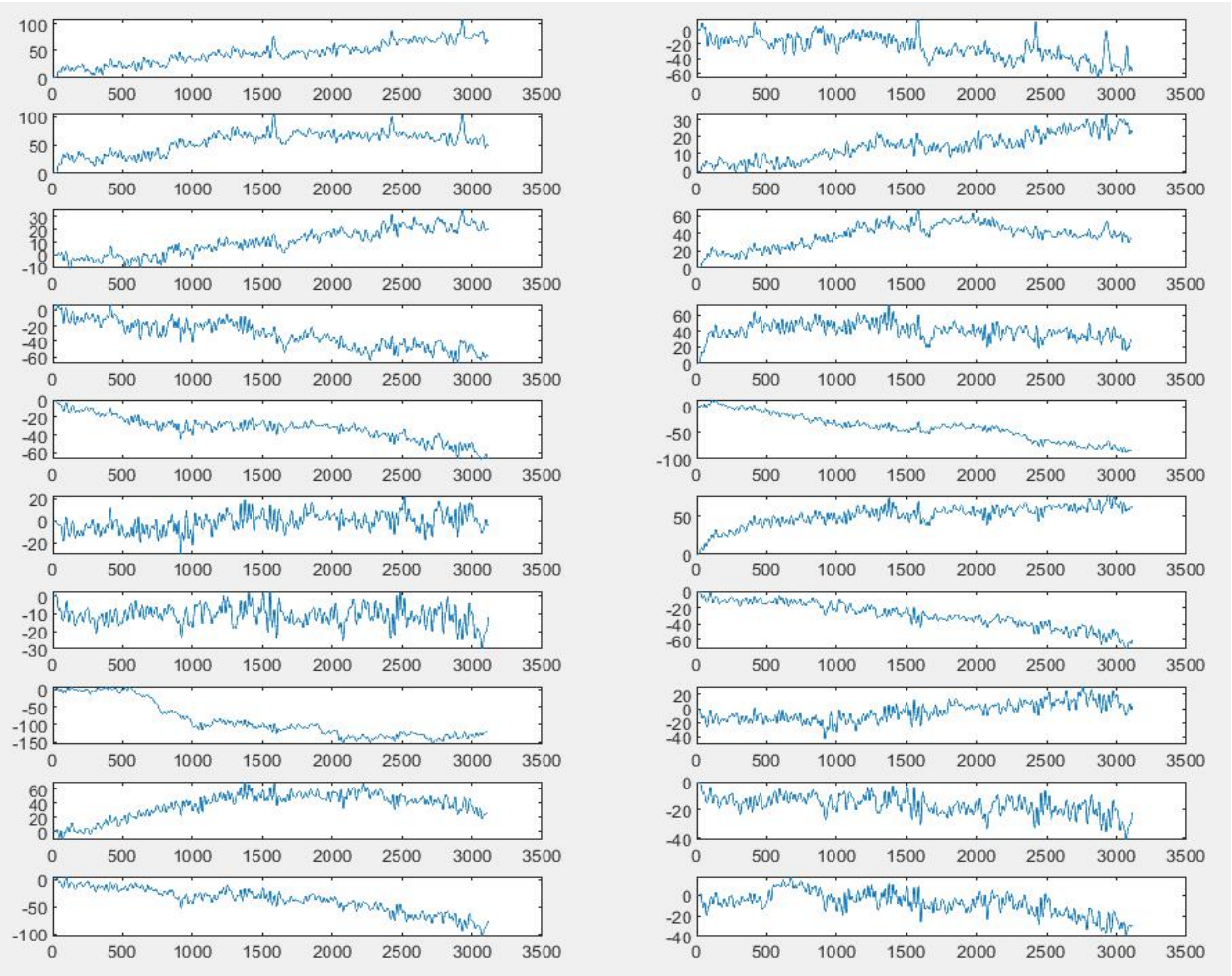


图 9 通道波形对比图

4.3、问题三

4.3.1 基础概述

基于数据的机器学习包括监督学习、无监督学习以及半监督学习。监督学习需要利用大量的有标记的数据集来训练模型,但实际生活中,有标记样数据集的采集和制作不仅费时而且费力，而实际上由于大数据时代的到来，存在这大量无标记的数据。无监督学习不使用先验信息,利用无标签样本自身特征，使具有相似特征的样本聚在一起，但模型准确性具有不稳定性，结果难以保证。半监督学习是一种无监督学习和监督学习相结合的一种学习方法，半监督学习只需选取适量的样本作为有标签的样本知道，同时使用大量训练样本作为无标签样本对模型的性能进行改进与优化，不仅突破传统方法只用一类样本类型的局限，又能够挖掘出大量无标签数据集中隐藏的信息。半监督的主要内容有半监督聚类，半监督分类、半监督回归与半监督降维等[8]。常见的半监督常见的半监督分类代表算法可以划分为四类，包括生成式方法、半监督支持向量机、半监督图算法和基于分歧的半监督方法。半监督支持向量机在大多数领域都比较适用，而且在小型样本的数据集中表现出色，它具有良好的泛化能力以及稳定性，适用于非线性数据集的建模。故选用半监督支持向量机。

4.3.2 实验结果

在实际的 P300 脑-机接口系统中，往往需要花费很长时间获取有标签样用来训练模型。为了减少样本制作时间和训练时间，决定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，在问题二所得一组最优通道组合的基础上，对模型进行训练，用测试数据（char13-char17）检验方法的有效性对数据集进行验证其有效性，本方法在 char13~char17 上进行验证，试验结果表明五个字符的准确率为 60%。在验证的同时找出测试集中其余分类待识别目标如下表所示:

表 5　通道选择后的分类结果

	S1	S2	S3	S4	S5
char13	M	Q	B	E	M
char14	R	L	A	Y	L
char15	5	F	5	K	I
char16	2	B	3	2	M
char17	U	Z	I	L	6
char18	B	U	C	M	P
