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）。

我们参赛选择的题号是（从组委会提供的赛题中选择一项填写）：A 题

我们的参赛编号（请填写完整参赛编号）：202418001016

所属学校（请填写完整的全名）：国防科技大学

参赛队员（打印后签名）：1. 

2. 

3. 

指导教师或指导教师组负责人（打印后签名）：

日期：2024 年 7 月 12 日

（请勿改动此页内容和格式。以上内容请仔细核对，如填写错误，论文可能被取消评奖资格。）

第九届湖南省研究生数学建模竞赛

题　目：智能手机传感器数据的分类与个体识别问题

摘要：

问题 1：针对智能手机传感器数据的无标签分类问题，提出了一种基于特征提取与聚类分析的方法。首先，通过对 3 名实验人员的 60 组加速度计和陀螺仪数据进行特征提取，构建特征向量，并使用 K-means 聚类算法将这些特征向量聚类为 12 个簇。为了评估聚类效果，本文使用了轮廓系数、卡尔金指数和戴维森堡丁指数，同时使用 t-SNE 进行聚类结果的可视化分析。结果表明，该方法能够较好地将不同活动状态的数据进行分类，但仍存在一定的改进空间。通过本文的研究，验证了特征提取和聚类分析在无标签数据分类中的有效性。

问题 2：针对问题 2，建立了有效的人员活动状态判别模型，并分析和比较了聚类算法和随机森林在不同运动状态下的分类性能。首先，通过提取附件 2 中 12 类活动数据的典型特征，构建基于随机森林的判别模型。然后，比较随机森林模型与问题 1 中的无标签分类模型在附件 2 数据中的分类准确率。最后，应用判别模型对附件 3 中的实验人员活动数据进行分类预测。数据增强方面，我们通过添加高斯噪声的方法生成多个增强数据样本，以提高模型的泛化能力。在特征提取过程中，分别从时间域和频域提取了多种统计特征，并计算了加速度和陀螺仪的模量和抖动特征。模型评估使用了准确率、精确率、召回率和 F1 分数等指标，并通过混淆矩阵和匈牙利算法优化模型的匹配关系。结果表明，随机森林模型在整体准确率和各类别准确率上均显著优于无标签分类模型，其中随机森林的整体准确率达到 0.96，而无标签分类模型最优匹配后的准确率为 0.56。基于随机森林模型的分类结果显示，模型在处理数据偏态和噪声方面具有更强的鲁棒性和一致性。此外，对附件 3 的数据分类预测也进一步验证模型的泛化能力

问题 3：问题 3 可以分解为四个步骤进行回答。首先，通过单因素方差分析和多元方差分析确定了不同人员在同一活动状态下的传感器数据存在显著差异。其次通过对传感器数据进行预处理、特征提取和相关性分析，计算并可视化展示活动状态数据与实验人员的生理特征之间的皮尔逊相关系数及显著性检验结果，确定了这些特征之间存在显著的统计关系。然后利用线性回归、随机森林回归、梯度提升回归和支持向量回归等模型，分析活动状态数据与实验人员生理特征（年龄、身高、体重）之间的关系从而说明可以使用活动传感器数据进行人员画像。最后，采用随机森林、XGBoost 和神经网络模型对未知实验人员进行识别。结果表明，神经网络模型的平均准确率最高，达到了 0.8208，明显优于其他模型，展示了其在复杂数据分类任务中的强大性能。综上所述，我们成功展示了利用活动传感器数据进行人员识别的可行性和有效性，最后对附件 5 进行了人员预测。

关键词：随机森林　神经网络　特征提取　人员识别　相关性分析　假设检验

智能手机传感器数据的分类与个体识别问题

摘要

随着智能设备的普及，特别是智能手机内置传感器技术的发展，人体活动监测与识别中的无标签数据分类、活动状态判别及个体识别成为了重要的研究课题。本文针对这些问题，以实现不同活动状态数据分类、提高分类准确率和个体识别为目标，建立了聚类分析模型、随机森林判别模型和多种回归模型，并使用 K-means 算法、随机森林算法、线性回归、梯度提升回归和支持向量回归等方法以及神经网络方法对模型进行求解。

问题 1：针对智能手机传感器数据的无标签分类问题，提出了一种基于特征提取与聚类分析的方法。首先，通过对 3 名实验人员的 60 组加速度计和陀螺仪数据进行特征提取，构建特征向量，并使用 K-means 聚类算法将这些特征向量聚类为 12 个簇。为了评估聚类效果，本文使用了轮廓系数、卡尔金指数和戴维森堡丁指数，同时使用 t-SNE 进行聚类结果的可视化分析。结果表明，该方法能够较好地将不同活动状态的数据进行分类，但仍存在一定的改进空间。通过本文的研究，验证了特征提取和聚类分析在无标签数据分类中的有效性。

问题 2：针对问题 2，建立了有效的人员活动状态判别模型，并分析和比较了聚类算法和随机森林在不同运动状态下的分类性能。首先，通过提取附件 2 中 12 类活动数据的典型特征，构建基于随机森林的判别模型。然后，比较随机森林模型与问题 1 中的无标签分类模型在附件 2 数据中的分类准确率。最后，应用判别模型对附件 3 中的实验人员活动数据进行分类预测。数据增强方面，我们通过添加高斯噪声的方法生成多个增强数据样本，以提高模型的泛化能力。在特征提取过程中，分别从时间域和频域提取了多种统计特征，并计算了加速度和陀螺仪的模量和抖动特征。模型评估使用了准确率、精确率、召回率和 F1 分数等指标，并通过混淆矩阵和匈牙利算法优化模型的匹配关系。结果表明，随机森林模型在整体准确率和各类别准确率上均显著优于无标签分类模型，其中随机森林的整体准确率达到 0.96，而无标签分类模型最优匹配后的准确率为 0.56。基于随机森林模型的分类结果显示，模型在处理数据偏态和噪声方面具有更强的鲁棒性和一致性。此外，对附件 3 的数据分类预测也进一步验证模型的泛化能力

问题 3：问题 3 可以分解为四个步骤进行回答。首先，通过单因素方差分析和多元方差分析确定了不同人员在同一活动状态下的传感器数据存在显著差异。其次通过对传感器数据进行预处理、特征提取和相关性分析，计算并可视化展示活动状态数据与实验人员的生理特征之间的皮尔逊相关系数及显著性检验结果，确定了这些特征之间存在显著的统计关系。然后利用线性回归、随机森林回归、梯度提升回归和支持向量回归等模型，分析活动状态数据与实验人员生理特征（年龄、身高、体重）之间的关系从而说明可以使用活动传感器数据进行人员画像。最后，采用随机森林、XGBoost 和神经网络模型对未知实验人员进行识别。结果表明，神经网络模型的平均准确率最高，达到了 0.8208，明显优于其他模型，展示了其在复杂数据分类任务中的强大性能。综上所述，我们成功展示了利用活动传感器数据进行人员识别的可行性和有效性，最后对附件 5 进行了人员预测。

综上所述，本文验证了特征提取和聚类分析在无标签数据分类中的有效性，展示了利用活动传感器数据进行人员识别的可行性和准确性，具有较高的实用价值和算法效率。

关键词：随机森林　神经网络　特征提取　人员识别　相关性分析　假设检验

目录

1 问题综述

1.1 问题背景

随着智能设备的普及，特别是智能手机[1-2]，内置的传感器如加速度计和陀螺仪，已成为监测人体活动状态的重要工具。这些传感器能够捕捉到细微的运动变化，从而为评估日常活动消耗的热量提供了一种新的方法。目前，市场上的运动健康软件，例如华为运动健康，正是基于这些传感器数据来计算用户的日常活动消耗。这一领域的研究不仅关系到个人健康监测，还涉及到大数据、机器学习等技术的应用，是当前科技和健康领域的热点话题。

尽管现有的研究在利用智能手机传感器数据进行人体活动识别方面取得了一定的进展，但仍存在一些不足之处。例如，现有算法可能在处理复杂或模糊的运动数据时准确度不高，或者在不同个体特征的适应性上存在局限。此外，大多数研究可能忽视了个体差异对活动识别准确性的影响，以及如何将这些数据与用户的生理特征相结合以提供更个性化的健康建议。针对这些不足，本文旨在通过建立更为精确的判别模型，不仅提高活动识别的准确度，还要探索如何将个体的生理特征纳入模型，以实现更全面的人员画像构建。通过这种方法，我们希望能够在现有研究的基础上进行创新，提供一种更为个性化和精准的人体健康监测解决方案。

1.2 问题提出

（1）无标签数据分类问题：如何对没有活动状态标签的传感器数据进行分类，识别出不同的活动状态。

（2）活动状态判别模型的构建和验证：建立一个判别模型，通过特征提取和分类算法，对实验人员的活动状态进行准确识别，并验证模型的性能。

（3）人员活动特征分析及个体识别：分析不同人员在同一活动状态下的差异，以及活动数据与生理特征（如年龄、身高、体重）之间的关系，进而探讨利用传感器数据进行人员识别的可行性。可以分解成下面四个问题。

（3.1）不同人员的同一活动状态是否存在差异？

（3.2）活动状态数据与实验人员的年龄、身高、体重有无关系？

（3.3）能否使用活动传感器数据进行人员画像？

（3.4）使用模型判断附件 5 中的未知实验人员分别最可能来源于问题 2 中的哪一名实验人员？

2 模型假设与符号说明

2.1 模型基本假设

（1）传感器数据准确性假设

假设智能手机内置的加速度计和陀螺仪能够准确地记录实验人员在各个活动状态下的线性加速度和角速度数据。传感器的数据采集没有显著的系统误差和随机误差。

（2）数据独立性假设

假设各组实验数据是独立的，即一个实验人员的一个活动状态数据不会受到其他活动状态数据的影响。每组数据能够独立反映相应的活动状态特征。

（3）统一性假设

假设所有实验人员在执行相同的活动状态时，数据记录的方式和标准是一致的。传感器的固定位置、采样率等实验条件在不同实验人员和不同活动状态下保持一致。

（4）特征稳定性假设

假设提取的特征（如时间域特征和频域特征）能够稳定地反映不同活动状态的区别。即同一活动状态在不同实验人员之间有一致的特征模式，而不同活动状态之间有显著的特征差异

2.2 符号说明

本文定义了如下 21 个使用次数较多的符号，其余符号在使用时注明，详情如表 1。

表 1 符号说明

序号	符号	含义
1	acc_x	X 方向上的加速度计读数
2	acc_y	Y 方向上的加速度计读数
3	acc_z	Z 方向上的加速度计读数
4	gyro_x	X 方向上的陀螺仪读数
5	gyro_y	Y 方向上的陀螺仪读数
6	gyro_z	Z 方向上的陀螺仪读数
7	μ	平均值
8	σ	标准差
9	Max	最大值
10	Min	最小值
11	RMS	均方根
12	FFT	快速傅里叶变换
13	PeakFreq	频谱峰值频率
14	acc_modulus	加速度模量
15	gyro_modulus	陀螺仪模量
16	jerk_x	X 方向上的加速度抖动
17	jerk_y	Y 方向上的加速度抖动
18	jerk_z	Z 方向上的加速度抖动
19	$\mathrm{jerk}_{\mathrm{gyro}_x}$	X 方向上的陀螺仪抖动
20	$\mathrm{jerk}_{\mathrm{gyro}_y}$	Y 方向上的陀螺仪抖动
21	$\mathrm{jerk}_{\mathrm{gyro}_z}$	Z 方向上的陀螺仪抖动

3 问题一的分析与建模

3.1 问题分析

附件 1 提供了 3 名实验人员的运动数据，每名实验人员完成了 12 种不同的活动，每种活动记录了 5 组加速度计和陀螺仪的数据。加速度计数据包括三个方向的加速度值，陀螺仪数据包括三个方向的角速度值。总共每人有 60 组数据，但是这些数据没有标注具体的活动状态。

我们需要对每位实验人员的 60 组数据进行分类，使其能够对应到 12 种不同的活动状态。首先，通过对每组数据提取特征，构建特征向量。这些特征包括加速度和陀螺仪数据的时间域特征和频域特征，以及加速度模量和陀螺仪模量。选择这些特征的依据是它们能全面反映不同活动状态下运动特性的变化。

时间域特征如均值、标准差、最大值、最小值和均方根能够捕捉到信号在不同活动状态下的幅度和变化范围；频域特征如频谱峰值频率则能够反映运动信号在频域上的特征，帮助识别周期性或特定频率的运动模式。加速度模量和陀螺仪模量是加速度和角速度在三维空间上的综合度量，能够反映整体运动强度。

对提取的特征进行标准化处理，使其在相同尺度上进行比较。接下来，使用 K-means 聚类算法[3]对标准化后的特征向量进行聚类，将数据分成 12 个簇。通过聚类结果，对每组数据进行分类，生成每位实验人员的活动状态分类结果，并将结果填入表 2 中。最后，通过计算聚类效果的评价指标，如轮廓系数、卡尔金指数和戴维森堡丁指数，验证聚类的效果和合理性。综上所述，问题 1 的思路流程图如图 1 所示：

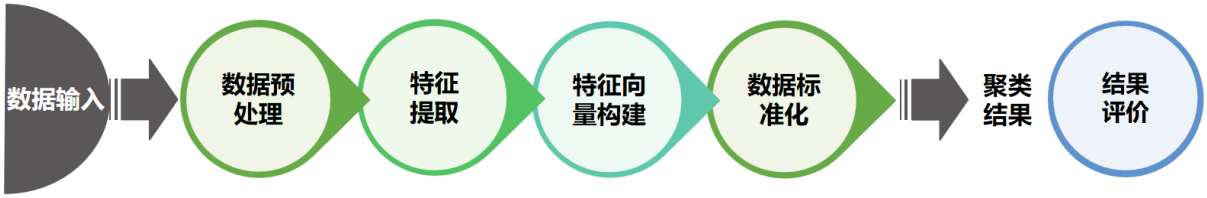


图 1 问题 1 的思路流程图

3.2 聚类分析模型构建

3.2.1 数据预处理

假设我们有一组传感器数据集$\{(\mathrm{acc}_{x,i}, \mathrm{acc}_{y,i}, \mathrm{acc}_{z,i}, \mathrm{gyro}_{x,i}, \mathrm{gyro}_{y,i}, \mathrm{gyro}_{z,i})\}_{i=1}^{n}$，其中$\mathrm{acc}_{x,i}, \mathrm{acc}_{y,i}, \mathrm{acc}_{z,i}$分别表示第$i$个样本在 X、Y 和 Z 方向上的加速度计读数，$\mathrm{gyro}_{x,i}, \mathrm{gyro}_{y,i}, \mathrm{gyro}_{z,i}$分别表示第 i 个样本在 X、Y 和 Z 方向上的陀螺仪读数。

