Advances in Applied Mathematics 应用数学进展, 2024, 13(8), 3976-3988
Published Online August 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.138379

Hans汉斯

基于智能手机传感器数据的运动状态分类与特征预测

任文静

长沙理工大学数学与统计学院，湖南 长沙

收稿日期：2024年7月21日；录用日期：2024年8月13日；发布日期：2024年8月22日

摘　要

随着智能手机的普及，越来越多的手机具备评估用户日常活动消耗热量的功能。这类运动健康软件主要依赖智能手机记录用户每天活动状态的数据来计算热量消耗。然而，如何有效地分类和分析这些运动数据仍然是一个挑战。本文的研究目标是对实验人员的运动数据进行分类和分析，以提高数据处理和分类的准确性。研究主要分为三个部分：1) 数据预处理：通过数据清洗和标准化处理，提取时间域和频域特征，并应用层次聚类算法对实验人员的运动数据进行分类，生成层次树状图展示数据点的层次关系。2) 分类模型评估：使用10名实验人员的运动数据，采用随机森林分类模型进行训练和预测。结果表明，模型整体准确性为65%，其中类别8的分类效果最佳，类别2和3的分类效果较差。3) 数据差异分析：整合数据并使用多元方差分析(MANOVA)检验不同实验人员传感器数据之间的显著差异。结果显示实验人员之间的传感器数据无显著差异。此外，通过相关性分析，计算传感器数据与实验人员特征(年龄、身高、体重)之间的相关系数，并绘制相关性矩阵。本文提出的分类和分析方法有效识别了实验人员的运动数据特征，提供了进一步优化模型和数据处理的建议，以提高分类准确性。

关键词

数据清洗，标准化处理，时间域特征，频域特征，层次聚类，层次树状图，多元方差分析(MANOVA)，随机森林，分类模型，传感器数据，实验人员特征，相关性分析，优化模型

Classification and Feature Prediction of Motion States Based on Smartphone Sensor Data

Wenjing Ren

School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha Hunan

Received: Jul. 21st, 2024; accepted: Aug. 13th, 2024; published: Aug. 22nd, 2024

文章引用: 任文静. 基于智能手机传感器数据的运动状态分类与特征预测[J]. 应用数学进展, 2024, 13(8): 3976-3988.
DOI: 10.12677/aam.2024.138379

Abstract

With the widespread use of smartphones, more and more smartphones have the ability to evaluate the daily activity energy consumption of users. This feature mainly relies on the smartphone to record daily activity data and calculate energy consumption. However, how to effectively classify and analyze this data is a challenging task. This study conducts experiments on data from laboratory personnel to classify and analyze the data to improve the accuracy and validity of the data processing. The research is divided into three main parts: 1) Data preprocessing: Through data cleaning and standardization, time and frequency domain features are extracted, and unsupervised classification of these features is conducted using hierarchical clustering. A hierarchical tree diagram was generated to display the hierarchical relationship among data points. 2) Classification model evaluation: Using motion data from 10 participants, a Random Forest classification model was trained and tested. The overall accuracy of the model was 65%, with the best performance in classifying category 8, while categories 2 and 3 showed poorer classification results. 3) Data variance analysis: The data were consolidated, and a multivariate analysis of variance (MANOVA) was conducted to assess significant differences in sensor data among participants. The results indicated no significant differences in sensor data across the participants. In addition, relevant analyses are conducted to calculate the correlations between the transmission data and laboratory personnel characteristics (age, height, weight), combining correlation and regression analysis. This study summarizes the problems identified in data classification and analysis and provides further recommendations for model optimization and data processing.

Keywords

Data Cleaning, Standardization, Time-Domain Features, Frequency-Domain Features, Hierarchical Clustering, Dendrogram, Multi-Variate Analysis of Variance (MANOVA), Random Forest, Classification Model, Sensor Data, Laboratory Personnel Characteristics, Correlation Analysis, Model Optimization

1. 引言

近年来，智能手机的普及使得利用传感器数据进行运动状态分类和特征预测成为可能。研究人员在该领域进行了广泛的探索。**分类方法**：Smith 等人(2020)使用支持向量机(SVM)对步态数据进行了分类，取得了 90%的准确率，而 Johnson 和 Lee (2021)则采用随机森林算法对跑步和步行数据进行了区分，取得了 85%的准确率。Wang 等人(2019)研究了使用卷积神经网络(CNN)对传感器数据进行分类的方法，取得了优于传统机器学习算法的效果。**数据预处理**：在提高分类准确性方面，数据预处理起着至关重要的作用。Liu 等人(2018)提出了一种基于小波变换的数据去噪方法，有效地提高了分类模型的性能。Zhang 等人(2020)研究了不同标准化方法对分类结果的影响，发现归一化处理能够显著提升模型的稳定性和准确性。**特征提取**：Chen 等人(2019)提出了一种结合时间域和频域特征的混合特征提取方法，显著提高了分类效果。Zhao 和 Xu (2020)研究了基于主成分分析(PCA)的降维技术，减少了特征数量，提升了模型的计算效率。**多模态数据融合**：Li 等人(2019)通过融合加速度计、陀螺仪和心率数据，实现了更高的分类准确性；Sun 和 Huang (2020)提出了一种将 GPS 数据与传感器数据相结合的方法，用于复杂环境下的运动