char19	H	A	Z	W	E
char20	L	8	Q	L	E
char21	A	K	G	U	L
char22	0			B	B

4.4、问题四

4.4.1 基础概述

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。自发脑电信号是指在没有外界特殊刺激下，人进行特定思维活动，大脑产生特定模式的自发脑电信号，故在睡眠过程中采集的脑电信号，属于自发型的脑电信号。自发型的脑电信号能够反映身体状态

的自身变化，其高精度的分类结果，不仅可以用来诊断和治疗相关疾病的重要依据，而且通过计算机将其翻译成预先设定的控制命令，实现人脑对计算机等外部设备的直接控制。

睡眠过程是一个动态变化的复杂过程，除去清醒期外，睡眠周期可分为睡眠 I 期，睡眠 II 期，睡眠 III 期和睡眠 IV 期；其中睡眠 III 期和睡眠 IV 期又可合并为深睡眠期。基于脑电信号进行自动分期，不仅可以减轻专家医师的人工负担，而且能够评估睡眠质量、诊断甚至治疗睡眠相关疾病的重要辅助工具。

4.4.2 实验说明

实验提供取自不同的健康成年人整夜睡眠过程的 3000 个睡眠脑电特征样本及其标签。目的在使用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。因为其数据集共三千个样本且特征只有四个，判断为小数据规模的分类预测。针对小数据规模的分类预测模型有支持向量机(SVM),K 近邻算法(KNN),朴素贝叶斯，决策树模型，BP 神经网络等等。总体实验过程如下:

- 1、针对每一个分类模型进行试验，训练数据集从 100 以每次 100 增加的速度进行 29 次实验，选取准确率较高，同时所需的训练样本数相对较少的分类模型。
- 2、选取最优的网络模型后，对模型进行误差分析。
- 3、选择适当的方法提高最优模型的准确率，并调节该网络模型的超参数。

4.4.3 实验分析

模型对比 对每一个模型进行如下描述的实验:共有 3000 个数据集，进行 29 次试验，第一次实验训练数据集个数 100，测试数据集 2900，第二次实验训练数据集个数 200，测试数据集 2800，以此类推。实验所得的数据画成折现图如下图 10 所示