数据集中的 n表示样本数量。由于每人有 60 组数据，每个数据文件记录了一段时间内的加速度和陀螺仪数据，每秒的数据点数量取决于采样频率（假设为100 Hz ），即每秒有 100 个数据点。因此，如果每组数据记录了 10 秒的数据，则每组数据包含 1000 个样本点。

3.2.2 特征提取

对每个样本 (i) 提取时间域特征、频域特征、加速度模量和陀螺仪模量三类特征：

1. **时间域特征：**

（1）平均值 (Mean)：

$$\mu_{\mathrm{acc}_x}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{acc}_{x,i},\quad \mu_{\mathrm{acc}_y}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{acc}_{y,i},\quad \mu_{\mathrm{acc}_z}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{acc}_{z,i} \tag{3-1}$$

$$\mu_{\mathrm{gyro}_x}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{gyro}_{x,i},\quad \mu_{\mathrm{gyro}_y}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{gyro}_{y,i},\quad \mu_{\mathrm{gyro}_z}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{gyro}_{z,i} \tag{3-2}$$

（2）标准差 (Standard Deviation)：

$$\sigma_{\mathrm{acc}_x}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{acc}_{x,i}-\mu_{\mathrm{acc}_x}\right)^2},\quad \sigma_{\mathrm{acc}_y}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{acc}_{y,i}-\mu_{\mathrm{acc}_y}\right)^2},\quad \sigma_{\mathrm{acc}_z}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{acc}_{z,i}^2} \tag{3-3}$$

（3）最大值 (Max) 和最小值 (Min)：

$$\begin{aligned}
\mathrm{Max}_{\mathrm{acc}_x}&=\max\mathrm{acc}_{x,i}, & \mathrm{Min}_{\mathrm{acc}_x}&=\min_i\mathrm{acc}_{x,i}\\
\mathrm{Max}_{\mathrm{acc}_y}&=\max_i\mathrm{acc}_{y,i}, & \mathrm{Min}_{\mathrm{acc}_y}&=\min_i\mathrm{acc}_{y,i}\\
\mathrm{Max}_{\mathrm{acc}_z}&=\max_i\mathrm{acc}_{z,i}, & \mathrm{Min}_{\mathrm{acc}_z}&=\min_i\mathrm{acc}_{z,i}\\
\mathrm{Max}_{\mathrm{gyro}_x}&=\max_i\mathrm{gyro}_{x,i}, & \mathrm{Min}_{\mathrm{gyro}_x}&=\min_i\mathrm{gyro}_{x,i}\\
\mathrm{Max}_{\mathrm{gyro}_y}&=\max_i\mathrm{gyro}_{y,i}, & \mathrm{Min}_{\mathrm{gyro}_y}&=\min_i gy_{y,i}\\
\mathrm{Max}_{\mathrm{gyro}_z}&=\max_i\mathrm{gyro}_{z,i}, & \mathrm{Min}_{\mathrm{gyro}_z}&=\min_i\mathrm{gyro}_{z,i}
\end{aligned} \tag{3-4}$$

（4）均方根 (RMS)：

$$\mathrm{RMS}_{\mathrm{acc}_x}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{acc}_{x,i}^2},\quad \mathrm{RMS}_{\mathrm{acc}_y}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{acc}_{y,i}^2},\quad \mathrm{RMS}_{\mathrm{acc}_z}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{acc}_{z,i}^2} \tag{3-5}$$

$$\mathrm{RMS}_{\mathrm{gyro}_x}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{gyro}_{x,i}^2},\quad \mathrm{RMS}_{\mathrm{gyro}_y}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{gyro}_{y,i}^2},\quad \mathrm{RMS}_{\mathrm{gyro}_z}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{gyro}_{z,i}^2} \tag{3-6}$$

以上公式同样适用于 $\mathrm{gyro}_x,\mathrm{gyro}_y,\mathrm{gyro}_z$ 数据。

2. **频域特征：**

对于每个样本的信号 acc_x，进行快速傅里叶变换（FFT）：

$$\mathrm{FFT}_{\mathrm{acc}_x}=\mathrm{fft}(\mathrm{acc}_x) \tag{3-7}$$

频谱峰值频率（Peak Frequency）：

$$\mathrm{PeakFreq}_{\mathrm{acc}_x}=\mathrm{argmax}\left|\mathrm{FFT}_{\mathrm{acc}_x}\right| \tag{3-8}$$

以上公式同样适用于 $\mathrm{acc}_y,\mathrm{acc}_z,\mathrm{gyro}_x,\mathrm{gyro}_y,\mathrm{gyro}_z$数据。

3. **加速度模量和陀螺仪模量：**

（1）加速度模量：

$$\text{acc\_modulus}_i = \sqrt{\text{acc}_{x,i}^2 + \text{acc}_{y,i}^2 + \text{acc}_{z,i}^2} \quad (3-9)$$

（2）陀螺仪模量：

$$\text{gyro\_modulus}_i = \sqrt{\text{gyro}_{x,i}^2 + \text{gyro}_{y,i}^2 + \text{gyro}_{z,i}^2} \quad (3-10)$$

3.2.3 特征向量构建

将上述特征组合形成特征向量 $\mathbf{f}_i$

$$\mathbf{f}_i = \left[\mu_{\text{acc}\ x}, \sigma_{\text{acc}_x}, \text{Max}_{\text{acc}_x}, \text{Min}_{\text{acc}_x}, \text{RMS}_{\text{acc}\ x} \cdots \text{acc\_modulus}, \text{gyro\_modulus}\right] \quad (3-11)$$

每个样本 (i) 形成一个特征向量$\mathbf{f}_i$，共有 (m = 38) 维度的特征。

3.2.4 数据标准化

将所有样本的特征向量进行标准化处理，使其在相同的尺度上：

$$\mathbf{f'}_i = \frac{\mathbf{f}_i - \mu_{\mathbf{f}}}{\sigma_{\mathbf{f}}} \quad (3-12)$$

其中，$\mu_{\mathbf{f}}$ 和 $\sigma_{\mathbf{f}}$分别是所有特征的均值和标准差。

3.2.5 聚类分析

使用 K-means 算法[6]对标准化后的特征向量进行聚类:

$$\text{K} - \text{means}(\mathbf{F}') \quad (3-13)$$

其中，$\mathbf{F}' = [\mathbf{f}'_1, \mathbf{f}'_2, \ldots, \mathbf{f}'_n]$，$n$ 为样本数量。K-means 算法的目标是最小化簇内平方和（Within-cluster sum of squares, WCSS）：

$$\text{WCSS} = \sum_{j=1}^{k} \sum_{\mathbf{f}'_i \in C_j} \left\| \mathbf{f}'_i - \mu_j \right\|^2 \quad (3-14)$$

其中，$k = 12$ 是簇的数量，C_j 是第 j 个簇，μ_j 是第 j 个簇的质心。

3.3 聚类效果评价标准

使用以下指标评价聚类效果:

1. 轮廓系数 (Silhouette Coefficient)：

$$\text{SC} = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{\max(a_i, b_i)} \quad (3-15)$$

其中，a_i 是样本 i 到其所在簇内其他样本的平均距离，b_i 是样本 i 到最近其他簇的平均距离。

2. 卡尔金指数 (Calinski-Harabasz Index)：

$$\text{CH} = \frac{B_k/(k-1)}{W_k/(n-k)} \quad (3-16)$$

其中， B_k 是簇间离散度， W_k 是簇内离散度。

3. 戴维森堡丁指数（Davies-Bouldin Index）：

$$\mathrm{DB}=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(\mu_i,\mu_j)}\right) \tag{3-17}$$

其中， σ_i 是第 i 个簇内样本到簇质心的平均距离， $d(\mu_i,\mu_j)$ 是簇 i 和簇 j 之间的质心距离。

3.4 模型求解

表 2 聚类效果评价指标表

分类	Person1	Person2	Person3
第 1 类	SY14，SY47，SY17，SY41，SY28	SY16，SY43，SY40，SY21，SY29	SY14，SY43，SY9，SY37，SY30
第 2 类	SY13，SY4，SY55，SY11，SY60	SY56，SY54，SY11，SY5，SY22	SY38，SY21，SY6，SY17，SY33
第 3 类	SY43，SY19，SY26，SY50，SY44	SY48，SY1，SY13，SY50，SY49	SY18，SY58，SY22，SY57，SY59
第 4 类	SY57，SY10，SY9，SY12，SY20	SY55，SY41，SY45，SY14，SY53	SY47，SY12，SY28，SY15，SY1
第 5 类	SY8，SY56，SY5，SY59，SY7	SY26，SY47，SY10，SY60，SY20	SY39，SY8，SY52，SY23，SY42
第 6 类	SY32，SY40，SY35，SY34，SY22	SY52，SY59，SY31，SY32，SY39	SY50，SY32，SY55，SY54，SY3
第 7 类	SY23，SY1，SY36，SY2，SY53	SY28，SY34，SY27，SY42，SY2	SY36，SY19，SY10，SY53，SY49
第 8 类	SY15，SY21，SY33，SY46，SY42	SY4，SY51，SY3，SY38，SY24	SY41，SY4，SY5，SY29，SY60
第 9 类	SY45，SY16，SY54，SY6，SY30	SY36，SY17，SY8，SY9，SY18	SY48，SY2，SY34，SY20，SY40
第 10 类	SY37，SY38，SY49，SY58，SY52	SY7，SY44，SY57，SY12，SY33	SY35，SY44，SY51，SY25，SY31
第 11 类	SY25，SY29，SY48，SY18，SY31	SY23，SY37，SY19，SY30，SY46	SY11，SY24，SY27，SY56，SY13
第 12 类	SY27，SY51，SY39，SY3，SY24	SY58，SY6，SY35，SY15，SY25	SY45，SY46，SY16，SY26，SY7

3.4.1 聚类结果的可视化分析

对聚类结果作 t-SNE 图的可视化分析，展示了高维特征数据在二维空间中的分布情况，每个点代表一组数据，不同颜色代表不同的聚类结果。其中：

1.簇的分布：

①不同颜色的点代表不同的簇。从图中可以看到，簇的分布有一定的重叠现象。部分簇之间的边界不清晰，这可能导致部分数据点被误分类。

② 一些簇如簇 0（红色）和簇 1（蓝色）分布相对集中，显示了较好的聚类效果。

2.簇的分离度：

①某些簇之间的分离度较好，如簇 3（紫色）和簇 4（橙色），它们在不同的区域有较为明显的分隔。

②另一些簇，如簇 5（黄色）和簇 6（暗红色），在图中有部分重叠，表明这些簇之间的分离度较差。

3.聚类密度：

①聚类密度较高的区域表示这些点在高维空间中非常相似。这可以从图中某些簇的中心区域看出，例如簇 0（红色）和簇 1（蓝色）。

②相对密度较低的区域，点之间的距离较大，表示这些点在高维空间中有更多的差异性。

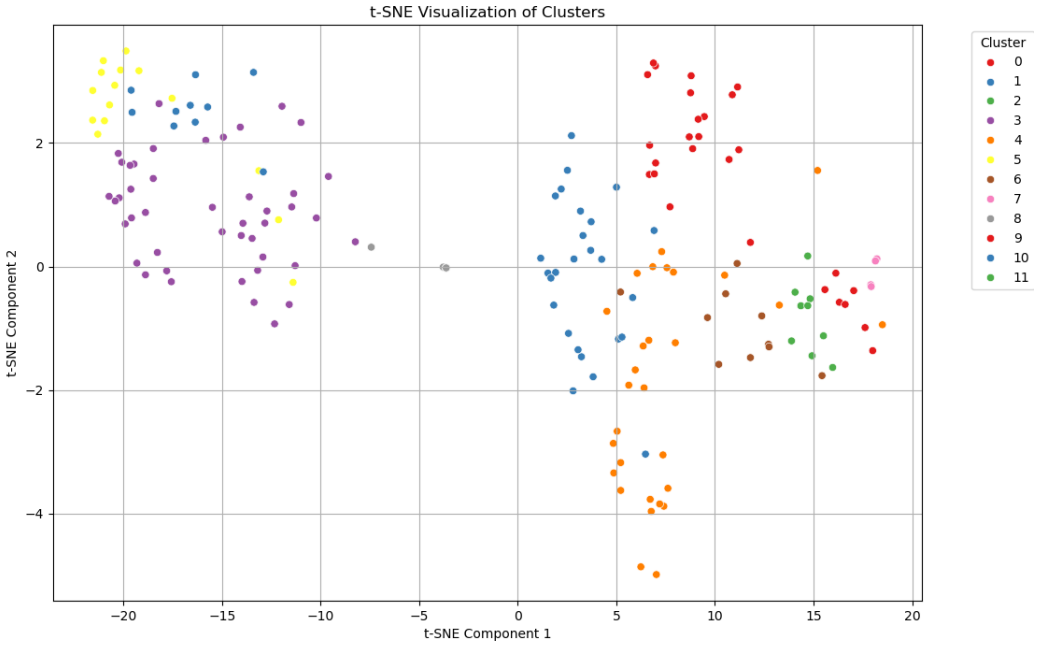


图 2 聚类可视化 t-SNE 图

3.4.2 聚类效果评价

根据 3.3 提出的聚类效果评价标准的评价指标，计算聚类结果的相应指标，结果如表 3 所示。

表 3 聚类效果评价指标表

评价指标	轮廓系数	卡尔金指数	戴维森堡丁指数
值	0.393	116.46	1.083

（1）轮廓系数为 0.393，轮廓系数的取值范围是[-1, 1]。值越接近 1，表示聚类效果越好；值越接近-1，表示聚类结果存在严重的重叠。0.393 的值表明，聚类效果还不错，但仍有改进的空间。