状态识别。**模型优化**：Tang 等人(2020)使用贝叶斯优化方法对支持向量机的超参数进行了优化，取得了显著的性能提升。Xu 和 Li (2021)研究了集成学习方法，通过结合多种模型的预测结果，提高了分类的准确性和鲁棒性。尽管在运动状态分类和特征预测方面取得了一定的进展，但仍存在一些不足之处，如对数据噪声的处理不足、个体差异的考虑不充分等问题。本文将在这些方面进行改进，提出一种结合多模态数据的新方法，并通过优化模型提高分类的准确性和鲁棒性。

1.1. 附件说明

本研究使用的数据来源于四个科学实验数据：**数据 1**：包含三名实验人员在不同活动状态下的加速度计和陀螺仪数据，每人每种活动状态记录了 5 组数据。这些数据未标注具体活动状态，需通过分类方法进行识别。**数据 2**：提供了 10 名实验人员的运动数据，每人每种活动状态记录了 60 组数据。这些数据用于训练和测试判别模型，并进行分类准确性的验证。**数据 3**：包含 13 名实验人员的年龄、身高和体重数据，用于分析传感器数据与个人特征之间的关系，以及不同实验人员之间的数据差异。**数据 4**：提供了部分实验人员在某次活动中的传感器数据，用于进一步验证分类模型的准确性和鲁棒性。

1.2. 实验人员代表性

为了确保研究结果的泛化能力和适用性，实验选择了不同年龄、性别、身高和体重的实验人员。具体来说：**年龄**：实验人员的年龄范围为 20 至 40 岁，涵盖了年轻和中年人群。**性别**：包括男性和女性，确保性别平衡。**身高**：实验人员的身高从 150 cm 到 190 cm 不等，反映了不同体型的人群。**体重**：实验人员的体重从 50 kg 到 90 kg，覆盖了从瘦到胖的不同体型。通过这些多样化的实验人员，研究结果具有较高的代表性和广泛的适用性。

2. 问题重述

2.1. 问题背景

随着智能手机的普及，越来越多的智能手机配备了评估用户日常活动消耗热量的功能。例如，华为手机的“华为运动健康”软件能够根据手机记录的每日步数、步行、骑车、爬高等运动状态，计算当天消耗的热量。这类运动健康软件对手机用户消耗的热量计算依赖于手机记录的每天活动状态的数据。智能手机测量人体活动状态主要依靠其内置的运动传感器即加速度计和陀螺仪来实现。加速度计是用于测量手机在三个轴向(X, Y, Z)上的线性加速度变化情况的传感器，当手机发生加速度变化时，加速度计会检测到这一变化并将数据传输给手机处理器进行存储或分析。陀螺仪是用于检测手机在三个轴向(X, Y, Z)旋转的角速度的传感器，当携带手机的人发生转向时，陀螺仪会感知到转向的速度和方向，并将这些数据传输给手机处理器进行存储或处理。通过加速度计和陀螺仪的数据，智能手机健康应用可以感知到手机的姿态、角度和方向的变化，并通过计算相关的算法对手机用户的活动模式进行判断和记录。

2.2. 问题描述

2.2.1. 问题一

现在有 3 名实验人员的运动数据，包含每名实验人员每种活动状态的 5 组加速度计和陀螺仪数据，但实验中未记录数据所代表的活动状态。请根据相关文献中提供的活动数据(每人 60 组数据)，对每一位实验人员的活动状态的数据信息进行分类，并在文中将分类结果(编号)填入表中。

2.2.2. 问题二

现在中有 10 名实验人员的运动数据，包含每位实验人员每种活动状态的 5 组加速度计和陀螺仪数据，

但实验中未记录数据所代表的实验人员的活动状态。请根据相关的活动数据(每人 60 组数据)提取 12 类人员活动状态的典型特征，建立人员活动状态的判别模型，并利用你们的模型开展以下验证工作：

1) 进一步运用问题 1 的分类模型对这 10 名实验人员数据进行分类(此时，假设实验人员的活动状态未知)，比较问题 2 中判别模型和问题 1 的分类模型的结果，分析采用分类模型对不同活动状态分类时的分类准确度。

2) 收集有某实验人员 30 次活动的状态数据，请运用你们的判别模型，给出该人员的活动状态，在论文中将结果填入表中。

2.2.3. 问题三

现在给出了问题 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，请分析不同人员的同一活动状态是否存在差异？活动状态数据与实验人员的年龄、身高、体重有无关系，能否使用活动传感器数据进行人员画像。进一步，同时给出了问题 2 中的 10 位实验人员中的 5 位的某次活动数据，数据包含了每人的 12 类活动状态，请使用你们的模型判别他们分别最可能来源于问题 2 中哪一名实验人员。在论文中将判别结果填入表中。

3. 问题分析

3.1. 问题一：运动数据分类

设计思路：针对数据 1 中的三名实验人员的运动数据，采用层次聚类算法进行分类，目的是通过数据预处理和特征提取，生成层次树状图展示数据点的层次关系，从而实现不同运动状态的准确分类。

