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、国防科技大学电子对抗学院

参赛队员（打印后签名）：1. [illegible]

2. [illegible]

3. [illegible]

指导教师或指导教师组负责人（打印后签名）：[illegible]

日期：2024年7月5

（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

基于机器学习的人体活动识别与人体特征刻画

摘要

随着智能手机的普及，评估手机使用者日常活动状态问题是运动健康应用中的重要研究课题。本文针对实验人员的运动数据分类问题，基于加速度计和陀螺仪传感器数据的分析和机器学习的思想，通过多向加速度、角速度等时间序列提炼指标参数，以活动识别为目标建立分类模型，并使用 Kmeans 算法、Random Forest 算法、XGBoost 算法和 ANOVA 算法对模型进行求解。

针对问题一: 已知数据集中有 12 种运动状态，每种状态含有五组数据样本。在无给定标签的基础上实现数据集各组数据运动状态的分类问题，即以聚类为目的、含约束条件的无监督学习问题。本文通过 IQR 以及均值插补法实现判明、处理异常数据，进一步通过高斯滤波实现数据去噪。进一步构建 K-means 模型，对该模型进行求解，求解结果详见 4.4 节。

针对问题二: 以 Random Forest 算法和 XGBoost 算法为基本框架，融合 PCA 降维算法、Pipeline 和 GridSearchCV 算法对所建立的模型进行优化，得到最佳判别模型；在无标签情况下对比分类模型和判别模型，得出基于 Random Forest 算法的判别模型性能较优，求解结果详见 5.4.2 节；运用判别模型，对 30 个活动样本进行状态判别。求解结果详见 5.4.3 节。

针对问题三: 基于平均合加速度和平均角速度指标，构建基于 OLS 的 ANOVA 模型，得出不同实验人员在同一活动状态下的传感器数据存在显著差异的结论；提出 BMI 指标，结合 XGBoost 模型，建立有监督人员刻画模型并对其准确性和鲁棒性进行分析；运用刻画模型，对五位无人员标签的活动数据进行刻画，求解结果详见 6.4.3 节。

最后, 我们对提出的模型进行全面的评价: 本文的模型贴合实际, 能合理解决提出的问题, 具有实用性强, 算法效率高等特点。

关键词：人体活动识别　Kmeans 聚类　PCA　Random Forest　XGBoost　ANOVA

目录

1 问题综述

1.1 问题背景

随着智能手机的普及，大多数智能手机具备评估用户日常活动消耗热量的功能，能够实现对持有者健康状态的监测。然而，对手机使用人消耗热量的计算往往依赖于智能手机对使用者每天活动状态的记录数据。智能手机测量人体的活动状态，主要依靠内置于其中的运动传感器。通过加速度计和陀螺仪的数据，实现感知手机的姿态、角度和方向变化，并通过相应算法判断识别用户的活动模式。这一实现过程通常被称作人体活动识别（Human Activity Recognition, HAR）[1,2]。然而，加速度计数据容易受到各种噪声的干扰，且不同用户的活动模式存在差异，例如步长、步频、运动习惯等，这都使得通用模型的准确性受到限制。因此需要寻找合适的人体活动识别方法来提高人体活动状态识别的可靠性和准确性。

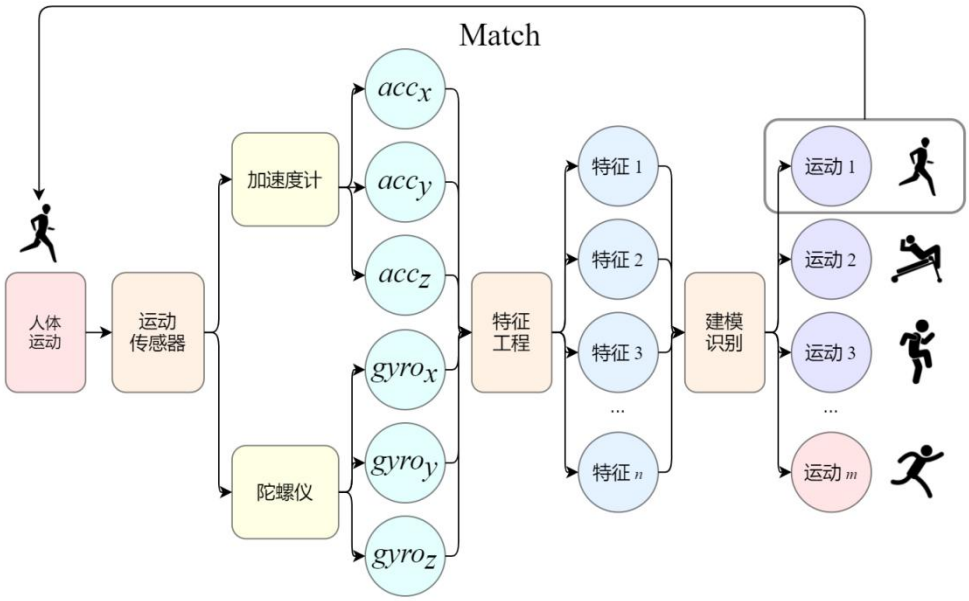

图 1 基于运动传感器的人体活动识别框架

1.2 问题提出

人体活动识别涉及多方面的问题, 往往由传感器数据获取，数据预处理，特征提取，特征选择与降维，模型训练与优化，活动分类与识别[3,4]，结果评估与验证这几个要素构成。论文中需要基于上述要素，解决以下 3 个问题：

(1) 问题 1：查阅相关文献，提取出基于三轴加速度计、陀螺仪的可佩戴运动传感器人体活动特征指标；再根据所提指标，对活动标签缺失的样本数据进行聚类分析，以判断实验人员的动作状态；最后进行特征指标的筛选、融合，剔除冗余特征，对所建立的分类模型进行优化，以提升模型的准确性和鲁棒性。

(2) 问题 2：根据所提取 12 类活动状态的典型特征，基于有活动标签下的样本建立运动状态判别模型，并使用该判别模型进行新样本分类；基于分类模型对实验人员数据进行无活动状态标签下的分类，分析采用分类模型对 12 种活动类型分类时的分类准确度。

(3) 问题 3：探究活动状态与不同人员的年龄、身高、体重特征的差异性和相关性，基于运动传感器数据和上一阶段所提特征，建立人员身体特征刻画模型；根据不同的活动类型，使用该模型对实验人员标签进行判别。

2 模型假设与符号说明

2.1 模型基本假设

(1)运动传感器设备佩戴位置固定且人员活动时无松动、滑落等意外情况；
(2)传感器设备已经过校准，能够准确测量加速度和角速度；
(3)采集的数据中噪声水平于可控范围内；
(4)参与实验的人员活动状态明确，不会出现多种活动交叉的情况；
(5)实验环境稳定，不会有外部干扰影响传感器数据的采集；
(6)特征提取仅基于题设指标，不考虑问题之外指标的过分影响。

2.2 符号说明

本文定义了如下 11 个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
acc_x	沿重力方向(X)的加速度	g
acc_y	沿前进方向(Y)的加速度	g
acc_z	沿身体一侧方向(Z)的加速度	g
$gyro_x$	沿重力轴向(X)的角速度	dps
$gyro_y$	沿前进轴向(Y)的角速度	dps
$gyro_z$	沿身体一侧轴向(Z)的角速度	dps
Range	$\text{Range}=\max(x_i)-\min(x_i)$ (极差)	与 x_i 同量纲
$\text{Kurt}(X)$	$\text{Kurt}(X)=E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]=\frac{E\left[(X-\mu)^4\right]}{\left(E\left[(X-,)^2\right]\right)^2}$ (峰度)	/
γ_1	$\gamma_1=\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]=\frac{\mu_3}{\sigma^3}$ (偏度)	/
ρ_{XY}	$\rho_{XY}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_X}=\frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sqrt{\sigma_X^2\sigma_Y^2}}$ (皮尔逊相关系数)	/
$\varsigma(x_i,y_i)$	$\varsigma(x_i,y_i)=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)\cdot\left(R(y_i)-\overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2\right)}}$ (斯皮尔曼相关系数)	/

3 数据预处理

3.1 指标提取

3.1.1 常规指标

将题供运动传感器、人员类型、人体特征等指标作为基本指标，同时将加速度计三轴进行欧几里得范数求和，得到合加速度，作为合成指标。常规指标如表所示。

表 4 常规指标提取

指标来源	具体指标
运动传感器	加速度计三轴测量的加速度值、 陀螺仪三轴测量的角速度值
类型标签	实验人员编号、动作状态、执行动作次数
人体特征	年龄、身高、体重
合成指标	合加速度

3.1.2 特征指标

根据运动传感器指标的时序特征、频域特征和统计特征，提取特征指标如表所示。

表 5 特征指标提取

度量维度	一级特征指标	二级特征指标
集中趋势	均值	对加速度、角速度、合成加速度指标进行维度特征提取
	四分位点（Q_1,Q_2,Q_3）	
离散程度	标准差	
	极差	
	四分位差	
分布形状	偏度	
	峰度	
相关性	皮尔逊相关系数	
	斯皮尔曼相关系数	
时间序列分析	峰密度	
	过零点数	
信号处理	频谱峰频点	对加速度、合成加速度指标进行维度特征提取
	频谱平均绝对离差	
	能量	
	频谱能量	

3.2 数据清洗

数据清洗需将提取指标数据中的错误、缺失值和不一致性进行识别和修正，以提高数据质量和分析结果的准确性。主要是对上一阶段提取的加速度序列、角速度序列、合加速度序列进行数据清洗。数据清洗将按照步骤进行。

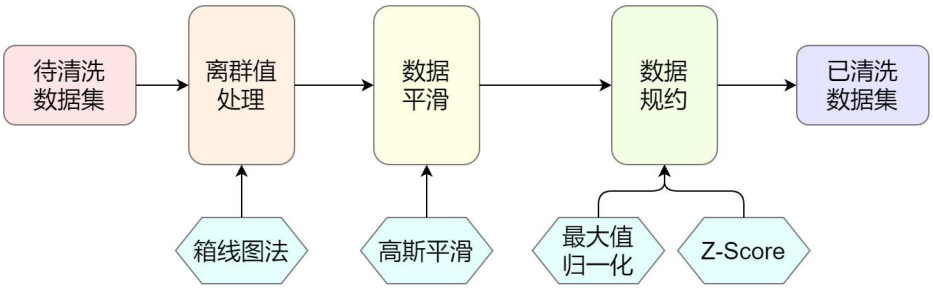

图 2 数据清洗流程

3.2.2 离群值处理

论文采用箱线图法识别和去除离群值。在箱线图中，离群值定义如下

$$IQR = Q_3 - Q_1 \tag{3-1}$$

$$Outliers < Q_1 - 1.5 \times IQR \text{ and } Outliers > Q_3 + 1.5 \times IQR \tag{3-2}$$