（2）卡尔金指数为 116.46，卡尔金指数用于衡量聚类的紧密度和分离度。值越大，聚类效果越好。116.46 的值在常见的范围内，表明聚类效果比较好。

（3）戴维森堡丁指数为 1.083，戴维森堡丁指数用于衡量聚类的相似性。值越小，聚类效果越好。1.083 的值表明，聚类结果具有一定的分离度，但还有改进的空间。

3.5 问题小结

通过 t-SNE 聚类可视化图和聚类效果评价指标的分析，我们得出结论：聚类效果整体较好，簇内紧密度和簇间分离度较高，这从轮廓系数和卡尔金指数中得到了体现。然而，戴维森堡丁指数指出部分簇之间存在重叠，表明聚类效果还有提升空间。当前特征选择，包括时间域特征、频域特征、加速度模量和陀螺仪模量，能够较好地反映数据的运动特性，且在不同活动状态下具有明显的区分度。为了进一步提高聚类效果，建议进一步优化特征工程，尝试引入更多或不同的特征，并使用统计特征、频域特征和时间序列分析等方法。同时，可以尝试其他聚类算法如 DBSCAN、层次聚类，或调整 K-means 算法的参数，观察其对聚类效果的影响。此外，数据预处理的进一步规范化和标准化也是提升聚类效果的关键步骤，这有助于在聚类过程中实现不同特征权重的均衡。

4 问题二的分析与建模

4.1 问题分析

问题 2 可以将问题分解为 3 个小问题，首先提取附件 2 中 12 类运动数据的典型特征，建立判别模型，在此基础上结合问题 1 得到的分类模型，比较两者在附件 2 数据中的准确率，最后运用建立的判别模型来对附件 3 中的数据进行分类。

（1）提取 12 类人员活动状态的典型特征，建立判别模型

•关键点：读取并整合数据，进行数据清洗和预处理，提取和选择有用特征，选择合适的分类模型并进行训练和评估。

•关系：这是基础步骤，决定了后续分类模型的准确性和可靠性。

（2）比较两个模型的分类结果和准确度

•关键点：使用判别模型和无标签的分类模型，对 10 名实验人员的数据进行分类，比较两个模型的分类结果和分类准确度。

•关系：通过比较两个模型的结果，可以验证和优化分类模型的性能，找出最优的分类方法。

（3）运用判别模型对附件 3 中某实验人员的 30 次活动状态数据进行分类

•关键点：读取和预处理附件 3 中的数据，使用判别模型对该数据进行分类预测，并记录结果。

•关系：通过对新数据进行分类预测，进一步验证模型的泛化能力。

4.2 模型建立

4.2.1 数据增强

在数据增强[7]过程中，主要通过添加高斯噪声的方法对原始数据进行处理。高斯噪声是一种常见的噪声形式，具有零均值和一定的标准差。具体步骤如下：

（1）噪声生成：为每个原始数据点生成一个服从高斯分布的噪声，噪声的均值为0，标准差为0.01。公式表示为noise~$\mathcal{N}(0,0.01)$，其中表示正态分布。

（2）数据增强：将生成的噪声添加到原始数据中，形成新的增强数据。公式表示为：augmented_data=original_data+noise。

（3）多次增强：对每组原始数据进行多次噪声添加操作，以生成多个增强数据样本。

4.2.2 特征提取

1. **时间域特征：**

（1）均值（Mean）：

$$\text{Mean}(x)=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{4-1}$$

（2）标准差（Standard Deviation）：

$$\text{Std}(x)=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2} \tag{4-2}$$

（3）中位数绝对偏差（Median Absolute Deviation, MAD）：

$$\text{MAD}(x)=\text{median}(|x_i-\text{median}(x)|) \tag{4-3}$$

（4）最大值（Max）和最小值（Min）：

$$\text{Max}(x)=\max(x_i),\text{Min}(x)=\min(x_i) \tag{4-4}$$

（5）四分位距（Interquartile Range, IQR）：

$$\text{IQR}(x)=Q3-Q1 \tag{4-5}$$

（6）熵（Entropy）：

$$\text{Entropy}(x)=-\sum_{i}p\,(x_i)\log p(x_i) \tag{4-6}$$

2. **频域特征：**

快速傅里叶变换（FFT）：对每个信号进行快速傅里叶变换（FFT）：$X(f)=\text{FFT}(x)$，取FFT结果的绝对值：$|X(f)|$

3. **模量特征：**

（1）加速度模量（Magnitude of Acceleration）：

$$|\mathrm{a}| = \sqrt{\mathrm{acc}_x^2 + \mathrm{acc}_y^2 + \mathrm{acc}_z^2} \quad (4-7)$$

（2）陀螺仪模量（Magnitude of Gyroscope）：

$$|\mathrm{g}| = \sqrt{\mathrm{gyro}_x^2 + \mathrm{gyro}_y^2 + \mathrm{gyro}_z^2} \quad (4-8)$$

4. **抖动特征：**

（1）加速度抖动：对每个方向上的加速度信号进行一阶差分得到抖动信号

$$\mathrm{jerk}_x = \frac{dx}{dt} \quad (4-9)$$

（2）陀螺仪抖动：对每个方向上的陀螺仪信号进行一阶差分得到抖动信号

$$\mathrm{jerk}_g = \frac{dg}{dt} \quad (4-10)$$

4.2.3 模型评估

使用测试集对模型进行评估，输出分类报告，包括准确率、精确率、召回率和 F1 分数。通过混淆矩阵进一步分析模型在不同活动状态下的分类效果。

1. 评估指标：

准确率（Accuracy）：

$$\mathrm{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (4\text{-}`11)$$

其中，TP 为真正例数，TN 为真负例数，FP 为假正例数，FN 为假负例数。

精确率（Precision）：

$$\mathrm{Precision}=\frac{TP}{TP+FP} \quad (4\text{-}12)$$

召回率（Recall）：

$$\mathrm{Recall}=\frac{TP}{TP+FN} \quad (4\text{-}13)$$

F1 分数（F1 Score）：

$$\text{F1 Score}=\frac{2\times \mathrm{Precision}\times \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \quad (4\text{-}14)$$

2. 混淆矩阵及优化

（1）混淆矩阵（Confusion Matrix）是一个方阵，用于评估分类模型的性能。矩阵的每一行表示实际类别，列表示预测类别。混淆矩阵的对角线元素表示正确分类的样本数量，非对角线元素表示错误分类的样本数量。

混淆矩阵 C 的元素 C_{ij} 表示实际类别为 i 而被预测为类别 j 的样本数量。

$$C_{ij} = \sum_{k=1}^{N} \delta\left(y_k = i, \hat{y}_k = j\right) \tag{4-15}$$

其中， δ 是指示函数，当条件为真时取值 1，否则取值 0 。

准确率表示模型预测正确的样本数量占总样本数量的比例。

$$\text{Accuracy}=\frac{\sum_i C_{ii}}{\sum_i \sum_j C_{ij}} \tag{4-16}$$

（2）匈牙利算法[7]（Hungarian Algorithm）是一种优化算法，用于解决分配问题。这里我们用匈牙利算法来优化混淆矩阵的匹配关系，使得总体准确率最大化。匈牙利算法用于优化混淆矩阵的匹配关系，目的是最大化对角线元素的和，即最大化准确率。

$$\max \sum_{i=1}^{n} C_{i,\sigma(i)} \tag{4-17}$$

其中， σ 是一个排列，表示类别的最佳匹配关系。具体步骤如下：

步骤 1 构建初始混淆矩阵：根据模型的预测结果和实际标签构建初始混淆矩阵。

步骤 2 使用匈牙利算法找到最佳匹配：通过匈牙利算法找到最优的类别匹配关系，重新排列混淆矩阵的列。

步骤 3 计算最佳匹配后的混淆矩阵：根据优化后的匹配关系，生成优化后的混淆矩阵。

4.3 模型求解

4.3.1 随机森林方法

随机森林（Random Forest）[4-6]是一种集成学习方法，用于分类、回归和其他任务。它通过构建多个决策树并将其集成来改善模型的性能和稳定性。每棵决策树是从数据的一个随机子集（样本和特征）中构建的，最终的预测结果是所有树的预测结果的平均或多数投票结果。基本步骤如下：

步骤 1 Bootstrap 采样：从数据集 D 中有放回地随机抽取 n 个样本，生成一个新的训练集 D_i 。

$$D_i = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \tag{4-18}$$

步骤 2 随机特征选择：对于每棵树的每个节点，随机选择 m 个特征（ $m < M$ ，其中 M 是总特征数）来决定最佳分割点。Select m features out of M

步骤 3 决策树训练：使用选择的特征训练决策树，生成树 T_i 。

步骤 4 多数投票法：对于一个新样本 x ，让所有树 T_i 进行分类，选择出现次数最多的类别作为最终预测结果。

$$\hat{y} = mode(\{T_1(x), T_2(x), \ldots, T_k(x)\}) \tag{4-19}$$

在这个题目中我们利用建立随机森林判别模型得到的结果下表，并得到了所有特征中最重要的 20 个特征。

表 4 随机森林性能评估表

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	33
2	1.00	1.00	1.00	27
3	1.00	1.00	1.00	30
4	1.00	1.00	1.00	31
5	1.00	1.00	1.00	26
6	1.00	1.00	1.00	27
7	1.00	1.00	1.00	25
8	1.00	1.00	1.00	43
9	0.94	1.00	0.97	29
10	1.00	1.00	1.00	28
11	0.89	0.65	0.75	37
12	0.66	0.88	0.75	24
Accuracy	-	-	0.96	360
Macro Avg	0.96	0.96	0.96	360
Weighted Avg	0.96	0.96	0.96	360

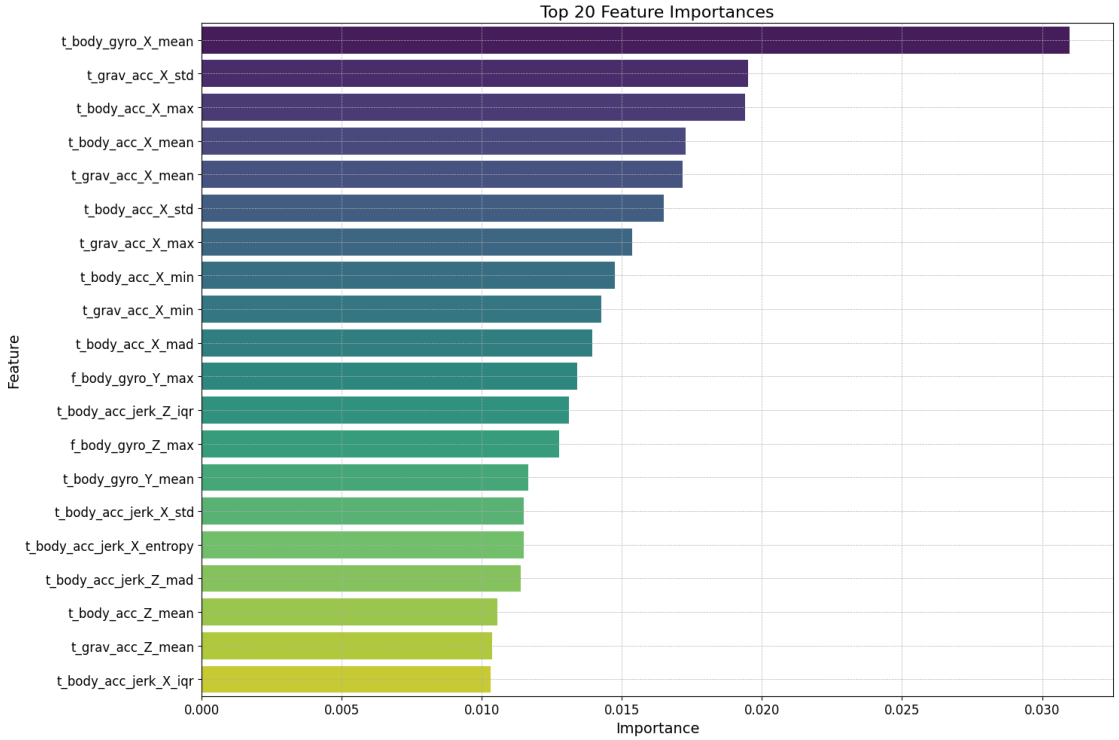


图 3 随机森林重要特征图

4.3.2 支持向量机方法

支持向量机（Support Vector Machine, SVM）是一种用于分类和回归分析的监督学习模型。SVM 的目标是找到一个最佳的超平面（或边界），能够将不同类别的样本进行

有效地分离。关键是找到一种映射关系，对于线性可分问题采用线性核，不可分问题采用多项式核、径向基函数核等。

线性核：线性核是最简单的一种核函数，适用于线性可分的数据。线性核不需要进行任何映射，直接在原始空间中找到一个超平面来分离数据。对于线性核，SVM 的决策函数为:

$$K(x_i, x_j) = x_i \cdot x_j \tag{4-20}$$

在使用线性核时，SVM 模型的优化目标和约束条件为:

$$\min_{w,b} \frac{1}{2} \| w \|^2 \tag{4-21}$$

满足约束条件:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1,2,\dots,n \tag{4-22}$$

当数据在原始空间中不可线性分离时，核函数通过将数据映射到高维空间，在高维空间中找到一个线性超平面进行分类。常用的核函数包括:

多项式核 (Polynomial Kernel) ：

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d \tag{4-23}$$

径向基函数核 (RBF Kernel)：

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{4-24}$$

表 5 支持向量机性能评估表

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	33
2	1.00	1.00	1.00	27
3	1.00	1.00	1.00	30
4	1.00	1.00	1.00	31
5	1.00	1.00	1.00	26
6	1.00	1.00	1.00	27
7	1.00	1.00	1.00	25
8	1.00	1.00	1.00	43
9	0.91	1.00	0.95	29
10	1.00	1.00	1.00	28
11	0.69	0.59	0.64	37
12	0.50	0.54	0.52	24
Accuracy	-	-	0.93	360
Macro Avg	0.92	0.93	0.93	360