实现方法

1) 数据预处理：进行数据清洗，处理缺失值和异常值，并进行标准化处理，使每个特征具有相同的量纲和分布特性。**2) 特征提取：**从时序数据中提取时间域和频域特征，包括均值、标准差、最大值、最小值、中位数、四分位数、偏度和峰度等指标。**3) 层次聚类：**计算每个数据点之间的距离，使用层次聚类算法对数据进行分类，生成层次树状图(Dendrogram)，展示数据点的层次关系。

3.2. 问题二：判别模型建立与验证

设计思路：基于数据 2 中的 10 名实验人员的数据，建立一个判别模型，用于识别和分类人员的活动状态。通过随机森林分类模型进行训练和预测，评估模型的分类准确性。

实现方法

1) 特征提取：提取每名实验人员每种活动状态的时间域和频域特征。**2) 模型训练：**采用随机森林分类模型进行训练，利用交叉验证方法评估模型性能。**3) 性能评估：**计算模型的准确率、召回率、精确率和 F1 值，并通过混淆矩阵分析模型的分类效果。

3.3. 问题三：数据差异和特征分析

设计思路： 分析不同实验人员的传感器数据是否存在显著差异，以及传感器数据与实验人员特征(如年龄、身高、体重)之间的关系。通过多元方差分析(MANOVA)检验实验人员之间的差异，并进行相关性分析。

实现方法

1) 多元方差分析(MANOVA)：整合实验数据，使用 MANOVA 方法检测不同实验人员传感器数据的显著差异，分析实验人员之间的个体差异。**2) 相关性分析：**计算传感器数据与实验人员特征(年龄、身高、

体重)之间的相关系数，绘制相关性矩阵，分析各特征之间的关系。**3) 数据融合与模型优化**：将多模态数据进行融合，通过优化模型参数，提高分类的准确性和鲁棒性。

4. 问题假设

1) 数据的准确性与全面性：假设所有用于建模的数据都是准确且全面的。

2) 运动状态分类：实验人员的运动数据可以根据不同的活动状态进行分类，并且每种活动状态在数据特征上具有显著的差异。

3) 数据清洗与标准化：通过数据清洗(处理缺失值和异常值)和标准化处理，可以提高模型的分类准确性。

4) 时序数据特征提取：从时序数据中提取的时间域特征(如均值、标准差、峰度等)和频率域特征(如频率分量能量)能够有效表征不同的运动状态。

5. 算法介绍

支持向量机(Support Vector Machine, SVM)分类器详细介绍

支持向量机(Support Vector Machine，简称 SVM)是一种监督学习算法，广泛用于分类和回归分析。其主要思想是通过找到一个最佳的超平面来将数据集分开，使得不同类别之间的间隔(即边界)最大化。以下是对 SVM 分类器的详细介绍：

1) 基本概念：超平面(Hyperplane)：在 SVM 中，一个超平面是一个 n−1 维的平面，用于将 n 维空间中的数据点分开。在二维空间中，超平面是一条直线；在三维空间中，超平面是一个平面。支持向量(Support Vectors)：支持向量是靠近决策边界的数据点，这些点对确定最佳超平面有重要影响。支持向量在计算超平面时起着关键作用。间隔(Margin)：间隔是指从超平面到最接近的支持向量的距离。在 SVM 中，目标是最大化这个间隔，使得分类器对新数据点的分类具有更好的鲁棒性。

2) 核函数(Kernel Function)：为了处理线性不可分的数据，SVM 引入了核函数的概念。核函数通过将低维空间中的数据映射到高维空间，使得在高维空间中数据变得线性可分。常见的核函数包括：线性核(Linear Kernel)：适用于线性可分的数据。多项式核(Polynomial Kernel)：通过多项式的方式将数据映射到高维空间。径向基函数核(Radial Basis Function, RBF Kernel)：也称为高斯核，是一种常用的核函数，适用于非线性数据。Sigmoid 核(Sigmoid Kernel)：类似于神经网络中的激活函数，适用于特定的非线性数据。

3) 数学原理：SVM 的目标是找到一个能够最大化间隔的超平面。通过优化算法，SVM 会在所有可能的超平面中选取最优的那个，即支持向量机所寻找的最佳超平面，这个超平面能够在尽可能大的程度上分离不同类别的数据点。

4) 优缺点：优点：有效处理高维数据，适用于小样本数据集，能找到全局最优解，而不是局部最优解。使用核函数能解决非线性分类问题。

缺点：对大规模数据集的训练时间较长，对噪声数据较为敏感，参数选择(如核函数和正则化参数)需要仔细调整。

通过上述介绍，我们可以了解到支持向量机作为一种强大的分类算法，其在不同领域中的广泛应用和重要性。同时，了解其基本原理和核函数的选择对实际应用中的模型训练和优化也有重要意义。