其中，Q_1 为第一四分位数，Q_3 为第三四分位数。

通过箱线图四分位数，将加速度、角速度和合加速度指标中的离群值剔除。

3.2.3 数据平滑

数据平滑可用于减少论文提取特征中数据的噪声，使上一阶段所提指标更容易分析和解释。由于加速度、角速度和合加速度指标为时间序列指标，且离群值在 3.2.1 节中已经被处理，论文选择高斯平滑进行数据平滑处理。

高斯平滑在平滑噪声中具有显著优势，并且有利于保持数据的整体趋势和本质形状的完整性。其公式如下：

$$\text{Gaussian}_t = \sum_{i=-k}^{k} G(i) x_{t+i} \tag{3-3}$$

其中，$G(i)$是高斯核函数，k 是窗口大小的一半。

通过箱线图离群值处理、高斯平滑，将加速度、角速度和合加速度指标中的离群值、噪声去除，并体现指标在时间序列上的本质特征。

3.2.4 数据规约

论文中主要采用聚类思想解决问题，即对于清洗后的数据，需要数据规约以消除不同量纲的影响，并加速模型收敛，改善模型的性能，提高模型的泛化能力。本章节从数据的归一化和标准化进行数据规约。

归一化方面，由于加速度、角速度和合加速度指标包含正负数值，论文采用最大绝对值归一化法，原理如下：

$$x' = \frac{x}{\max(|x|)} \tag{3-4}$$

通过最大绝对值归一化法，将指标数据缩放到[1, 1]范围内。

在 Kmeans 聚类和主成分分析中需要用到数据标准化规约，论文中采用 ZScore 法进行数据标准化，其原理如下

$$x' = \frac{x - \mu}{\sigma} \tag{3-5}$$

其中，μ是均值，σ是标准差。

部分特征提取与清洗数据如表所示

表 6 部分特征提取与清洗数据

acc_x_g__mean	acc_x_g__std	acc_x_g__range	acc_x_g__Q1	acc_x_g__Q2
0.901904663	0.301378078	2.348278349	0.703058909	0.850540135
0.900874668	0.295318929	2.51686869	0.693843491	0.847384891
0.900942681	0.308868741	2.427982121	0.690496638	0.840420212
0.911048003	0.30101901	1.990213065	0.700141027	0.850859996
0.911759034	0.317998802	2.355496526	0.693401867	0.840989436
0.911733034	0.172052254	1.167350722	0.789849304	0.899495906
0.916178978	0.192315287	1.325407638	0.794675631	0.907052682
0.911372448	0.143745392	0.971378464	0.813188231	0.904402333
0.911347283	0.180604884	1.11992639	0.785942189	0.889001117
0.913256264	0.240965499	1.428347068	0.762420774	0.899186519
0.957894138	0.286334756	1.978277914	0.799107119	0.957454988
0.954941916	0.289382329	2.289393438	0.816485355	0.954015668
0.971747074	0.237271755	1.710341622	0.837727633	0.965177308
0.966136105	0.244839481	2.402772506	0.844528799	0.960705995
0.961814066	0.262637398	2.081134695	0.821786734	0.962652727
0.913727751	0.259535426	2.290062298	0.720635461	0.89885014

4 问题一分析与建模求解

4.1 问题分析

题目一提供了 3 名实验人员在完成 12 类规定动作后的运动数据。每类运动数据包含 5 组加速度计和陀螺仪的测量值，即每位实验人员共有 60 组数据。需将上述 60 组数据进行分类。

针对该问题，本章基于 3.1 节提取的特征指标，进行多维度 12 簇 Kmeans 聚类，并根据每位实验人员 60 组数据的组成设置约束条件，结合 PCA 进行特征降维和模型优化，将每名人员的 60 组运动数据分类至 12 簇。

4.2 模型准备

4.2.1 Kmeans 聚类算法

Kmeans 聚类是一种常用的无监督学习算法[5]，用于将数据集分成 K 个簇，每个簇由其中心表示。给定一个待聚类的数据集

$$\{x_1, x_2, \ldots, x_n\} \tag{4-1}$$

其中每个数据 x_i 是 d 维向量。Kmeans 聚类的的目标是将这些数据点划分为 K 个簇 $C_1, C_2, \ldots, C_k$，以最小化簇内数据点到簇中心的距离总和，以实现数据集聚类。

(1) Kmeans 聚类步骤

Step 1 初始化

随机选择 K 个初始簇中心

$$\{\mu_1, \mu_2, \ldots, \mu_K\} \tag{4-2}$$

Step 2 簇分配

对于每个数据点 x_i，计算其到每个簇中心的距离，并将其分配到最近的簇。这个过程可以表示为

$$C_k = \{x_i : \| x_i - \mu_k \|^2 \le \| x_i - \mu_j \|^2 \quad \forall j, 1 \le j \le K\} \tag{4-3}$$

其中 $\|\cdot\|$ 表示欧几里得距离。

Step 3 更新簇中心

对于每个簇 C_k，计算其新的簇中心

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{4-4}$$

其中 $|C_k|$ 表示簇 C_k 中的数据点数量。

重复步骤 2 和步骤 3，直到簇中心不再发生显著变化，即 4.2.1.2 节中损失函数最小，或者达到预定的迭代次数。

(2) Kmeans 聚类目标

Kmeans 聚类的目标是最小化以下损失函数：

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2 \tag{4-5}$$

即所有数据点到其所属簇中心的距离平方和最小。

4.2.2 PCA 算法

主成分分析是一种用于降维的统计方法[6]。它通过将数据投影到一个新的坐标系中，使得投影后的数据在新的坐标系中具有最大的方差。具体实现步骤如下：

Step 1 标准化数据

对于给定的数据矩阵 X（大小为 $n\times d$，其中 n 是样本数，d 是特征数），对每个特征进行标准化，使其均值为 0，方差为 1。

$$X_{\text{std}} = \frac{X - \mu}{\sigma} \tag{4-6}$$

其中，μ是每个特征的均值，σ是每个特征的标准差。

Step 2 计算协方差矩阵

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2 \tag{4-7}$$

Step 3 计算协方差矩阵的特征值和特征向量

设协方差矩阵 $\sum$ 的特征值为 $\lambda_1, \lambda_2, \ldots, \lambda_d$，对应的特征向量为 $v_1, v_2, \ldots, v_d$

Step 4 选择主成分

选择前 k 个最大的特征值对应的特征向量，形成特征向量矩阵 V_k (大小为 $d\times k$)。

Step 5 转换数据到新的空间

$$Z = X_{\text{std}} V_k \tag{4-8}$$

4.2.3 t-SNE 算法

t-SNE 是一种非线性降维算法，主要用于高维数据的可视化。它通过最小化高维空间和低维空间中相似度分布之间的 KL 散度，将高维数据嵌入到低维空间中（通常是二维或三维），以便更直观地观察数据的结构和分布。t-SNE 算法实现步骤如下：

Step 1 计算高维空间中点对的相似度

对于每一对高维数据点 x_i 和 x_j，计算条件概率 $p_{j|i}$，表示在给定点 x_i 的情况下选择点 x_j 的概率。使用高斯分布计算 $p_{j|i}$

$$p_{j|i} = \frac{\exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\| x_i - x_k \|^2}{2\sigma_i^2}\right)} \tag{4-9}$$

其中，σ_i 是点 x_i 的高斯分布的标准差。

Step 2 对称化相似度

计算联合概率 $p_{i|j}$，表示选择点 x_i 和 x_j 的联合概率

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{4-10}$$

其中，n 是数据点的总数。

Step 3 计算低维空间中点对的相似度

对于每一对低维数据点 y_i 和 y_j，计算相似度 q_{ij}，使用 t 分布计算

$$q_{ij}=\frac{\left(1+\| \mathbf{y}_i-\mathbf{y}_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\| \mathbf{y}_k-\mathbf{y}_l\|^2\right)^{-1}} \tag{4-11}$$

Step 4 最小化 KL 散度

最小化高维空间和低维空间中相似度分布之间的 KL 散度

$$\mathrm{KL}(P\parallel Q)=\sum_{i\neq j}p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{4-12}$$

使用梯度下降法优化低维嵌入 y_i 。

4.3 模型建立

为解决问题一，建立 Kmeans-PCA 降维聚类模型，以将 3 名实验人员的各 60 组数据进行 12 种运动状态的分类。从题设角度而言，每名实验人员的 60 组数据分为 12 类动作，每类动作 5 组数据，即为模型求解的最优效果。

Kmeans-PCA 降维聚类模型的目标是通过逐步增加 PCA 降维后的主成分数量，结合 Kmeans 聚类，找到一个最优的特征空间，使得聚类后的样本在二维空间中的分散效果最均匀。模型求解步骤如下：

Step 1 模型初始化

设定初始样本分类的总体方差 s_{initial}，m 表示特征数据维度，对应指标集为 M。维度寄存器 $\mathrm{M_{reg}}$ 为原始维度特征指标 $\mathrm{M}^{(1)}$。循环次数 i=1，总循环次数为 100，$m^{(1)}$=102

Step 2 输入 $m^{(i)}$维特征指标 $\mathrm{M}^{(i)}$

Step 3 Kmeans 聚类

将 $m^{(i)}$维特征原始数据进行簇为 12 的 Kmeans 聚类，每簇所包含样本数量为 N_{ci}。得到簇和簇分配为

$$\{C_1^{(i)},C_2^{(i)},...,C_{12}^{(i)}\} \tag{4-13}$$

$$\{N_{C_1}^{(i)},N_{C_2}^{(i)},...,N_{C_{12}}^{(i)}\} \tag{4-14}$$

Step 4 t-SNE 降维

使用 t-SNE 降维算法将原始数据降维到二维空间，生成二维图，观察样本的分散效果，并通过总体样本方差评价所分 12 类的样本数量分散程度。

$$s^{(i)}=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(N_{C_j}^{(i)}-\overline{N^{(i)}})^2} \tag{4-15}$$

Step 5 样本分散程度对比

若方差 $s^{(i)}$大于 s_{initial}，则直接执行 Step6；若 $s^{(i)}$小于 s_{initial}，则 s_{initial}= $s^{(i)}$，将该特征维度 $\mathrm{M}^{(i)}$计入 $\mathrm{M_{reg}}$，执行 Step 6。

Step 6 PCA 降维

对 $m^{(i)}$维特征原始数据 $M^{(i)}$进行 PCA 降维至 $m^{(i+1)}$维 $M^{(i+1)}$。若 i=100，跳出循环；否则，i=i+1，进行 Step 2 操作。