Class	Precision	Recall	F1-Score	Support
Weighted Avg	0.93	0.93	0.93	360

4.3.3 结果比较

1.总体准确率：随机森林的准确率为 0.96，高于 SVM 的 0.93。

2.精度、召回率和 F1-得分：随机森林在几乎所有类别上的精度、召回率和 F1-得分都优于 SVM。特别是在类别 11 和类别 12 上，随机森林的性能明显优于 SVM，显示了更强的鲁棒性和一致性。

3.类别表现：对于精度和召回率均较高的类别（如类别 1-10），两种模型表现都很好。对于类别 11 和 12，随机森林的召回率和 F1-得分显著高于 SVM，表明其在处理数据偏态和噪声方面表现更佳。

4.4 对比分析

上述过程已经完成了问题 1 判别模型的建立，下面将通过附件 2 的数据对比分析问题 1 中建立的分类模型和判别模型的结果，其实质是无监督学习和有监督学习的比较。

4.4.1 分类模型结果

在问题 1 中得到分类模型，利用有标签的数据集附件 2 来计算分类模型的准确率，由于分类模型的标签是随机的，我们利用匈牙利算法来找到随机标签与实际标签的最优匹配，使得准确率最高。

表 6 准确率变化表

状态	准确率
调整前整体准确率	0.07
调整后整体准确率	0.56

表 6 展示了分类模型在调整前后的整体准确率对比情况。模型在未进行优化调整时的整体分类准确率为 0.07。模型在进行优化调整后的整体分类准确率为 0.56。反映了模型在优化调整前后的性能提升情况，准确率从 0.07 提升到 0.56，可以视为当匹配成功时的准确率，分类模型能达到的最大准确率。

表 7 不同类别准确率变化表

类别	调整前准确率	调整后准确率
类别 1	0.20	0.50
类别 2	0.00	0.62
类别 3	0.00	0.76
类别 4	0.00	1.00
类别 5	0.20	0.62
类别 6	0.00	0.38

类别	调整前准确率	调整后准确率
类别 7	0.00	0.88
类别 8	0.00	0.54
类别 9	0.00	0.16
类别 10	0.00	0.22
类别 11	0.24	0.56
类别 12	0.20	0.50

表 7 展示了分类模型在调整前后各个类别的准确率对比情况。调整后，所有类别的准确率均有不同程度的提升，其中类别 4 的准确率从 0.00 显著提升至 1.00，类别 7 的准确率从 0.00 提升至 0.88，显示了显著的改进。大多数类别在调整后准确率达到了 0.50 以上，表明模型优化显著提高了分类效果。

表 8 不同人员准确率变化表

实验人员	调整前准确率	调整后准确率
Person4	0.07	0.48
Person5	0.10	0.37
Person6	0.05	0.77
Person7	0.07	0.48
Person8	0.10	0.37
Person9	0.05	0.77
Person10	0.07	0.48
Person11	0.10	0.37
Person12	0.05	0.77
Person13	0.05	0.77

表 8 展示了分类模型在调整前后各个实验人员的准确率对比情况。调整后，每个实验人员的准确率均有显著提升，其中 Person6、Person9、Person12 和 Person13 的准确率从 0.05 提升至 0.77，提升幅度最大。Person4、Person7 和 Person10 的准确率从 0.07 提升至 0.48，而 Person5、Person8 和 Person11 的准确率从 0.10 提升至 0.37。整体来看，模型优化显著提高了各实验人员的分类效果。表 9 给出分类模型预测与实际类别匹配关系表。

表 9 分类模型预测与实际类别匹配关系表

Predicted Category	Actual Category
1	2

Predicted Category	Actual Category
2	11
3	6
4	10
5	7
6	1
7	3
8	4
9	5
10	12
11	8
12	9

图 4 展示了分类模型在初始状态和优化后的混淆矩阵。第一张图（初始混淆矩阵）显示了模型在未经优化时的分类结果，可以看到许多误分类的情况。第二张图（优化后的混淆矩阵）展示了经过优化调整后的分类结果，误分类的情况明显减少，各类别的分类准确率显著提高，尤其是类别 1、类别 2、类别 4 等，显示了优化后的模型在分类性能上的显著改进。

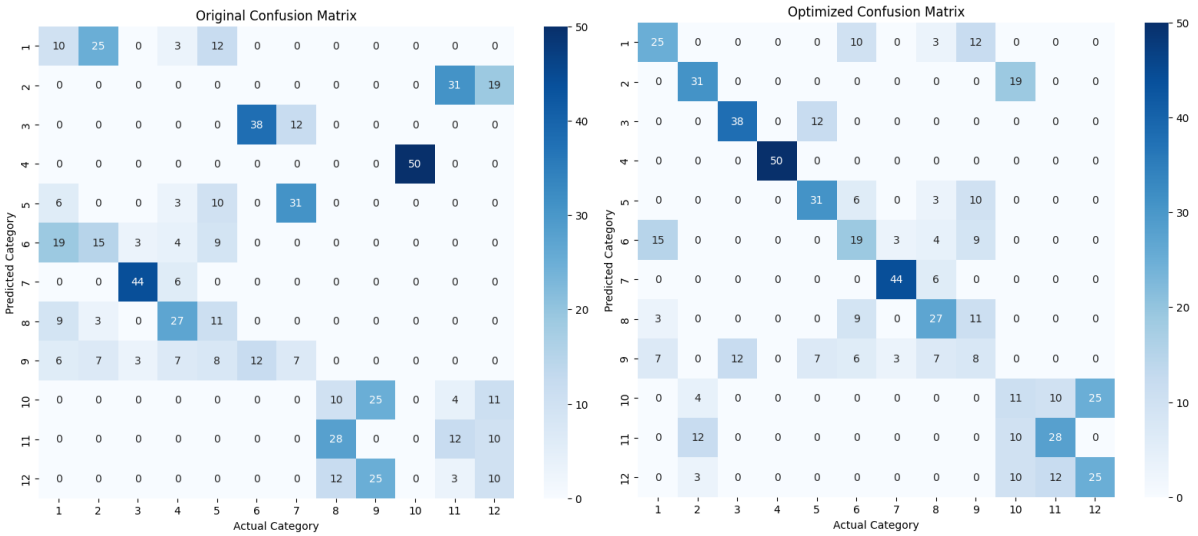


图 4 初始混淆矩阵和优化后混淆矩阵

4.4.2 结果对比

通过对比问题 1 优化后的结果和问题 2（随机森林）的结果，可以看出：

（1）整体准确率：随机森林模型（0.96）明显优于问题 1 优化后的模型（0.56）。

（2）每个类别的准确率：随机森林模型在每个类别的分类准确率上均优于或接近问题 1 优化后的模型。

总体来说，随机森林模型在分类任务中的表现优于问题 1 的分类模型。

4.4.3 结果解释

1.有监督学习的优势

（1）随机森林是有监督学习模型，在训练过程中使用了带标签的数据。这使得模型能够学习到数据和标签之间的关系，能够更好地进行分类。

（2）K-Means 聚类是无监督学习模型，不使用标签信息，仅依赖于数据的相似性进行分类，因此在复杂任务中的分类效果往往不如有监督学习模型。

2.模型复杂性和鲁棒性

（1）随机森林通过构建多个决策树并将它们集成起来进行分类。每棵决策树在不同的子样本上训练，模型的多样性降低了单个决策树过拟合的风险，提高了整体模型的泛化能力和准确性。

（2）K-Means 仅通过简单的簇心和数据点之间的距离来进行分类，模型的表达能力有限，尤其是在数据分布复杂的情况下，无法很好地捕捉数据的内在结构。

3.特征利用

（1）随机森林能够评估每个特征的重要性，自动选择对分类有贡献的特征，从而更有效地利用数据中的重要信息进行分类。这一过程显著提升了模型的分类能力。

（2）K-Means 在特征选择上没有专门的机制，所有特征在聚类过程中被平等对待，无法有效区分重要和次要特征，可能导致分类效果不佳。

4.数据使用效率

（1）随机森林在训练过程中能够充分利用所有标记的数据，每个决策树在不同的样本和特征上进行训练，从而在整体上增强模型的学习能力。

（2）K-Means 则依赖于簇心的初始化和更新，对初始簇心的选择敏感，可能导致聚类结果不稳定。

5.适应性和泛化能力

（1）随机森林通过引入随机性（如随机选择特征和样本）来训练多个决策树，减少了模型的过拟合风险，提高了泛化能力。

（2）K-Means 虽然也可以通过不同的初始化方法和迭代策略来优化，但在处理复杂分类问题时，效果可能不如随机森林。

4.5 结果预测

附件 3 的预测结果如下

表 10 附件 3 预测结果

活动类型	判别状态
SY1	5
SY2	10
SY3	8
SY4	7
SY5	4
SY6	3

活动类型	判别状态
SY7	4
SY8	1
SY9	4
SY10	5
SY11	10
SY12	1
SY13	8
SY14	6
SY15	1
SY16	10
SY17	5
SY18	1
SY19	8
SY20	10
SY21	5
SY22	6
SY23	7
SY24	5
SY25	8
SY26	7
SY27	10
SY28	1
SY29	6
SY30	7

5 问题三的分析与建模

问题 3 可以拆解成以下 4 个小问题：

1.不同人员的同一活动状态是否存在差异？

2.活动状态数据与实验人员的年龄、身高、体重有无关系？

3.能否使用活动传感器数据进行人员画像？

4.使用模型判断附件 5 中的未知实验人员分别最可能来源于问题 2 中的哪一名实验人员？

5.1 不同人员的同一活动状态是否存在差异

5.1.1 问题分析

在本研究中，我们希望探究不同实验人员在进行同一活动时，其传感器数据是否存在显著差异。实验人员通过佩戴加速度计和陀螺仪等传感器进行一系列活动，记录下多个方向的加速度和角速度数据。我们通过统计学方法对这些数据进行分析，以确定不同实验人员在相同活动状态下的行为是否存在显著差异。

步骤 1 数据预处理：首先，对原始传感器数据进行预处理，包括清洗缺失值和异常值。接着，对每个实验人员在每种活动状态下的多次实验数据提取一系列统计特征，如均值、标准差、最大值、最小值和均方根值。这些特征将用于后续的统计分析。

步骤 2 单因素方差分析：为了初步判断各实验人员在同一活动状态下是否存在差异，我们对每种活动状态的各个统计特征进行单因素方差分析。通过计算 F 值和 p 值，我们可以判断在特定特征下，不同实验人员之间是否存在显著差异。

步骤 3 多元方差分析：考虑到各个特征之间可能存在相关性，我们进一步进行多元方差分析 (MANOVA)。通过 MANOVA 分析，我们可以综合多个特征，判断不同实验人员在整体传感器数据上的差异是否显著。我们使用 Wilks' Lambda 统计量和相应的 p 值来评估显著性。

步骤 4 可视化与结果解读：为了更直观地展示分析结果，我们对 ANOVA 和 MANOVA 的结果进行可视化。通过柱状图和箱线图等方式，展示各活动状态下的统计显著性差异。这些图表有助于我们更直观地理解不同实验人员在同一活动状态下的表现差异。

5.1.2 数学建模

5.2.2.1 数据预处理

1.读取实验数据：设实验数据矩阵 $\mathrm{X_{ijt}}$ 表示第 i 个实验人员在第 j 个活动状态下第 t 次试验的传感器数据，包含以下 6 个传感器值:

$$\mathrm{X}_{ijt} = \left(\mathrm{acc}_x, \mathrm{acc}_y, \mathrm{acc}_z, \mathrm{gyro}_x, \mathrm{gyro}_y, \mathrm{gyro}_z\right) \quad (5-1)$$

2.处理缺失值和异常值:若数据中存在缺失值或异常值，则对其进行处理。设清理后的数据为 $\mathrm{X'_{ijt}}$ ：

$$\mathrm{X}'_{ijt} = \mathrm{X}_{ijt} \setminus \mathrm{NaN}, \text{异常值} \quad (5-2)$$

3.提取统计特征:

对每个传感器数据提取以下统计特征:

均值: $\mathrm{m}ean_{acc_x}, mean_{acc_y}, mean_{acc_z}, mean_{gyro_x}, mean_{gyro_y}, mean_{gyro_z}$

标准差: $\mathrm{std_{acc_x}}, \mathrm{std_{acc_y}}, \mathrm{std_{acc_z}}, \mathrm{std_{gyro_x}}, \mathrm{std_{gyro_y}}, \mathrm{std_{gyro_z}}$

最大值: $\mathrm{max_{acc_x}}, \mathrm{max_{acc_y}}, \mathrm{max_{acc_z}}, \mathrm{max_{gyro_x}}, \mathrm{max_{gyro_y}}, \mathrm{max_{gyro_z}}$

最小值: $\mathrm{min_{acc_x}}, \mathrm{min_{acc_y}}, \mathrm{min_{acc_z}}, \mathrm{min_{gyro_x}}, \mathrm{min_{gyro_y}}, \mathrm{min_{gyro_z}}$

均方根值: $\mathrm{rms_{acc_x}}, \mathrm{rms_{acc_y}}, \mathrm{rms_{acc_z}}, \mathrm{rms_{gyro_x}}, \mathrm{rms_{gyro_y}}, \mathrm{rms_{gyro_z}}$

具体计算公式如下:

$$\begin{aligned}\text{mean}_i &= \frac{1}{N}\sum_{j=1}^{N} X_{i,j} \\ \text{std}_i &= \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(X_{i,j} - \text{mean}_i\right)^2} \\ \max_i &= \max\left(X_{i,1}, X_{i,2}, \ldots, X_{i,N}\right) \\ \min_i &= \min\left(X_{i,1}, X_{i,2}, \ldots, X_{i,N}\right) \\ \text{rms}_i &= \sqrt{\frac{1}{N}\sum_{j=1}^{N} X_{i,j}^2}\end{aligned} \quad (5-3)$$