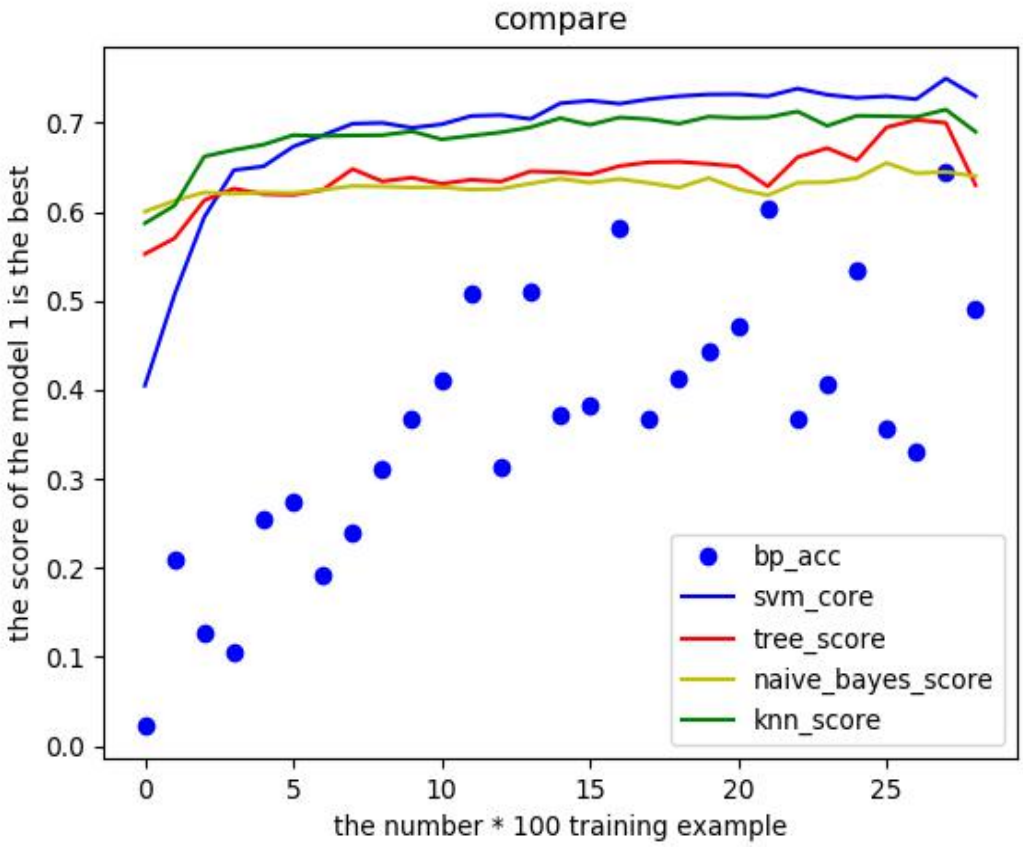


图 10 几种方法结果对比

通过观察可得:BP 神经网络效果最差，模型很不稳定，且准确率偏低；决策树模型和朴素贝叶斯训练所需样本最少，数据集的大小对模型的准确率影响较低，模型比较稳定，但是准确率偏低；支持向量机，和 KNN 分类精度较高，其中支持向量机分类精度最高，可达到 71%，且支持向量机在只需 1400 个训练样本后即可训练出相对稳定的模型。K 近邻算法(KNN)虽训练出稳定的模型所需样本数比支持向量机低，但精度相比更低一点，且数据集的大小对模型准确率的影响比支持向量机较小一点，故综合考虑选择支持向量机。

误差分析 为了进一步分析实验结果，实验中还通过观察支持向量机的混淆矩阵和分类报告，并将结果汇总为表格便于对比，结果如下表 6、表 7 所示:

表 6 支持向量机的混淆矩阵

类别	深睡眠期	睡眠 II 期	睡眠 II 期	快速动眼期	清醒期
深睡眠期	229	37	4	2	0
睡眠 II 期	31	117	34	24	0
睡眠 I 期	3	35	130	49	30
快速动眼期	1	17	58	226	21
清醒期	0	0	48	1	256

符号说明—— precision:精准率，分类器正确分为正例的个数(TP)占被分类器分为正例的样本(TP+FP)的比重;recall:召回率,分类器正确分为正例的个数(TP)占原始数据中全部正例(TP+FN)的比重；F1: 调和平均数; support：支持度，指原始的真实数据中属于该类的个数；micro avg:宏平均，将各类的 precision 先算好再对它们求算术平均； micro avg：微平均，是将所有类中中真阳例先加起来，再除以所有类中的(真阳例+假阳例);weighted avg: 加权平均,加上每个类的权重，即可得到它的 support 的值。