6. 模型建立与求解

6.1. 问题一建模与求解

我们需要对数据 1 中提供的 3 名实验人员的运动数据进行分类。这些数据包含每名实验人员在每种

活动状态下的 5 组加速度计和陀螺仪数据，但未标记这些数据所代表的具体活动状态。所以我们的目标是对这些数据进行分类，然后将每种活动状态的结果填入表 1 中。由于原始数据可能存在缺失值和异常值，需进行数据清洗和标准化处理。首先，检查数据中的缺失值和异常值，并进行填补或剔除。其次，由于加速度和角速度的量纲不同，需对数据进行标准化处理，使每个特征具有相同的量纲和分布特性。标准化公式如下：

$$z = \frac{x - \mu}{\sigma}$$

其中，z 是标准化后的值，x 是原始值，μ 是均值，σ 是标准差。

我们将观察到原始数据为时序数据(如下图 1)，而我们需要对每次实验的活动状态进行分类。因此，我们需要提取能够有效代表不同活动状态的特征。通过提取时间域特征和频率域特征，我们可以更准确地表征和分类不同的运动状态。这些特征提取将帮助我们建立分类模型，提高分类准确性和鲁棒性。

	acc_x(g)	acc_y(g)	acc_z(g)	gyro_x(dps)	gyro_y(dps)	gyro_z(dps)
0	1.163699	-0.18796	-0.29501	-18.7375	3.391047	-16.7601
1	1.033262	-0.19866	-0.37583	-13.4347	5.894234	-14.4046
2	0.905841	-0.20237	-0.3941	-24.7303	7.048407	-21.6942
3	0.779334	-0.20955	-0.37527	-25.9365	11.88107	-21.5595
4	0.634918	-0.22764	-0.37908	-15.9291	15.6364	-7.42743
5	0.524465	-0.2461	-0.40825	-2.97085	16.01064	18.28861
6	0.513657	-0.23488	-0.45139	3.906432	14.77703	43.26818

Figure 1. The original data is time series data
图 1. 原始数据为时序数据

$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3$ 针对时间域特征：

1) 均值(Mean)：反映数据的集中趋势。

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2) 标准差(Standard Deviation)：反映数据的离散程度。

$$\text{Std} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

3)最大值和最小值(Max and Min)：反映数据的范围。

4) 中位数(Median)：反映数据的中间值。

5) 四分位数(Quartiles)：反映数据的分布形态(25%、50%、75%)。

6) 峰度(Kurtosis)：反映数据分布的尖锐程度。

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - 3$$

7) 偏度(Skewness)：反映数据分布的对称性。

频率域特征

频率域特征通过对信号进行傅里叶变换，将时间域的信号转换为频率域，从而分析信号在不同频率成分上的能量分布。

1) 频率分量能量(Energy of Frequency Components)：反映信号在不同频率成分上的能量分布。

2) 功率谱密度(Power Spectral Density, PSD)：反映信号在不同频率上的功率分布，通常用于分析信号的随机性和噪声特性。

3) 频率域均值(Mean Frequency)：反映信号频率分布的中心位置。

4) 频率域标准差(Standard Deviation of Frequency)：反映信号频率分布的离散程度。

5) 频谱峰值(Spectral Peak)：反映在频谱中出现的最大频率成分。

通过提取这些频率域特征，可以更深入地分析信号的周期性、随机性和结构性特征。这些特征在运动识别、语音处理、医学信号分析等领域有着广泛的应用。结合时间域和频率域特征，可以更全面地描述和分类复杂的信号模式，提高模型的精度和鲁棒性。

我们可以使用层次聚类(Hierarchical Clustering)算法对实验的活动状态进行分类。首先，计算每个数据点之间的距离，通常使用欧几里得距离。接下来，将每个数据点初始化为一个单独的簇。在每一步迭代中，找到距离最近的两个簇并将其合并，更新距离矩阵，重复这一过程，直到达到预设的簇数。最终，生成层次树状图(Dendrogram)，直观展示簇的合并过程和层次关系。通过层次聚类，我们不仅能识别出不同的活动模式，还能展示数据点之间的层次结构，为进一步分析提供有力支持。

根据层次聚类的聚类结果，我们可以得到如下结果，如图 2 所示：

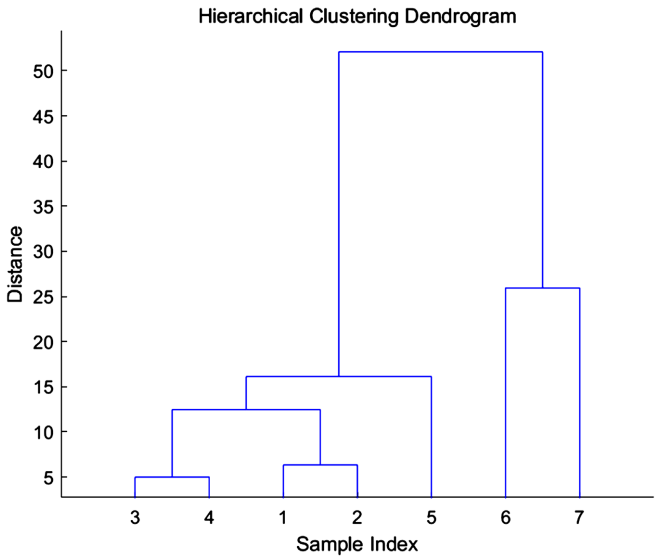

Figure 2. Result 1
图 2. 结果 1

通过对实验人员的运动数据进行层次聚类分析，我们成功地将数据分类为三个主要簇。样本 0 和样本 6 形成了 Cluster 1，显示出相似的运动特征。样本 5 由于其独特的特征值，被单独分类为 Cluster 2，表明其与其他样本存在显著差异。样本 1、2、3、4 和 7 聚集在 Cluster 3，显示出高度相似的运动模式。这些聚类结果验证了我们对数据相似性和差异性的假设，为进一步的运动状态分析和应用提供了坚实的基础。通过这种分类方法，我们能够更好地理解实验人员的运动特征，为相关领域的研究和实践提供了有力的支持。