5.2.2.2 特征合并

合并实验数据特征：将每个实验人员的实验数据特征合并成一个综合数据集，记为 $\mathbf{F}_{ijt}$。

$$\mathbf{F}_{ijt} = (\text{mean}_{\text{acc}_x}, \text{mean}_{\text{acc}_y}, \text{mean}_{\text{acc}_z}, \text{mean}_{\text{gyro}_x}, \ldots \quad \text{rms}_{\text{gyro}_y}, \text{rms}_{\text{gyro}_z}) \quad (5-4)$$

加入实验人员信息，将实验人员的编号和活动状态的信息加入到综合数据集中:

$$\mathbf{F}_{ijt} = \left(\mathbf{F}_{ijt}, \text{person}_i, \text{activity}_j, \text{trial}_t\right) \quad (5-5)$$

5.2.2.3 单因素方差分析

单因素方差分析 (ANOVA) 旨在比较多个组的均值是否存在显著差异。对于本题中的活动数据，我们希望检验不同实验人员在相同活动状态下是否存在显著差异。

（1）定义 ANOVA 模型:

$$\begin{aligned}&\text{H}_0: \mu_1 = \mu_2 = \cdots = \mu_{10} \\ &\text{H}_\text{a}: \exists \text{i}, \text{jsuch that} \mu_\text{i} \neq \mu_\text{j}\end{aligned} \quad (5-6)$$

其中， μ_i 表示实验人员 i 的特征均值。

（2）方差和自由度计算

组间自由度 $\text{df}_{\text{between}}$:

$$\text{df}_{\text{between}} = \text{k} - 1 \quad (5-7)$$

组内自由度 $\text{df}_{\text{within}}$:

$$\text{df}_{\text{within}} = \text{k}(\text{n} - 1) \quad (5-8)$$

总自由度 df_{total}:

$$\text{df}_{\text{total}} = \text{kn} - 1 \quad (5-9)$$

组间方差 $\text{MS}_{\text{between}}$:

$$\mathrm{MS}_{\mathrm{between}} = \frac{\mathrm{SS}_{\mathrm{between}}}{\mathrm{df}_{\mathrm{between}}} \quad (5-10)$$

组内方差 $\mathrm{MS}_{\mathrm{within}}$:

$$\mathrm{MS}_{\mathrm{within}} = \frac{\mathrm{SS}_{\mathrm{within}}}{\mathrm{df}_{\mathrm{within}}} \quad (5-11)$$

（3）F 值计算

$$\mathrm{F} = \frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}} \quad (5-12)$$

（4）显著性判断

$$\alpha = 0.05 \quad (5-13)$$

①临界 F 值法:

计算临界 F 值:

$$\mathrm{F}_{\mathrm{critical}} = \mathrm{F}_{\alpha}(\mathrm{df}_{\mathrm{between}}, \mathrm{df}_{\mathrm{within}}) \quad (5-14)$$

若计算得到的 F 值大于临界 F 值，则拒绝原假设，认为不同实验人员的均值存在显著差异。

②p 值法:

计算 p 值:

$$\mathrm{p} = \Pr(\mathrm{F} > \mathrm{F}_{\mathrm{observed}}) \quad (5-15)$$

若 p 值小于显著性水平 α，则拒绝原假设，认为不同实验人员的均值存在显著差异。

5.2.2.4 多因素方差分析 (MANOVA)

为了综合多个特征，使用多元方差分析 (MANOVA)。对于每种活动状态 j，构建多元特征矩阵 $M_{i,j}$：

$$M_{i,j} = (F_{i,j,1} \quad F_{i,j,2} \quad \cdots \quad F_{i,j,T}) \quad (5-16)$$

其中，$F_{i,j,\mathrm{t}}$ 表示第 i 个实验人员在第 j 个活动状态下第 t 次试验的特征向量。

进行多元方差分析，测试多特征组合的差异：

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_P \\ H_1: \exists i, j \text{ such that } \mu_i \neq \mu_j \end{cases} \quad (5-17)$$

其中，μ_{i} 表示第 i 个实验人员的多特征均值向量。

使用 Wilks' Lambda 统计量进行检验：

$$\Lambda = \frac{\det(\mathrm{W})}{\det(\mathrm{W}+\mathrm{B})} \quad (5-18)$$

其中，W 为组内协方差矩阵，B 为组间协方差矩阵。

（1）计算组内协方差矩阵 W 和组间协方差矩阵 B：

组内协方差矩阵 W：

$$W = \sum_{i=1}^{P}\sum_{j=1}^{k}\left(M_{i,j} - \bar{M}_i\right)^T\left(M_{i,j} - \bar{M}_i\right) \quad (5-19)$$

其中，$\bar{M}_i$ 为第 i 个实验人员的多特征均值矩阵。

组间协方差矩阵 B：

$$B = \sum_{i=1}^{P} n_i\,(\bar{M}_i - \bar{M})^T(\bar{M}_i - \bar{M}) \quad (5-20)$$

其中，$\bar{M}$ 为总体多特征均值矩阵，n_i 为第 i 个实验人员的样本量。

（2）显著性检验的 p 值计算：

$$\text{p-value}=\Pr(\Lambda > \text{observed}\,\Lambda) \quad (5-21)$$

显著性水平 $\alpha = 0.05$。如果 p 值小于 α，则拒绝原假设，认为不同实验人员在多特征组合上存在显著差异。

5.1.3 结果分析

5.1.3.1 单因素方差分析

对于每个特征，计算 ANOVA 的 p 值。如果 p 值小于显著性水平 α（通常为 0.05），则认为该特征在不同实验人员之间存在显著差异。

从实验数据文件中提取特征并保存到 all_features DataFrame，对每个活动状态的每个特征进行单因素方差分析 (ANOVA)。

利用 P 值进行判断,统计每个活动状态下存在显著差异的特征数量，得到表，表的内容如下，表格显示，活动 1 到活动 6 的所有 30 个特征均存在显著差异（显著特征数为 30）。活动 7 有 29 个显著特征，活动 8 有 30 个显著特征，活动 9 有 28 个显著特征。活动 10、11 和 12 分别有 27、21 和 20 个显著特征。

临界 F 值：

1. 对于每个特征，计算 ANOVA 的 F 值，并与临界 F 值进行比较。如果 F 值大于临界 F 值# 计算临界 F 值
2. k = 10 # 实验人员数量
3. n = 5 # 每个实验人员的试验次数
4. df_between = k - 1
5. df_within = k * (n - 1)
6. alpha = 0.05
7. critical_f_value = f.ppf(1 - alpha, df_between, df_within)
8. **print**(f"Critical F-value: {critical_f_value}")
9. Critical F-value: 2.124029264016696

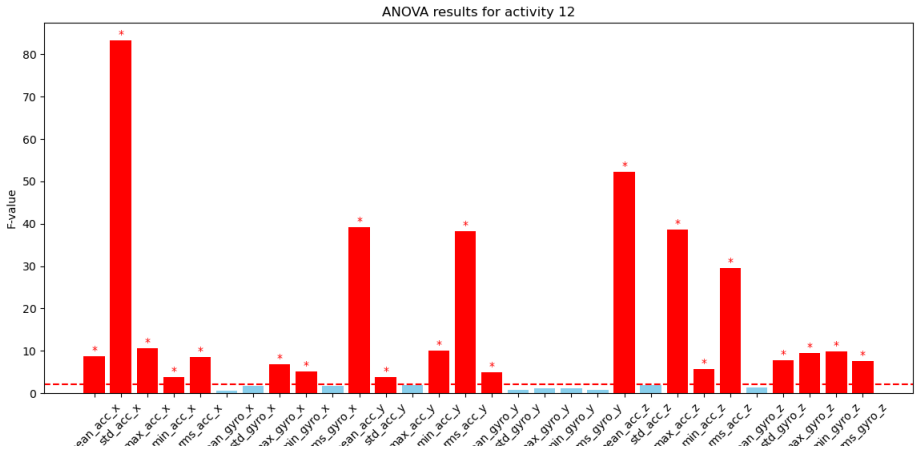
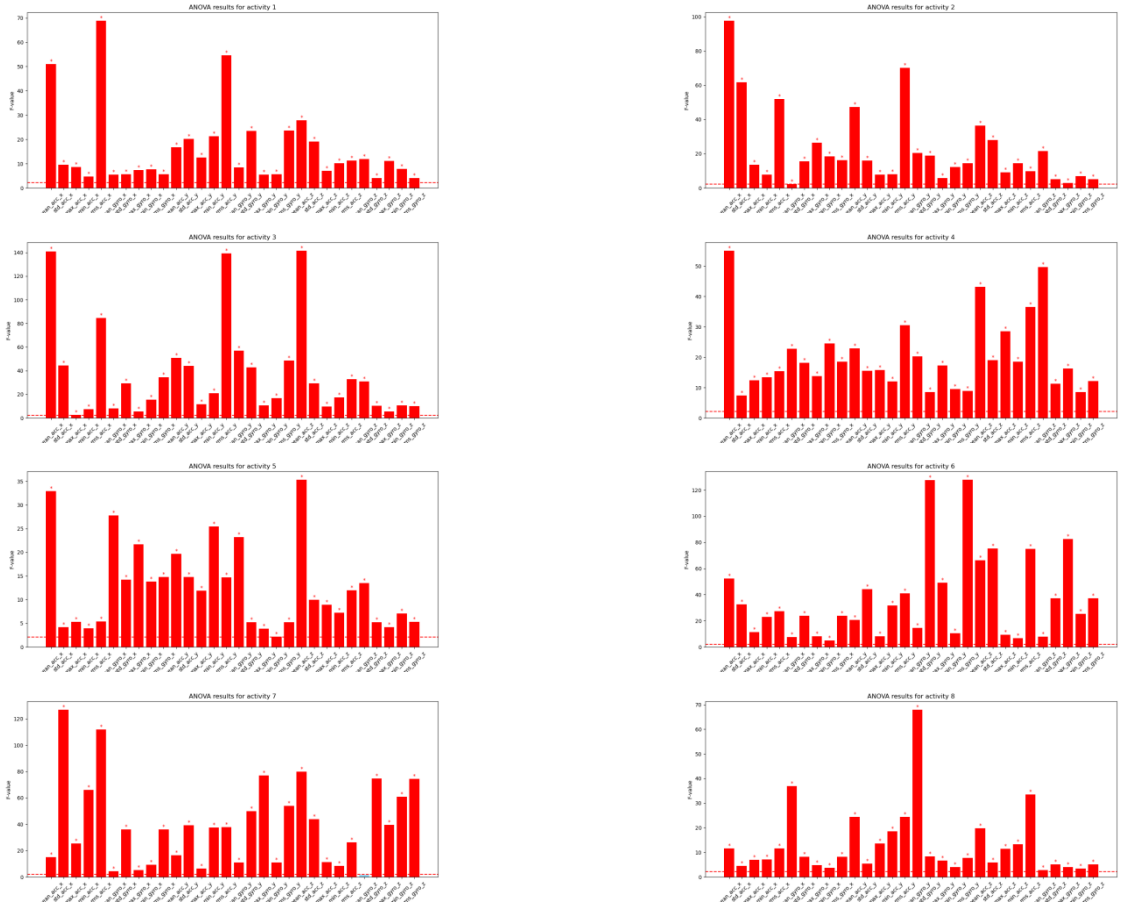


图 5 单因素方差分析 F 值结果图 – 活动 12

这张图展示了 Activity 12 的 ANOVA 检验结果。图中的柱状图代表了每个特征的 F 值，红色星号标注了显著差异的特征。红色虚线代表了临界 F 值，用于显著性判断。从图 5 中，可以看到红色柱状图上的红色星号标记了 20 个显著特征。这与表格中活动 12 的显著特征数量一致，均为 20 个。这说明在活动 12 中，有 20 个特征的 F 值超过了临界 F 值，且这些特征的 p 值也小于显著性水平 α，表示这些特征在不同实验人员之间存在显著差异。这些显著特征在单因素方差分析（ANOVA）图中得到了可视化展示，通过红色柱状图和红色星号标记了显著特征的位置。



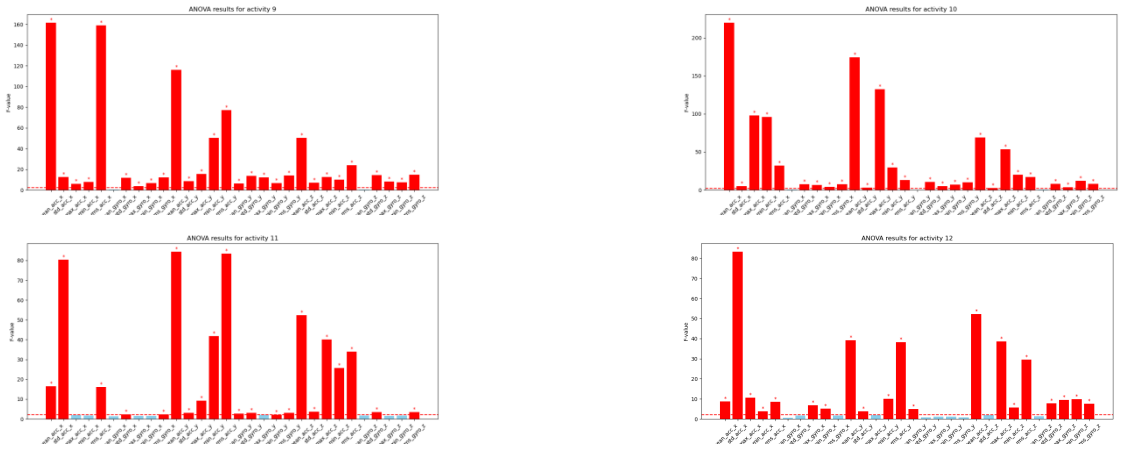


图 6 单因素方差分析 F 值结果图集 – 12 种活动

5.1.3.2 多因素方差分析

在进行 MANOVA 分析后，我们获得了每个活动的 p 值和 Wilks' Lambda 统计量。p 值和 Wilks' Lambda 统计量有助于判断不同实验人员在同一活动状态下是否存在显著差异。通过观察图 7 和表 10，可以看出活动 1 到 8 的 p 值非常小，接近于零，显示出非常显著的差异。活动 9 到 12 的 p 值稍微大一些，但仍然远小于 0.05，同样表明显著差异。这意味着我们可以拒绝原假设，认为在所有活动中，不同实验人员之间的差异是显著的。在图 8 中，每个活动的 Wilks' Lambda 值如图所示。Wilks' Lambda 值越小，表示不同组之间的差异越显著。Wilks' Lambda 值接近 1 表示组间差异不显著，接近 0 表示组间差异显著。下面对各活动的 Wilks' Lambda 值进行详细分析：