表 7 支持向量机的分类报告

类别	precision	recall	f1-score	support
深睡眠期	0.87	0.84	0.85	272
睡眠 II 期	0.67	0.67	0.67	266
睡眠 I 期	0.47	0.53	0.50	247
快速动眼期	0.74	0.75	0.75	304
清醒期	0.89	0.84	0.8	311
micro avg	0.73	0.73	0.73	0.71400
macro avg	0.73	0.72	0.73	1400
weighted avg	0.74	0.73	0.73	1400

发现其算法对睡眠 I 期，睡眠 II 期这二类数据的分类效果不算特别好，其中对睡眠 I 分类效果最差。通过进行特征的组合实验，训练数据集 1500 份 测试数据集 1500 份，其实验结果如下表 8 所示。

通过观察表 8[2]发现第四个特征 E:delta 对模型分类准确率的贡献较大，其余特征的组合都相对全部特征值的精度低一些。同时通过观察睡眠 I 和睡眠 II 的折线图，不同睡眠分期对应的脑电信号时序列。根据观察到脑电信号在不同睡眠分期所呈现的特点和折线图的曲折程度，决定添加一个额外的特征:反应数据离散程度的的特征 avedev。

[2] 参数说明: B：Alpha C: Beta D: Theta E: Delta

表 8　特征组合试验结果

特征参 dt	B	C	D	E	BC	BCD	BCDE	BCE
模型评价 (1 为最好)	0.44	0.51	0.421	0.645	0.594	0.691	0.701	0.683
特征参 dt	BD	BDE	BE	CD	CDE	CE	DE	
模型评价 (1 为最好)	0.605	0.643	0.669	0.597	0.673	0.671	0.651	

模型优化　用二种数据集:1）原始数据集；2）添加 Avedev 特征的数据集分别对支持向量机模型进行训练。同时对不同数量的训练数据集，训练网络模型，进行实验精度的比较，并调节模型参数，选取最优的支持向量机的模型参数，结果如图 11 所示。

结果表明第五个特征选取对模型分类准确度的提高有一定的帮助，大概是 5 个百分点，未加入第五个特征值大概需要 1400 个训练样本即可训练出相对稳定的支持向量机模型，当加入第五个特征值是，大概需要 1800 个训练样本可训练出相对稳定的模型，此时模型的参数为: {'C': 19, 'gamma': 0.1, 'kernel': 'rbf'}。其中 C 是惩罚系数，即对误差的宽容度;; gamma 是选择 RBF 函数作为 kernel[3]后，该函数自带的一个参数。隐含地决定了数据映射到新的特征空间后的分布，gamma 越大，支持向量越少，gamma 值越小，支持向量越多，而支持向量的个数会影响训练与预测的速度。

