# 第九届湖南省研究生数学建模竞赛承诺书

我们仔细阅读了湖南省高校研究生数学建模竞赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

**我们完全清楚，在竞赛中必须合法合规地使用文献资料和软件工具，不能有任何侵犯知识产权的行为。否则我们将失去评奖资格，并可能受到严肃处理。**

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们授权湖南省研究生数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

所属学校和学院（请填写完整的全名）：国防科技大学电子科学学院

参赛队员（打印后签名）：1. 方丽萌

2. 刘琪琪

3. 耿敏娟

指导教师或指导教师组负责人（打印后签名）：刘鹏跃

日期：　2024 年 7 月 3 日

# 智能纪元下的动作解码：智能手机感知人体活动状态的模型之旅

## 摘要

随着智能手机等电子设备的发展，人体活动状态感知与识别问题是人机交互中的重要研究课题。本文针对基于传感器数据的分类和判决人体活动状态的问题，对数据进行特征提取，基于聚类和决策树思想，以活动状态归类和识别为目标分别建立了分类模型和判别模型，并使用聚类算法和随机森林算法对模型进行求解。

**针对问题一：**对于原始数据进行分析，进行活动状态分类的关键在于数据挖掘和特征提取，因此考虑了多域的特征指标，随后基于聚类思想，建立了分类模型，并进行了模型的优化与评价。本文使用谱聚类算法对该模型进行求解，求解结果可视化截图如下，具体见正文。

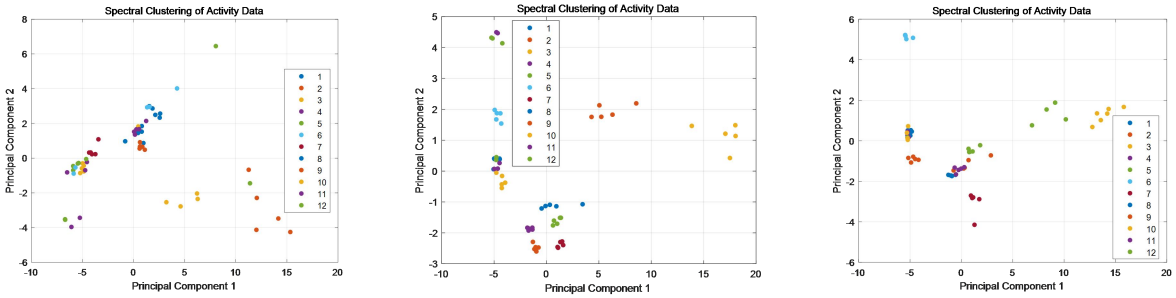



图 1　问题一可视化结果

**针对问题二：**首先提出了以随机森林作为判别器的判别模型，数据集的处理同问题一，该模型的准确率达到 84.17%，与问题一准确率只有 79.9%的分类模型相比，效果有所提升；其次，在该判别模型的基础上，进一步提出双层判别模型，使用随机森林和 LSTM 的加权输出作为最终的判别结果，受到测试集数量和特征值的限制，准确率达到 85.83%，相比最初的判别模型准确率提升了 1.66%。最后使用双层判别模型对附件 3 的特征集进行活动判别。

**针对问题三：**对于不同实验人员的活动数据进行灰色关联分析，同时利用斯皮尔曼（spearman）相关系数分析人体年龄、身高、体重属性因子对活动数据的影响，建立了未知实验人员情况下的活动数据个性化分类判别模型。首先需对问题二的特征集进行优化，然后选取随机森林作为个性化判别模型；经测试，该模型可对附件 5 中 5 类未知活动数据正常判别。

最后我们对模型建立过程中的优化策略，改进效果等方面进行总结与分析。

**关键词：数据挖掘　人体活动识别　特征提取　随机森林算法　相关系数**

# 目录

# 1 问题综述

## 1.1 问题背景

智能手机作为 21 世纪最具革命性的科技产品之一广泛应用于人们的生活之中，其内置的高精度传感器和先进的计算能力，不仅改变了人们的通信方式，还极大地拓展了手机在健康管理、日常活动监测等领域的应用。同时，近年来随着人工智能，大数据分析等技术的飞速发展，智能手机对人体活动状态的识别能力也得到了显著提升。

智能手机通过其中的传感器测量数据进行姿态、角度和方向的变化分析人体态势从而进行活动状态的识别和判断。人体活动状态识别的应用前景广阔，如智能人机交互、智能视频监控、运动员辅助训练、健康监测等。它不仅能为个人提供定制化的健康建议和运动指导，帮助用户更好地管理生活习惯和提升生活质量，还能在医疗监测、老年人护理、职业健康评估等领域发挥重要作用。例如，通过分析用户的日常活动量，智能手机应用可以提醒久坐的用户适时休息，预防健康问题；对于需要进行康复训练的患者，精准的活动监测则能帮助医生调整治疗方案，加速康复进程。因此关于人体活动状态的研究对于提高生活质量具有重要意义。


图 2 智能手机的各类人体状态检测

基于传感器的人体运动姿态识别研究涉及较多方面的知识，如传感器、传输技术、信号处理、机器学习与模式识别[1]。该领域的研究过程主要是通过可穿戴传感器或环境传感器等设备收集人体活动的数据，并利用机器学习和数据分析方法对这些数据进行处理，以实现对人体活动的准确识别。该文献[2]通过由 30 位志愿者在智能手机上执行六项动作（行走、上楼梯、下楼梯、坐、站立、躺下）时产生的加速度传感器和陀螺仪数据集，设置传感器频率为 50Hz，对人体活动识别进行研究。随着深度学习技术的发展，卷积神经网络（CNN）、循环神经网络（RNN）、长期短期记忆（LSTM）等技术在人体活动状态识别中取得了显著进展。这些方法能够自动学习并提取出有效的特征，避免了传统方法中手工设计特征的繁琐和局限性。例如，一些研究利用 3D CNN 来捕获时空特征，从而对人体动作进行识别[3]。此外，基于骨骼信息的识别方法也受到了广泛关注。这类方法通常从视频中提取出人体骨骼信息[4]，然后利用图卷积网络（GCN）等模型对骨骼信息进行建模，以识别出不同的动作。

总之，人体活动状态的识别可采用多种方法，我们根据实际情况进行合理的模型和算法选择并从多方面考量进行模型优化，分析改进模型效果。

## 1.2 问题提出

人体活动状态的感知问题，涉及多方面的问题，主要由数据采集与预处理, 特征提取与选择, 数据聚类与识别算法这几个要素构成，其关键在于数据挖掘与分析。本文需要从数据处理、模型选择、算法实现方面考虑, 解决以下 3 个问题：

(1) 问题 1: 分别完成 3 名实验人员的 12 种活动状态划分，无需对应具体状态类别，只实现归类。
(2) 问题 2: 首先根据带有标签的 10 名实验人员数据集建立识别模型，将分类结果与问题 1 作比较，其次通过该模型完成对其他一名实验人员具体活动状态的判断。
(3) 问题 3:根据以上实验人员的条件数据，对比分析不同人的同一活动状态情况，并建立判别模型，利用给出的活动数据完成 5 名实验人员的活动数据与提供该数据人员的对应。

## 1.3 资料条件

附件 1-5 文件提供了用于求解分类和识别问题的原始测量数据，每个活动状态对应一个包含测试样本的 Excel 文件，其中每个文件的样本数各不相同。各文件的详细说明如下:

- SY*.xlsx 和 a*t*.xlsx 文件提供了加速度计在三个轴向(X，Y，Z)上的线性加速度变化数据和陀螺仪在三个轴向(X，Y，Z)上的角速度的数据, 正是通过这些测量数据来感知手机等设备的姿态、角度和方向的变化，并设计相应的算法分析处理数据识别手机使用人的活动模式，因此对这些原始数据的处理与挖掘是建立模型解决问题的重要基础。
- 附件 4.xlsx 文件介绍了参与测试人员的年龄、身高、体重等数据，利用这些数据可以分析不同人员的活动状态情况。

# 2 符号说明

本文定义了如下表 2 所示使用次数较多的符号，其余符号在使用时注明。

表 2 符号说明

| 符号 | 含义 | 单位 |
|---|---|---|
| $X_i$ | 加速度/角速度在 X 轴方向的所有样本数据 | 无 |
| $Y_i$ | 加速度/角速度在 Y 轴方向的所有样本数据 | 无 |
| $Z_i$ | 加速度/角速度在 Z 轴方向的所有样本数据 | 无 |
| $\bar{X}$ | 加速度/角速度在 X 轴方向的所有样本数据的平均值 | 无 |

# 3 数据预处理

## 3.1 附件 1235 数据处理

本文在数据采集阶段采用了 100Hz 的传感器采样率。为了防止在采集过程中由于轻微的抖动或设备自身的不稳定因素引入噪声，对后续数据处理造成不利影响，我们对原始数据实施了滤波降噪处理，以去除冗余信息，从而有效提升了动作状态的识别准确率。

进一步地，通过细致分析 person1 的加速度计（acc_y）数据的散点图，我们观察到该序列信号显示出一定的周期性特征。尽管如此，序列中仍然混杂着轻微的噪声干扰。

这种噪声如果不加以处理，可能会对信号的周期性特征分析造成干扰。因此，我们的滤波降噪工作不仅提高了数据质量，也为准确地捕捉和分析动作周期性提供了有力保障。

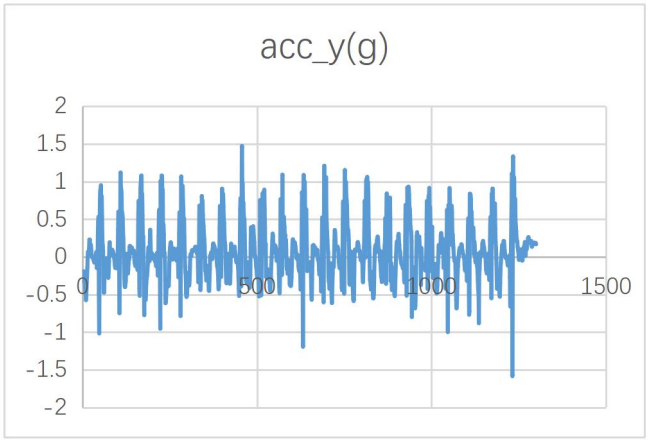



图 3　数据散点图

滤波方法有滑动平均滤波法、加权递推平均滤波法、Savitzky - Golay 滤波算法、小波滤波算法等。通过查找文献[5-6]，可知从傅里叶变换的幅度特征分析得出 0-10hz 频率的幅值特征的重要性远大于 10-20hz 频率的幅值特征。因此这里我们采用截止频率为 10Hz 的三阶巴特沃斯低通滤波器对所有原始的一组采集数据进行滤波处理，滤波效果如下图所示，结果显示滤波后的数据更加平滑且没有过分失真，同时低频信号特征更加突出。