1.活动 1 到 8：

这些活动的 Wilks' Lambda 值较小，均在 0.02 到 0.06 之间。这表明这些活动中，不同实验人员之间的差异较为显著。

2.活动 9 到 12：

这些活动的 Wilks' Lambda 值相对较高，但依然在 0.1 左右，表示这些活动中，不同实验人员之间的差异依然存在。

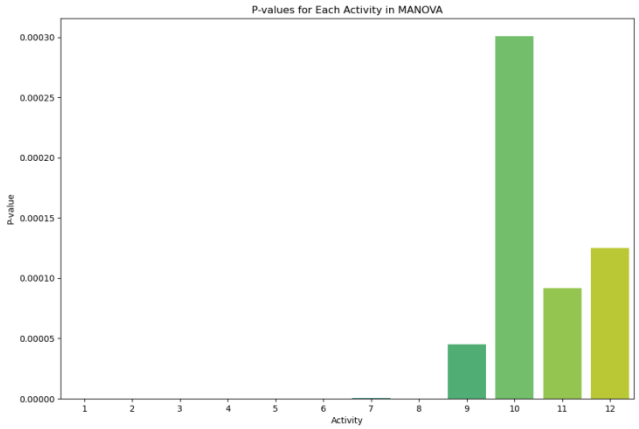


图 7 多因素方差分析 P 值结果图 - 12 种活动

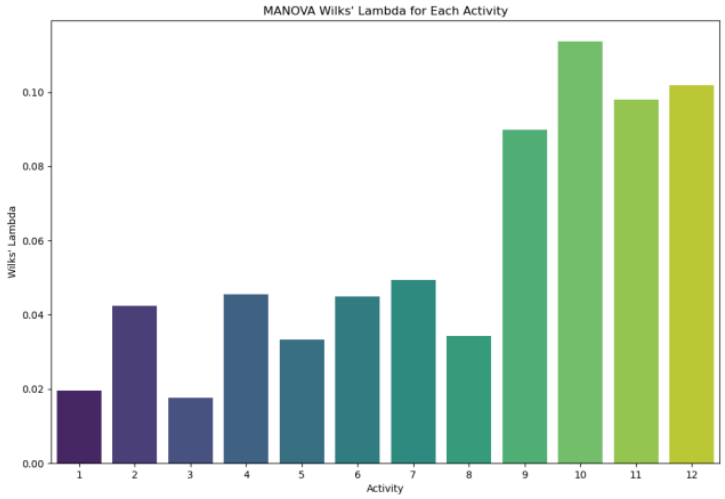


图 8 12 种活动的 Wilks' Lambda

表 11 多因素方差分析 P 值 – 12 种活动

活动	1 – 8	9	10	11	12
P 值	非常小	0.00004	0.00030	0.00009	0.00012

图 9，这张箱线图显示了在活动 8 中，不同实验人员的 x 轴加速度平均值（mean_acc_x）的分布情况。从箱线图可以看出，不同实验人员在活动 8 中的 mean_acc_x 值分布差异明显。这种差异说明不同实验人员在进行相同活动时的加速度特征存在显著不同。数据分布较广的实验人员（如人员 5 和 9）与数据分布较集中的实验人员（如人员 7 和 8）之间的差异尤为明显。通过多元方差分析（MANOVA），我们可以进一步验证这种差异的显著性。

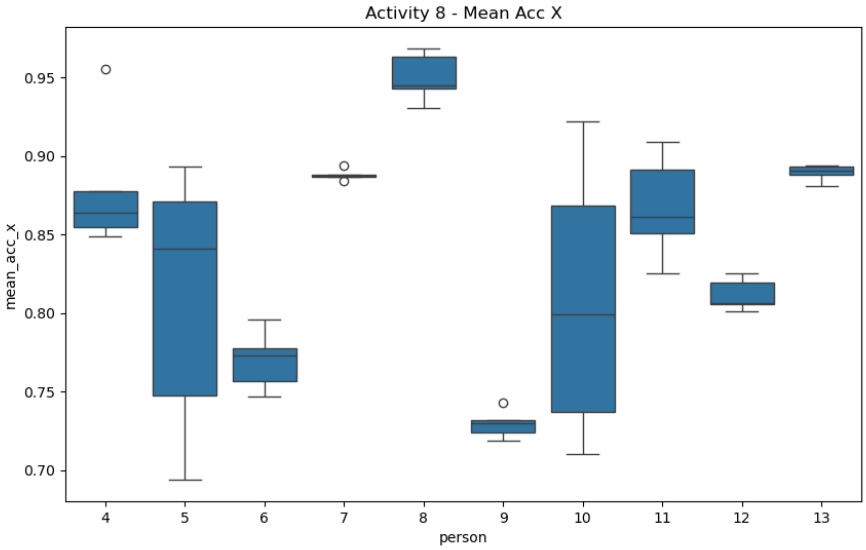


图 9 活动 8-x 轴加速度平均值箱线图

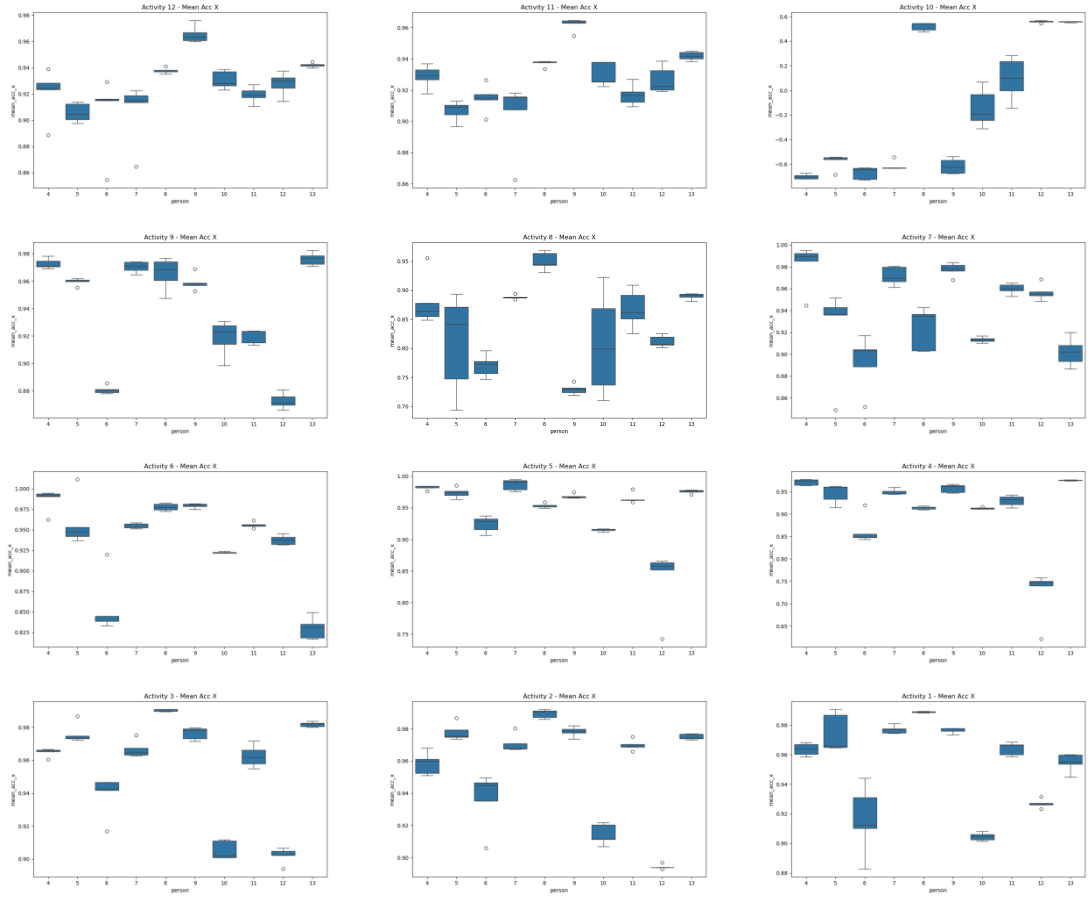


图 10 12 个活动状态下-x 轴加速度平均值箱线图

5.2 活动状态数据与实验人员的年龄、身高、体重有无关系

5.2.1 问题分析

智能手机依靠内置的加速度计和陀螺仪来测量和记录人体的活动状态数据。通过这些传感器的数据，可以捕捉到用户在不同活动状态下的动态变化。本次实验收集了 10 位实验人员的日常活动数据，包括 12 种不同的活动状态（例如步行、跑步、跳跃等），每种活动状态下记录了多组实验数据。此外，实验人员的年龄、身高、体重等基本信息也被记录下来。我们的目标是分析这些活动状态数据与实验人员的生理特征（年龄、身高、体重）之间的关系，判断它们之间是否存在显著的统计相关性。

问题分析步骤：

步骤 1 数据预处理：读取实验人员的基本信息和传感器数据，对传感器数据进行必要的清洗和格式转换，确保数据的一致性和完整性。考虑到不同实验数据的时间长度不一致，采用线性插值方法将所有时间序列数据归一化到统一长度（2400 个数据点）。

步骤 2 特征提取：从归一化后的时间序列数据中提取统计特征和时间序列特征，包括均值、标准差、最大值、最小值、中位数、能量、波峰数量等。这些特征能够有效描述不同活动状态下传感器数据的变化模式。

步骤 3 相关性分析：计算每个提取特征与实验人员年龄、身高、体重之间的皮尔逊相关系数，并进行显著性检验，计算对应的 p 值。通过相关系数和 p 值判断每个特征与年龄、身高、体重之间的关系强度和显著性。

步骤 4 可视化展示：使用热图展示相关系数和 p 值，将具有显著相关性的特征以直观的方式呈现，帮助识别出哪些特征与实验人员的生理特征存在显著关系。

步骤 5 结果解释：通过分析显著相关性的特征，解释活动状态数据与实验人员年龄、身高、体重之间的关系。这将帮助我们了解哪些活动特征可以反映实验人员的生理特征，从而支持进一步的建模和分析工作。

通过这些步骤，我们可以系统地分析和回答“活动状态数据与实验人员的年龄、身高、体重有无关系”这一问题，并为进一步的个性化健康监测和活动识别提供基础。

5.2.2 数学建模过程

1.问题定义

我们希望确定实验人员的活动状态数据（由传感器测得的加速度和陀螺仪数据）与实验人员的年龄、身高、体重之间的关系。具体来说，通过统计方法找出显著相关的特征，并验证这些特征与实验人员的年龄、身高、体重之间是否存在显著的统计关系。

2.数据预处理

假设有 $N = 10$ 个实验人员，编号分别为 $i \in \{4,5,\dots,13\}$ 。每个实验人员完成 $M = 12$ 种不同的活动，每种活动进行 $K = 5$ 次实验。实验数据包括三轴加速度计和陀螺仪数据:

- 加速度计数据: $\mathrm{acc}_x(t), \mathrm{acc}_y(t), \mathrm{acc}_z(t)$
- 陀螺仪数据: $\mathrm{gyro}_x(t), \mathrm{gyro}_y(t), \mathrm{gyro}_z(t)$

每次实验数据的采样率为 $f_s = 100\ \mathrm{Hz}$ ，时间长度不一致。

3. 时间归一化

为了将不同长度的时间序列数据归一化到相同的长度 $T = 2400$ ，我们采用线性插值方法。设原始数据长度为 n ，归一化后的时间序列长度为 T ，时间序列为 $x(t)$ ，则:

$$t \in 0,1,\dots,n-1$$
$$t' \in \left\{0, \frac{n-1}{T-1}, 2\frac{n-1}{T-1}, \dots, n-1\right\} \qquad (5-22)$$

使用线性插值函数 $\mathrm{interp}\big(t, x(t)\big)$ 计算归一化后的时间序列 $x'(t')$ 。

选择 $T = 2400$ 的理由是：假设在所有实验中，最长的数据长度不超过 2400 个数据点。这意味着在 100 Hz 采样率下，最长的实验时间为 24 秒。这个值可以确保所有实验数据都能被归一化到相同的长度，而不会丢失信息。

4.特征提取

从归一化后的时间序列数据中提取以下统计特征和时间序列特征:

设 $\mathrm{x(t)}$ 为时间序列数据，则提取特征如下:

均值:

$$\mu_{\mathrm{x}} = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=0}^{\mathrm{T}-1} \mathrm{x}\,(\mathrm{t}) \qquad (5-23)$$

标准差:

$$\sigma_x = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}(x(t) - \mu_x)^2} \quad (5-24)$$

最大值:

$$x_{max} = \max_t x(t) \quad (5-25)$$

最小值:

$$x_{min} = \min_t x(t) \quad (5-26)$$

中位数: median (x)

第 25 百分位数： $x_{25} =$ quantile $(x, 0.25)$

第 75 百分位数： $x_{75} =$ quantile $(x, 0.75)$

偏度： skew(x)

峰度: kurt(x)

能量:

$$E_x = \sum_{t=0}^{T-1} x(t)^2 \quad (5-27)$$

波峰数量: peaks(x)

波谷数量: valleys (x)

将这些特征应用于所有传感器数据，构建特征矩阵 X ：

$$X = \begin{bmatrix} \mu_{acc_x} & \sigma_{acc_x} & x_{acc_x,max} & x_{acc_x,min} & \cdots \\ \mu_{acc_y} & \sigma_{acc_y} & x_{acc_y,max} & x_{acc_y,min} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ \mu_{gyro_z} & \sigma_{gyro_z} & x_{gyro_z,max} & x_{gyro_z,min} & \cdots \end{bmatrix} \quad (5-28)$$