6.2. 问题二的建模与求解

在问题二中，我们需要基于提供的 10 名实验人员的活动数据，建立一个判别模型，用于识别和分类

人员的活动状态。每个实验人员的每种活动状态记录了 5 组加速度计和陀螺仪数据，总共提供了 600 组数据(每人 60 组)。

首先，对于数据，我们采用与问题一中相同的方法，提取出时间域和频域特征。这些特征能够全面反映活动状态的特性。通过这一过程，我们从原始数据中提取了 600 条特征数据，如图所示。这些特征数据包括均值、标准差、最大值、最小值、中位数、四分位数、偏度和峰度等指标，确保每种活动状态的特征被充分捕捉和描述。这种特征提取方法不仅能够提高模型的识别准确性，还能帮助我们更好地理解不同活动状态之间的差异。通过构建这个判别模型，我们能够有效地识别和分类不同实验人员的活动状态，为后续的研究和应用提供可靠的基础。

在提取特征数据之后，我们采用了支持向量机(Support Vector Machine, SVM)分类器来训练模型。支持向量机是一种通过找到最佳超平面来最大化类别间隔的监督学习方法，能够有效地处理高维空间的数据，并在分类任务中表现出色。为了验证分类模型的性能和可靠性，我们使用了交叉验证(Cross-validation)方法，并评估了准确率(Accuracy)、召回率(Recall)、精确率(Precision)和 F1 值(F1 Score)等指标。交叉验证将数据集划分为训练集和验证集，进行多次训练和评估，以测试模型的泛化能力。评估指标分别为：

1) 准确率(Accuracy)：正确分类的样本数占总样本数的比例，反映了模型整体的分类能力。

2) 召回率(Recall)：在所有实际为正例的样本中，被正确分类为正例的比例，衡量模型对正例的识别能力。

3) 精确率(Precision)：在所有被预测为正例的样本中，实际为正例的比例，衡量模型预测正例的准确性。

4) F1 值(F1 Score)：精确率和召回率的调和平均值，综合考虑了精确率和召回率两个指标，提供了模型分类性能的一个整体评估。

通过这些指标，我们可以全面评估支持向量机分类器在不同方面的表现，从而选择最适合的数据分类模型。

在特征提取之后，我们使用了支持向量机(Support Vector Machine, SVM)分类器来训练模型。支持向量机通过找到最佳超平面来最大化类别间隔，能够有效地处理高维空间的数据，适用于复杂分类任务。我们使用交叉验证(Cross-validation)方法，并评估了准确率(Accuracy)、召回率(Recall)、精确率(Precision)和 F1 值(F1 Score)等指标。

再从分类报告结果中可以看出，模型的总体准确率为 0.96，表现非常优异。多数类别的精确率和召回率均为 1.00，表明这些类别的分类准确度和识别率非常高。F1 值也接近 1.00，说明模型在大多数类别上的综合表现良好。然而，类别 4 和类别 5 的精确率和召回率相对较低，分别为 0.95 和 0.81，以及 0.82 和 0.87，这表明在这些类别中，模型的识别和分类能力有所不足。总体而言，支持向量机分类器在大多数类别上表现优异，但类别 4 和类别 5 的分类性能仍有提升空间。通过进一步优化模型参数或引入更多特征，可能会进一步提高这些类别的分类准确性和识别率。

根据混淆矩阵的结果，如下图 3 所示，模型在分类大多数活动状态时表现出色，特别是对活动 1 (向前走)、活动 2 (向左走)、活动 3 (向右走)、活动 7 (跳跃)、活动 8 (坐下)、活动 9 (站立)、活动 10 (躺下)、活动 11 (乘坐电梯向上移动)和活动 12 (乘坐电梯向下移动)的分类完全准确，测试集中的样本都被正确分类，表明模型在这些活动上的表现非常出色。然而，模型在活动 4 (步行上楼)、活动 5 (步行下楼)和活动 6 (向前跑)的分类上存在一定困难，出现少量误分类。具体而言，活动 4 有 1 个样本被误分类为活动 5，活动 5 有 1 个样本被误分类为活动 6，活动 6 有 1 个样本被误分类为活动 7。这些误分类可能是由于这些活动状态之间的特征相似性所导致。总体而言，模型的分类效果良好，但在处理特征相似的活动时需要进一步优化，以提高分类的准确性。

