



**数据挖掘技术**

**专题作业**

学 院 计算机学院

班 级

姓 名

学 号

**2024年12月26日**

目录

[1. ECG睡眠分期综述 3](#_Toc186144290)

[2. 基于AASM标准的5分类模型 3](#_Toc186144291)

[3. 任务描述、模型选择说明、拟用算法步骤 3](#_Toc186144292)

[4. 数据描述 4](#_Toc186144293)

[5. 数据预处理与特征提取 5](#_Toc186144294)

[6. 调试分析与模型评价 6](#_Toc186144295)

[7. 结论分析 7](#_Toc186144296)

# 1. ECG睡眠分期综述

心电图（ECG）信号在睡眠医学中的应用已成为近年来的研究热点。ECG信号可以提供丰富的生理信息，例如心率变异性（HRV）、R波间期等，这些信息与睡眠的不同阶段密切相关。与传统的多导睡眠记录（PSG）相比，ECG信号具有采集设备简单、成本低和患者舒适度高的优点。

基于ECG信号的睡眠分期研究主要集中在以下几个方面：

1) 信号处理技术：包括噪声去除、特征提取和信号分解等。

2) 分类模型：从早期的线性模型（如逻辑回归）到复杂的深度学习模型（如卷积神经网络和循环神经网络）。

3) 性能评估：使用准确率、敏感性、特异性和F1评分等指标评估模型性能。

近年来的研究表明，结合机器学习技术和临床诊断标准（如AASM标准）进行自动化的睡眠分期，可以显著提高分期效率。然而，基于ECG信号的睡眠分期仍面临诸多挑战，例如信号特征的高维性、数据标注的不一致性以及跨个体差异等问题。

# 2. 基于AASM标准的5分类模型

美国睡眠医学会（AASM）标准是当前睡眠医学领域的主要诊断依据之一，将睡眠阶段分为5个类别：觉醒（Wake）、浅睡（N1）、中度睡眠（N2）、深度睡眠（N3）和快速眼动睡眠（REM）。在本实验中，我们使用AASM标准对睡眠分期进行5分类。

为实现此目标，我们设计了如下工作流程：

1) 数据预处理：清洗和过滤原始ECG信号，提取特征用于分类；

2) 分类模型：使用深度学习模型（如卷积神经网络）和传统机器学习算法（如支持向量机）进行实验对比；

3) 模型评估：通过交叉验证确保模型的鲁棒性，并使用多种指标（如准确率和F1评分）评估性能。

实验中严格遵循5折交叉验证：数据集被随机分为5个不重叠的子集，每次使用其中一个子集作为测试集，其他4个作为训练集。重复5次，每次评估模型在测试集上的性能，并最终计算平均性能指标。