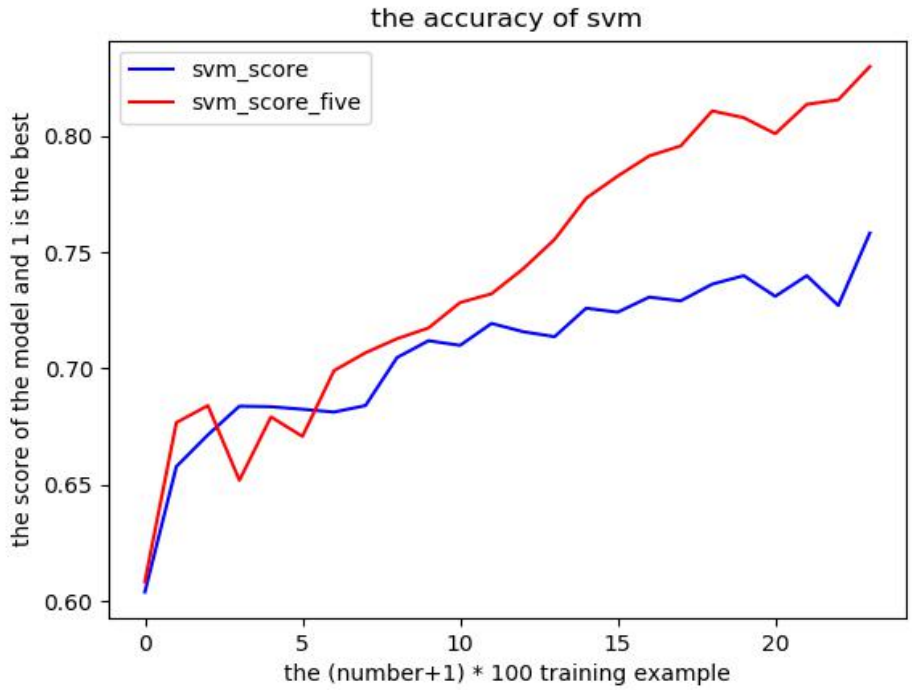


图 11　SVM 模型精度折线图

最后，从实验的结果来看，本论文在一定程度上完成了四个问题的求解并具体分析了从实验开始到结束过程中具体步骤的意义以及其相应的实施过程。以下内容是对本次实验的一些客观评价和总结：

本实验的核心内容是问题一，采取一种合适的分类判别方法可以大大提高工作效率，后续的工作都要在问题一的基础上进行，例如问题二是使用问题一的分类精度作为判别标准，这也是在实验过程中花

[3] Kernel：核函数，选取径向基核函数。

费大量时间和精力去最大可能的优化问题一的分类结果的主要原因，虽然在某些方面的确达到了实验的目的。但就从实验结果来看，精度始终存在一定的缺陷。在部分实验中甚至出现了识别结果重复的结果，从另一方面证明了模式内在存在明显缺陷的问题，而单单从训练的结果来看几种模型对于训练集的准确率较高，但在测试集上却展现出极差的表现，经过分析可以确定主要原因在于训练集上的两类标签在数量上差异巨大，造成类别不均衡问题，最终虽然其在训练集上表现良好，但却无法在泛化数据上产生相同的结果，这也是本模型的一个缺陷和改进方向。

本题目的求解过程中含有大量的相关数据及代码，为了更好地展示实验结果且便于理解和阅读，将其以附件 1 的形式给出。

参考文献

[1] Ruilong Zhang,Qun Zong,Liqian Dou,Xinyi Zhao,Yifan Tang,Zhiyu Li. Hybrid deep neural network using transfer learning for EEG motor imagery decoding[J]. Biomedical Signal Processing and Control,2021,63.
[2] 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.
[3] Bauke van der Velde,Caroline Junge. Limiting data loss in infant EEG: putting hunches to the test[J]. Developmental Cognitive Neuroscience,2020,45.
[4] 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.
[5] Yang You,Wanzhong Chen,Tao Zhang. Motor imagery EEG classification based on flexible analytic wavelet transform[J]. Biomedical Signal Processing and Control,2020,62.
[6] Lee Yuhyun,Lee Kyeong Jae,Jang Jae-Won,Lee Sang-Im,Kim Sohee. An EEG system to detect brain signals from multiple adult zebrafish.[J]. Biosensors & bioelectronics,2020,164.
[7] Biotechnology - Biomedical Engineering; New Findings in Biomedical Engineering Described from Stanford University (Power-saving Design Opportunities for Wireless Intracortical Brain-computer Interfaces)[J]. Computer Technology Journal,2020.
[8] 韩嵩,韩秋弘.半监督学习研究的述评[J].计算机工程与应用,2020,56(06):19-27