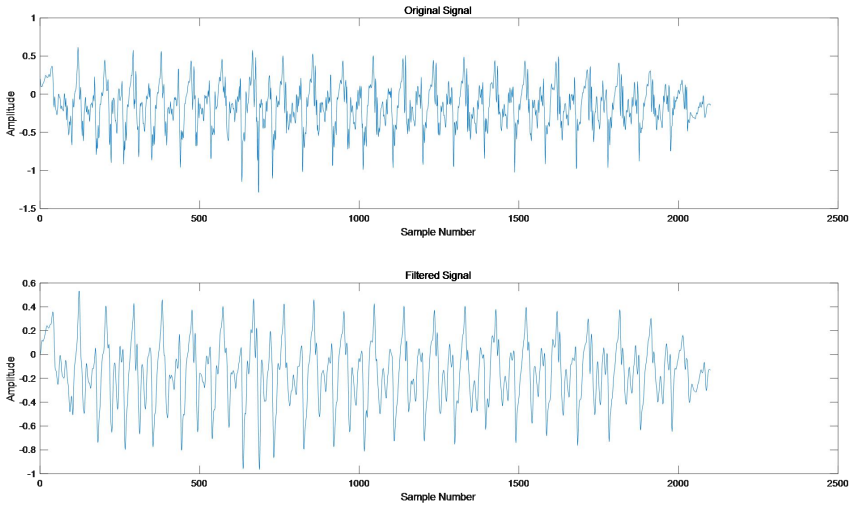



图 4　数据滤波效果

# 4 问题一

## 4.1 问题分析与模型建立

题目以智能手机的健康软件为背景，介绍了其监测人体活动状态通过处理传感器测得数据感知人体姿态的原理，问题 1 的核心在于如何将 60 组数据准确地划分为 12 类，这些数据分别对应 12 种不同的活动状态，每种状态均由五组数据组成。大致解决思路是：首先对现有的海量数据进行深入的数据挖掘，以提取反映数据本质的特征信息，再根据信息运用聚类分析技术，将特征相似度高的数据样本聚合在一起，形成不同的类别。

### 4.1.1 分析与求解思路

在所给的每个测试者的数据中共有 12 种活动状态，每种活动状态有五组的加速度计和陀螺仪分别在 X、Y、Z 轴上的六列数据，也就是 60 组数据。60 组数据分别包含数量不一的测试样本，而一组数据对应一个活动状态也就包含了该状态所具有的信息，反映其运动规律，因此首先对一组数据进行处理与分析。

将随机选取某组数据的第一列样本可视化，得到如下图所示的结果，可以发现，有的数据具有周期性有的则起伏波动很小，所以若对每列数据进行某一种统计处理即可得到相对应的特征指标值。

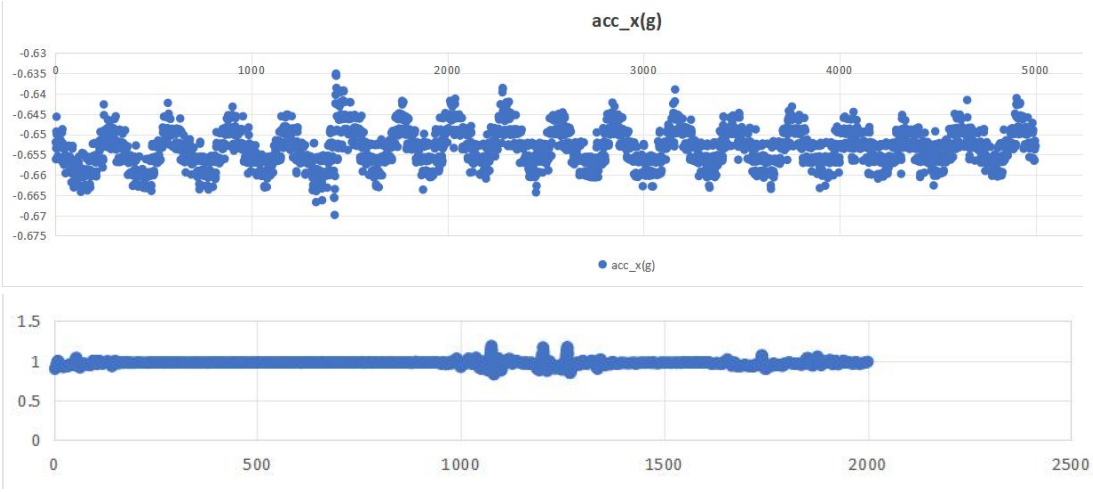



图 5　acc_x 样本数据可视化结果

随机选取的一列数据绘制其频率直方图和频谱图如下所示，由此频域特征也可体现。

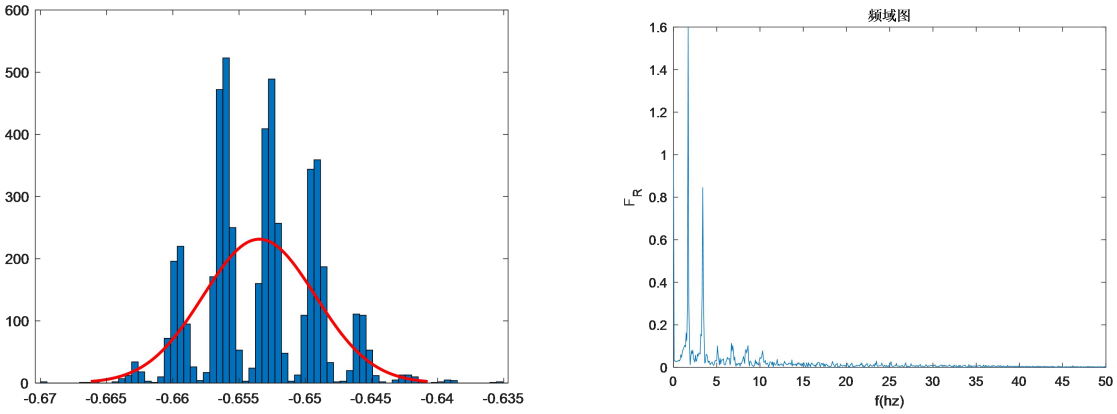



图 6　（a）样本数据频率直方图　（b）样本数据频域图

不同的活动状态具有不同的运动特点，测量的数据也具有差异，因此根据所提取的特征信息，选取合适的分类模型方法即可实现对活动状态的类别划分。

### 4.1.2 模型建立——基于特征提取与聚类算法的人体活动状态分类

（1）基本框架

根据上述分析，求解该分类问题的主要步骤有数据处理、特征提取与选择、聚类算法，模型建立流程如下图所示。

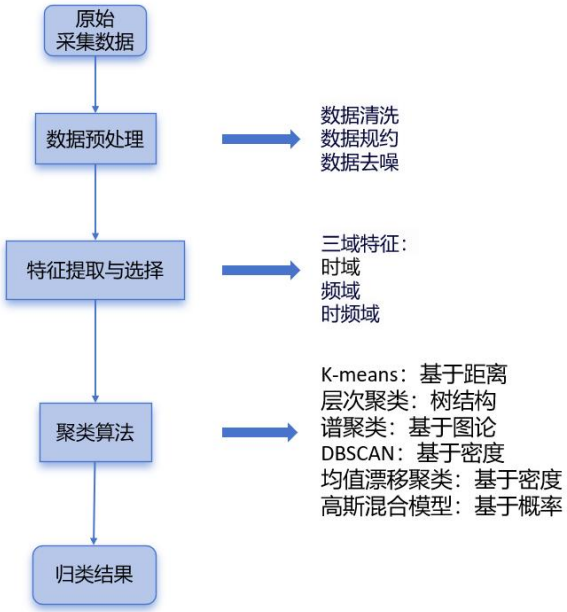



图 7　模型建立框架

（2）特征提取

在一个人体运动姿态识别系统的构建中，特征提取与筛选扮演着连接数据采集与姿态识别两大核心环节的关键角色。该过程普遍会对初始的加速度和角速度数据进行深入分析，旨在精准捕捉并表达人体运动过程中的物理特性与识别目标的本质特征。其核心在于，从纷繁复杂的原始数据中提炼出对运动姿态识别最具判别力的特征向量，这一举措直接增强了系统分类与识别的准确性，从而对整体分类与识别的性能产生深远的影响，确保了更高的识别效率与可靠性。

由此，人体运动状态识别的关键一点是提取有效特征表示，所以通过对原始数据的抽象与统计获取可以准确描述数据的特征指标值。特征提取类别主要有时域、频域以及时频域，下面是较为广泛应用的特征值。

表 3　常用特征列表

| 特征类别 | 特征名称 | 计算方法 |
|---|---|---|
| 时域 | 平均值（mean） | $\frac{1}{n}\sum_{i=1}^{n}X_i$ |
| | 方差（STD） | $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ |
| | 均方根（RMS） | $\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}$ |
| | 最大值 最小值 | $\max(X_i)\ \ \min(X_i)$ |
| | 峰值 | $\max\{\lvert X_i\rvert\}$ |
| | 信号幅值面积（SMA） | $\frac{1}{t}(\int_0^t\lvert X(t)\rvert dt+\int_0^t\lvert Y(t)\rvert dt+\int_0^t\lvert Z(t)\rvert dt)$ |
| | 信号能率 | $\sum_{i=1}^{n}X_i^2$ |

| 特征类别 | 特征名称 | 计算方法 |
| --- | --- | --- |
| | 能量对数 | $\sum_{i=1}^{n}\log(X_i^2)$ |
| | 偏度 | $E[(\frac{X_i-\overline{X}}{\sigma})^3]$ |
| | 四分位差（IQR） | 75%和 25%位置差的平均值 |
| | FFT 系数 | |
| 频域 | 频域熵（FDE） | |
| | 能谱密度（PSD） | |
| 时频特征 | 小波系数（Wavelets） | |
| Counts | 合速度 VM | $\sum_{i=1}^{n}\sqrt{X_i^2+Y_i^2+Z_i^2}$ |