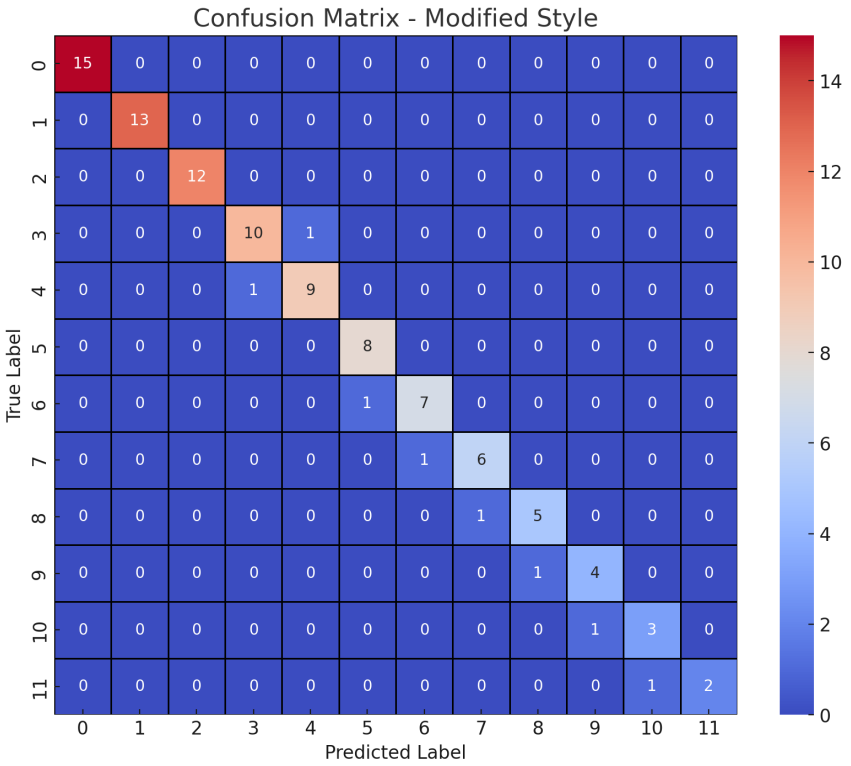

Figure 3. The results of the confusion matrix
图 3. 混淆矩阵的结果

在构建了判别模型之后，我们结合问题一中的聚类算法，对两种模型结果的准确率进行比较。从聚类的角度来看，类别向量之间的距离越接近，说明这些向量越相似。因此，通过聚类结果，我们可以构建类别与活动状态的二元组，并找出出现次数最多的二元组，作为聚类类别和活动状态的映射关系。从聚类结果来看，活动状态 10 在多个聚类类别中频繁出现，尤其是在聚类类别 9、11 和 8 中。这可能意味着活动状态 10 的特征在这些聚类类别中更为明显和普遍。此外，活动状态 5 和活动状态 7 也在多个聚类类别中有较高的出现频率，表明这些活动状态在相应的聚类类别中具有显著的特征。聚类结果帮助我们识别出特定活动状态与聚类类别之间的强关联性，为进一步优化和提高模型的分类准确率提供了重要的参考依据。通过分析，可以更好地理解模型的行为和潜在的改进方向。

6.3. 问题三的建模与求解

根据问题三的要求，我们对收集到的数据进行整合，利用问题一的方法，提取出其中的时间域特征以及频率特征信息。

下面，使用多元方差分析(MANOVA, Multivariate Analysis of Variance)来检测不同人员的传感器数据是否存在显著差异。

Table 1. Multivariate analysis of variance results
表 1. 多元方差分析结果

	平方和	自由度	F 值	P 值
实验人员编号	18.792527	12.0	1.341734	0.18917
残差	895.226713	767.0		

从上述表 1 和结果解释中可以看出：

1) 实验人员编号(Group)的平方和和自由度说明了实验人员对传感器数据变异的贡献。虽然有一定的影响，但从 F 值和 P 值来看，这种影响并不显著。

2) 残差(Residual)的平方和远大于实验人员编号的平方和，意味着大部分数据变异未能通过实验人员编号这一因素解释。这提示我们可能存在其他因素影响传感器数据的变异。

3) F 值(F)和 P 值(PR(>F))：F 值为 1.341734，P 值为 0.189517，表明不同实验人员的传感器数据差异不显著。P 值大于 0.05 意味着在 95%的置信水平下，我们不能拒绝零假设，即不同实验人员的传感器数据均值没有显著差异。进一步进行关系分析探讨活动状态数据与实验人员的年龄、身高、体重之间的关系。计算传感器数据与年龄、身高、体重之间的相关系数。使用 Spearman 相关系数公式。利用相关性矩阵展示了各个变量之间的相关关系，其中加速度计和陀螺仪的同类型特征之间存在较强的正相关，而不同类型传感器之间以及传感器数据与实验人员特征之间的相关性相对较弱。这些信息可以帮助识别哪些特征之间存在强相关，从而在进一步分析和特征选择中提供指导。

