基于多种机器学习方法对人体行为识别研究

摘　　要

本题是基于手机传感器对人体行为识别的分类问题。本文通过多种机器学习方法对观测信息的分析，一方面实现了对 561 个特征信息的降维筛选，这有利于高效完成预测。另一方面构建了稳健的人体行为的识别模型，该模型可以实现，在智能手机佩戴在腰部时，对常见的人体动作有较为准确的判断。同时，本文提出了两种对智能手机佩戴在其他位置时的解决方案。这些在人工智能领域具有重要的现实意义。

首先，针对**高维数据**的处理，应聚焦于降维方法的选择。已有数据集中，我们无法精确获得每一个变量数值的物理意义，因此在数据降维时，传统的降维方法会降低模型的可解释性。本文通过机器学习选择变量实现对重要观测特征的筛选。

针对问题一，本文采用随机森林算法决策出显著影响类别结果的十个特征，发现其中 9 个指标与身体加速度或重力加速度相关。用非参数核密度估计图对离散数据近似拟合，比较不同行为类别下特征的差异。

针对问题二，考虑模型实际意义，采用多个**支持向量机**，先对动态、静态姿势二分类，后对二者具体活动做进一步细致分类。在静态姿势识别中，两次分类的精度高于文献中采用多分类方法的分类精度。

针对问题三，考虑实际人体行为中，静态姿势转换与动态姿势产生的动作类似，且是非线性分类问题。故使用 k 近邻法进行多分类，参数调优后，发现对静态姿势转换的识别准确度较低，本文通过人体运动和数据本身的两个角度给出原因。在上文中准确预测静态姿势的基础上，通过**随机森林**重新筛选变量改进模型。

针对问题四，查阅相关文献后，归纳出 8 类对位置敏感的指标与模型中的指标对比，论证关键特征不适用。对改变位置的预测，本文采用两种解决方案：一是在文献中归纳对位置不敏感的特征，通过**决策树**改进人体行为识别模型。二是先识别位置，在具体部位构建出不同的模型，本文设计了此构想的实验。

总之，本文采用多种机器学习方法良好完成了人体行为识别的任务，并从两个方面提出了“如何在佩戴位置发生改变时，对人体行为识别有大致正确的判断”这一问题的解决方案。另外，由于数据高维且缺少数值的物理意义，本文没有使用复杂或解释性差的模型，而是注重结果的可解释性和现实意义。**以上是本文的核心优点。**

关键词：高维数据　支持向量机　k 近邻法　随机森林　决策树

一、问题重述

1.1 问题背景

近年来，人们对智能交互、健康监护以及智慧医疗等方面的需求日益迫切，应用也越来越广泛。因此，人体行为识别就变得越来越重要。例如：开发运动检测的应用插件，计算卡路里消耗，给出健身提示；识别病人的行为信息，开展康复辅助；开发基于手机的危险行为检测和报警系统。目前，智能手机已经成为我们生活中不可或缺的物品。随着科技产品变得更小、更节能，以及我们的手腕上有更多的传感器，可穿戴设备越来越受欢迎。通过每天佩戴这些设备，我们可以很容易地每天收集百万字节的数据。尽管从这些传感器上有大量的可用数据，但我们并不能从这些原始数据中得知我们每天所做的事情。在这个项目中，我们的目标是从这些原始数据中识别模式，并提取有关用户日常活动的有用信息。

1.2 具体问题重述

经过试验采集，去噪、加窗、分割处理，分为训练集和测试集，建立数学模型，解决以下问题：

问题 1：针对附件所给出的 7767 行、561 列的观测特征信息及其对应的行为类别信息，分析不同行为类别下特征的差异，确定显著影响行为类别结果的几个特征；结合适当的分析和检验，说明所选取特征的合理性。

问题 2：结合附件数据，针对三种动态活动（步行、下楼和上楼）和三种静态姿势（站立、坐、卧）共 6 种行为类别，构建基于特征信息的人体行为识别模型，并评估特征选择、参数调优等对识别结果的影响。