### 4.1.3 聚类算法

常见的聚类算法有 K-means 聚类算法、系统（层次）聚类算法、谱聚类算法、DBSCAN 算法等[7-9]。其中，K-means 聚类通过指定聚类数 K，随机选择 K 个初始聚类中心，通过迭代优化聚类中心，直至聚类中心不再变化。虽然该算法实现简单效率高，但缺点是需要预先指定 K 值，对初始聚类中心和噪声数据敏感，大量数据时可能无法实现结果收敛；DBSCAN 算法是基于密度的聚类算法，虽然能够处理任意形状和大小的聚类，可以发现异常点，但是对参数敏感，难以确定最佳参数值，同时在求解本文问题的实际实验过程中我们发现由于参数值确定的困难聚类效果也不如意。同时由于谱聚类算法在处理非线性可分数据和高维数据方面具有优势，因此我们舍弃前两种算法采用谱聚类和层次聚类的方法。

谱聚类算法是一种基于图论和矩阵特征值分解的聚类方法，它通过构建相似度矩阵、选择特征向量和聚类等步骤来实现对数据集的聚类分析。其关键在于其划分准则，即如何确定最佳的聚类划分，常见的划分准则包括 RatioCut、Ncut 等，这些准则都旨在最大化子图内部的相似度，同时最小化子图之间的相似度。该算法在处理非线性可分数据和高维数据方面具有优势。


图 8　谱聚类算法步骤

层次聚类算法（Hierarchical Clustering Algorithm）是一种无监督学习算法，将数据集划分成多个不同层次的簇。其原理主要基于样本之间的相似度或距离来构建一个层次结构，从而得到样本之间的聚类关系。优点是距离和规则的相似度容易定义，限制少。

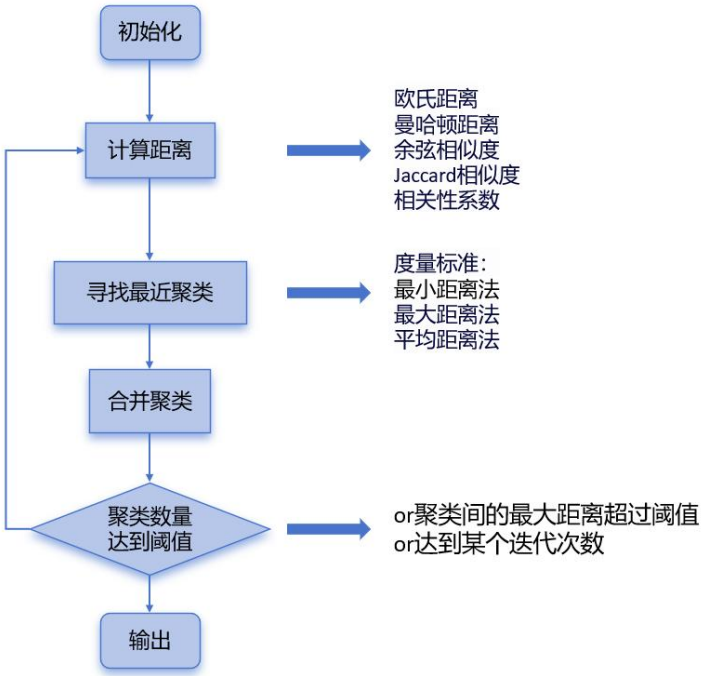



图 9　层次聚类算法步骤

### 4.1.4 分类结果

基于谱聚类算法的分类模型的分类结果如下表。

表 4　问题 1 结果

| 分类 | Person1 | Person2 | Person3 |
|---|---|---|---|
| 第 1 类 | SY27 SY31 SY37 SY49 SY52 | SY16 SY21 SY29 SY40 SY43 | SY32 SY3 SY50 SY54 SY55 |
| 第 2 类 | SY15 SY19 SY26 SY43 SY44 | SY15 SY25 SY35 SY58 SY6 | SY19 SY24 SY35 SY36 SY49 |
| 第 3 类 | SY11 SY12 SY18 SY20 SY9 | SY24 SY33 SY44 SY57 SY7 | SY15 SY28 SY41 SY45 SY46 |
| 第 4 类 | SY13 SY32 SY39 SY40 SY51 | SY1 SY48 SY49 SY4 SY50 | SY10 SY4 SY53 SY5 SY7 |
| 第 5 类 | SY10 SY21 SY24 SY3 SY55 | SY11 SY22 SY54 SY56 SY5 | SY11 SY25 SY31 SY44 SY60 |
| 第 6 类 | SY29 SY48 SY4 SY53 SY57 | SY13 SY27 SY28 SY34 SY42 | SY14 SY30 SY37 SY43 SY9 |
| 第 7 类 | SY16 SY30 SY45 SY54 SY6 | SY10 SY12 SY26 SY2 SY47 | SY17 SY21 SY33 SY38 SY6 |
| 第 8 类 | SY25 SY46 SY56 | SY20 SY38 SY3 | SY13 SY27 SY29 |

| 分类 | Person1 | Person2 | Person3 |
|---|---|---|---|
| 第 9 类 | SY5 SY8<br>SY33 SY38 SY42<br>SY58 SY7 | SY51 SY60<br>SY17 SY41 SY52<br>SY59 SY8 | SY51 SY56<br>SY16 SY1 SY20<br>SY26 SY40 |
| 第 10 类 | SY1 SY23 SY2<br>SY36 SY59 | SY19 SY23 SY30<br>SY37 SY46 | SY18 SY22 SY57<br>SY58 SY59 |
| 第 11 类 | SY14 SY17 SY28<br>SY41 SY47 | SY31 SY32 SY39<br>SY45 SY9 | SY23 SY39 SY42<br>SY52 SY8 |
| 第 12 类 | SY22 SY34 SY35<br>SY50 SY60 | SY14 SY18 SY36<br>SY53 SY55 | SY12 SY2 SY34<br>SY47 SY48 |

## 4.2 模型改进

为了提高分类准确度，我们对分类模型的改进主要从三个维度开展，一是特征提取方面，增加更多有用的特征：不仅包括时域、频域和时频这三域特征，还可以包括统计特征、信号能量；二是对于提取的特征数据的处理，采用主成分分析方法对特征数据进行降维处理以减小噪声，提高聚类效果；三是针对所采取的具体聚类算法进行改进，比如，在谱聚类算法中优化特征向量选择，改进相似度矩阵的构造，在层次聚类算法中优化约束和聚类策略，若之后进行进一步研究，或可采取机器学习的方法来学习最佳的距离度量，以更好地捕捉样本之间的相似性，比如孪生网络（Siamese Network）。在实际改进过程中主要完成了前两个维度的优化。

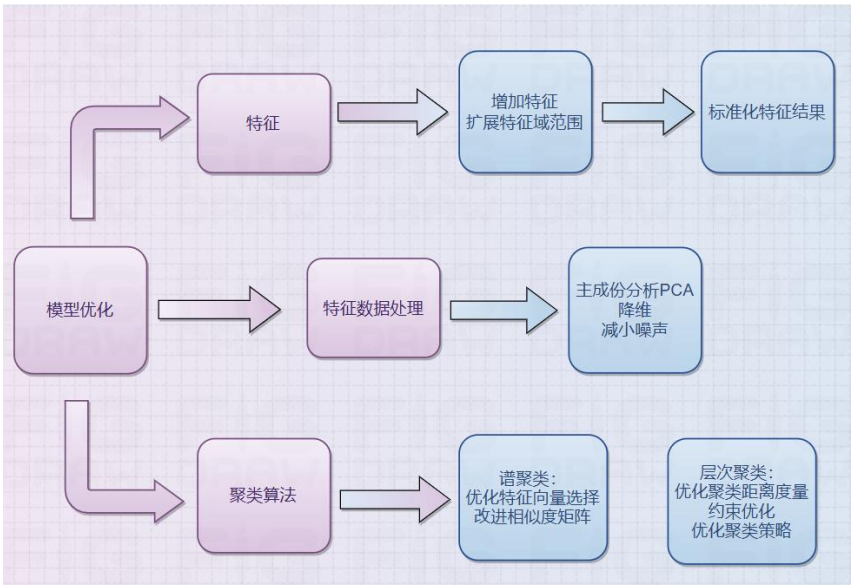
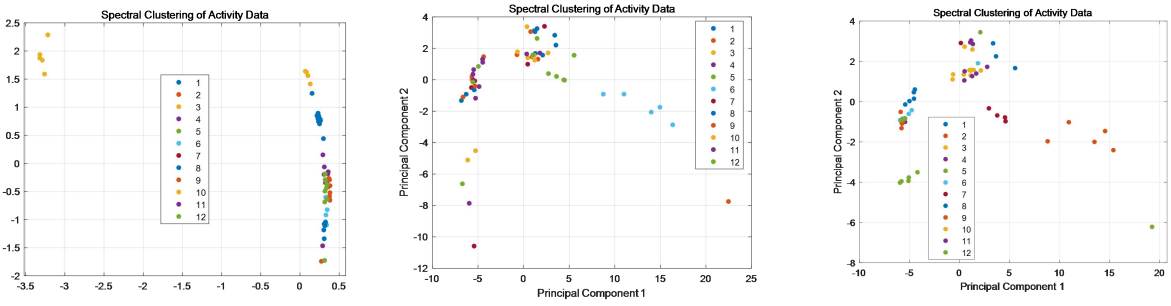



图 10　模型优化策略

除此之外，在模型的具体优化过程中，我们还通过标准化特征结果以保证特征值的有效性，以及增加平衡每个聚类大小的算法，确保每个类别中有 5 组数据，以解决聚类结果中分布不均的问题，改善分类效果。优化前后的效果对比如下图所示，图（a）是

未平衡聚类大小之前的效果，每个类别中的数据数量相差较大，分布不均匀，改进之后聚类大小一致均为 5 组数据，结果如图（b）、（c）所示。




（a）　　（b）　　（c）

图 11　基于谱聚类算法的聚类结果图

其次在准确度方面，以 person4 数据为例，在扩展特征提取范围和采用主成分分析进行特征数据降维之前的准确率为 43%，改进之后准确率为 72%。分类结果部分截图如下图所示。

```
Cluster 3:
    {'a6t1_filtered.xlsx'}
    {'a6t2_filtered.xlsx'}
    {'a6t3_filtered.xlsx'}
    {'a6t4_filtered.xlsx'}
    {'a6t5_filtered.xlsx'}

Cluster 4:
    {'a3t1_filtered.xlsx'}
    {'a3t2_filtered.xlsx'}
    {'a3t3_filtered.xlsx'}
    {'a3t4_filtered.xlsx'}
    {'a3t5_filtered.xlsx'}
```

```
Cluster 5:
    {'a10t1_filtered.xlsx'}
    {'a10t3_filtered.xlsx'}
    {'a10t4_filtered.xlsx'}
    {'a10t5_filtered.xlsx'}
    {'a8t5_filtered.xlsx' }

Cluster 6:
    {'a8t3_filtered.xlsx'}
    {'a8t4_filtered.xlsx'}
    {'a9t1_filtered.xlsx'}
    {'a9t3_filtered.xlsx'}
    {'a9t5_filtered.xlsx'}
```