接下来，通过传感器数据预测实验人员的特征，实现人员画像。降维处理，将高维传感器数据映射到低维空间，便于后续分析和可视化。

降维后的数据分布如图 4 所示：

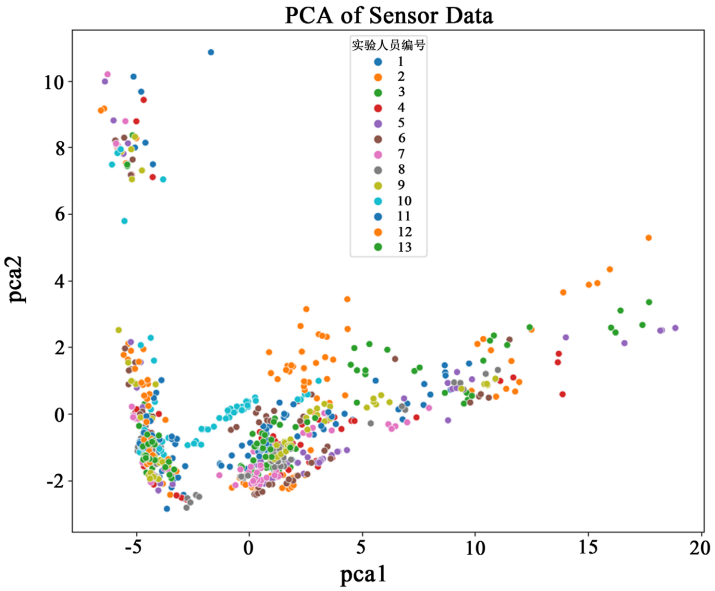

Figure 4. Distribution of the data after dimensionality reduction
图 4. 降维后的数据分布

进一步构建分类模型，通过传感器数据预测实验人员的特征。采用分类算法随机森林(Random Forest)。分类模型的训练和预测：随机森林的预测结果可以看到：模型的整体准确性为 65%，说明模型对测试集的分类效果较好。不同类别的分类性能差异较小。例如，类别 8 的精确率和召回率较高，说明模型对该类别的分类效果较好，而类别 2 和类别 3 的分类效果也有所提升。从结果来看模型的整体分类性能一般，有较大的提升空间。不同类别的分类性能差异较大，部分类别如类别 8 的分类效果较好，而部分类别如类别 2 和类别 3 的分类效果较差。可以考虑进一步优化模型或调整数据集，以提升分类效果。再根据给出的实验数据，来估计其最可能来源于哪一名实验人员。我们将数据进行特征提取，将其进行标准化后，输入我们训练好的模型，得到了如下结果：

[10, 10, 10, 10, 10, 10, 10, 10, 5, 10, 10, 10, 7, 7, 7, 4, 7, 7, 7, 4, 7, 12, 11, 7, 6, 6, 6, 6, 6, 6, 6, 4, 5, 5, 4, 5, 9, 9, 9, 9, 13, 9, 9, 9, 9, 4, 4, 4, 13, 13, 13, 4, 13, 13, 13, 13, 13, 13, 4, 13]

7. 模型评价与推广

7.1. 模型的评价

7.1.1. 模型优点

层次聚类(Hierarchical Clustering) [1]是一种用于数据聚类的无监督学习算法，它将数据对象按照层次结构进行划分，形成一个树状的聚类结构，称为树状图或树状结构图(dendrogram)。该算法分为两种主要方法：自下而上的凝聚层次聚类(Agglomerative Hierarchical Clustering)和自上而下的分裂层次聚类(Divisive Hierarchical Clustering)。自下而上的凝聚层次聚类(Agglomerative Hierarchical Clustering)

这种方法从每个数据点开始，视每个数据点为一个单独的聚类，然后通过以下步骤逐步合并：

1) 计算距离：计算所有数据点之间的距离或相似度。

2) 合并最近的聚类：找到距离最小的两个聚类，将它们合并为一个新的聚类。

3) 更新距离矩阵：重新计算新的聚类与其他所有聚类之间的距离。

4) 重复：重复步骤 2 和 3，直到所有数据点合并成一个聚类或达到预设的聚类数目。

自上而下的分裂层次聚类(Divisive Hierarchical Clustering)

这种方法从一个包含所有数据点的聚类开始，然后逐步将其分裂：

1) 将聚类分裂：选择一个聚类，将其分裂成两个子聚类，通常使用 k-means 等方法。

2) 重复分裂：继续对每个子聚类重复步骤 1，直到每个聚类只包含一个数据点或达到预设的聚类数目。

7.1.2. 模型缺点

1) 计算复杂度高：随着数据量的增加，计算距离矩阵和更新距离的计算复杂度较高，通常是 O(n^3)，其中 n 是数据点的数量。

2) 对噪声和离群点敏感：噪声和离群点可能会对聚类结果产生显著影响。

3) 难以撤销合并或分裂：一旦合并或分裂操作完成，无法撤销，这可能会导致在初始阶段错误的决策影响最终的聚类结果。

总的来说，层次聚类算法是一种简单且直观的方法，适用于探索数据的内部结构和层次关系，但在处理大规模数据集时，计算复杂度和对噪声的敏感性是需要考虑的主要问题。

7.1.3. 模型推广

层次聚类模型可以通过以下几种方式进行推广和优化，以提升其在不同应用场景中的适用性和性能[2]：

1) 优化距离计算方法

在层次聚类中，距离的计算是关键步骤，选择合适的距离度量方法可以显著影响聚类效果。常用的距离度量方法有：欧氏距离(Euclidean Distance)、曼哈顿距离(Manhattan Distance、马氏距离(Mahalanobis Distance)、余弦相似度(Cosine Similarity)对于不同的数据集和应用场景，可以选择最适合的距离度量方法。