最终，得出样本总方差最小的分类结果，即结果接近最优解：每名实验人员的 60 组数据分为 12 类动作，每类动作 5 组数据。

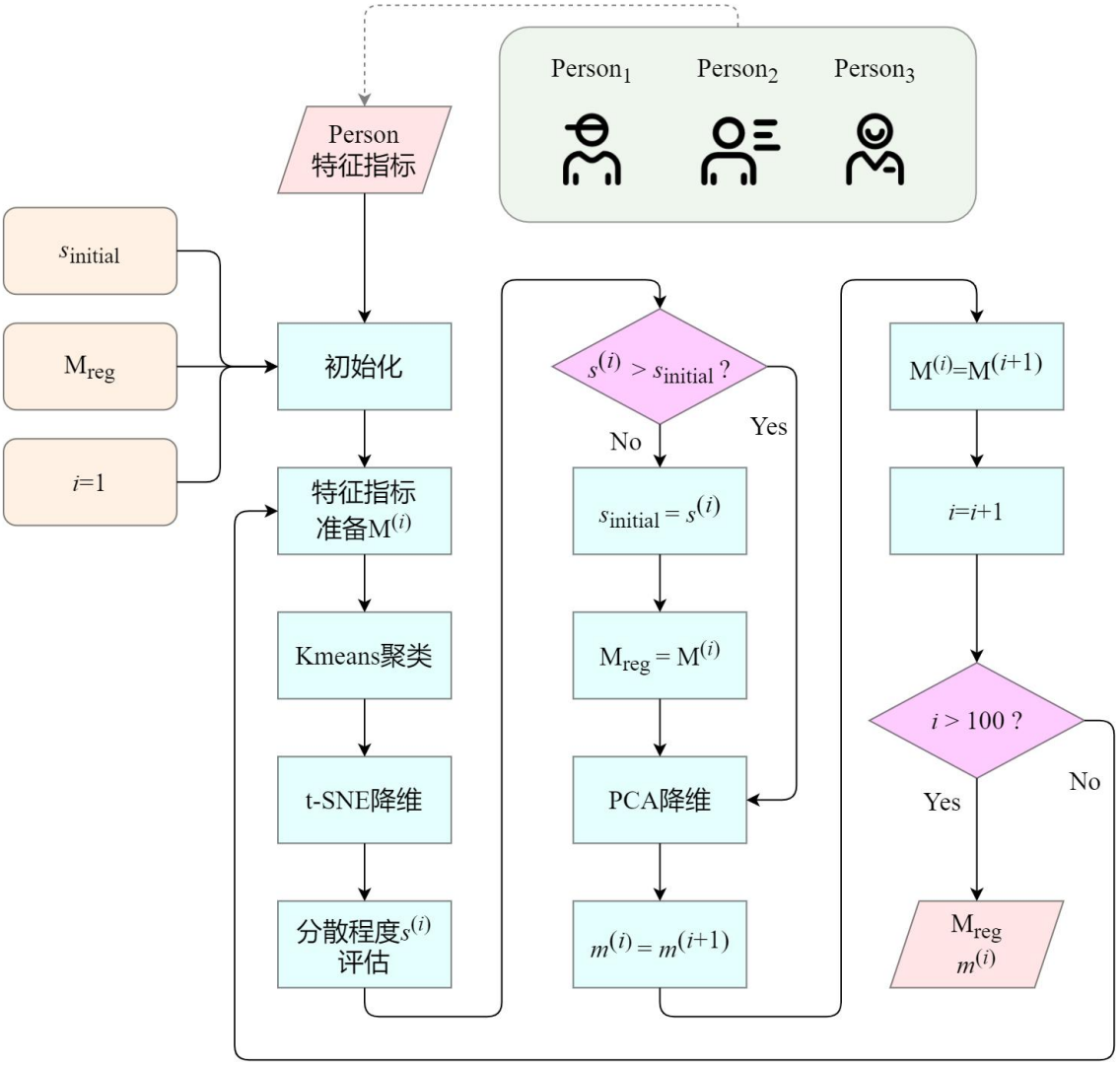

图 3 Kmeans-PCA 降维聚类模型

4.4 模型求解

根据第 3 章处理数据及所提取指标，分别将每名实验人员的 60 组数据按照所提取 102 维指标处理，作为输入端分别输入 Kmeans-PCA 降维聚类模型，得到 3 名实验人员的期望迭代次数、最优主成分维度和聚类结果。

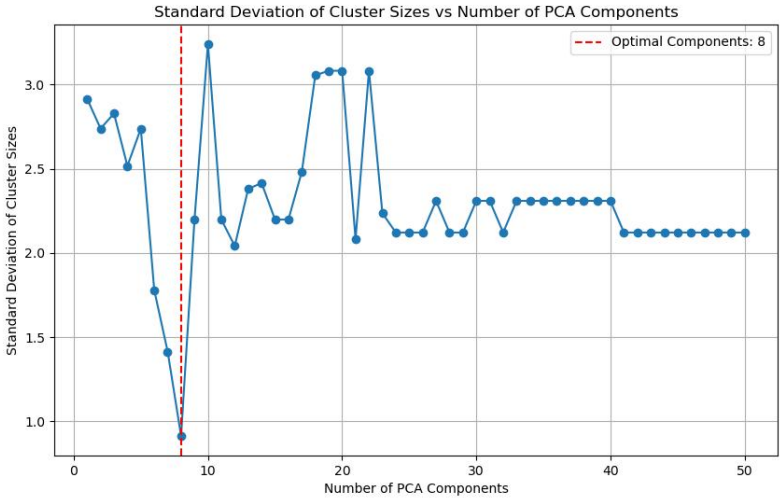

图 4 Person1 降维过程

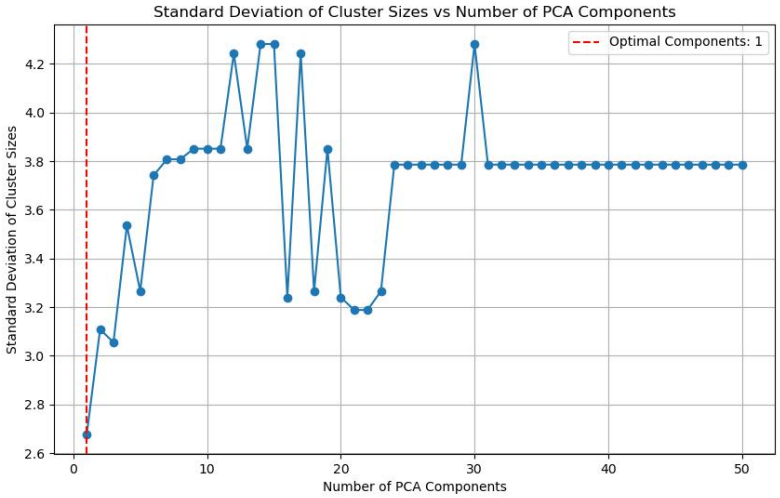

图 5 Person2 降维过程

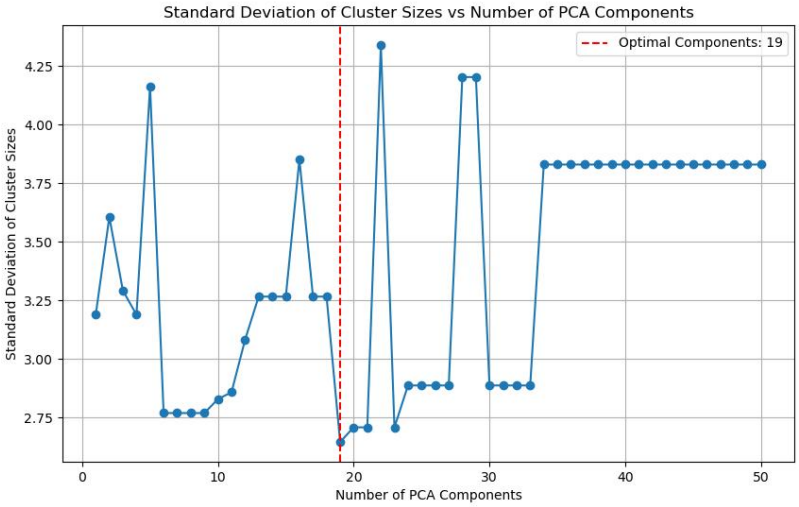

图 6 Person3 降维过程

由图可知，对于 Person 1、Person 2、Person 3 的聚类，分别选择 8 个、1 个、19 个 PCA 成分时，簇大小的标准差最小，即期望值最高。