图 12　分类结果部分展示

## 4.3 模型评价

为对所构建的模型进行包含准确度等方面的评价，使用该模型对附件 2 中带有类别标签的数据进行分类，并分析模型效果。

首先若采用层次聚类算法，以 person4 数据的分类结果为例，在改进算法以保证结果中每类有 5 组数据之前的聚类结果如下图所示，可以发现虽然每一类包含数量不同，但是可以将大多相同活动状态归为一类。

```
Cluster 2:
    {'a3t1_filtered.xlsx'}
    {'a3t2_filtered.xlsx'}
    {'a3t3_filtered.xlsx'}
    {'a3t4_filtered.xlsx'}
    {'a3t5_filtered.xlsx'}
    {'a4t1_filtered.xlsx'}
    {'a4t2_filtered.xlsx'}
    {'a4t3_filtered.xlsx'}
    {'a4t5_filtered.xlsx'}
```

```
Cluster 9:
    {'a5t1_filtered.xlsx'}
    {'a5t2_filtered.xlsx'}
    {'a5t3_filtered.xlsx'}
    {'a5t4_filtered.xlsx'}
    {'a5t5_filtered.xlsx'}

Cluster 10:
    {'a6t1_filtered.xlsx'}

Cluster 11:
    {'a7t2_filtered.xlsx'}
    {'a7t3_filtered.xlsx'}
    {'a7t4_filtered.xlsx'}
    {'a7t5_filtered.xlsx'}

Cluster 12:
    {'a10t1_filtered.xlsx'}
    {'a10t2_filtered.xlsx'}
    {'a10t3_filtered.xlsx'}
    {'a10t4_filtered.xlsx'}
    {'a10t5_filtered.xlsx'}
```

图 13　基于层次聚类算法的分类结果

说明该算法在此次实验中虽然没有实现最终的 5 组数据一类的效果，可能是算法设计中迭代次数未达到最优导致聚类没有继续进行，但是将相似特征数据进行归类的聚类是有效的，而根据该算法的原理，其通过不断更新距离计算将数据集划分为树形结构，因此后续或将改进算法中的停止条件以便完成最终聚类。该算法的树形图聚类结果如下图所示。

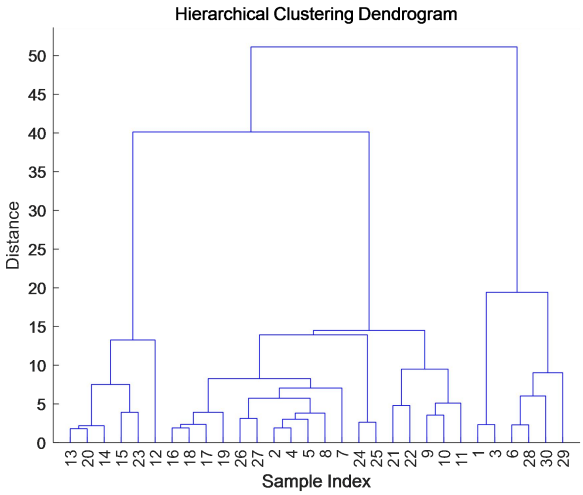



图 14　层次聚类树状图结果

但是在增加平衡聚类大小的操作之后基于层次聚类算法的准确度不高，person4 的准确率为 68%，优化模型之前为 60%（具体优化措施为增加特征提取和使用主成分分析对特征数据降维），而根据 4.2 内容可知，基于谱聚类的分类准确率优化后由 43%变为 72%。因此，进行特征维度和特征数据降维的优化措施对于谱聚类模型的改进效果明显，而对于层次聚类模型几乎没有影响，若要改进层次聚类模型则需从聚类算法的维度开展，对算法进行具体的优化。这也印证了谱聚类算法在处理非线性可分数据和高维数据方面具有优势，故最终确定选用谱聚类算法求解该模型。

# 5 问题二

## 5.1 问题分析与模型建立

### 5.1.1 分析与求解思路

根据问题 2 的描述，我们首先需要建立一个人体活动判别模型，利用附件 2 的 10 名实验人员的活动数据进行训练和测试，再将该训练好的判别模型与问题一的分类模型进行准确度的比较，最后利用该模型预测附件 3 中的人员活动类别。

### 5.1.2 模型建立

数据集的处理同问题一类似，首先将 1 名实验人员的活动数据进行特征值的计算，区别是需要添加标签列，根据实验人员活动的编号进行活动标签的添加，故特征集是一个 $60\times45$ 的 excel（除去首行的文字说明），其他实验人员的处理相同。

通过大量的资料和文献查询，我们选择随机森林作为分类器，随机森林算法是一种基于决策树的集成学习算法[12-13]，流程如图 15 所示，通过构建多个决策树并将它们的结果进行汇总（通常是投票或平均），来得到最终的预测结果，这样可提高模型的准确性和稳定性。同时这种方法抗过拟合能力强，且对异常值和噪声具有一定的鲁棒性，能够有效地处理非线性问题，并且擅长处理大量样本和特征。

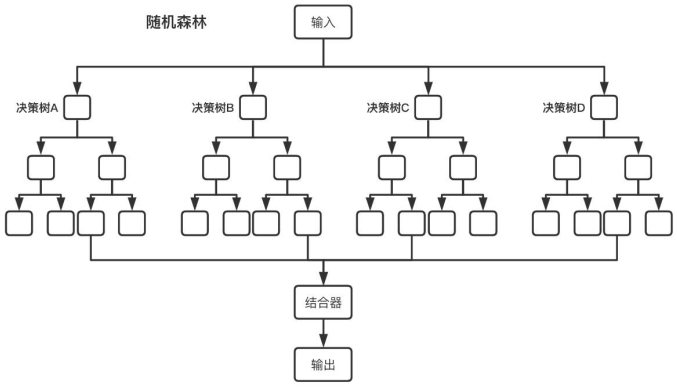



图 15　随机森林算法流程图

## 5.2 模型应用与分析

将处理好的 10 名实验人员的特征 excel 输入到随机森林分类器中，可以得到如图 1、图 2、图 3 所示的结果。由图 16(a)和图 16(b)可以看出，该判别模型对为向左走、步行下楼、跳跃、坐下、站立、躺下的活动区分度高达 100%，对向前走、向右走、步行上楼、向前跑的活动判别欠佳，向右走有被判别成向前走的风险，步行上楼可能被判别为向右走和向前跑，步行下楼会被误判为步行上楼，坐下会被判别成乘坐电梯向上移动，极其容易混淆乘坐电梯向上移动、乘坐电梯向下移动，对乘坐电梯向上移动的正确率只有 25%，最后得出总的准确率有 84.17%，模型整体的判别效果良好。

分类报告:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.90 | 1.00 | 0.95 | 9 |
| 2 | 1.00 | 1.00 | 1.00 | 10 |
| 3 | 0.93 | 0.93 | 0.93 | 14 |
| 4 | 0.93 | 0.87 | 0.90 | 15 |
| 5 | 1.00 | 0.89 | 0.94 | 9 |
| 6 | 0.88 | 1.00 | 0.93 | 7 |
| 7 | 1.00 | 1.00 | 1.00 | 10 |
| 8 | 1.00 | 0.83 | 0.91 | 6 |
| 9 | 1.00 | 1.00 | 1.00 | 8 |
| 10 | 1.00 | 1.00 | 1.00 | 9 |
| 11 | 0.25 | 0.57 | 0.35 | 7 |
| 12 | 0.62 | 0.31 | 0.42 | 16 |
| accuracy | | | 0.84 | 120 |
| macro avg | 0.88 | 0.87 | 0.86 | 120 |
| weighted avg | 0.87 | 0.84 | 0.85 | 120 |

准确率:
0.8416666666666667

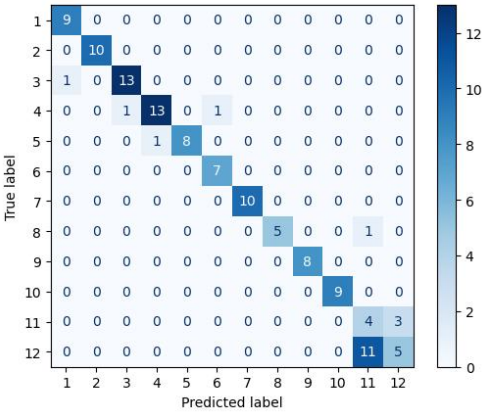



图 16　（a）分类报告　　（b）混淆矩阵

图 17 是一个各个特征值的重要性排序，通过该图可以看出加速度均值 _x 、角速度均值_x 、加速度四分位差_x 对以随机森林作为分类器的判别模型影响更大，对后续工作有指导意义。

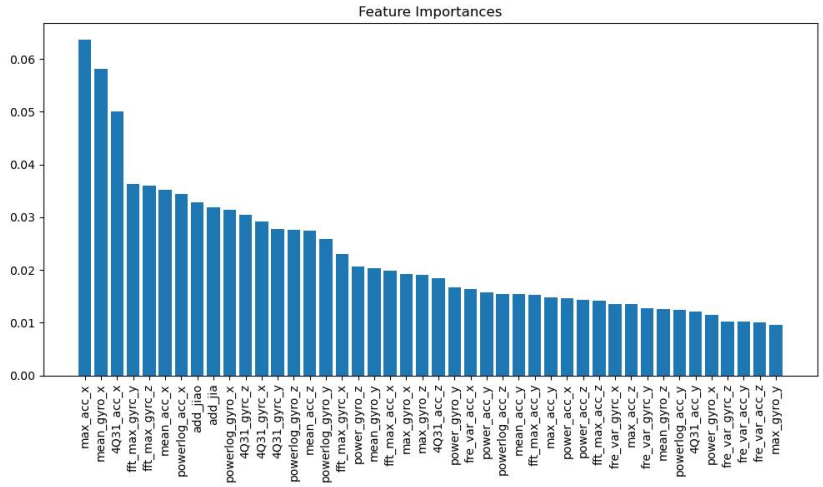



图 17　各特征的重要性分析

## 5.3 模型改进

LSTM（长短期记忆网络）与随机森林的优势对比：LSTM 在处理时间序列数据和序列化数据方面具有显著优势，能够捕捉数据中的时间依赖关系；LSTM 能够自动从数据中学习特征，特别是在处理高维和复杂时序数据时，而随机森林则需要手工设计和提取特征。LSTM 具有强大的非线性建模能力，能够处理数据中的复杂非线性关系。而随机森林虽然也能处理非线性关系，但其主要依赖于大量的决策树来进行集成学习，效果可能不如 LSTM 显著。