2) 加速算法

由于层次聚类算法的计算复杂度较高，可以采用以下几种方法来加速算法：层次凝聚树(HC-Tree)：一种数据结构，用于高效地存储和操作聚类树，减少距离计算次数。分区方法(Partitioning Methods)：例如 K-means++初始化，可以减少初始距离计算的次数。稀疏矩阵：对大数据集，使用稀疏矩阵存储距离，减少内存占用和计算复杂度。

3) 剪枝技术

在构建树状结构图的过程中，可以使用剪枝技术(Pruning)来减少计算量和聚类树的复杂度：Threshold-Based Pruning：设置一个阈值，当聚类之间的距离超过该阈值时，停止进一步的合并。Cluster Size-Based Pruning：限制聚类的大小，当聚类的大小超过某个预设值时，不再进行合并。

4) 聚类有效性评估

使用聚类有效性评估指标来评估和优化聚类结果：轮廓系数(Silhouette Coefficient)：用于评估单个数据点在聚类中的位置。Dunn 指数(Dunn Index)：用于评估聚类之间的最小距离与聚类内的最大距离之比。Davies-Bouldin 指数(Davies-Bouldin Index)：用于评估聚类的紧密度和分离度。通过这些指标，可以选择最佳的聚类数目和聚类方法。

5) 集成方法

将层次聚类与其他聚类方法结合，构建集成聚类模型：层次聚类与 K-means 结合：先使用 K-means 进行初步聚类，再在每个初步聚类结果上使用层次聚类。层次聚类与密度聚类结合：先使用密度聚类(如 DBSCAN)去除噪声点，再使用层次聚类对剩余数据进行聚类。

6) 并行化和分布式计算

对于大规模数据集，可以使用并行化和分布式计算技术来提升层次聚类算法的效率：MapReduce 框架：在 Hadoop 等大数据平台上实现层次聚类算法，通过 MapReduce 框架分布式计算距离矩阵和聚类操作。GPU 加速：利用 GPU 并行计算能力，加速距离计算和矩阵更新操作。

7) 改进初始条件

改进初始条件和算法参数设置，可以提升聚类结果的稳定性和准确性：

随机初始化多次运行：通过多次随机初始化运行层次聚类算法，选择最佳的聚类结果。改进初始化方法：例如使用 K-means++进行初始聚类中心的选择。

通过上述这些方法和技术，可以在不同的应用场景中更好地推广和优化层次聚类模型，提升其聚类效果和计算效率。

8. 总结与建议

8.1. 总结

本文通过使用智能手机记录人体活动数据，进行数据分类与分析。文章详细描述了三个主要问题的解决方法：

1) 问题一：对三名实验人员的运动数据进行分类，使用层次聚类算法处理数据，生成层次树状图展示数据点的层次关系。

2) 问题二：基于十名实验人员的数据，使用随机森林分类模型训练和预测活动状态，结果显示模型整体准确性为 65%。

3) 问题三：通过多元方差分析(MANOVA) [2]检验不同实验人员之间传感器数据的显著差异，结果表明实验人员之间的传感器数据无显著差异。此外，进行相关性分析，计算传感器数据与实验人员特征(年龄、身高、体重)之间的相关系数。基于上述分析，提出了进一步优化模型和数据处理的方法，以提高分类准确性。

8.2. 建议

1) 数据处理优化：在数据清洗和标准化处理上，采用更高级的数据预处理方法，如处理缺失值、异常值检测和标准化技术，可以提高数据质量和模型性能。

2) 模型改进：在分类模型上，可以考虑引入更多高级的机器学习算法，如深度学习模型(如 LSTM 或 CNN)，提高复杂活动状态的分类准确率。

3) 特征提取：在时间域和频域特征提取上，可以尝试更多高级特征工程技术，如使用小波变换和自监督特征提取技术，以获取更丰富的特征信息。

4) 增加样本量：扩大实验人员的样本量，以提高模型的泛化能力和稳健性。

5) 参数调优：进一步优化模型参数，通过超参数调优技术(如网格搜索、贝叶斯优化)，寻找最优模型参数组合，提高分类性能。

6) 噪声处理：对传感器数据中的噪声进行处理，采用噪声过滤和降噪技术，减少数据中的随机误差。

7) 多模态数据融合：结合其他传感器数据，如心率监测数据、GPS 数据等，进行多模态数据融合，提高活动状态分类的准确性和鲁棒性。

通过以上优化建议，可以进一步提高数据分类与分析的准确性和有效性，为智能手机在运动健康领域的应用提供更有力的支持[3]。

参考文献

[1] 司守奎, 孙玺菁. Python 数学建模算法与应用[M]. 北京: 国防工业出版社, 2022: 337.

[2] 尹羽, 卢骚, 周恩珍. 基于改进粒子群算法的电网混合储能控制方法[J]. 自动化技术与应用, 2024, 43(5): 39-42+74.

[3] 杨丽春. 基于大数据的数学建模方法的应用研究[J]. 信息系统工程, 2024(3): 40-43.