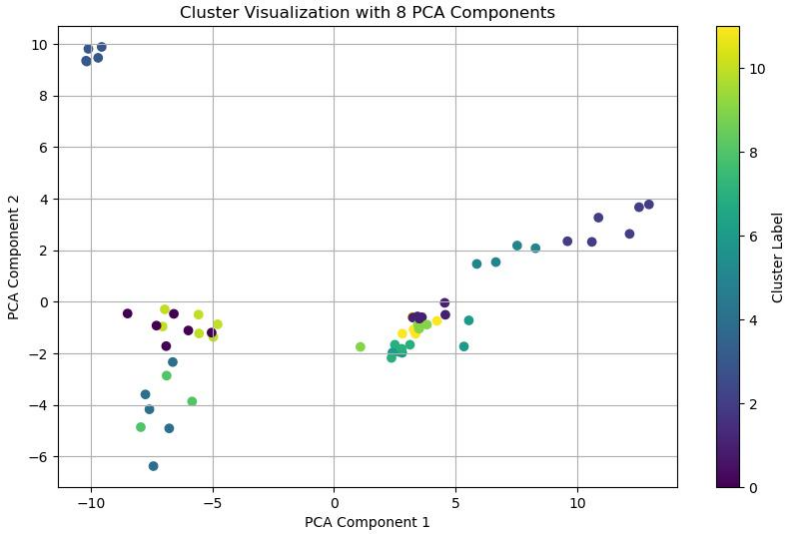

图 7 Person1 聚类效果

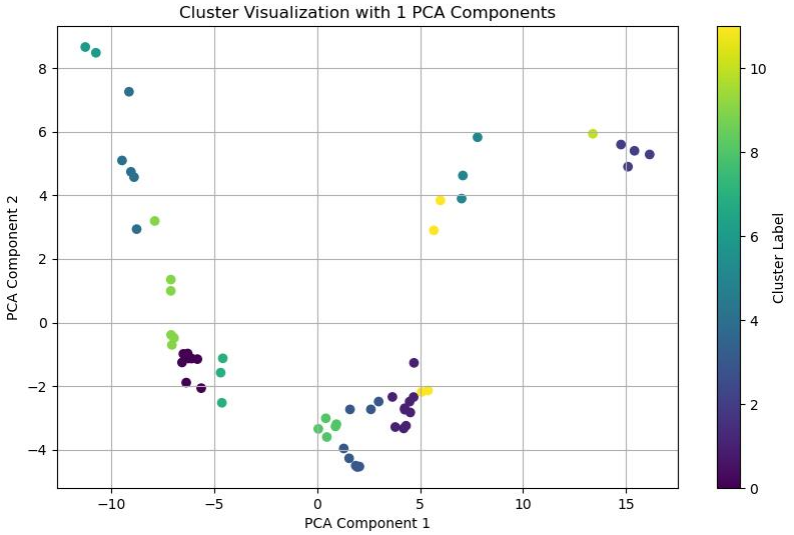

图 8 Person2 聚类效果

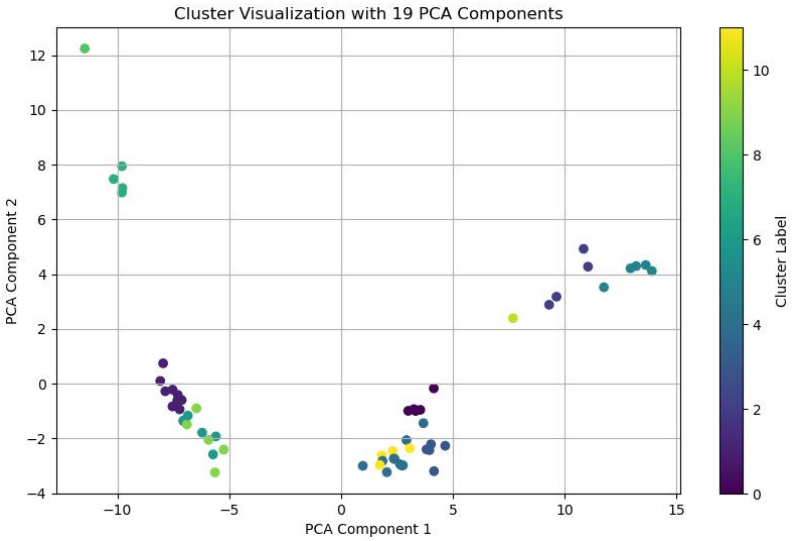

图 9 Person3 聚类效果

由图可知，对于 Person 1、Person 2、Person 3 的聚类，分别选择 8 个、1 个、19 个 PCA 成分时，基于该点处的聚类效果最好，且不同簇在前两个主成分的平面上有良好的分离效果。

综上，得出 3 位实验人员的动作聚类情况，如表所示

表 7 实验人员动作聚类结果

分类	Person 1	Person 2	Person 3
第 1 类	33/38/42/48/58	7/24/33/44/57	10/13/27/29/51
第 2 类	5/8/25/56/59	3/20/38/51/60	8/23/39/42/52
第 3 类	14/17/28/41/47	16/21/29/40/43	9/14/30/37/43
第 4 类	21/29/12/46/53	47/26/10/2/12	56/35/44/49/60
第 5 类	27/31/37/49/52	19/23/30/37/46	11/19/24/25/31
第 6 类	19/26/43/44/50	1/4/48/49/50	18/22/57/58/59
第 7 类	1/2/15/23/36	13/27/28/34/42	4/5/36/41/53
第 8 类	4/13/18/20/39	5/11/22/54/56	2/6/12/21/48
第 9 类	7/9/10/24/57	6/15/25/35/58	38/17/33/34/47
第 10 类	22/32/34/35/40	31/32/39/52/59	3/32/50/54/55
第 11 类	3/6/11/16/55	8/9/36/41/55	1/7/26/40/45
第 12 类	30/45/51/54/60	14/17/18/45/53	15/16/20/28/46

5 问题二分析与建模求解

5.1 问题分析

题目二由三个子问题构成，一是要求提取 12 类人员活动状态的典型特征，建立判别模型；二是通过问题 1 和建立的模型，对 10 名实验人员数据进行无活动状态标签下的分类，并进行比较，分析采用分类模型对 12 种活动类型分类时的分类准确度；三是结合一、二子问题的结论，采用建立的判别模型对附件 3 中的 30 个活动样本进行状态类型判别。

针对第一个子问题，首先选择第三章提出的所有指标，分别运用 XGBoost 算法、随机森林算法和 PCA 进行有监督分类模型的建立，并通过 Pipeline 和 GridSearchCV 对所建立的模型进行优化，得到最佳分类模型。

针对第二个子问题，将 10 名实验人员数据的活动状态标签剔除，提取特征指标，输入模型一进行分类。对比模型一分类结果和真实标签，得出对 12 种活动类型分类时的分类准确度，并与判别模型对比性能。

针对第三个子问题，提取附件 3 中的 30 种活动样本的特征指标，输入到第一子问题建立的判别模型，得出活动样本的分类标签。

5.2 模型准备

5.2.1 Random Forest 算法

随机森林（Random Forest）是一种集成学习方法，通过构建多个决策树在训练时进行合并来改进预测的准确性和控制过拟合。随机森林的基本原理是每个决策树都尽可能的大程度上独立，并从每个树中得到一个预测结果，然后通过投票的方式得到最终的分类结果。

在有 B 棵决策树组成的随机森林中，对于输入样本 x，第 k 棵树的预测结果为 $h_k(x)$。随机森林的最终预测结果为所有决策树预测结果的投票分类。具体实现步骤如下：

Step 1 自助采样

从数据集中使用自助采样的方式选取 n 个样本，允许重复选择相同样本。

Step 2 建立决策树

• Start →对于每一个自助采样得到的样本集，建立一棵决策树。在建立决策树的过程中，引入随机属性选择。

• Case →当节点分裂时，从属性的一个随机子集中选择一个最优属性。

• End →分裂过程一直进行，直到满足既定停止条件。

Step 3 结果预测

在分类问题中，选择多数分类结果作为最后的结果，即随机森林的预测结果是所有决策树预测结果的众数。

$$\hat{y} = \text{MajorityVote}(h_1(x), h_2(x), \ldots, h_B(x)) \tag{5-1}$$

5.2.2 Pipeline 算法

Pipeline 算法是机器学习工作流中常用的一种技术，它能够将多个处理步骤串联在一起，形成一个有序的流程。如有 n 个待处理函数 $f_1, f_2, ..., f_n$，则 Pipeline 处理结果为

$$\text{Pipeline}(f_1, f_2, ..., f_n) = f_1(f_2(...(f_n))) \tag{5-2}$$

其中，x 是输入数据。

建模中使用 Python 中的 scikit-learn 库实现 Pipeline 算法。

5.2.3 GridSearchCV 算法

GridSearchCV 通过穷举搜索的方法，遍历所有可能的超参数组合，并使用交叉验证来评估每组超参数组合的表现，从而选择出最佳的超参数。

输入模型 M 和一组超参数θ后，GridsearchCV 的目标是通过交叉验证来找到最优的参数组合θ^*。其优化目标可以表示为

$$\theta^* = \arg\max_\theta \frac{1}{k}\sum_{i=1}^{k}\text{Score}(M_\theta, X_i, y_i) \tag{5-3}$$

其中，k 是交叉验证的折数，$\text{Score}(M_\theta, X_i, y_i)$ 是模型在第 i 折上的评分。

5.2.4 XGBoost 算法

XGBoost 的核心思想是通过集成多个弱分类器，每一步都专注于修正前一步的错误，从而逐步提升整体模型的准确性。具体步骤如下：

Step 1 数据准备

首先，我们需要准备好训练数据集

$$\{(x_i, y_i)\}_{i=1}^{n} \tag{5-4}$$

其中 x_i 是特征向量，y_i 是标签。

Step 2 初始化模型

设定 XGBoost 初始值

$$\hat{y}_i^{(0)} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5-5}$$

Step 3 定义目标函数

目标函数由损失函数和正则化项组成

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t}\Omega(f_k) \tag{5-6}$$

其中，L 是损失函数是正则化项，定义为：

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \tag{5-7}$$

这里，T 是树的叶子节点数，w_j 是叶子节点的权重，γ 和 λ 是正则化参数。

Step 4 梯度提升

在第 t 轮迭代中，我们构建一个新的决策树 f_t，其目标是最小化当前模型的损失。我们使用梯度提升的方法，计算每个样本的梯度。

$$g_i^{(t)} = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \tag{5-8}$$

$$h_i^{(t)} = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} \tag{5-9}$$

Step 5 构建决策树

构建决策树时，我们通过贪心算法选择分裂点，最大化增益：

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i^{(t)})^2}{\sum_{i\in I_L} h_i^{(t)} + \lambda} + \frac{(\sum_{i\in I_R} g_i^{(t)})^2}{\sum_{i\in I_R} h_i^{(t)} + \lambda} - \frac{(\sum_{i\in I} g_i^{(t)})^2}{\sum_{i\in I} h_i^{(t)} + \lambda}\right] - \gamma \tag{5-10}$$

其中，I_L 和 I_R 分别是左子节点和右子节点的样本集合。

Step 6 更新预测值

对于每个叶子节点，我们计算其权重:

$$w_j = -\frac{\sum_{i\in I_j} g_i^{(t)}}{\sum_{i\in I_j} h_i^{(t)} + \lambda} \tag{5-11}$$

然后更新预测值:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{5-12}$$

其中，η是学习率。

Step 7 重复迭代至模型结束

重复 Step 4 到 Step 6，直到达到预定的迭代次数或其他停止条件.

5.3 模型建立

利用 GridSearchCV 和 Pipeline 自动化地找到最优的 PCA 成分，并分别与随机森林分类器、XGBoost 组合，建立分类模型，从而提高模型的预测准确性和泛化能力。基于 Python 的具体实现步骤如下：

Step 1 划分数据集

将数据集区分特征指标X和目标变量Y，并划分为训练集(X_{train}, Y_{train})和测试集(X_{test}, Y_{test})，并保证结果的可重复。其中测试集占比 20%。

Step 2 定义 PCA 和随机森林分类器并设置网格搜索参数

PCA 保留的成分比例参数分别为 [0.7, 0.8, 0.9, 0.95]，随机森林分类器的估计器数量参数分别为 [50, 100, 200]。

Step 3 使用 Pipeline 组合 PCA 和随机森林分类器（或 XGBoost）

Step 4 模型优化

将数据集分成 5 个互斥的子集，即进行 5 折交叉验证，并将准确率作为评价指标，利用网格搜索优化模型，以得最佳参数和最佳分数。

Step 5 预测评估

得到分类报告，包括精确率、召回率、F1 分数，对模型预测进行评估，并输出混淆矩阵，显示预测结果的详细信息。同时，判断随机森林模型与 XGBoost 模型优劣。

通过上述步骤，建立 12 类人员活动状态的判别模型。

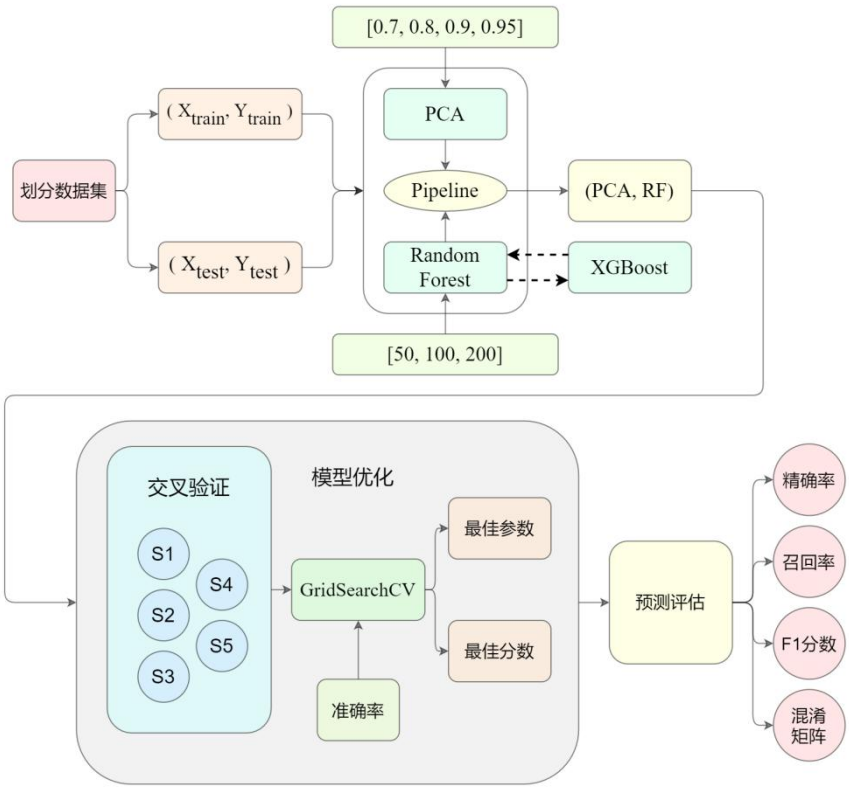

图 10 判别模型流程图

5.4 模型求解

5.4.1 子问题一

根据第 3 章所提特征指标和清洗数据，对数据进行有监督下的 Random Forest 和 XGBoost 模型搭建，并基于 PCA 降维、GridSearchCV 和 Pipeline 得到最佳判别模型。