支持向量机（SVM）相比随机森林的优势有以下优势：LSTM 在处理时间序列数据和序列化数据方面具有显著优势，能够捕捉数据中的时间依赖关系。而随机森林在时间序列数据处理上通常需要额外的特征工程，并且不具备捕捉长时间依赖关系的能力；LSTM 能够自动从数据中学习特征，特别是在处理高维和复杂时序数据时，而随机森林则需要手工设计和提取特征；LSTM 具有强大的非线性建模能力，能够处理数据中的复杂非线性关系。而随机森林虽然也能处理非线性关系，但其主要依赖于大量的决策树来进行集成学习，效果可能不如 LSTM 显著。

实验中我们对比了单个 LSTM 和单个 SVM 的分类效果，具体如图 18 和图 19 所示，发现单个 LSTM 比单个 SVM 的准确率更高。

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.89 | 0.94 | 9 |
| 1 | 1.00 | 1.00 | 1.00 | 10 |
| 2 | 0.86 | 0.86 | 0.86 | 14 |
| 3 | 0.88 | 0.93 | 0.90 | 15 |
| 4 | 1.00 | 0.89 | 0.94 | 9 |
| 5 | 0.88 | 1.00 | 0.93 | 7 |
| 6 | 1.00 | 1.00 | 1.00 | 10 |
| 7 | 1.00 | 0.83 | 0.91 | 6 |
| 8 | 1.00 | 1.00 | 1.00 | 8 |
| 9 | 1.00 | 1.00 | 1.00 | 9 |
| 10 | 0.31 | 0.71 | 0.43 | 7 |
| 11 | 0.75 | 0.38 | 0.50 | 16 |
| accuracy | | | 0.85 | 120 |
| macro avg | 0.89 | 0.87 | 0.87 | 120 |
| weighted avg | 0.89 | 0.85 | 0.85 | 120 |

准确率:
0.85

（a）LSTM 分类报告

分类报告:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 1.00 | 1.00 | 1.00 | 9 |
| 2 | 0.91 | 1.00 | 0.95 | 10 |
| 3 | 1.00 | 1.00 | 1.00 | 14 |
| 4 | 0.93 | 0.93 | 0.93 | 15 |
| 5 | 1.00 | 0.89 | 0.94 | 9 |
| 6 | 1.00 | 1.00 | 1.00 | 7 |
| 7 | 1.00 | 1.00 | 1.00 | 10 |
| 8 | 1.00 | 0.83 | 0.91 | 6 |
| 9 | 1.00 | 1.00 | 1.00 | 8 |
| 10 | 1.00 | 1.00 | 1.00 | 9 |
| 11 | 0.21 | 0.57 | 0.31 | 7 |
| 12 | 0.40 | 0.12 | 0.19 | 16 |
| accuracy | | | 0.83 | 120 |
| macro avg | 0.87 | 0.86 | 0.85 | 120 |
| weighted avg | 0.86 | 0.83 | 0.83 | 120 |

准确率:
0.8333333333333334

（b）SVM 分类报告

图 18

通过大量查找资料和文献，发现文献[12]中使用随机森林与 SVM 构成双层分类器，最后识别的准确率高达 90%，受此启发，我们使用随机森林与 LSTM 构成双层分类器，经过反复试验，对参数进行调整优化，最终识别的准确率 85.83%，对比单个的随机森林有所提高。由图 19 可以看出，LSTM 在训练过程中的准确度还是很高的，都在 90%以上，但是测试的效果不尽如人意，有可能是测试数据过少的缘故，而且，由图 20 可以发现，在向前走、向左走、步行上楼、跳跃、坐下、站立、躺下上的判别正确率达到 100%，对向右走、步行下楼、向前跑分辨欠佳，对坐电梯向上移动、乘坐电梯向下移动的分辨率依旧很差，总体虽然有所提高，但是提高不大，猜测可能是受到特征值的限制。

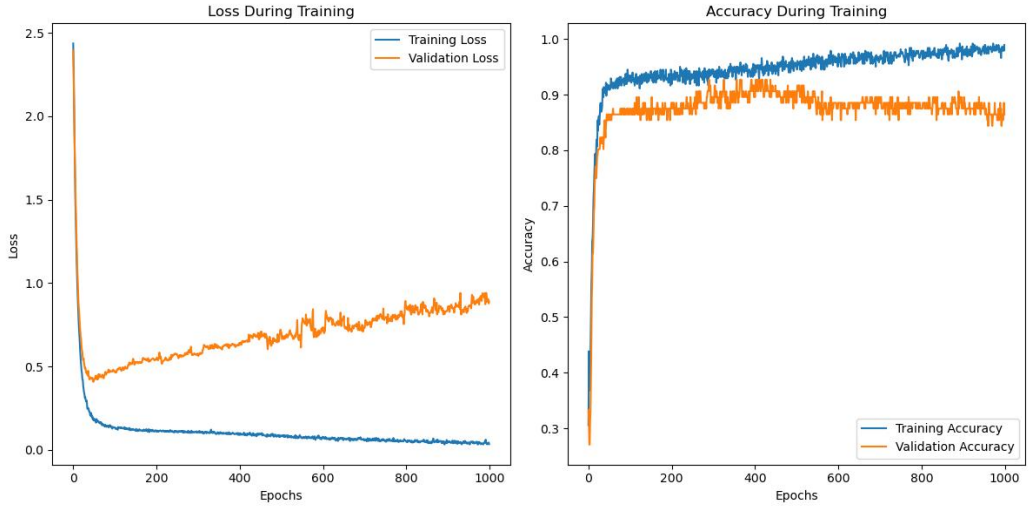



图 19　LSTM 训练过程数据

分类报告:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 1.00 | 1.00 | 1.00 | 9 |
| 2 | 1.00 | 1.00 | 1.00 | 10 |
| 3 | 0.93 | 1.00 | 0.97 | 14 |
| 4 | 1.00 | 0.87 | 0.93 | 15 |
| 5 | 0.89 | 0.89 | 0.89 | 9 |
| 6 | 0.88 | 1.00 | 0.93 | 7 |
| 7 | 1.00 | 1.00 | 1.00 | 10 |
| 8 | 1.00 | 1.00 | 1.00 | 6 |
| 9 | 1.00 | 1.00 | 1.00 | 8 |
| 10 | 1.00 | 1.00 | 1.00 | 9 |
| 11 | 0.29 | 0.71 | 0.42 | 7 |
| 12 | 0.67 | 0.25 | 0.36 | 16 |
| accuracy | | | 0.86 | 120 |
| macro avg | 0.89 | 0.89 | 0.87 | 120 |
| weighted avg | 0.89 | 0.86 | 0.86 | 120 |

准确率:
0.8583333333333333

图 20　双层判别模型分类报告

表 5 是问题 1 的基于谱聚类算法的分类模型对附件 2 的 10 名实验人员数据进行分类的结果准确率。

表 5　分类模型准确度结果

| 实验人员 | 准确率 |
|---|---|
| Person4 | 72% |
| Person5 | 70% |
| Person6 | 60% |
| Person7 | 78% |
| Person8 | 83% |
| Person9 | 90% |
| Person10 | 83% |
| Person11 | 80% |
| Person12 | 88% |
| Person13 | 95% |
| Mean（10） | 79.9% |

将 10 组数据的准确率求平均作为分类模型整体的准确率，其结果是 79.9%。同时结果表明，该分类模型缺乏稳定性，其准确率随着数据的不同而具有差异，最低时为 60%，最高为 95%（效果如下图所示），也就是说该模型的效果对数据有一定程度的依赖，不具有较好的鲁棒性。

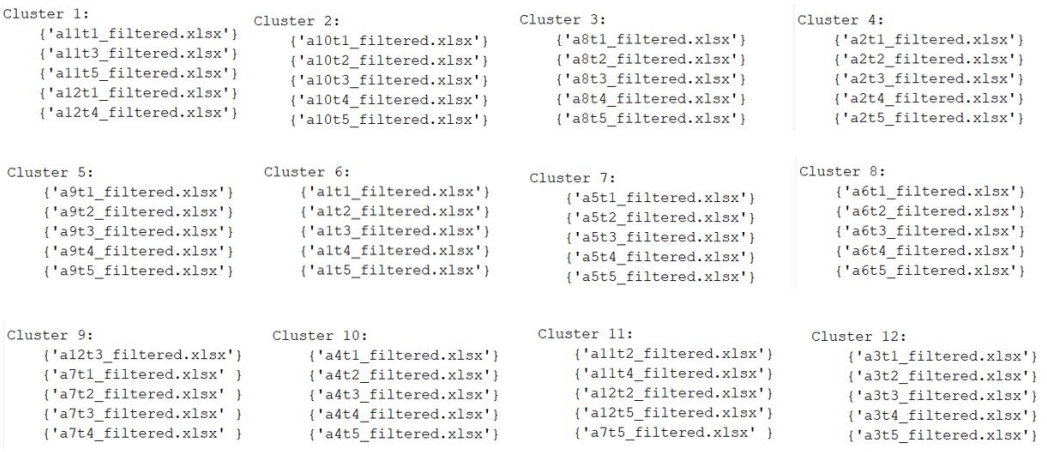



图 21　Person13 分类模型结果

综合对比问题一的分类模型和问题二提出的两种判别模型，从准确率的角度出发，提出的两种判别模型判别准确率分别是84.17%和 85.83%，明显由于问题一的分类模型；从稳定性的角度可以看出，问题一的分类模型对数据有一定程度的依赖，不具有较好的鲁棒性，而我们的双层判别模型在随机森林的基础上加入了 LSTM，能够自动从数据中学习特征，特别是在处理高维和复杂时序数据时，对稳定性有了巨大提升。