5.相关性分析

设 $y_{age}, y_{height}, y_{weight}$ 分别为实验人员的年龄、身高、体重向量:

$$\begin{cases} y_{age} = [y_{age,4}, y_{age,5}, \cdots, y_{age,13}] \\ y_{height} = [y_{height\,4}, y_{height,5}, \cdots, y_{height,13}] \\ y_{weight} = [y_{weight\,4,}, y_{weight,5}, \cdots, y_{weight,13}] \end{cases} \quad (5-29)$$

计算每个特征 x_j 与年龄、身高、体重之间的皮尔逊相关系数:

$$r_{age,j}=\frac{\sum_{i=1}^{N}\left(x_{i,j}-\bar{x}_j\right)\left(y_{age,i}-\bar{y}_{age}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_{i,j}-\bar{x}_j\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_{age,i}-\bar{y}_{age}\right)^2}}$$
$$r_{height,j}=\frac{\sum_{i=1}^{N}\left(x_{i,j}-\bar{x}_j\right)\left(y_{height,i}-\bar{y}_{height}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_{i,j}-\bar{x}_j\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_{height,i}-\bar{y}_{height}\right)^2}} \quad (5-30)$$

$$r_{weight,j}=\frac{\sum_{i=1}^{N}\left(x_{i,j}-\bar{x}_j\right)\left(y_{weight,i}-\bar{y}_{weight}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_{i,j}-\bar{x}_j\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_{weight,i}-\bar{y}_{weight}\right)^2}} \quad (5-31)$$

其中， $x_{i,j}$ 为第 i 个实验人员的第 j 个特征， $\bar{x}_j$ 和 $\bar{y}$ 分别为特征和目标变量的均值。

6.统计检验

对每个相关系数进行显著性检验，计算对应的 p 值 $p_{age,j}$， $p_{height,j}$， $p_{weight,j}$ ，使用以下公式:

$$t_{age,j}=r_{age,j}\sqrt{\frac{N-2}{1-r_{age,j}^2}}$$
$$p_{age,j}=2\left(1-T_{dist}\left(t_{age,j},N-2\right)\right)$$
$$t_{height,j}=r_{height,j}\sqrt{\frac{N-2}{1-r_{height,j}^2}}$$
$$p_{height,j}=2\left(1-T_{dist}\left(t_{height,j},N-2\right)\right) \quad (5-32)$$
$$t_{weight,j}=r_{weight,j}\sqrt{\frac{N-2}{1-r_{weight,j}^2}}$$
$$p_{weight,j}=2\left(1-T_{dist}\left(t_{weight,j},N-2\right)\right)$$

其中， $T_{dist}(t,N-2)$ 为自由度为 $N-2$ 的学生 t 分布的累积分布函数。显著性水平设为 $\alpha=0.05$ ，如果 $p<\alpha$ ，则认为特征与目标变量之间存在显著关系。

7.显著性特征分析

根据显著性检验结果，找出具有显著性关系的特征:

如果 $p_{age,j}<0.05$ ，则认为第 j 个特征与年龄之间存在显著关系。

如果 $p_{height,j}<0.05$ ，则认为第 j 个特征与身高之间存在显著关系。

如果 $p_{weight,j}<0.05$ ，则认为第 j 个特征与体重之间存在显著关系。

5.2.3 结果分析

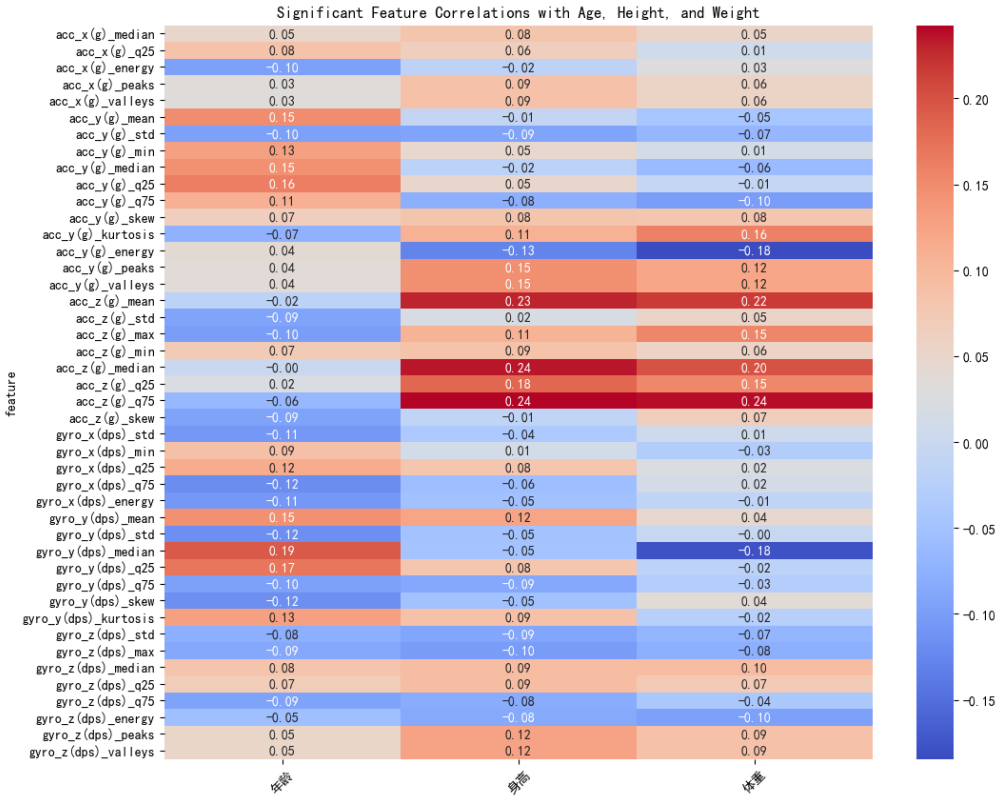


图 11 相关系数热图

图 11 和图 12 中显示的是所有具有显著相关性的特征，筛选标准是特征与年龄、身高或体重的 p 值小于 0.05。图 1 展示了各个特征与年龄、身高和体重的相关系数（Correlation Coefficients），具体特征名称在左侧，相关系数值显示在单元格中。颜色映射从蓝色到红色，表示负相关到正相关。比如说 acc_z(g)_median 与身高、体重之间有显著的正相关。acc_y(g)_energy 与身高、体重之间有较强的负相关。

图 12 展示了各个特征与年龄、身高和体重相关性的 p 值，反映了相关性的统计显著性。颜色映射（viridis）从黄色到深蓝色，表示从高 p 值（不显著）到低 p 值（显著）。通常，p 值小于 0.05 被认为是显著的。图中深蓝色单元格表示显著相关的特征。

图 11 展示了活动状态数据中的特征与实验人员的年龄、身高和体重之间的相关性，颜色和数值表示了这种线性关系的强度和方向，从而表明哪些特征与这些人体指标存在显著的相关性。图 2 展示了各个特征与年龄、身高、体重之间相关性的显著性水平，帮助我们识别出哪些特征与这些人体指标之间的相关性在统计上显著，从而支持活动状态数据与实验人员的年龄、身高、体重之间存在显著关系的结论。

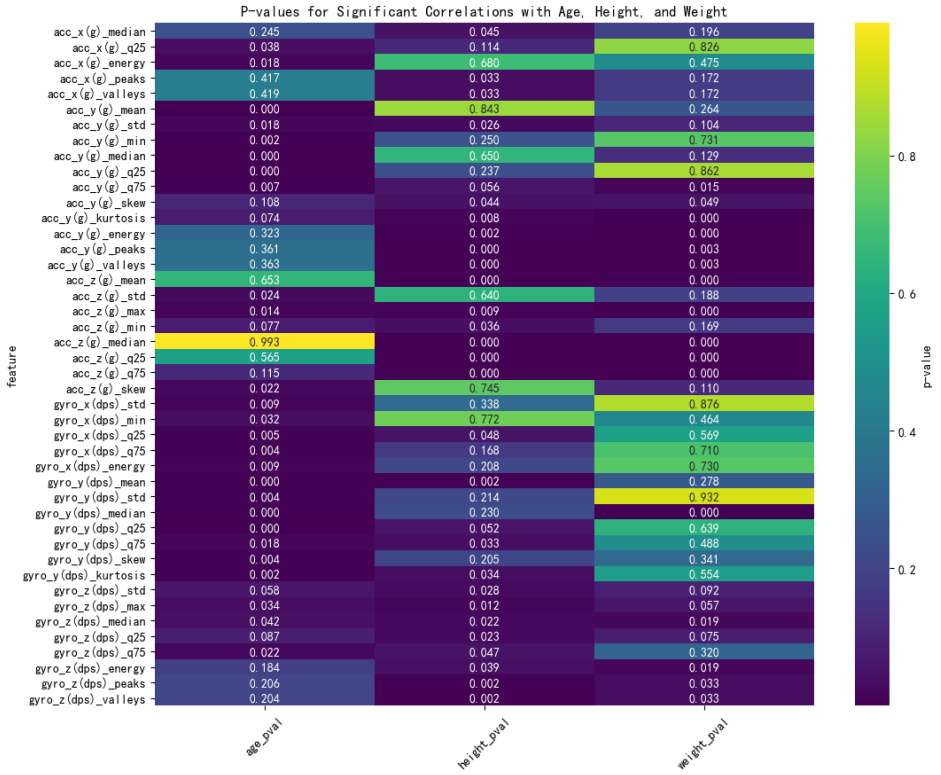


图 12 p 值热图

5.3 能否使用活动传感器数据进行人员画像

5.3.1 问题分析

从每个人员的体重、身高、年龄与传感器活动数据的相关性分析中，我们可以得到结论，生理特征与传感器活动数据是存在相关性的，基于上面的研究，下面我们将利用已有的数据建立模型来进一步验证可以通过活动传感器数据来对人员进行画像。

5.3.2 模型建立

1.模型选择

（1）线性回归 (Linear Regression)

简单易懂：线性回归是最基本的回归方法，容易理解和实现。

基准模型：线性回归可以作为一个基准模型，用于与其他更复杂的模型进行性能比较。

假设线性关系：如果数据中的特征与目标变量之间存在线性关系，线性回归可以提供良好的预测结果。

（2）随机森林回归 (Random Forest Regressor)

高准确性：通过集成多棵决策树，随机森林模型通常具有较高的预测准确性。

鲁棒性强：对异常值和噪声具有较强的鲁棒性，不容易过拟合。

非线性关系：能够捕捉特征与目标变量之间的复杂非线性关系。

特征重要性：能够评估每个特征的重要性，有助于特征选择和理解模型。

（3）梯度提升回归 (Gradient Boosting Regressor)

高准确性：通过逐步构建多个弱学习器（通常是决策树），逐步减小预测误差，具有很高的预测精度。

灵活性强：可以调整多种参数（如学习率、树的数量和深度）来控制模型复杂度和防止过拟合。

处理复杂关系：能够捕捉特征与目标变量之间的复杂非线性关系。

高精度需求：当对预测精度要求较高时，梯度提升模型通常能提供更好的性能。

（4）支持向量回归 (SVR)

高准确性：对于小样本、高维数据具有良好的泛化能力和预测性能。

非线性关系：通过使用核函数（如 RBF 核）处理特征与目标变量之间的非线性关系。

小样本问题：SVR 在小样本、高维数据集上表现优异，适合样本数量较少的情况。

通过结合模型特点和问题特性，选择了线性回归、随机森林回归、梯度提升回归和支持向量回归，确保能够全面评估和比较不同模型在进行人员画像时的表现，最终选出最优的模型。

2.模型架构

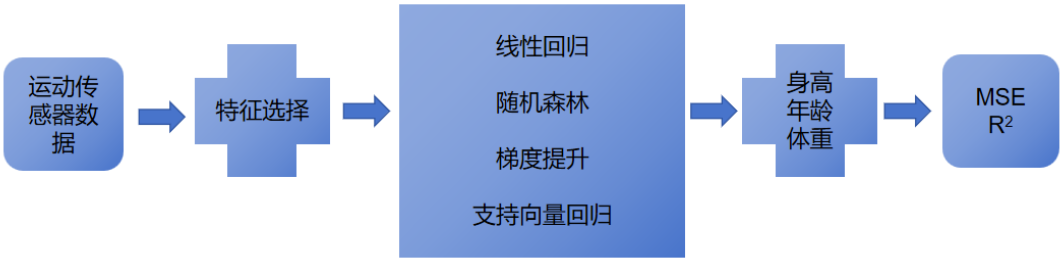


图 13 人物画像流程图

这张流程图展示了利用运动传感器数据进行人员画像的整体流程：首先，通过加速度计和陀螺仪获取运动数据，然后进行特征选择，提取时间域和频域特征。接着，利用四种不同的回归模型（线性回归、随机森林、梯度提升和支持向量回归）进行多目标预测，预测个体的生理特征（身高、年龄、体重）。最后，通过均方误差（MSE）和决定系数（R^2）评估模型性能，选择最优模型用于人员画像。

MSE (均方误差)：衡量预测值与实际值之间的平均平方误差，值越小表示预测效果越好。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5-33)$$

$\mathbf{R}^2$ (决定系数)：衡量模型对目标变量变异的解释能力，值越接近 1 表示模型解释能力越强。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (5-34)$$

y_i ：实际值，第 i 个样本的真实值。

$\hat{y}_i$ ：预测值，第 i 个样本的预测值。

$\bar{y}$：实际值的均值，表示所有样本真实值的平均值。

5.3.3 模型结果

表 12 回归模型性能比较

模型	MSE（年龄）	MSE（身高）	MSE（体重）	R^2
LinearRegression	57.21	19.02	135.45	0.073
RandomForest	16.99	5.86	59.74	0.638
GradientBoosting	19.30	9.52	67.22	0.580
SVR	44.08	8.91	85.40	0.394

从上面表格中得到的指标，我们可以得出以下结论：

1. **高相关性**：运动传感器数据与人员的生理特征（年龄、身高、体重）存在显著的相关性，尤其在随机森林回归模型中表现最佳，说明通过这些数据可以较准确地预测出人员的生理特征。

2. **模型性能优异**：随机森林回归模型在多目标回归任务中表现优异，具有较低的均方误差和较高的决定系数（R^2），能够解释大部分的目标变量变异。这证明了使用活动传感器数据进行人员画像是可行的。