(1) 基于对 Random Forest 优化的判别模型

将特征指标作为自变量，活动标签作为因变量，构成数据集基本元素。数据划分后输入判别模型，得到 PCA 降维效果图、混淆矩阵和模型准确率评价表。

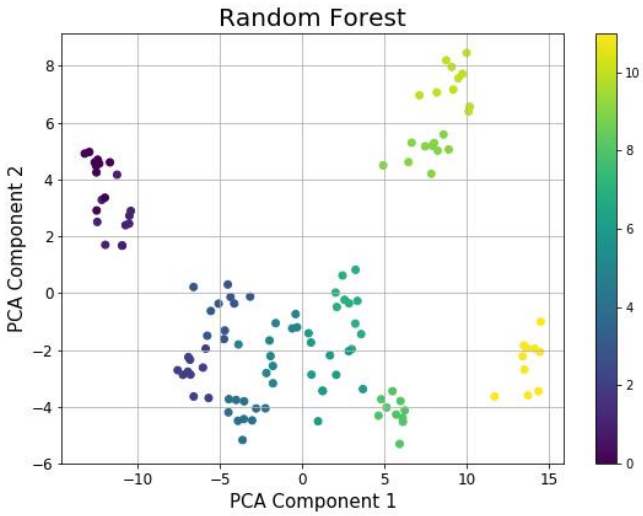

图 11 随机森林分类结果

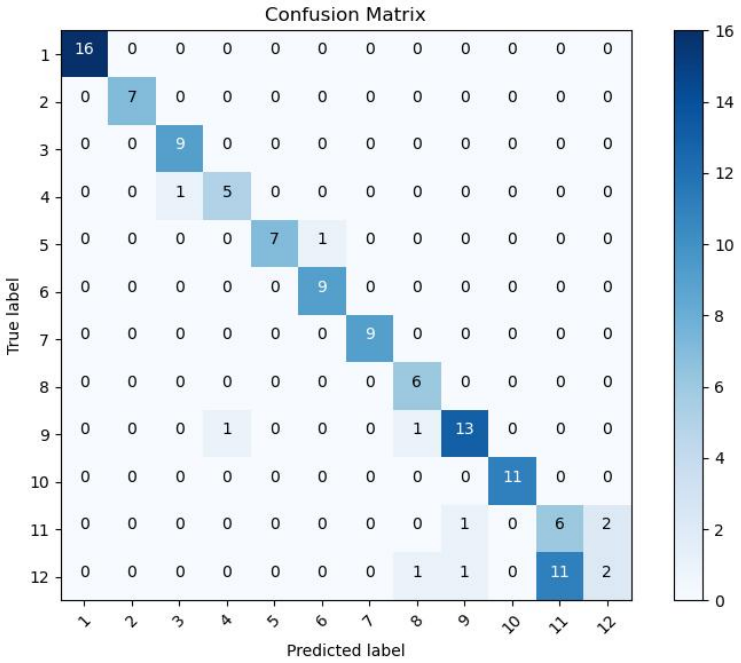

图 12 随机森林分类混淆矩阵

基于 PCA 降维、GridSearchCV 和 Pipeline 的优化，Random Forest 模型在混淆矩阵中，大多数类别的分类结果集中在对角线上，说明模型在这些类别上的分类效果较好，类别 1、2、3、6、7、8、10 的分类结果完全正确。

但类别 11 和 12 的误分类现象较为明显，特别是类别 12，有多个样本被误分类为其他类别，表明模型在这两个标签上的分类能力较弱。

表 8 随机森林模型效果

指标	最佳效果
Best parameters	pca_n_components: 0.95　rf_n_estimators: 200
Best cross-validation accuracy	0.8958333
Accuracy	0.8666667

从参数调优和模型评估表中可以看出：

A. 参数调优过程中，进行了 5 折交叉验证，共测试了 12 种不同的参数组合，总共进行了 60 次拟合；

B. 最佳参数搜索中，找到的最佳参数组合是 PCA 组件数为 95%，随机森林的树的数量为 200；

C. 最佳交叉验证准确率表明，最佳参数组合下的交叉验证准确率约为 89.58%；

D. 测试准确率表明，在独立测试集上的准确率约为 86.67%。

表 9 随机森林各指标判别效果 1

类别	精确度	召回率	F1 分数	支持度
1	1.00	1.00	1.00	16
2	0.88	1.00	0.93	7
3	1.00	1.00	1.00	9
4	0.86	1.00	0.92	6
5	1.00	0.88	0.93	8
6	1.00	1.00	1.00	9
7	1.00	1.00	1.00	9
8	1.00	1.00	1.00	6
9	0.88	0.93	0.90	15
10	1.00	1.00	1.00	11
11	0.38	0.67	0.48	9
12	0.67	0.27	0.38	15

表 10 随机森林各指标判别效果 2

类别	精确度	召回率	F1 分数	支持度
accuracy	0.87	-	-	120
macro avg	0.89	0.90	0.88	120
weighted avg	0.88	0.87	0.86	120

从分类报告表中可以看出：

A. 类别 1、3、6、7、8、10 的性能较好，精确度、召回率和 F1 分数都达到了 100%。

B. 类别 2、4、5、9 的性能较优，但分数略有不足。

C. 类别 11 和 12 的性能较差，尤其是类别 12 的召回率只有 27%，表明模型在这两个类别上的识别能力较弱。

综上，基于 PCA 降维、GridSearchCV 和 Pipeline 优化的 Random Forest 模型已经达到较为理想的判别效果。

(2) 基于对 XGBoost 优化的判别模型

同理(1)，将数据集划分后输入判别模型，得到 PCA 降维效果图、混淆矩阵和模型准确率评价表。

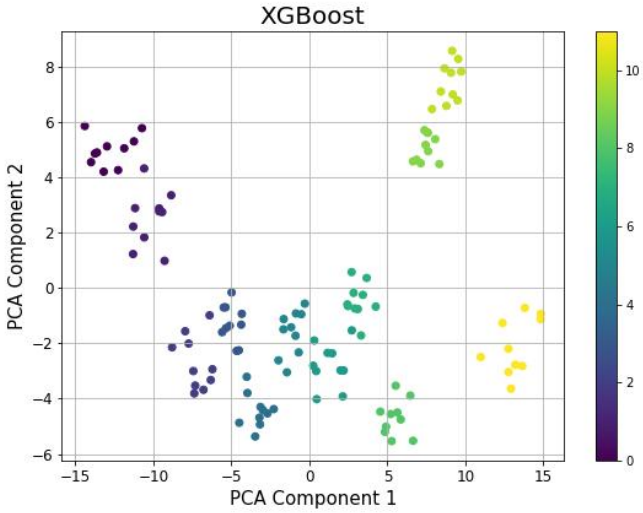

图 13 XGBoost 分类结果

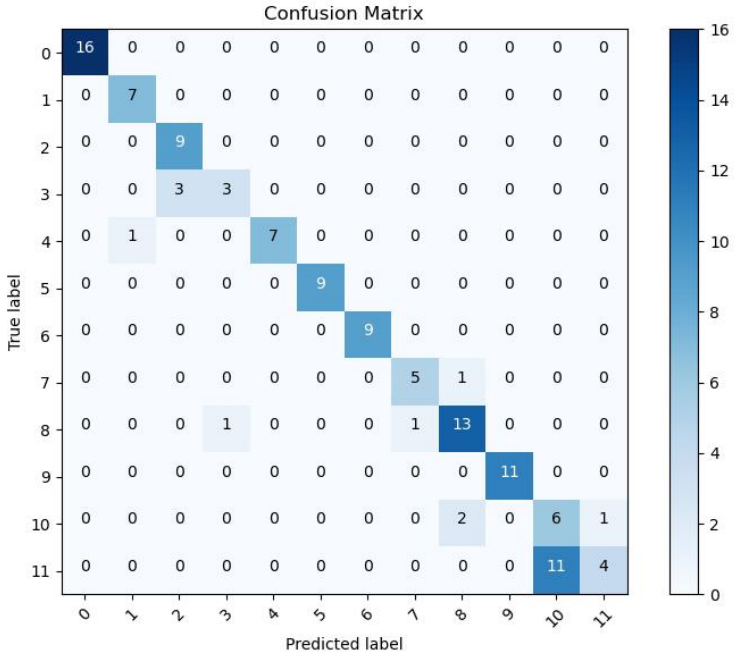

图 14 XGBoost 分类混淆矩阵

基于 PCA 降维、GridSearchCV 和 Pipeline 的优化，XGBoost 模型在混淆矩阵中，根据对角线分析，大多数类别上的分类效果很好，特别是类别 1、2、3、6、7、10 的分类效果。

类别 11 和 12 的分类效果较差，需要进一步改进模型。

表 11 XGBoost 模型效果

指标	最佳效果
Best parameters	pca_n_components: 0.95　xgb_n_estimators: 200
Best cross-validation accuracy	0.8778333
Accuracy	0.8250000