## 5.4 数据集的预测

对附件 3 的数据集进行同问题一的特征提取处理，使用改进后的双层判别模型对附件 3 的数据进行判别分类，分类结果如表 2 所示。

表 6 问题 2 结果

| 活动类型 | 判别状态 |
|---|---|
| SY1 | 5 |
| SY2 | 1 |
| SY3 | 7 |
| SY14 | 11 |
| SY5 | 7 |
| SY6 | 10 |
| SY7 | 2 |
| SY8 | 6 |
| SY9 | 7 |
| SY10 | 10 |
| SY11 | 9 |
| SY12 | 7 |
| SY13 | 3 |
| SY14 | 3 |
| SY15 | 4 |

| | |
|---|---|
| SY16 | 1 |
| SY17 | 4 |
| SY18 | 5 |
| SY19 | 8 |
| SY20 | 8 |
| SY21 | 6 |
| SY22 | 2 |
| SY23 | 8 |
| SY24 | 5 |
| SY25 | 2 |
| SY26 | 9 |
| SY27 | 8 |
| SY28 | 5 |
| SY29 | 6 |
| SY30 | 5 |

# 6 问题三

## 6.1 对不同实验人员活动数据差异性分析

### 6.1.1 特征提取与灰色关联分析配置

对问题三：分析附件 4 给出的问题 1 和问题 2 中参与实验的 13 位实验人员的年龄、身高、体重等数据，探究不同人员的同一活动状态是否存在差异？考虑到在问题一的前三位实验人员的分类结果和判别结果对本小问的分析存在影响，故而我们选择了后十位的数据进行特征分析，特征提取方法同问题一和问题二，特征处理后得到 12 组活动数据，每组活动数据为$60\times44$ excel 表格（除去第一行的文字说明），行数代表了十位实验人员的对同个活动的 5 个样本数据，列数代表了特征量。由于对同一实验人员的同一活动数据具有共性，则对 5 个样本数据进行均值处理获得 12 类活动的$10\times44$ 的 excel 表格组，在 matlab 中编写程序对差异性进行可视化分析，结果如图 1 所示。

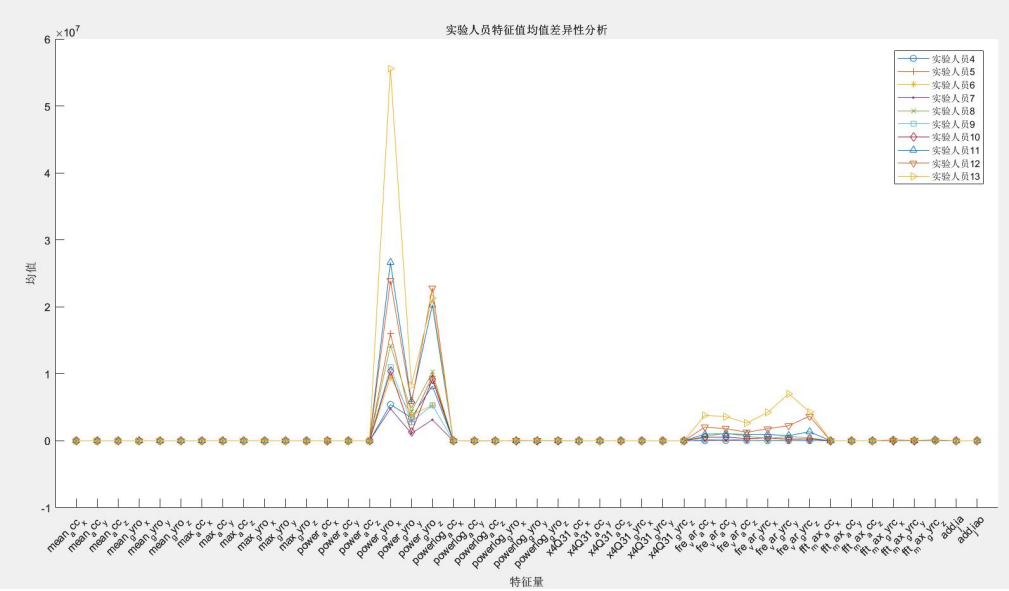



图 22　不同实验人员的特征值均值差异性图

对差异性结果进行分析可得在角速度信号能率_x、角速度信号能率_y、角速度信号能率_z，加速度频率方差_x、加速度频率方差_y、加速度频率方差_z、角速度频率方差_x、角速度频率方差_y、角速度频率方差_z 这 9 类特征值上具有明显差异性。

为了进一步对这种差异性进行定量分析，我们考虑使用灰色关联分析的方法。灰色关联分析是灰色系统理论中十分活跃的一个分支，其基本思想是根据序列曲线几何形状的相似程度来判断不同序列之间的联系是否紧密。具体来说，它通过确定参考数据列和若干个比较数据列的集合形状相似程度来判断其联系是否紧密，反映了曲线间的关联程度。在系统发展过程中，若两个因素变化的趋势具有一致性，即同步变化程度较高，则两者关联程度较高；反之则较低。灰色关联分析法是根据因素之间发展趋势的相似或者相异程度，即灰色关联度，作为衡量因素间关联程度的一种方法。灰色关联分析的对象往往是一个系统，系统的发展会受到多个因素的影响。该方法通常用于分析复杂系统或过程，其中因素之间的关联关系不是完全清晰或无法用精确的数学模型描述。它旨在通过定量描述和比较系统发展变化的态势，找出影响系统行为的主要因素和次要因素，以及它们对系统行为的影响程度和方向。我们在 SPSSPRO 平台的综合性分析窗口对 12 组数据进行分析，配置参数如下：

表 7　灰色关联算法配置参数

| 算法 | 灰色关联分析 |
|---|---|
| 特征序列变量 | 实验人员 5~13 |
| 母序列 | 实验人员 4 |
| 参数 | 无量纲处理方式：{均值化} |
| 分辨系数 | 0.5 |

然后针对数据进行无量纲化处理（均值化、初值化）。同时求解母序列（对比序列）和特征序列之间的灰色关联系数值。然后对灰色关联度值进行排序，得出结论。经分析，12 类活动数据的灰色关联分析相似，下面对编号为 1 的活动 2 数据进行分析，其余 11 类活动数据分析结果可在附件中查看。

### 6.1.2 灰色相关系数结果分析

输出结果 1：灰色关联系数

| 索引项 | 实验人员5 | 实验人员6 | 实验人员7 | 实验人员8 | 实验人员9 | 实验人员10 | 实验人员11 | 实验人员12 | 实验人员13 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.99999984 | 1 | 0.9999997 | 0.9999998 | 1 | 0.9999999 | 0.9999997 | 0.99999971 | 0.99999967 |
| 2 | 0.99999996 | 1 | 1 | 1 | 0.9999999 | 0.9999999 | 1 | 0.99999993 | 0.99999998 |
| 3 | 0.99999996 | 1 | 0.9999999 | 0.9999999 | 0.9999999 | 1 | 0.9999999 | 0.99999994 | 0.99999994 |
| 4 | 0.99999992 | 0.9999993 | 0.999997 | 0.9999995 | 0.999999 | 0.9999996 | 0.9999997 | 0.99999983 | 0.99999995 |
| 5 | 0.99999937 | 0.9999983 | 0.9999999 | 0.9999991 | 0.9999994 | 0.9999982 | 0.999999 | 0.99999895 | 0.99999915 |
| 6 | 0.9999996 | 0.9999999 | 0.9999991 | 0.9999998 | 0.9999998 | 0.9999993 | 0.9999999 | 0.99999981 | 0.99999979 |
| 7 | 0.99999975 | 1 | 0.9999994 | 0.9999997 | 0.9999999 | 0.9999999 | 0.9999994 | 0.99999947 | 0.99999932 |
| 8 | 0.99999985 | 0.9999999 | 0.9999998 | 0.9999997 | 0.9999999 | 0.9999999 | 0.9999997 | 0.9999997 | 0.99999966 |
| 9 | 0.99999994 | 0.9999999 | 0.9999998 | 0.9999997 | 0.9999998 | 0.9999998 | 0.9999996 | 0.99999968 | 0.99999963 |
| 10 | 0.99999681 | 0.9999868 | 0.9999774 | 0.9999534 | 0.9999923 | 0.99998 | 0.9999523 | 0.99995984 | 0.99995553 |
| 11 | 0.99997731 | 0.9999823 | 0.9999951 | 0.9999642 | 0.9999842 | 0.9999637 | 0.999956 | 0.99996517 | 0.99996246 |
| 12 | 0.99997627 | 0.9999892 | 0.9999896 | 0.9999459 | 0.9999825 | 0.9999662 | 0.9999447 | 0.99996663 | 0.99994015 |
| 13 | 0.99968762 | 0.9998663 | 0.9989626 | 0.998976 | 0.9999058 | 0.9995112 | 0.9998325 | 0.9999785 | 0.99978029 |
| 14 | 0.99997142 | 0.9998985 | 0.9999882 | 0.9999677 | 0.9999887 | 0.9998231 | 0.9999755 | 0.99989076 | 0.99996224 |
| 15 | 0.99999537 | 0.9999877 | 0.9999258 | 0.9999969 | 0.9999651 | 0.9999413 | 0.9999843 | 0.99999724 | 0.99997208 |
| 16 | 0.36717161 | 0.4236244 | 0.4170227 | 0.5380144 | 0.3678794 | 0.5064411 | 0.4984359 | 0.71344361 | 0.42137589 |
| 17 | 0.57614934 | 0.9800112 | 0.6277493 | 0.6793009 | 0.7142044 | 0.4918754 | 0.5845042 | 0.55023475 | 0.52854154 |
| 18 | 0.50193299 | 0.4220478 | 0.498671 | 0.4656249 | 0.404297 | 0.6429554 | 0.5223235 | 0.53681378 | 0.33333333 |
| 19 | 0.99997699 | 0.9999589 | 0.9999663 | 0.9999466 | 0.9999874 | 0.9998646 | 0.9999897 | 0.99995425 | 0.9999702 |
| 20 | 0.99937595 | 0.9995095 | 0.9975483 | 0.9977834 | 0.9999093 | 0.9998063 | 0.9994426 | 0.99937141 | 0.99993769 |
| 21 | 0.99945015 | 0.9999052 | 0.9982391 | 0.998119 | 0.9996199 | 0.9990714 | 0.9997917 | 0.99985238 | 0.99965154 |
| 22 | 0.99926401 | 0.9999076 | 0.9972605 | 0.997777 | 0.999643 | 0.9985336 | 0.9995064 | 0.9999376 | 0.99960043 |
| 23 | 0.99895854 | 0.9998619 | 0.9984589 | 0.9985019 | 0.9998863 | 0.9996673 | 0.99997 | 0.99956067 | 0.99917042 |
| 24 | 0.99891414 | 0.9996386 | 0.997851 | 0.9984317 | 0.9997563 | 0.9983467 | 0.9999256 | 0.99970806 | 0.99913608 |
| 25 | 0.9999999 | 1 | 0.9999999 | 0.9999999 | 0.9999999 | 0.9999999 | 0.9999998 | 0.99999982 | 0.9999998 |
| 26 | 0.99999993 | 1 | 1 | 0.9999999 | 0.9999999 | 0.9999999 | 0.9999999 | 0.99999988 | 0.99999986 |
| 27 | 0.99999996 | 1 | 0.9999999 | 0.9999999 | 1 | 0.9999999 | 0.9999999 | 0.99999991 | 0.9999999 |
| 28 | 0.99999453 | 0.9999952 | 0.9999787 | 0.9999835 | 0.9999976 | 0.9999847 | 0.9999782 | 0.99997731 | 0.99997477 |
| 29 | 0.99998882 | 0.9999981 | 0.9999982 | 0.999985 | 0.9999915 | 0.9999842 | 0.9999816 | 0.9999809 | 0.99997976 |
| 30 | 0.99998486 | 0.9999887 | 0.9999946 | 0.9999771 | 0.999985 | 0.9999859 | 0.9999724 | 0.99997191 | 0.99996965 |
| 31 | 0.99648462 | 0.9951072 | 0.9730997 | 0.8877239 | 0.961854 | 0.8728748 | 0.9034026 | 0.82571871 | 0.81056141 |
| 32 | 0.99711192 | 0.9977025 | 0.9539166 | 0.8178901 | 0.9314205 | 0.87031 | 0.9062263 | 0.84838072 | 0.82587762 |
| 33 | 0.99464086 | 0.9914308 | 0.9651107 | 0.8884654 | 0.9858268 | 0.9256284 | 0.9168794 | 0.89594075 | 0.87140931 |
| 34 | 0.99769982 | 0.9886927 | 0.9941589 | 0.9497913 | 0.9987891 | 0.8956104 | 0.9123539 | 0.84891692 | 0.79648903 |
| 35 | 0.99161317 | 0.9925067 | 0.9826783 | 0.8857009 | 0.9953235 | 0.9467078 | 0.9302649 | 0.81169279 | 0.69535815 |
| 36 | 0.99686199 | 0.9892899 | 0.9782869 | 0.9266939 | 0.9853089 | 0.9374828 | 0.8708737 | 0.72033361 | 0.79188459 |
| 37 | 0.99969601 | 0.9999165 | 0.9990091 | 0.9991482 | 0.9999235 | 0.9994394 | 0.9998139 | 0.99998519 | 0.99984508 |
| 38 | 0.99995722 | 0.9998053 | 0.9999692 | 0.9999881 | 0.9998981 | 0.9994683 | 0.9999532 | 0.99973144 | 0.99998604 |
| 39 | 0.99992001 | 0.9999387 | 0.9998308 | 0.9999084 | 0.9998653 | 0.999821 | 0.9999528 | 0.99996405 | 0.99999101 |
| 40 | 0.99758142 | 0.9931117 | 0.9784704 | 0.9853808 | 0.9923541 | 0.9952084 | 0.9985724 | 0.99052592 | 0.9911217 |
| 41 | 0.99395827 | 0.9993412 | 0.9970576 | 0.9984891 | 0.9996998 | 0.9937958 | 0.9938604 | 0.99213327 | 0.9902583 |
| 42 | 0.99531675 | 0.9949493 | 0.9871023 | 0.9936837 | 0.9979342 | 0.9941583 | 0.9964877 | 0.99211517 | 0.9914575 |
| 43 | 0.99999983 | 1 | 0.9999997 | 0.9999998 | 0.9999999 | 0.9999999 | 0.9999997 | 0.99999969 | 0.99999964 |
| 44 | 0.99998961 | 0.9999982 | 0.9999908 | 0.99998 | 0.9999941 | 0.9999849 | 0.999976 | 0.99997523 | 0.99997276 |