相关代码

```
%%
%代码 1：滤波处理
clc;
clear;
%引入 8 阶巴特沃斯滤波器
[b,a]=butter(8,30/(250/2),'low');
[b2,a2]=butter(8,15/(250/2),'low');
[b3,a3]=butter(8,5/(250/2),'low');
k =1;
    data = xlsread('data/S1/S1_train_data.xlsx',k);
figure
    picture1=filter(b,a,data(:,1)-data(1,1));
    picture2=filter(b2,a2,data(:,1)-data(1,1));
    picture3=filter(b3,a3,data(:,1)-data(1,1));
    subplot(4,1,1);
    plot(1:length(data(:,1)),data(:,1)-data(1,1));xlabel("原图");
    subplot(4,1,2);
    plot(1:length(data(:,1)),picture1);xlabel("30Hz 低通滤波");
    subplot(4,1,3);
    plot(1:length(data(:,1)),picture2);xlabel("15Hz 低通滤波");
    subplot(4,1,4);
    plot(1:length(data(:,1)),picture3);xlabel("5Hz 低通滤波");


%%
%代码 2:事件对比图
clc;
clear;
z = 1;
picture1 = zeros(12,1000);
picture2 = zeros(12,1000);
%引入滤波器
[b,a]=butter(8,10/(250/2),'low');
M = ['A','B','C','D','E','F';
     'G','H','I','J','K','L';
     'M','N','O','P','Q','R';
     'S','T','U','V','W','X';
     'Y','Z','1','2','3','4';
     '5','6','7','8','9','0'];
k =4;
    data = xlsread('data/S1/S1_train_data.xlsx',k);
```

```
m = 1;
        data1 = data(:,m);
        event = xlsread('data/S1/S1_train_event.xlsx',k);
        [status,chart,xlsFormat] = xlsfinfo('data/S1/S1_train_event.xlsx');
        zimus = char(chart(k));
        zimu = zimus(8);
        [num1,num2] = find(M==zimu);
        event(event(:,1)==100+6*(num1-1)+num2 |event(:,1)==100,:)=[];
        for i =1:60
            picture2(event(i,1),200*(z-1)+1:200*z)=
filter(b,a,data1(event(i,2):event(i,2)+199)-data1(event(i,2)));
            picture1(event(i,1),200*(z-1)+1:200*z)=
data1(event(i,2):event(i,2)+199)-data1(event(i,2));
            if mod(i,12)==0
                z = z+1 ;
            end
        end

figure
j=1;
for i = 1:12
    subplot(12,2,j);j=j+1;
    if i==num1 ||i==num2+6
        plot(1:1000,picture1(i,:),"r");
        hold on ;
        plot([200,200],[-30,30],'k',[400,400],[-30,30],'k',...
        [600,600],[-30,30],'k',[800,800],[-30,30],'k','linewidth',1)
        subplot(12,2,j);j=j+1;
        plot(1:1000,picture2(i,:),"r");
    else
        plot(1:1000,picture1(i,:));
        hold on ;
        plot([200,200],[-30,30],'k',[400,400],[-30,30],'k',...
        [600,600],[-30,30],'k',[800,800],[-30,30],'k','linewidth',1)
        subplot(12,2,j);j=j+1;
        plot(1:1000,picture2(i,:));
    end
    hold on ;
    plot([200,200],[-30,30],'k',[400,400],[-30,30],'k',...
        [600,600],[-30,30],'k',[800,800],[-30,30],'k','linewidth',1)
end
%title("10Hz_12 事件对比图",'position',[0,0]);
%%
%代码 21:事件对比图
```

```
clc;
clear;
z = 1;
picture3 = zeros(12,200);
%引入滤波器
[b,a]=butter(8,10/(250/2),'low');
M = ['A','B','C','D','E','F';
     'G','H','I','J','K','L';
     'M','N','O','P','Q','R';
     'S','T','U','V','W','X';
     'Y','Z','1','2','3','4';
     '5','6','7','8','9','0'];
k =1;
    data = xlsread('data/S1/S1_train_data.xlsx',k);
m = 2;
        data1 = data(:,m);
        event = xlsread('data/S1/S1_train_event.xlsx',k);
        [status,chart,xlsFormat] = xlsfinfo('data/S1/S1_train_event.xlsx');
        zimus = char(chart(k));
        zimu = zimus(8);
        [num1,num2] = find(M==zimu);
        event(event(:,1)==100+6*(num1-1)+num2 |event(:,1)==100,:)=[];
        for i =1:60
            picture3(event(i,1),1:200)=picture3(event(i,1),1:200)...

+filter(b,a,data1(event(i,2):event(i,2)+199)-data1(event(i,2)))';
        end
picture3 = picture3/5;