从参数调优和模型评估表中可以看出：

A. 参数调优过程中，进行了 5 折交叉验证，共测试了 12 种不同的参数组合，总共进行了 60 次拟合；

B. 最佳参数搜索中，找到的最佳参数组合是 PCA 组件数为 95%，XGBoost 的 Cart 树数量为 200；

C. 最佳交叉验证准确率表明，最佳参数组合下的交叉验证准确率约为 89.78%；

D. 测试准确率表明，在独立测试集上的准确率为 82.5%。

表 12 XGBoost 各指标判别效果 1

类别	精确度	召回率	F1 分数	支持度
0	1.00	1.00	1.00	16
1	0.88	1.00	0.93	7
2	0.75	1.00	0.86	9
3	0.75	0.50	0.60	6
4	1.00	0.88	0.93	8
5	1.00	1.00	1.00	9
6	1.00	1.00	1.00	9
7	0.83	0.83	0.83	6
8	0.81	0.87	0.84	15
9	1.00	1.00	1.00	11
10	0.35	0.67	0.46	9
11	0.80	0.27	0.40	15

表 13 XGBoost 各指标判别效果 2

类别	精确度	召回率	F1 分数	支持度
accuracy	0.82	-	-	120
macro avg	0.85	0.83	0.82	120
weighted avg	0.86	0.82	0.82	120

从分类报告表中可以看出：

D. 类别 0、1、2、4、5、6、9 的分类效果较好，精确率、召回率和 F1-Score 较高

E. 类别 3、10 和 11 的分类效果较差，特别是类别 11，召回率只有 27%。

综上，基于 PCA 降维、GridSearchCV 和 Pipeline 优化的 XGBoost 模型判别效果较好。综合考量后，选择基于 Random Forest 算法的判别模型。

5.4.2 子问题二

利用所提特征指标和问题一模型，对 Person4 至 Person13 进行 PCA-Kmeans 降维聚类分析，分别得到的聚类效果如图所示

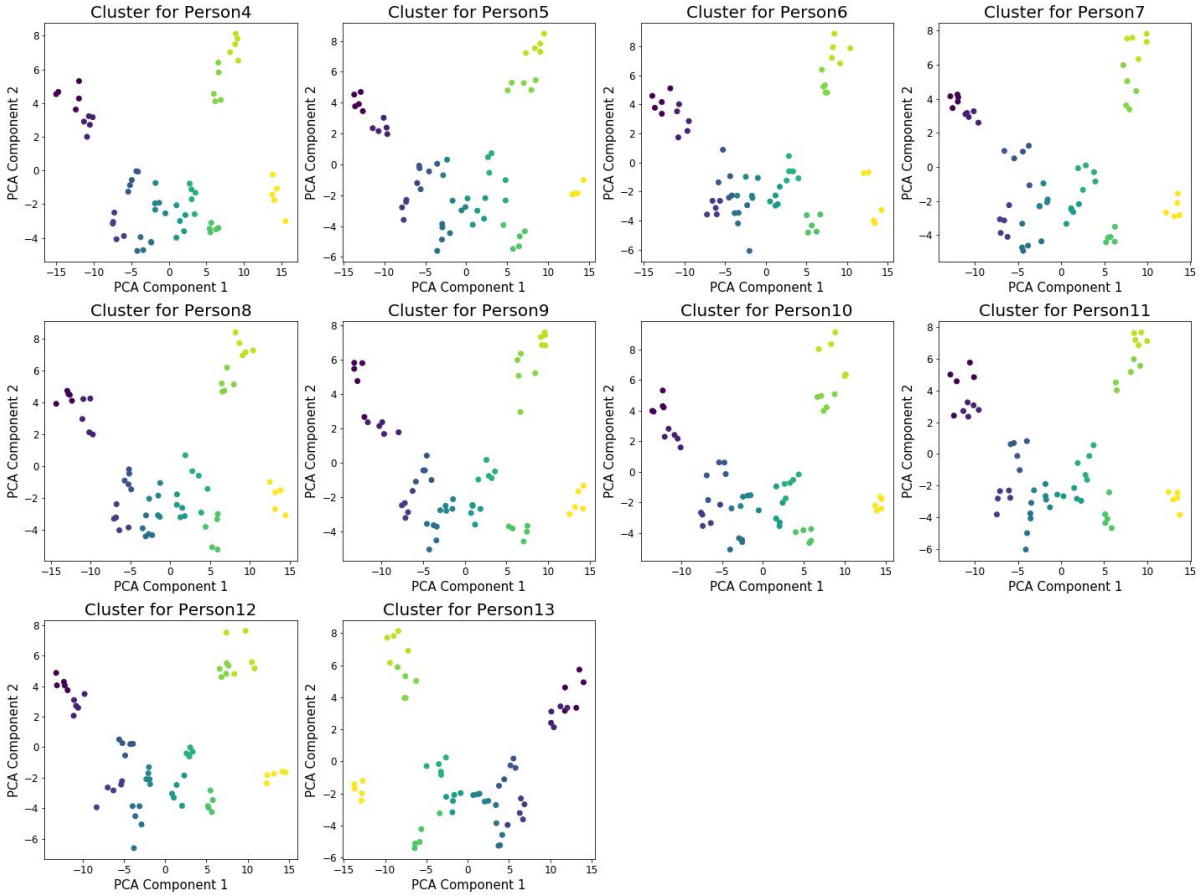

图 15 PCA-Kmeans 降维聚类效果

无监督聚类后，对应样本活动标签，取每类中占比最高样本为该活动分类，计算基于分类模型的准确率，表格内为某人员在某类中的正确个数。结果如表所示。

表 14 PCA-Kmeans 正确值统计

分类\Person	4	5	6	7	8	9	10	11	12	13
第 1 类	2	3	4	5	4	4	3	5	4	2
第 2 类	3	2	5	4	2	3	2	4	5	3
第 3 类	4	4	2	3	2	3	4	5	1	4
第 4 类	5	5	3	2	4	4	5	4	3	5
第 5 类	4	4	4	5	4	4	4	5	4	4
第 6 类	5	5	5	4	2	3	5	4	5	5
第 7 类	4	4	4	5	2	3	4	5	3	4
第 8 类	5	5	5	1	4	4	5	4	5	5
第 9 类	4	4	4	5	4	4	4	5	1	4
第 10 类	5	3	5	4	4	4	5	4	5	5
第 11 类	4	4	4	5	4	4	4	5	2	4
第 12 类	2	3	4	5	2	3	3	5	4	2

由聚类图得出，运用分类模型对 Person4 至 Person13 进行分类，不同簇之间有明显的分离，部分簇之间存在重叠的情况。聚类效果总体良好。

由表得出，各人员准确率分布如下

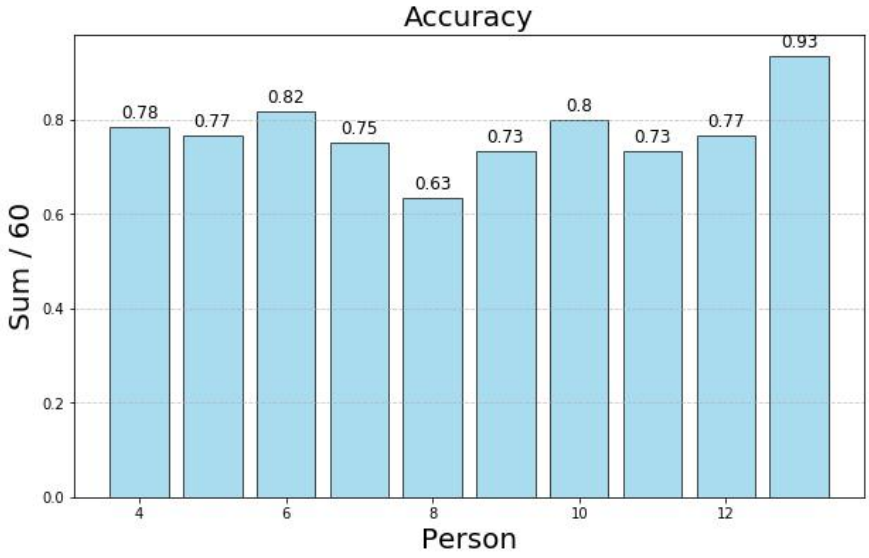

图 16 各类人员准确率柱状图

表 15 各类人员准确率

人员	准确率
Person 4	0.78333333
Person 5	0.76666667
Person 6	0.81666667
Person 7	0.75
Person 8	0.63333333
Person 9	0.73333333
Person 10	0.8
Person 11	0.73333333
Person 12	0.76666667
Person 13	0.93333333

分析可知，聚类效果除 Person 13 外，准确率均低于 Random Forest 模型和 XGBoost 模型，且准确率因人而异，波动较大。故判别模型准确率要高于分类模型，且 Random Forest 模型要优于 XGBoost 模型。

5.4.3 子问题三

基于子问题一建立的两类判别模型，分别对附件 3 中的 30 种活动样本进行判别，判别结果如下表

表 16 判别结果

SY	Random Forest	XGBoost
1	5	5
2	1	1
3	7	7
4	10	9
5	7	7
6	10	10
7	2	2
8	6	6
9	7	7
10	10	10
11	10	9
12	7	7
13	4	4
14	3	3
15	4	4
16	3	3
17	4	4
18	5	5
19	8	10
20	8	8
21	6	6
22	2	2
23	8	10
24	5	2
25	2	2
26	9	8
27	8	8
28	5	2
29	6	6
30	5	5

6 问题三分析与建模求解

6.1 问题分析

问题三由三个子问题构成，一是比较不同实验人员在同一活动状态下的传感器数据，判断是否存在显著差异；二是加入实验人员的身体特征指标，评估年龄、身高、体重对活动状态数据的影响；三是基于运动传感器数据和论文所提特征指标，建立人员身体特征刻画模型，并用此刻画模型，针对未知人员样本的特征指标进行实验人员的溯源。

针对第一个子问题，首先选取合加速度均值与合角速度均值作为特征指标，对不同编号的实验人员采用方差分析、置信区间分析、柱状图效果可视化来评估不同实验人员之间的差异性。

针对第二个子问题，首先根据参考资料，从 13 位实验人员的身高和体重数据中提取 BMI（身体质量指数）作为高级特征。接下来，对年龄和 BMI 两个特征进行分箱操作，最后得到每位实验人员的二维向量标签，为后续的数据分析和模型训练提供便利。运用 XGBoost 算法进行有监督分类模型的建立，并通过 Pipeline 和 GridSearchCV 找到最佳参数的 XGBoost 回归器来预测年龄和 BMI 标签。同时使用 PCA 降维减少数据的维度。最后计算训练集和测试集的均方误差（MSE）和 R2 得分来评估模型表现。