图 23　实验人员 5~13 与实验人员 4 之间的灰色关联系数

从上表可知，针对 9 个评价项（实验人员 5-13）以及 44 项数据进行灰色关联度分析，并且以实验人员 4 作为“参考值”（母序列），研究 9 个评价项（实验人员 5-13）与实验人员 4 的关联关系（关联度），并基于关联度提供分析参考，使用灰色关联度分析时，分辨系数取 0.5，结合关联系数计算公式计算出关联系数值，并根据关联系数值，然后计算出关联度值用于评价判断。分辨系数 $\rho$ 越小，分辨力越大，一般 $\rho$ 的取值区间

为（0，1），具体取值可视情况而定。当 $\rho \le 0.5463$ 时，分辨力最好，通常取 $\rho = 0.5$ 。

输出结果 2：关联系数图 ；关联系数代表着该子序列实验人员 5-13 对与母序列对应维度上的关联程度值（数字越大，代表关联性越强）。


图 24　不同实验人员的特征值均值差异性图

经对 12 类活动数据分析，不同实验人员仅在角速度信号能率_x、角速度信号能率_y、角速度信号能率_z、加速度频率方差_x、加速度频率方差_y、加速度频率方差_z、角速度频率方差_x、角速度频率方差_y、角速度频率方差_z 九类特征值上具有明显的差异性，在其他特征值上差异不明显，也就是说在对不同实验人员进行分类时，这九类特征对分类准确率贡献度较大。输出结果如表 8 所示的灰色关联度排序表。

表 8　关联度排序表

| 评价项 | 关联度 | 排名 |
|---|---|---|
| 实验人员 6 | 0.972 | 1 |
| 实验人员 5 | 0.964 | 2 |
| 实验人员 7 | 0.962 | 3 |
| 实验人员 9 | 0.962 | 4 |
| 实验人员 10 | 0.956 | 5 |
| 实验人员 11 | 0.955 | 6 |
| 实验人员 8 | 0.955 | 7 |
| 实验人员 12 | 0.948 | 8 |
| 实验人员 13 | 0.933 | 9 |

结合上述关联系数结果进行加权处理，最终得出关联度值，使用关联度值对 9 个评价对象进行评价排序；关联度值介于 0~1 之间，该值越大代表其与“参考值”（母序列）之间的相关性越强，也即意味着其评价越高。从上表可以看出：针对本次 9 个评价项，实验人员 6 评价最高（关联度为：0.972），其次是实验人员 5（关联度为：0.964）。

## 6.2 活动数据对年龄、身高、体重关联性分析

对问题三：分析附件 4 给出的问题 1 和问题 2 中参与实验的 13 位实验人员的年龄、身高、体重等数据，探究活动状态数据与实验人员的年龄、身高、体重有无关系？数据处理流程同 3.1 节，在特征提取后增加标签为 year、weight、height 的三类属性数据，然后在 SPSS 软件内利用斯皮尔曼（spearman）相关系数分析不同列的相关性。斯皮尔曼等级相关是根据等级资料研究两个变量间相关关系的方法。它是依据两列成对等级的各对等级数之差来进行计算的，所以又称为“等级差数法” 斯皮尔曼等级相关对数据条件的要求没有积差相关系数严格，只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级资料，不论两个变量的总体分布形态、样本容量的大小如何，都可以用斯皮尔曼等级相关来进行研究。这使得它在处理样本量较小的情况下，具有其他相关系数估计所不具备的高可靠性优势。导入 12 类活动数据后，其数据分析结果见附件。

通过对编号为 1（即“向前走”）的数据结果进行分析可见，针对 year 标签，加速度均值_x、加速度均值_z、角速度均值_z、角速度四分位差_x、角速度四分位差_y、合角速度特征值与实验人员的年龄呈现较强负相关性；加速度信号能率_y、加速度信号能率_z、加速度频域峰值_z 特征值呈现弱正相关性；其他特征值相关性极弱，这里不再详述。

针对 weight 标签，加速度能量对数_y、加速度能量对数_z、加速度四分位差_y、角速度四分位差_z 特征值与实验人员的体重呈现较强负相关性；加速度均值_x 特征值呈现强正相关性；加速度信号能率_x、加速度频率方差_x、加速度频率方差_y、加速度频率方差_z、加速度频域峰值_x、角速度频域峰值_x、角速度频域峰值_z 特征值呈现中相关性；角速度信号能率_x、角速度能量对数_y、角速度能量对数_z、角速度频域方差_y、角速度频域峰值_y 与体重呈现弱相关性。其他特征值相关性极弱，这里不再详述。

针对 height 标签，加速度能量对数_y、加速度能量对数_z、加速度四分位差_x、加速度四分位差_y、角速度四分位差_z 特征值与实验人员的身高呈现较强负相关性；加速度信号能率_x、角速度能量对数_x、角速度能量对数__y、加速度频率方差_x、加速度频率方差_y、加速度频率方差_z、角速度频率方差_y、加速度频域峰值_x、角速度频域峰值_z 特征值呈现中相关性；角速度频率方差_x、角速度频率方差_z、角速度频域峰值_x 与身高呈现弱相关性。其他特征值相关性极弱，这里不再详述。

其余 11 类活动数据结果见附件，分析原理及分析方法同活动 1。

## 6.3 对未知实验人员的活动数据个性化分类判别

### 6.3.1 数据集优化与分类判别模型

无论哪种类型的传感器，人类进行的活动都具有很强的主观性，这与年龄、性别、体重、身高、和生活方式等不同因素有关。在 3.1 节及 3.2 节中我们验证了活动数据与实验人员的个人属性具有极大的相关性。为了考虑这些主观因素，需要研究实验人员的个性化模型。

首先对特征集的优化部分，我们做出以下调整：对特征集的提取方法同 3.1.1 节，为满足对实验人员的个性化判别要求，增加了实验人员的编号和属性特征（即年龄、身高、体重）。

其次，对分类判别模型，由于随机森林具有有效处理非线性问题和擅长处理大量样本与特征的优势，在本节中继续采用随机森林作为个性化判别器。

### 6.3.2 活动数据个性化判别结果

经过个性化判别模型后的结果如图 1 和图 2 所示。由图 1 和 2 可知，该模型对编号为 12、13、6 的实验人员区分只有 70%多，判别结果欠佳，对其余实验人员的判别准确率都在 90%以上，特别对实验人员编号为 8、11 的判别准确率高达 100%总的判别准确率有 81.67%，模型效果良好。