figure
j=1;
for i = 1:12
    subplot(6,2,i);
    if i==num1 ||i==num2+6
        plot(1:200,picture3(i,:),"r");
    else
        plot(1:200,picture3(i,:));
    end
end
%title("10Hz_12 事件对比图",'position',[0,0]);



%%
```

```
%代码 3:通道对比图
clc;
clear;
z = 1;
picture1 = zeros(12,1000);
picture2 = zeros(12,1000);
%引入滤波器
[b,a]=butter(8,10/(250/2),'low');
k =2;
    data = xlsread('data/S1/S1_train_data.xlsx',k);
figure
for m = 1:20
    subplot(10,2,m);
    picture1=filter(b,a,data(:,m)-data(1,m));
    plot(1:length(data(:,m)),picture1);
end

%%
%代码 4:事件对比图
clc;
clear;
z = 1;
picture3 = zeros(12,200);
%引入滤波器
[b,a]=butter(8,5/(250/2),'low');
M = ['A','B','C','D','E','F';
     'G','H','I','J','K','L';
     'M','N','O','P','Q','R';
     'S','T','U','V','W','X';
     'Y','Z','1','2','3','4';
     '5','6','7','8','9','0'];
k =1;
    data = xlsread('data/S1/S1_train_data.xlsx',k);
m = 2;
        data1 = data(:,m);
        event = xlsread('data/S1/S1_train_event.xlsx',k);
        [status,chart,xlsFormat] = xlsfinfo('data/S1/S1_train_event.xlsx');
        zimus = char(chart(k));
        zimu = zimus(8);
        [num1,num2] = find(M==zimu);
        event(event(:,1)==100+6*(num1-1)+num2 |event(:,1)==100,:)=[];
        for i =1:60
            picture3(event(i,1),1:200)=picture3(event(i,1),1:200)...
```

```
+filter(b,a,data1(event(i,2):event(i,2)+199)-data1(event(i,2)))';
        end
picture3 = picture3/5;
%代码 2:事件对比图
z = 1;
picture1 = zeros(12,1200);
picture2 = zeros(12,1000);
M = ['A','B','C','D','E','F';
     'G','H','I','J','K','L';
     'M','N','O','P','Q','R';
     'S','T','U','V','W','X';
     'Y','Z','1','2','3','4';
     '5','6','7','8','9','0'];
k =1;
    data = xlsread('data/S1/S1_train_data.xlsx',k);
m = 1;
        data1 = data(:,m);
        event = xlsread('data/S1/S1_train_event.xlsx',k);
        [status,chart,xlsFormat] = xlsfinfo('data/S1/S1_train_event.xlsx');
        zimus = char(chart(k));
        zimu = zimus(8);
        [num1,num2] = find(M==zimu);
        event(event(:,1)==100+6*(num1-1)+num2 |event(:,1)==100,:)=[];
        for i =1:60
            picture2(event(i,1),200*(z-1)+1:200*z)=
filter(b,a,data1(event(i,2):event(i,2)+199)-data1(event(i,2)));
            picture1(event(i,1),200*(z-1)+1:200*z)=
data1(event(i,2):event(i,2)+199)-data1(event(i,2));
            if mod(i,12)==0
                z = z+1 ;
            end
        end
 picture2(:,1001:1200)=picture3;

figure
j=1;
for i = 1:12
    subplot(6,2,i);
    if i==num1 ||i==num2+6
        plot(1:1200,picture2(i,:),"r");
    else
        plot(1:1200,picture2(i,:));
    end
    hold on ;
```

```
    plot([200,200],[-30,30],'k',[400,400],[-30,30],'k',...

[600,600],[-30,30],'k',[800,800],[-30,30],[1000,1000],[-30,30],'k','linewidth
',1)
end

%准备数据集
clc;clear;
[b,a]=butter(8,10/(250/2),'low');
z = 1;
traindata= zeros(201,2880);
M = ['A','B','C','D','E','F';
     'G','H','I','J','K','L';
     'M','N','O','P','Q','R';
     'S','T','U','V','W','X';
     'Y','Z','1','2','3','4';
     '5','6','7','8','9','0'];
for k =1:12
    data = xlsread('data/S1/S1_train_data.xlsx',k);
    for m = 1:20
        data1 = data(:,m);
        event = xlsread('data/S1/S1_train_event.xlsx',k);
        [status,chart,xlsFormat] = xlsfinfo('data/S1/S1_train_event.xlsx');
        zimus = char(chart(k));
        zimu = zimus(8);
        [num1,num2] = find(M==zimu);