针对第三个子问题，首先提取第三章所选的所有指标，然后输入第二个子问题所建立的模型以进行实验人员的判别，输出结果。

6.2 模型准备

6.2.1 特征选取

(1) 子问题一

综合考虑以下因素，选择合加速度和合角速度作为分析不同人员同一活动状态是否存在差异的指标

A. 合加速度和合角速度是对各个方向上加速度和角速度的综合度量，能有效地反映整体运动强度和旋转情况。该指标结合了三个轴向的数据，避免单个轴向可能出现的偏差和不完整性。

B. 合加速度和合角速度将三轴数据降维为单一指标，对模型效率进行优化，同时保留了主要信息。

C. 人体活动识别的研究和实际应用中，合加速度和合角速度被广泛使用并证明了其有效性，并在多种活动识别任务中显示出较好的性能。

(2) 子问题二

综合考虑以下因素，将每位实验人员的由年龄和 BMI 进行分箱操作，得到二维向量标签

A. 分箱将连续的年龄和 BMI 数据转换为离散的类别或区间，使得数据更加抽象和简化。减少模型复杂性，降低因数据细节带来的噪声影响，同时便于模型处理和理解。

B. 通过合理的分箱策略，可以将原始数据转换为更有信息量的特征，提升模型的预测能力和泛化能力。

C. 当数据被分箱后，模型更倾向于学习每个分箱内的通用模式，而不是过度依赖于个别数据点，从而提高模型在新数据上的泛化能力。

6.2.2 方差分析

方差分析（ANOVA）是一种统计方法，用于比较两个或多个样本的平均值是否有显著差异。论文中通过分解实验人员基于提取特征指标的总方差，将样本分为不同来源的方差成分，从而评估样本间和样本内的变异程度，并基于此推断出样本间差异的统计显著性。具体实现步骤如下：

Step 1 设定假设

设置原假设和备择假设，其中 H_0 表示各样本之间的平均值没有显著差异。H_1 表示至少有两组的平均值存在显著差异。

Step 2 计算变异值

该步骤中需要计算样本总变异、样本间变异及样本内变异。总变异即为样本数据的总方差，即所有观测值与总平均值之间的差异。

$$\mathrm{SST} = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 \tag{6-1}$$

其中，X_{ij} 是第 i 个组中第 j 个观测值。

样本间变异即不同样本之间的平均值与总平均值之间的差异。

$$\mathrm{SSB} = \sum_{i=1}^{n} n_i(\bar{X}_i - \bar{X})^2 \tag{6-2}$$

其中，n_i 是第之组中的样本数量。

样本内变异即每个样本内观测值与其样本内平均值之间的差异的总和。

$$\mathrm{SSW} = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2 \tag{6-3}$$

Step 3 计算 F 统计量

方差分析的核心是计算 F 值，用于比较样本间变异与样本内变异的比例。

$$F = \frac{\mathrm{SSB}/(n-1)}{\mathrm{SSW}/(N-n)} \tag{6-4}$$

其中，n 是样本数，N 是总样本数。分子是样本间的均方，分母是样本内的均方。

Step 4 判断显著性

根据 F 值和自由度，计算出的 p 值与显著性水平 0.05 比较，以决定是否拒绝 H_0。

6.2.3 最小二乘法

最小二乘法（OLS）是一种用于估计线性回归模型参数的统计方法。其主要目标是通过最小化预测值与实际观测值之间的误差平方和，找到最佳拟合线性模型。对于给定的一组观测数据点，OLS 试图找到线性关系：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{6-5}$$

其中，y 是因变量，$x_1, x_2, \ldots, x_p$ 是自变量。$\beta_0, \beta_1, \ldots, \beta_p$ 是待估计的回归系数。ϵ 是误差项，假设其均值为零且方差为常数。

OLS 的核心是最小化目标函数

$$\mathrm{RSS}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{6-6}$$

从而找到最佳拟合线性模型。

6.3 模型建立

为比较不同实验人员在同一活动状态下是否存在显著差异，根据所提平均合加速度、平均合角速度表征活动状态，实验人员标签指标为其变量，建立基于 OLS 的 ANOVA 统计模型模型，从而判断不同实验人员对其活动状态的影响。

Step 1 特征数据准备

Step 2 定义 OLS 模型

$$y_{ij}=\mu+\tau_i+\epsilon_{ij} \tag{6-7}$$

其中，y_{ij} 是第 i 组第 j 个总体加速度，μ是总体均值。τ_i 是第 i 组的实验人员的效应，ϵ_{ij} 是误差项，假设 $\epsilon_{ij}\sim \mathrm{N}(0,\sigma^2)$

Step 3 方差分析

求取 SST、SSB、SSW、F 统计量

Step 4 结果解释

针对所得结果的平方和、自由度、F 统计量和 P 值进行分析和模型解释。

6.4 模型求解

6.4.1 子问题一

首先选取合加速度均值与合角速度均值作为特征指标，对不同编号的实验人员采用及置信区间分析，得到其柱状图。

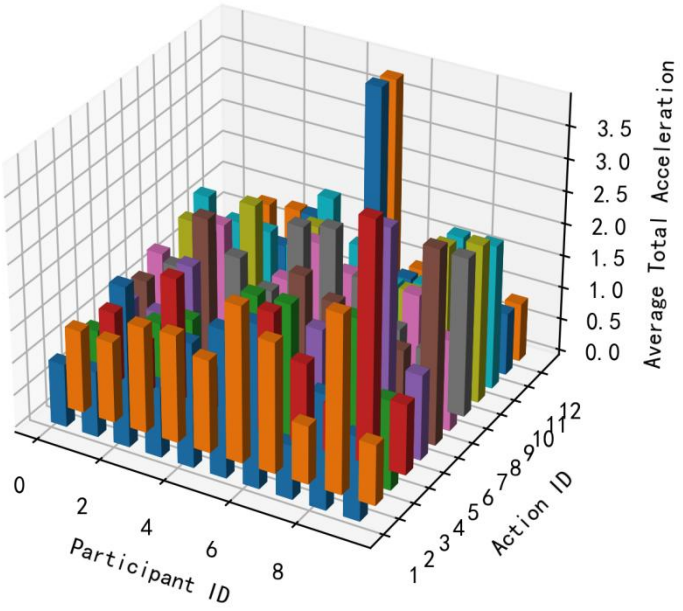

图 17 平均合加速度置信区间分析

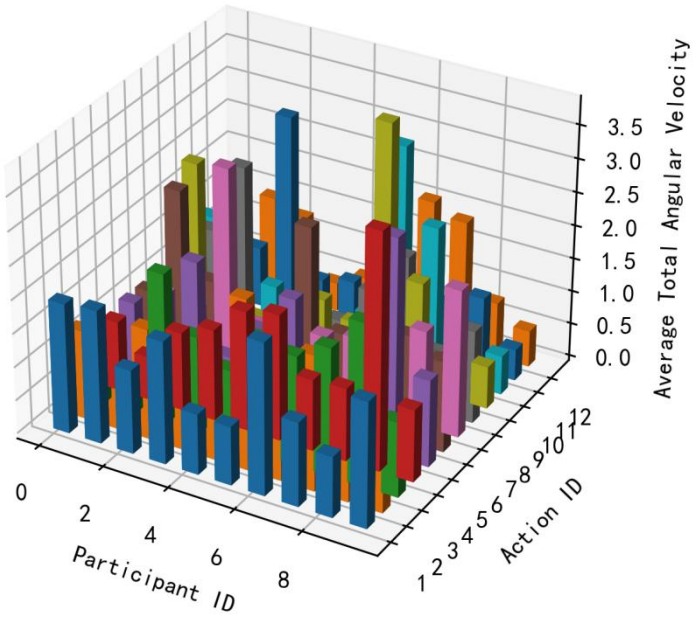

图 18 平均合角速度置信区间分析

由图可知，同一动作在不同实验人员之间的平均合加速度、平均合角速度差异较大。

表 17 ANOVA 分析结果

Action	$F_{totalacc}$	$PR_{totalacc}(>F)$	$F_{totalgyro}$	$PR_{totalgyro}(>F)$
1	36.756507	1.386924e-16	2.324021	0.032766
2	33.335925	7.569131e-16	3.183199	0.005415
3	32.879621	9.599456e-16	3.682214	0.001966
4	31.414599	2.099393e-15	10.026184	7.412121e-08
5	14.654239	4.021238e-10	6.847158	0.000007
6	61.646946	1.263090e-20	7.968862	0.000001
7	6.007267	0.000027	2.921342	0.009319
8	6.512096	0.000012	6.898855	0.000006
9	14.817441	3.420641e-10	15.678862	1.486880e-10
10	29.271024	6.990464e-15	3.953966	0.001148
11	19.178115	6.893607e-12	2.71924	0.014229
12	9.893479	8.793410e-08	2.432904	0.026027

(1) 对于平均合加速度：

所有 action 的 $PR_{totalacc}(>F)$均小于 0.05 这表明不同动作之间的加速度差异是显著的。其中，action 6 的 F 值最高，为 61.646946，表明该动作的加速度差异最显著。

(2) 对于平均合角速度：

所有动作的 $PR_{totalgyro}(>F)$均小于 0.05，这表明不同动作之间的角速度差异是显著的。其中，action 9 的 F 值最高，为 15.678862，表明该动作的角速度差异最显著。

6.4.2 子问题二

将身高和体重降维为 BMI 指标，基于 5.2.4 节 XGBoost 模型，预测年龄和 BMI 标签。同时基于 PCA 降维、Pipeline 和 GridSearchCV 对模型进行优化，并计算训练集和测试集的均方误差（MSE）和 R2 得分来评估模型表现。根据下图进行分析

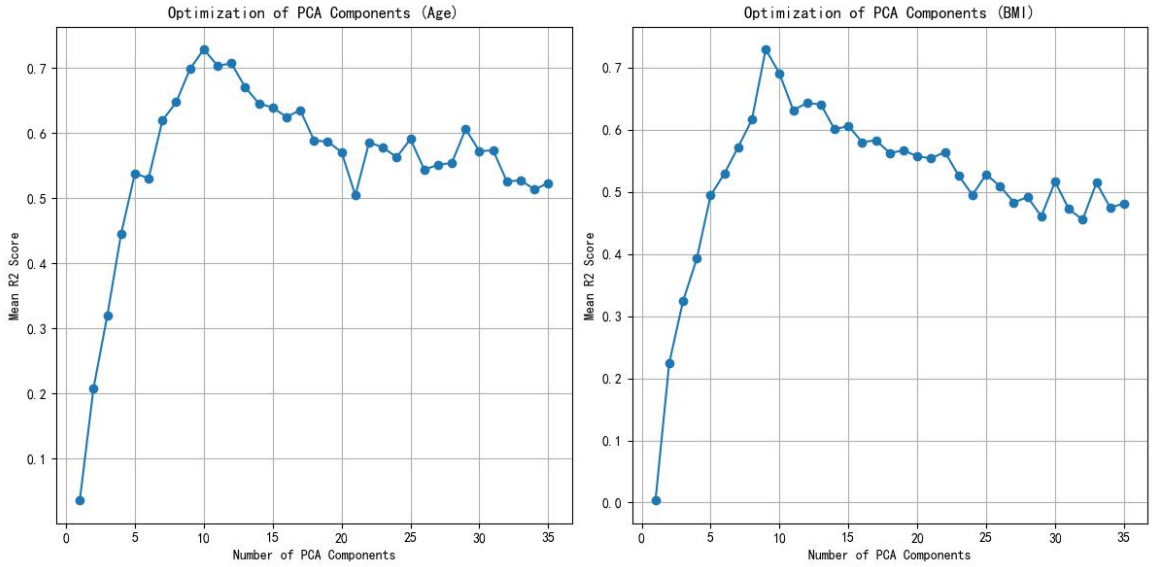

图 19 特征成分维度选取 (R^2)