问题 3：在问题 2 的基础上加入静态姿势转换的 6 种行为类别，重新构建基于特征信息的人体行为识别模型，并评估特征选择、参数调优等对判别结果的影响；判断哪些人体行为识别误判率较高，并分析其原因；尝试构建改进的人体行为识别模型以提高智能识别效果，并给出模型结果及评价。

问题 4：如果智能手机佩戴位置发生改变，例如拿在手中，上述所构建的模型及其选出的关键特征是否依然适用？讨论选择的特征有效性是否发生变化，是否可以扩充其他的特征来提高分类准确率？

二、问题分析

2.1 对问题的整体分析

本题是一个关于智能交互、信号处理，以及人体行为识别研究的问题。

从分析目的来看，需要对去噪和加窗分割处理后的数据，构建人体行为识别模型，并探讨设备佩戴在人体不同部位对识别结果的影响。

从数据特征来看，附件中所给的数据量丰富。但高维数据在机器学习中往往会带来维度灾难。即在高维空间中，由于维度的增加导致空间增大的速度远远快于数据点增加的速度，因此大多数数据点都集中在远离原点的边界区域，而在原点附近很难找到有意义的数据点。

为了应对维度灾难，需要对数据进行降维处理，使其能够在低维空间中更加有效地表示和分析。

从模型选择来看，由于智能交互、人体行为识别具有很强的现实意义，且本题的分析目的与实际情景息息相关。因此构建的模型应当追求模型的可解释性，不宜使用过于复杂且难以解释的模型。

2.2 对问题一的分析

题目要求：（1）确定显著影响行为类别结果的几个特征。（2）分析不同行为类别下特征的差异。（3）对特征选择的结果进行适当的分析和检验，说明所选取特征的合理性。

对于第一部分，通过随机森林，引入人体行为分类数据，决策出显著影响类别结果的几个特征。

对于第二部分，对显著影响分类结果的特征，比较其差异。本文绘制多个行为类别下的非参数核密度估计图，对离散数据去极限近似拟合。从曲线的相对一致程度，直观化判断选取的特征信息在多个行为特征下的表现，以此展现不同行为类别下特征的差异。

对于第三部分，利用选择出的特征变量训练模型，根据测试集中的数据，对模型的准确性进行分析检验。比较特征提取前后模型精度的差异，以此证明特征提取的合理性。

2.3 对问题二的分析

题目要求：（1）构建对动态、静态姿势共 6 种具体行为的人体行为识别模型。（2）评估特征选择、参数调优等对识别结果的影响。

对于第一部分，考虑到模型的实际意义，首先对动态、静态姿势的分类，通过支持向量机（SVM）进行二分类鉴别。然后再使用 SVM 模型分别对动态、静态活动中具体行为活动做进一步鉴别分类。通过两次识别，提高模型的准确性。

对于第二部分，通过对 SVM 以及决策树算法的参数调优，进一步优化模型。并将未筛选的 561 个指标和筛选后的指标进行模型分类精度比较，证明特征选择对模型的识别结果具有显著影响。

2.4 对问题三的分析

题目要求：（1）在问题二的基础上，加入静态姿势转换的 6 种行为类别，重新构建基于特征信息的人体行为识别模型。评估特征选择、参数调优等对重构模型判别结果的影响。判断哪些人体行为识别误判率较高，并分析其原因。（2）尝试构建改进的人体行为识别模型，对改进后的模型，进行结果评价。

对于第一部分，考虑到在实际人体行为中，静态姿势转换与动态姿势所产生的动作类似，故可将 12 种行为分类，通过同一模型进行判别，提升模型的应用性。相比于问题二中的二分类与三分类，本问题需要建立多分类模型。本文使用适合多分类的 k 近邻法（KNN）重新构建基于特征信息的人体行为识别模型。通过对 KNN 算法的参数调优，进一步优化模型。并将未筛选的 561 个指标和筛选后的指标进行模型分类精度比较，证明特征选择对模型的

识别结果具有显著影响。绘制 KNN 分类结果的混淆矩阵，将预测结果进行可视化，给出较高误判率的人体行为类别，并以人体运动和数据本身的两个角度给出其原因，加强模型的可解释性。