        event(event(:,1)==100+6*(num1-1)+num2 |event(:,1)==100,:)=[];
        for i = 1:60
            if event(i,1) == num1 || event(i,1)==num2+6
                traindata(1,z-1+event(i,1)) = 5;
            else
                traindata(1,z-1+event(i,1)) = -5;
            end

traindata(2:201,z-1+event(i,1))=traindata(2:201,z-1+event(i,1))...

+filter(b,a,data1(event(i,2):event(i,2)+199)-data1(event(i,2)));
        end
        z = z+12;
    end
end
traindata = traindata/5;

%%生成测试数据
```

```
%%
z = 1;
testdata= zeros(201,12000);
for k =1:10
    data = xlsread('data/S1/S1_test_data.xlsx',k);
    for m = 1:20
        data1 = data(:,m);
        event = xlsread('data/S1/S1_test_event.xlsx',k);
        event(event(:,1)==666 |event(:,1)==100,:)=[];
        for i =1:60
            testdata(1,z) = event(i,1);

testdata(2:201,z)=filter(b,a,data1(event(i,2):event(i,2)+199)-data1(event(i,2)
));
            z = z+1;
        end
    end
end

%%
%训练
model =
fitcsvm(traindata(2:end,:)',traindata(1,:)','KernelScale','auto','Standardize
',true,...
    'OutlierFraction',0.05);


test = testdata;
[label,score] = predict(model,test(2:end,:)');
res = [test(1,:)',label];

aa=zeros(12,10);bb=zeros(10,2);
for k = 1:10
    res1 = res(1+1200*(k-1):1200*k,:);
    for i = 1:12
        aa(i,k) = length(find(res1(:,1)==i & res1(:,2)==1));
    end
    [m1,n1]=max(aa(1:6,k));[m2,n2]=max(aa(7:12,k));
    bb(k,:)=[n1,n2+6];
end
bb
%%
xlswrite('new_data/5_jun/traindata1.xlsx',traindata);
xlswrite('new_data/5_jun/testdata1.xlsx',testdata);
```

```
import xlrd
import numpy as np
import random
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import train_test_split

model = MLPClassifier(alpha=1e-4, solver='adam',activation='relu',max_iter=300,
                  hidden_layer_sizes=(12,12), random_state=12)

#from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
#model = LinearDiscriminantAnalysis()

#from sklearn import svm
#model =svm.SVC()


tt = 0
xl = xlrd.open_workbook(r'new_data/5_jun/traindata1.xlsx')
lables = [];datas = [];
datas1= [];datas0= [];
table = xl.sheets()[tt]

for j in range(table.ncols):
    col = table.col_values(j)  # 读取第一列的数据
    if col[0]==1:
        datas1.append(col)
    elif col[0]==-1:
        datas0.append(col)


b = np.vstack((datas1,datas1))
for i in range(5-1):
    b = np.vstack((b,datas1))

#all_data = np.vstack((datas0[:480],datas1))
all_data = np.vstack((datas0,b))
#random.shuffle(all_data)
for m in all_data:
    lables.append(m[0])
    datas.append(m[1:])
```

```
X_train,X_test,y_train,y_test=train_test_split(datas,lables,test_size=0.1,ran
dom_state=10)
model.fit(X_train,y_train)
predicted = model.predict(X_test)

print(sum(predicted))
print(sum(y_test))
print('正确率为:',np.sum(predicted==y_test)/len(y_test))

xl = xlrd.open_workbook(r'new_data/5_jun/testdata.xlsx')
table = xl.sheets()[tt]

for i in range(10):
    event = []
    datas = []
    t=[]
    zz=[]
    res=[]
    a=[];b=[]
    for j in range(i*240,(i+1)*240):
        col = table.col_values(j)  # 读取 第一列的数据 table.ncols
        event.append(col[0])
        #datas.append([i*2.5 for i in col[2:]])
        datas.append(col[1:])
    t = model.predict(datas)
    zz = np.vstack((event,t))
    for k in range(1,13):
        res.append(sum(zz[1][zz[0]==k]))
    print('有效数据:',sum(res)/240)
    a=res[:6];b=res[6:];
    #print(a,b)
    print( (a.index(max(a))+1,b.index(max(b))+7))
```