(1) 最佳成分数

对于年龄，最佳 PCA 成分数目为 7 个。对于 BMI，最佳的 PCA 成分数目为 10；

(2) 模型优化

在这些最佳成分数目下，模型对数据的解释能力最强，平均 R^2 得分最高；

(3) 过拟合风险

增加成分数目超过最佳点后，模型的性能开始下降，这表明更多的成分并不总是更好，可能会导致过拟合。

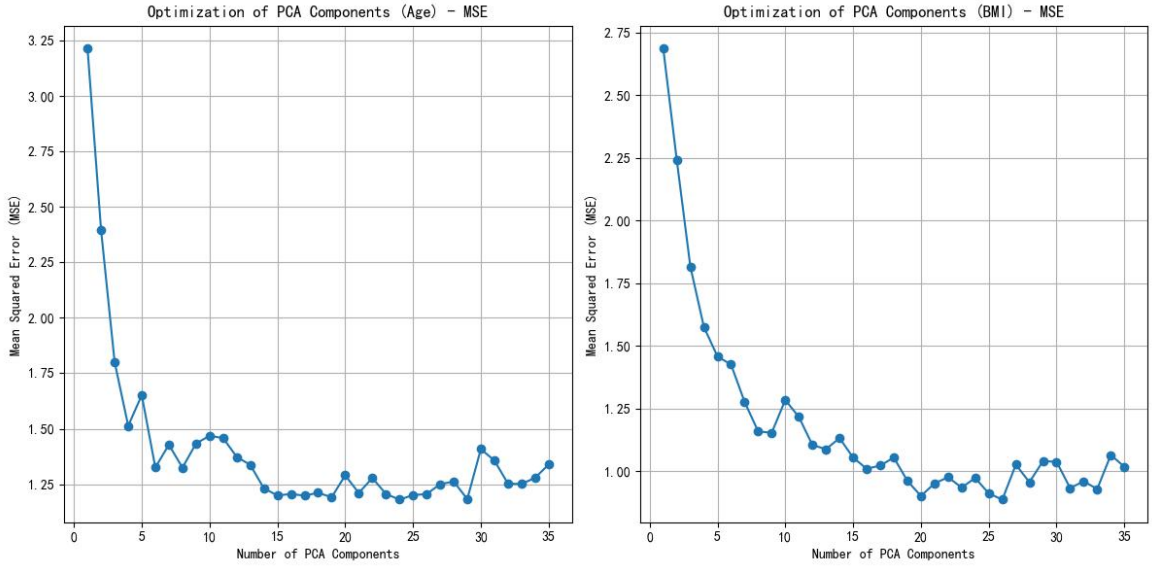

图 20 特征成分维度选取 (MSE)

(1) 最佳成分数

对于年龄，最佳 PCA 成分数目为 7 个。对于 BMI，最佳的 PCA 成分数目为 15；

(2) 模型优化

在这些最佳成分数目下，MSE 最低。模型的预测性能最佳；

(3) 平稳期

在达到最低点后，增加更多的成分并不会显著改善模型性能，MSE 保持相对稳定。

综上，对于年龄选择最佳 PCA 成分数目为 7 个，对于 BMI 最佳的 PCA 成分数目为 15 个。并根据相关指标提取、变量相关性分析和模型建立，可以使用活动传感器数据进行人员画像。

6.4.3 子问题三

将所选特征指标输入第二个子问题所建立的模型以进行实验人员的判别，输出结果。

表 18 人员刻画结果

序号	活动类型	判别结果
0	Unkonw1	2
1	Unknow2	12
2	Unknow3	7
3	Unknow4	9
4	Unknow5	10

7 模型评价与改进

7.1 模型的优点

(1) 模型充分结合 12 中运动状态各自的特点, 有针对性的提取特征参数, 如: 向前跑以及跳跃两个动作相对剧烈，本身 acc_x 的起伏变化相对明显;向左走以及向右走两个运动状态能够通过 gyro_x 的均值绝对值的大小来区别其他动作状态,通过其本身的正负性实现两个运动状态的区别刻画等等. 这样得到的特征参数紧密联系实际, 能较好的解决实际问题；

(2) XGBoost 采用了多种优化技术，如并行计算、缓存优化和数据压缩，使得它在训练速度和预测性能上都十分高效, 同时，其通过逐步构建和优化决策树，能够很好地捕捉高维数据中的复杂特征；

(3) Pipeline 可以与网格搜索（Grid Search）或随机搜索（Random Search）等调参方法结合使用，极大程度上简化了 Random Forest 以及 XGBoost 的超参数调优过程,提高了模型运行的效率。

7.2 模型的不足

(1) 在进行第一问的求解时采用的 Kmeans 算法对初始聚类中心的选择极为敏感，同时，Kmeans 算法在迭代过程中采用贪心策略，每一步都试图找到当前最优解。然而，这种策略可能导致算法陷入局部最优解，而无法达到全局最优；

(2) 本文提出的模型对于现有条件使用效果较好, 由于数据集有限的问题，并没有对模型进行普适性的检验. 对于具有孤僻动作人员运动状态的识别, 可能无法达到较好的效果。

7.3 模型的改进

由于在 K-Means 算法中，k 个初始化的质心的位置选择对最后的聚类结果和运行时间都有很大的影响，因此需要选择合适的 k 个质心。如果仅仅是完全随机的选择，有可能导致算法收敛很慢。因此对于其质心位置的选择进行如下的改进：

a) 从输入的数据点集合中随机选择一个点作为第一个聚类中心

b) 对于数据集中的每一个点，计算它与已选择的聚类中心中最近聚类中心的距离

c) 选择一个新的数据点作为新的聚类中心，选择的原则是：较大的点，被选取作为聚类中心的概率较大

d) 重复 b 和 c 直到选择出 k 个聚类质心

e) 利用这 k 个质心来作为初始化质心去运行标准的 K-Means 算法

参考文献

[1] 祝秀萍;吴学毅;刘文峰;人脸识别综述与展望[J];计算机与信息技术;2008(4):53-56

[2] F.Galton .Personal indentification and description[J].Nature,1888：173-177.

[3] 冯国双. 白话统计[M]. 电子工业出版社, 2018.

[4] 张良均. Python 数据分析与挖掘实战[M]. 机械工业出版社, 2016.

[5] 茆诗松, 程依明, 濮晓龙, 等. 概率论与数理统计教程第二版[M]. 北京: 高等教育出版社, 2011

[6]《运筹学》教材编写组. 运筹学.第 4 版[M]. 清华大学出版社, 2012.

[7] 周志华. 机器学习[M]. 清华大学出版社, 2016.

[8] Rachel Schutt, Cathy O'Neil. 数据科学实战[M]. 人民邮电出版社, 2015.

[9] 姜启源, 谢金星, 叶俊. 数学模型.第 4 版[M]. 高等教育出版社, 2011.

[10] 韩中庚. 数学建模方法及其应用-第 2 版[M]. 高等教育出版社, 2009.

附　录

附录 1：支撑材料列表

支撑材料列表

序号	文件名	材料说明
1	Data_visualization.py	数据可视化 python 代码
2	Data_merge.py	数据表格融合代码
3	RandomForestModel.py	随机森林模型代码
4	XGBoost.py	XGBoost 模型代码
5	q3q1.py	问题三第一小问求解代码
6	q3q2.py	问题三第二小问求解代码

附录 2：主要程序/关键代码

代码环境	操作系统: macOS Mojave (Version 10.14.3) 编程语言: Python 3.12.0 (Anaconda Navigator 1.9.2) 编辑器: VS Code 代码详见:

代码清单 1XGBoost.py

```
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.pipeline import Pipeline
from xgboost import XGBClassifier  # 导入 XGBoost 分类器

# 读取数据
data_path = './merged_data_Ptotal.csv'
data = pd.read_csv(data_path)
# 准备特征和目标数据
activity_type = data['action']
```

```
scaler = StandardScaler()
scaled_features = scaler.fit_transform(data.drop(columns=['action', 'person']))
# 将标签转换为从 0 开始的连续整数
activity_type = activity_type - 1
# 设置 PCA 和 XGBoost 分类器
pca = PCA()
xgb = XGBClassifier()
# 设置网格搜索参数
param_grid = {
    'pca__n_components': [0.7, 0.8, 0.9, 0.95],
    'xgb__n_estimators': [50, 100, 200],
    'xgb__learning_rate': [0.01, 0.1, 0.2],
}
# 使用管道来组合 PCA 和分类器
pipe = Pipeline([
    ('pca', pca),
    ('xgb', xgb)
])
# 设置网格搜索和交叉验证
grid_search = GridSearchCV(pipe, param_grid=param_grid, cv=5, verbose=1, scoring='accuracy')
# 分割数据集为训练集和测试集
X = scaled_features
y = activity_type.values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# 运行网格搜索
grid_search.fit(X_train, y_train)
# 输出最佳参数和最佳分数
print("Best parameters found:")
print(grid_search.best_params_)
print("Best cross-validation accuracy:")
print(grid_search.best_score_)
# 使用最佳参数的模型进行预测
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test)
# 输出评估结果
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
print('Classification Report:')
print(classification_report(y_test, y_pred))
print('Confusion Matrix:')
print(confusion_matrix(y_test, y_pred))
```

代码清单 2RandomForestModel.py

```
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.pipeline import Pipeline  # 导入 Pipeline

data_path = './merged_data_Ptotal.csv'
data = pd.read_csv(data_path)
activity_type = data['action']
scaler = StandardScaler()
scaled_features = scaler.fit_transform(data.drop(columns=['action','person']))
# 设置 PCA 和随机森林分类器
pca = PCA()
rf = RandomForestClassifier()
# 设置网格搜索参数
param_grid = {
    'pca__n_components': [0.7, 0.8, 0.9, 0.95],  # 尝试不同的 PCA 成分保留比例
    'rf__n_estimators': [50, 100, 200],  # 尝试不同的随机森林分类器的估计器数量
}
# 使用管道来组合 PCA 和分类器
pipe = Pipeline([
    ('pca', pca),
    ('rf', rf)
])
# 设置网格搜索和交叉验证
grid_search = GridSearchCV(pipe, param_grid=param_grid, cv=5, verbose=1, scoring='accuracy')
# 准备特征和目标数据
X = scaled_features
y = activity_type.values
# 分割数据集为训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# 运行网格搜索
grid_search.fit(X_train, y_train)
# 输出最佳参数和最佳分数
print("Best parameters found:")
print(grid_search.best_params_)
print("Best cross-validation accuracy:")
print(grid_search.best_score_)
# 使用最佳参数的模型进行预测
best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test)
# 输出评估结果
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
print('Classification Report:')
print(classification_report(y_test, y_pred))
print('Confusion Matrix:')
print(confusion_matrix(y_test, y_pred))
```