# 3. 任务描述、模型选择说明、拟用算法步骤

任务描述：本实验的目标是基于临床诊断标准（AASM）构建一个能自动识别睡眠阶段的分类模型，以便在临床诊断和家庭健康监测中应用。

模型选择说明：考虑到ECG信号的高维特性和时间序列特点， 我选择长短期记忆网络（LSTM）作为模型，因为其针对时间序列数据建模效果优异。

拟用算法步骤：

1) 数据加载与预处理；

2) 提取时域和频域特征；

3) 使用训练数据训练模型，并调整超参数以优化性能；

4) 通过交叉验证评估模型的稳定性；

5) 对测试集进行预测，并分析模型的性能表现。

# 4. 数据描述

#### 4.1 数据概况

本实验数据来自公开的睡眠数据库，该数据集包含多导睡眠记录（PSG）和专家标注的睡眠阶段标签。每个受试者的记录包含以下内容：

**信号类型**：ECG，EEG等信号。

**采样频率**：200Hz，30秒为一个Epoch。

**标签标准**：睡眠阶段按照AASM标准标注为五类（Wake、N1、N2、N3、REM）

#### 4.2 数据结构与属性

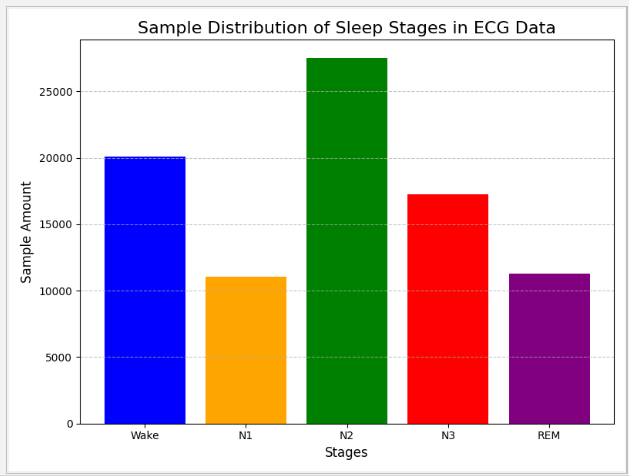
**记录数量**：包含100名受试者的完整夜间睡眠记录，每名受试者的数据时长约为7-8小时。

**特征属性**：以时间序列形式记录的ECG信号，单位为毫伏（mV）。

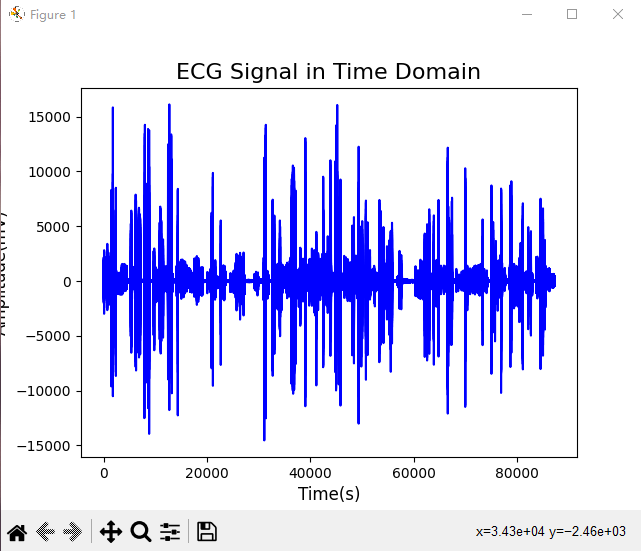
**标签分布**：不平衡分布特征明显，其中N2阶段样本最多，N1阶段样本最少。

#### 4.3 数据分布可视化

原始数据的信号图以及标签分布图如下所示：



4-1 ECG数据标签分布图



4-2 ECG数据的信号强度图

# 5. 数据预处理与特征提取

为了保证实验数据质量，需对原始ECG信号进行以下预处理，本次实验采用的是.mat文件中X2变两个对应的数据，后经过对比，发现.mat文件中的数据是滤波处理之后的，故而也可以省略掉滤波这步：

1. **噪声处理**

使用带通滤波器（Bandpass Filter）去除工频干扰（50Hz）和基线漂移（低频噪声）。滤波范围设定为0.5Hz-50Hz，以保留ECG信号的主要频段信息。

1. **缺失值填补**

缺失值比例较低时，使用线性插值法对信号缺失段进行补全。

若缺失值过多，则直接剔除受损片段，保证数据完整性。

1. **信号分段**

按照30秒为单位对ECG信号进行分段，每段对应一个睡眠阶段标签。这是因为睡眠分期专家标注通常以30秒为时间窗口。每行数据6000条。

1. **标准化处理**

为消除不同个体间信号幅度的差异，将信号幅度归一化到[-1, 1]区间。

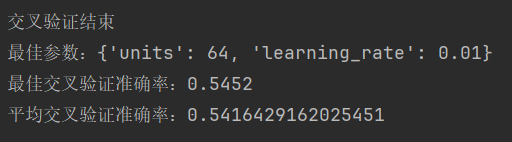
1. **特征提取**

选取均值，标准差，Petrosian分形维数，中值，偏度，峰度，排列熵，svd分解熵，吸引子维数等9个特征作为训练的基本特征。每行数据6000条，以200条数据为单位计算特征，一个Epoch数据会产生30组基本特征。设置时间步为30，每组训练特征会有30\*9条特征数据，代入LSTM模型中进行训练。提取的特征汇总成excel存放在压缩包中。