分类报告:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 4 | 0.80 | 0.57 | 0.67 | 7 |
| 5 | 0.90 | 0.75 | 0.82 | 12 |
| 6 | 0.72 | 0.93 | 0.81 | 14 |
| 7 | 0.83 | 0.91 | 0.87 | 11 |
| 8 | 1.00 | 0.75 | 0.86 | 12 |
| 9 | 0.91 | 1.00 | 0.95 | 10 |
| 10 | 0.71 | 1.00 | 0.83 | 15 |
| 11 | 1.00 | 0.55 | 0.71 | 20 |
| 12 | 0.73 | 0.89 | 0.80 | 9 |
| 13 | 0.75 | 0.90 | 0.82 | 10 |
| accuracy | | | 0.82 | 120 |
| macro avg | 0.84 | 0.82 | 0.81 | 120 |
| weighted avg | 0.85 | 0.82 | 0.81 | 120 |

准确率:
0.8166666666666667

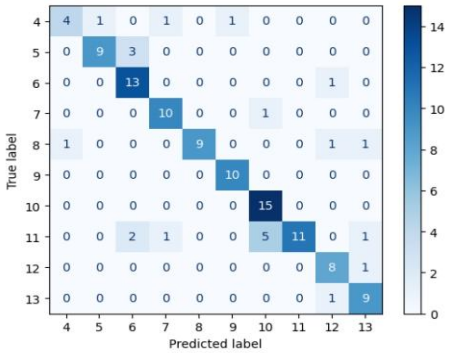



图 25 （a）个性化判别模型分类报告 （b）混淆矩阵

最后使用该个性化判别模型对附件 5 的未知实验人员编号进行判别预测，结果如表 9 所示。

表 9 判别结果

| 活动类型 | 判别结果 |
|---|---|
| Unkonw1 | 10 |
| Unknow2 | 7 |
| Unknow3 | 6 |
| Unknow4 | 9 |
| Unknow5 | 13 |

# 7 总结

在本文中，我们针对各类问题采取了一系列综合方法，其创新之处主要体现在以下三个方面：首先，在数据预处理阶段，我们精心设计了滤波器，对原始数据执行了降噪和平滑处理，有效强化了低频信号的特征，并进行了降维处理。随后，我们对滤波后的数据进行了深入的多维度特征提取，成功捕捉到了关键的特征信息。这两个步骤为后续的人体活动识别与分析打下了坚实的基础。其次，针对第二个问题，我们提出了一个新的双层判别模型。通过反复试验和调参，我们提升了模型在整体及各个活动上的判别精度。最后，为了避免特征值的分布形态和样本容量对分析结果产生显著偏差，我们采用了斯皮尔曼相关系数来分析年龄、身高、体重等人体属性因子对活动特征数据的影响。

尽管我们的双层分类模型在准确率上已达到 85.83%，但在区分“乘坐电梯向上移动”与“乘坐电梯向下移动”这两个动作时仍存在一定的混淆。为了进一步提高模型性能，我们可以考虑引入新的有效特征、扩充测试样本量，或者探索其他先进的深度学习分类模型进行优化。

# 参考文献


[1] 周博翔. 基于加速度传感器的人体运动姿态识别[D]. 湖南:长沙理工大学,2014. DOI:10.7666/d.Y2756180.

[2] Ankita,Jain,Vivek,et al.Human Activity Classification in Smartphones Using Accelerometer and Gyroscope Sensors[J].IEEE Sensors Journal, 2017.DOI:10.1109/JSEN.2017.2782492.

[3] 郭志涛,曹小青,王宝珠,等.3D CNN 人体动作识别中的特征组合优选[J].河北工业大学学报, 2021, 50(1):7.DOI:10.14081/j.cnki.hgdxb.2021.01.006.

[4] 张良,钱毅敏.基于深度图像和骨骼信息的人体动作识别方法[J].中国民航大学学报, 2021.DOI:10.3969/j.issn.1674-5590.2021.02.011.

[5] 李辉;李瑞祥;张耀威;乐燕芬;施伟斌. 多层分类器模型的相似人体活动识别[J]. 小型微型计算机系统, 2021(第 4 期): 861－867.

[6] 林海波;李扬;张毅;罗元. 基于时序分析的人体运动模式的识别及应用[J]. 计算机应用与软件, 2014(第 12 期): 225－228

[7] 杨俊闯,赵超.K-Means 聚类算法研究综述[J].计算机工程与应用, 2019, 55(23):9.DOI:10.3778/j.issn.1002-8331.1908-0347.

[8] 蔡元萃,陈立潮.聚类算法研究综述[J].科技情报开发与经济, 2007, 17(1):145-146.DOI:10.3969/j.issn.1005-6033.2007.01.086.

[9] 吕纪荣,王士虎.数据挖掘中聚类算法研究综述[J].魅力中国, 2014(3):2.

[10] 白璐,赵鑫,孔钰婷,等.谱聚类算法研究综述[J].计算机工程与应用, 2021, 57(14):12.DOI:10.3778/j.issn.1002-8331.2103-0547.

[11] 姬强,孙艳丰,胡永利,等.深度聚类算法研究综述[J].北京工业大学学报, 2021, 47(8):13.DOI:10.11936/bjutxb2021010013.

[12] 李辉,李瑞祥,张耀威,等.多层分类器模型的相似人体活动识别[J].小型微型计算机系统, 2021.DOI:10.3969/j.issn.1000-1220.2021.04.032

[13] 王肇宇.层级随机森林算法及其在人体活动识别应用研究[D].电子科技大学[2024-07-12].DOI:CNKI:CDMD:2.1014.137907.

[14] Ferrari A , Micucci D , Mobilio M ,et al.On the Personalization of Classification Models for Human Activity Recognition[J].IEEE Access, 2020, PP(99):1-1.DOI:10.1109/ACCESS.2020.2973425.

[15] 殷晓玲,夏启寿,陈晓江,等.基于智能手机感知的人体运动状态深度识别[J].北京邮电大学学报, 2019(3):8.DOI:10.13190/j.jbupt.2018-221.

[16] 郭毅博,孟文化,范一鸣,等.基于可穿戴传感器数据的人体行为识别数据特征提取方法[J]. 计算机辅助设计与图形学学报, 2021.DOI:10.3724/SP.J.1089.2021.18690.

# 附　录

## 附录 A：支撑材料列表

所有支撑材料文件包括主要代码和数据均在附件文件夹中，其中的内容如下表所示。

支撑材料列表

| 序号 | 文件名 | 材料说明 |
|---|---|---|
| 1 | 1_问题 1 代码 | 分类模型代码 |
| 2 | 2_特征提取数据结果 | 支撑建模的处理数据 |
| 3 | 3_filter_person | 数据预处理（滤波）代码 |
| 4 | 4_灰色关联系数 | |
| 5 | contest.ipynb | 判决模型代码 |
| 6 | 附件 2 特征集 | 用于模型训练及识别预测的数据集 |
| 7 | 附件 4 特征集 | |
| 8 | 未知标签实验人员特征集合 | |

## 附录 B：关键数据 特征提取数据结果

该文件包含对所有实验人员的样本数据进行特征提取的处理结果，每个人的结果对应一个 Excel 表格，每个表格由 60×44 的数据构成，60 即代表所有活动状态（12×5），44 是所有特征值维度，分别是加速度均值_x、加速度均值_y、加速度均值_z、角速度均值_x、角速度均值_y、角速度均值_z、加速度峰值_x、加速度峰值_y、加速度峰值_z、角速度峰值_x、角速度峰值_y、角速度峰值_z、加速度信号能率_x、加速度信号能率_y、加速度信号能率_z、角速度信号能率_x、角速度信号能率_x、角速度信号能率_x、加速度能量对数_x、加速度能量对数_y、加速度能量对数_z、角速度能量对数_x、角速度能量对数_y、角速度能量对数_z、加速度四分位差_x、加速度四分位差_y、加速度四分位差_z、角速度四分位差_x、角速度四分位差_y、角速度四分位差_z、加速度频率方差_x、加速度频率方差_y、加速度频率方差_z、角速度频率方差_x、角速度频率方差_y、角速度频率方差_z、加速度频域峰值_x、加速度频域峰值_y、加速度频域峰值_z、角速度频域峰值_x、角速度频域峰值_y、角速度频域峰值_z、合加速度、合角速度。

## 附录 C：主要程序/关键代码

| | |
|---|---|
| 代码环境 | 操作系统: Windows (Version 11)<br>编程语言: Python 3.7.1 (Anaconda Navigator 1.9.2)　Matlab2022a<br>编辑器: PyCharm 2018.3.2 (Professional Edition)<br>代码详见: contest.ipynb　SPecl_class*.m(0 1 2)　H_class*.m(0 1 2) |

**代码清单 1 分类模型优化**

```
% 标准化特征数据
features = zscore(features);

% 使用 PCA 进行降维
[coeff, score, ~] = pca(features);
reducedFeatures = score(:, 1:10); % 选择前 10 个主成分

% 设置聚类数量
numClusters = 12;

% 使用谱聚类
idx = spectralcluster(reducedFeatures, numClusters);

% 平衡每个聚类的大小，确保每个类别中有 5 组数据
clusterCounts = histcounts(idx, numClusters); targetCount = 5;
for i = 1:numClusters
    if clusterCounts(i) > targetCount % 随机选择多余的数据点
        extraIndices = find(idx == i);
        extraCount = clusterCounts(i) - targetCount;
        removeIndices = randsample(extraIndices, extraCount);
        idx(removeIndices) = 0; % 将多余的数据点标记为未分配
    end
end
% 重新分配未分配的数据点
unassignedIndices = find(idx == 0);
for i = 1:length(unassignedIndices)
    % 找到当前最小的聚类
    [~, minCluster] = min(clusterCounts);
    idx(unassignedIndices(i)) = minCluster;
    clusterCounts(minCluster) = clusterCounts(minCluster) + 1;
end
% 显示聚类结果
for i = 1:numClusters fprintf('Cluster %d:\n', i);
    clusterFiles = {files(idx == i).name};
    disp(clusterFiles');
end
% 绘制聚类结果
```

```
figure;
gscatter(reducedFeatures(:, 1), reducedFeatures(:, 2), idx);
xlabel('Principal Component 1');
ylabel('Principal Component 2');
title('Spectral Clustering of Activity Data');
legend('Location', 'best');
    grid on;
```

**代码清单 2 问题二双层分类器的权重部分设置**

```
    lstm_pred = lstm_model.predict(X_test_lstm)

    rf_clf = RandomForestClassifier(n_estimators=100, random_state=42)
    rf_clf.fit(X_train, y_train)

    rf_pred = rf_clf.predict_proba(X_test)

    # 设置权重
    rf_weight = 0.8
    lstm_weight = 0.2

    avg_pred = (rf_weight * rf_pred) + (lstm_weight * lstm_pred)
```

**代码清单 3 问题三个性化分类模型的算法核心部分**

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)
```