3. **数据预处理的有效性**：标准化和特征选择等数据预处理步骤显著提升了模型的性能，进一步验证了通过精细化的数据处理和模型调优，可以提高人员画像的准确性和可靠性。

综上所述，通过活动传感器数据进行人员画像不仅是可行的，而且在模型性能和数据处理上都有很好的实现效果。

5.4 利用活动传感器数据进行人员识别

5.4.1 问题分析

我们的目标是建立模型，利用活动传感器数据进行人员识别。数据来源于 13 名实验人员的运动传感器记录，包括每名实验人员 12 类活动状态的 5 组数据。在附件 4 中给出了对应人员的生理特征数据，由于前 3 位人员不知道活动类型，数据不完整，所以我们只用到后面 10 个人的数据。

5.4.2 数据处理

将附件 2 和附件 4 的人员一一对应，从而得到数据的实际标签。将数据划分为训练集和测试集，比例为 8：2。为了增加训练数据的多样性，使用数据增强技术（添加噪声）对原始数据进行扩展，具体方法与问题 2 中的数据增强方法一致。

5.4.3 特征提取

对每个样本 i 提取以下特征：

1. **时域特征**：

平均值（Mean）：

$$\mu_{X_i} = \frac{1}{N}\sum_{t=1}^{N} X_i\,[t] \qquad (5-35)$$

标准差（Standard Deviation）：

$$\sigma_{X_i} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(X_i[t] - \mu_{X_i}\right)^2} \qquad (5-36)$$

最大值（Max）和最小值（Min）：$\max(X_i)$ 和 $\min(X_i)$

偏度（Skewness）：$\text{Skewness}(X_i)$

峰度（Kurtosis）：$\text{Kurtosis}(X_i)$

2. **频域特征：**

快速傅里叶变换（FFT）：

$$\mathbf{F}_i = \text{FFT}(\mathbf{X}_i) \qquad (5-37)$$

FFT 的均值和标准差：

$$\mu_{\mathbf{F}_i} = \frac{1}{N}\sum_{t=1}^{N} |\mathbf{F}_i[t]| \qquad (5-38)$$

$$\sigma_{\mathbf{F}_i} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(|\boldsymbol{F}_i[t]| - \mu_{\mathbf{F}_i}\right)^2} \qquad (5-39)$$

3. **其他特征：**

能量（Energy）：

$$\text{Energy}(X_i) = \sum_{t=1}^{N} X_i\,[t]^2 \qquad (5-40)$$

峰值计数（Peak Count）：数据中峰值的个数。

自相关（Autocorrelation）：自相关函数值。

5.4.4 模型构建

1. 神经网络模型架构

采用神经网络模型（NeuralNet），输入层维度为传感器数据的特征数 ，输出层维度为实验人员类别数。模型的具体架构如下：

输入层：输入特征维度为66。

隐藏层 1：全连接层，输出维度为 256，激活函数为 LeakyReLU，Dropout 率为 0.5。

隐藏层 2：全连接层，输出维度为 128，激活函数为 LeakyReLU，Dropout 率为 0.5。

隐藏层 3：全连接层，输出维度为 64，激活函数为 LeakyReLU，Dropout 率为 0.5。

输出层：全连接层，输出维度为实验人员类别数 10，激活函数为 Softmax。

2. 模型训练

损失函数与优化器：采用交叉熵损失函数和 Adam 优化器。

交叉摘损失函数用于衡量模型预测结果与真实标签之间的差异。它通过计算预测概率分布与真实标签分布之间的差异来衡量模型的性能。交叉摘损失函数的公式如下:

$$Cross - EntropyLoss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\, log(\hat{y}_{ij}) \tag{5-41}$$

通过最小化交叉熵损失函数，模型能够更好地拟合训练数据，从而提高分类准确率。

采用 Adam 优化器，学习率设置为 0.001，Adam 优化器结合了动量和自适应学习率的优点，一阶动量（类似于动量优化）通过指数加权平均，平滑了梯度，使得更新更加稳定。二阶动量 (类似于 RMSProp) 通过自适应学习率，根据梯度的变化动态调整每个参数的学习率。

训练过程：共训练 50 个 Epoch，记录每个 Epoch 的训练损失、验证损失、训练准确率和验证准确率，并得到随训练次数的变化准确率的变化情况。

5.4.5 模型比较

1. XGBoost（Extreme Gradient Boosting）

XGBoost 是一种提升树（Boosting Tree）方法，通过逐步添加新的树来校正前一个模型的错误，提高整体模型的预测性能。它结合了梯度提升和决策树的优点，使用加权平均法来生成最终预测结果。

2. 比较结果

表 13 模型比较结果

方法	预测结果	平均准确率
随机森林	[10, 7, 6, 9, 13]	0.4217
XGBoost	[10, 7, 6, 9, 13]	0.4433
神经网络	[10, 7, 6, 9, 13]	0.8208

从表中我们从一下几个方面对模型进行比较：

1. 预测结果：在附件 4 的测试集中随机森林、XGBoost 和神经网络的预测结果一致均为[10, 7, 6, 9, 13]。
2. 平均准确率：神经网络的平均准确率最高，达到了 0.8208，表明其在数据分类上的表现最优。XGBoost 的平均准确率为 0.4433，略高于随机森林的 0.4217，但两者都明显低于神经网络。
3. 模型表现：

神经网络：表现最佳，验证准确率最高，显示了强大的特征提取和学习能力。模型在复杂数据上的表现优越。

XGBoost：次优，平均准确率较高，说明在处理分类任务上具有较好的性能，但仍不及神经网络。

随机森林：表现相对较差，但仍具有一定的分类能力。其随机性和树的数量对模型的影响较大。

综上所述，神经网络在本实验中的表现最为优异，其验证准确率明显高于随机森林和 XGBoost，显示出更好的分类能力和泛化能力。XGBoost 作为一种强大的提升方法，其表现优于随机森林，但仍无法达到神经网络的性能水平。随机森林虽然简单易实现，但在本实验中的表现不如 XGBoost 和神经网络。

5.4.6 结果分析

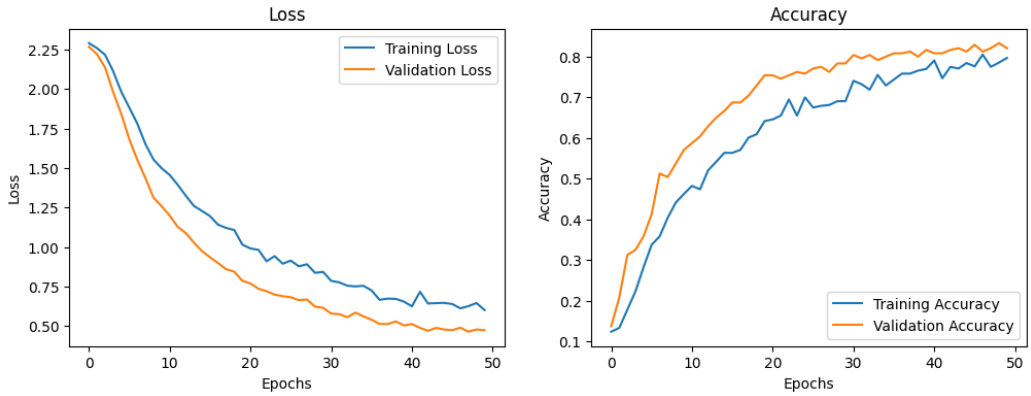


图 14 训练损失和验证损失随 Epoch 变化图

1. 损失函数分析（Loss）

损失函数图显示了训练损失（Training Loss）和验证损失（Validation Loss）在 50 个 Epochs 中的变化情况：

训练损失（Training Loss）：训练损失在整个训练过程中不断下降，表明模型在不断学习并优化其参数。最终训练损失趋于稳定，表明模型达到了一个较好的收敛状态。

验证损失（Validation Loss）：验证损失也在不断下降，且在前几个 Epochs 下降较快。最终验证损失趋于稳定，表明模型在验证集上的表现也达到了较好的状态。训练损失和验证损失的趋同性表明模型没有明显的过拟合现象。

2. 准确率分析（Accuracy）

准确率图显示了训练准确率（Training Accuracy）和验证准确率（Validation Accuracy）在 50 个 Epochs 中的变化情况：

训练准确率（Training Accuracy）：训练准确率在整个训练过程中不断上升，最终达到了接近 0.80 的水平，表明模型在训练集上的表现非常好。

验证准确率（Validation Accuracy）：验证准确率在前几个 Epochs 快速上升，随后稳定在 0.82 左右。验证准确率的上升趋势和训练准确率一致，表明模型在验证集上的表现也较好。

该表格显示了使用神经网络模型对未知实验人员进行识别的结果

3. 结果预测

表格 14 显示了使用神经网络模型对未知实验人员进行识别的结果

表 14 问题 3 结果

活动类型	判别结果
Unknow1	Person10
Unknow2	Person7
Unknow3	Person6
Unknow4	Person9
Unknow5	Person13

6 模型评价与改进

6.1 模型评价

在本研究中，我们提出了一种基于智能手机传感器数据的特征提取与聚类分析方法，用于无标签分类问题，并建立了人员活动状态的判别模型。通过对 3 名实验人员的 60 组加速度计和陀螺仪数据进行特征提取，利用 K-means 聚类算法将这些数据聚类为 12 个簇，并使用轮廓系数、卡尔金指数和戴维森堡丁指数对聚类效果进行了评估。此外，通过 t-SNE 进行聚类结果的可视化分析，展示了聚类效果。

在构建判别模型时，我们采用了随机森林模型，并与无标签分类模型进行了比较。结果表明，随机森林模型的整体准确率为 0.96，显著优于无标签分类模型的 0.56，展示了其在处理数据偏态和噪声方面的鲁棒性和一致性。对于人员画像问题，我们进一步验证了活动状态数据与实验人员生理特征（年龄、身高、体重）之间的显著相关性，并使用线性回归、随机森林回归、梯度提升回归和支持向量回归等模型进行了分析。结果表明，随机森林回归模型在多目标回归任务中表现最佳，具有较低的均方误差和较高的决定系数。

6.2 模型改进

尽管我们的方法在无标签数据分类和人员活动状态判别方面取得了良好的效果，但仍存在一些改进空间：

（1）特征选择与扩展：目前的特征提取主要集中在时间域和频域特征上。可以进一步尝试引入更多的特征，例如基于小波变换的特征、熵类特征或非线性动力学特征，以捕捉不同活动状态下更加细微的特征变化，提高模型的区分能力。

（2）聚类算法优化：虽然 K-means 算法在本研究中表现良好，但其对初始簇心选择较为敏感。未来可以尝试其他聚类算法，如层次聚类、密度聚类（如 DBSCAN）或自组织映射（SOM），以增强聚类的稳定性和效果。

（3）数据增强技术：在数据增强方面，目前仅采用了添加高斯噪声的方法。可以进一步引入更多的数据增强技术，如时间序列数据的平移、缩放、扰动等，以增加模型的泛化能力。

（4）深度学习模型的应用：本研究中采用了传统机器学习算法（如随机森林和支持向量机），未来可以尝试引入深度学习模型（如卷积神经网络、长短期记忆网络等）进行特征提取和分类。这些模型在处理复杂、高维数据时，可能具有更强的表现力和更高的分类准确率。

（5）个性化建模：为了进一步提高活动识别的准确性，可以考虑针对不同个体进行个性化建模，结合个体的生理特征和历史数据，建立更加精准的个体活动状态判别模型。

通过以上改进，我们有望进一步提升模型的性能和实用性，为智能手机传感器数据的无标签分类和人员活动状态识别提供更加有效的解决方案。

参考文献

[1] 殷晓玲，夏启寿，陈晓江，何娟，陈峰.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,第 42 卷(3): 43-50

[2] 何鹏,陈跃跃,扈啸.基于智能手表加速度传感器的人体行为识别[J].电脑与信息技术, 2015, 23(5):4.DOI:10.3969/j.issn.1005-1228.2015.05.002.

[3] 殷瑞刚，魏帅，李晗，于洪.深度学习中的无监督学习方法综述[J].计算机系统应用,2016,第 25 卷(8): 1-7

[4] Gi-Wook Cha, Hyeun-Jun Moon, Young-Chan Kim.Comparison of Random Forest and Gradient Boosting Machine Models for Predicting Demolition Waste Based on Small Datasets and Categorical Variables[J].International journal of environmental research and public health,2021,Vol.18(16): 8530

[5] Sidra Abbas;Shtwai Alsubai;Muhammad Ibrar Ul Haque;Gabriel Avelino Sampedro;Ahmad Almadhor;Abdullah Al Hejaili;Iryna Ivanochko.Active Machine Learning for Heterogeneity Activity Recognition Through Smartwatch Sensors[J].IEEE Access,2024,Vol.12: 22595-22607

[6] 向进勇，王振华，邓芸芸.基于随机森林算法的机器学习分类研究综述[J].人工智能与机器人研究,2024,(1): 143-152

[7] Sara Ashry;Tetsuji Ogawa;Walid Gomaa.CHARM-Deep: Continuous Human Activity Recognition Model Based on Deep Neural Network Using IMU Sensors of Smartwatch[J].IEEE Sensors Journal,2020,Vol.20(15): 8757-877

附录 I: 主要程序/代码名称和

代码环境	操作系统: Windows 11 编程语言: Python 3.9.0 编辑器: PyCharm Community Edition 2022.3.3)＆jupyterlab: 3.5.3 代码详见: 支撑材料

代码和结果清单 1 问题 1 代码名称

问题1（分类模型）.ipynb 结果：表1：问题1结果.csv

代码和结果清单 2 问题 2 代码

问题2（1）.ipynb 问题2（2）.ipynb 问题2（判别模型）.ipynb 结果：表2_问题2结果.xlsx

代码结果清单 3 问题 3 代码

问题3（1）.ipynb 问题3（2）.ipynb 问题3（3）.ipynb 问题3（XGboost）.ipynb 问题3（XGboost-GPU）.ipynb 问题3（神经网络）.ipynb 问题3（随机森林）.ipynb 表3：问题3结果.xlsx