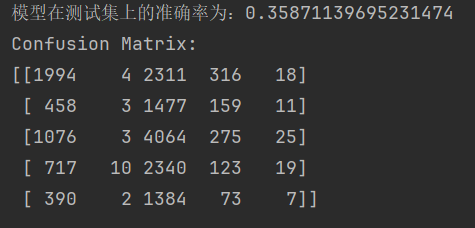
# 6.调试分析与模型评价

**调试分析**：本次实验的报错主要集中在环境版本不兼容上，前期由于驱动的缘故，只能安装较低版本的tensorflow，导致经常和包起版本冲突。之后更新显卡驱动，安装新版tensorflow，能够正常执行。

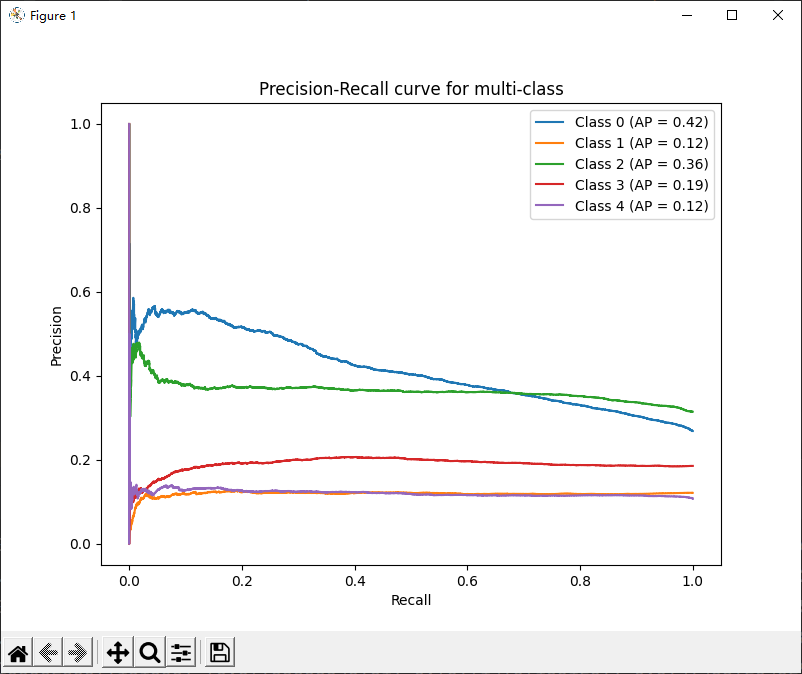
**模型评价：**使用训练集上交叉验证平均准确率，测试集上准确率、混淆矩阵、P-R曲线、精确率和召回率、F1 score等；



6-1 交叉验证平均准确率



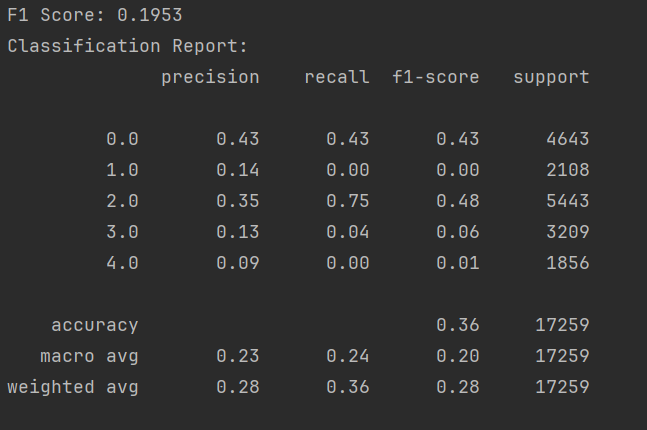
6-2 测试集上准确率与混淆矩阵



6-3 模型分类P-R曲线

# 7. 结论分析

本模型基于LSTM使用ECG数据进行睡眠分期，准确率超过1/3，仍然存在优化的空间，可作为划分睡眠分期的一个参考；



7-1 F1 score与分类报告