对于第二部分，考虑到问题二已对动态姿势与静态姿势，以及具体的静态姿势的判别有较高的精度。本问题可用适合用于变量选择的随机森林模型，将六个具体的静态姿势转换拆解为，由站着、坐着、躺着发起的三组二分类。重新在 561 个特征值中进行筛选，构建模型。本文首先运用测试集，测试模型训练的结果，对模型的准确性进行评估；然后采取 ROC 受试者工作曲线评价所构建的模型。

2.5 对问题四的分析

题目要求：（1）讨论当智能手机佩戴位置发生改变，例如拿在手中，上述所构建的模型及其选出的关键特征是否依然适用。（2）讨论选择的特征有效性是否发生变化。（3）讨论是否可以扩充其他的特征来提高分类准确率。

对于第一部分，查阅相关文献后，归纳出 8 类对位置敏感的指标，对比前文所选取的指标，给出关键特征是否适用的结论。

对于第二部分，查阅相关文献后，归纳出 9 类对位置不敏感的指标，通过选取 9 个位置不敏感的指标构建前文所述的模型，比较模型精确性的变化，进而说明特征有效性的变化。

对于第三部分，考虑到此问题是对模型现实意义的推广，本文采用设计实验的方法。设计多组实验，实验思路为：更改智能手机佩戴的位置；新增除加速度、角速度外的其他物理量进行实验观测。

三、模型假设

本文在分析过程中进行如下假设：

1、假设实验中仅有不同姿势会对各项指标的实验结果造成差异，其他因素如实验者的不同、手机佩戴位置的微小变化不对统计数据产生影响。

2、假设除题中所给的指标外，其他指标对智能手机放在腰部的人体行为识别结果无影响。

3、假设智能手机在实验中的运动情况可抽象为刚体运动。

四、符号说明

符号	说明
T	核函数密度估计中样本数
h	核密度估计曲线带宽
$K(X_j)$	关于特征信息x_j的核函数
$Gain(M,X)$	特征X的信息增益

注：其余表中未给出符号在相应部分会有说明。

五、模型的建立与求解

5.1 问题一 基于随机森林对观测特征信息的重新筛选

问题一分为三部分求解，第一部分为确定显著影响行为类别结果的几个特征。本文通过随机森林，引入人体行为分类数据实现。第二部分为分析不同行为类别下特征的差异，本文通过非参数核密度估计图，对数据极限近似拟合，判断不同行为类别下特征的差异实现。第三部分为对特征选择的结果进行适当的分析和检验，说明所选取特征的合理性。本文通过测试集中的数据，对模型的准确性进行分析检验，比较特征提取前后模型精度的差异，以此证明特征提取的合理性。

5.1.1 模型的建立

（一）第一部分：随机森林筛选重要观测特征

Step1：数据集的考察

考察数据集，数据集共包含 561 个变量，且所有变量的值均通过去噪、加窗、分割、归一化等处理，这导致我们无法精确获得每一个变量数值的物理含义。考虑到数据集的变量众多，以及物理含义的缺失，**使用传统的降维方法，如主成分分析、因子分析等，不易对选出的变量进行合理的解释。**从而大大降低后续构建模型的可解释性，失去模型的实用价值。本文通过降维方法中的变量选择实现对重要观测特征的筛选。变量选择主要是通过统计方法从繁多的变量中选出对响应变量有很大影响的解释变量，它是统计分析和推断的重要环节。变量选择的结果的好坏影响所建模型的质量，进而对统计预测精度产生较大的影响。[4]查阅相关文献，以及通过机器学习方法的对比，分析模型的实用价值，本文最终采用随机森林法筛选出了 10 个重要观测特征。

Step2：随机森林模型

随机森林，是指利用多棵树对样本数据进行训练和预测的一种分类器。对于每棵树，其使用的训练集从总训练集中有放回采样选出，这意味着总训练集中有些样本可能多次出现在一棵树的训练集中，也可能从未出现在一棵树的训练集中。在训练每棵树的节点时，使用的特征从所有特征中按照一定比例进行随机无放回抽取而产生。

随机森林的训练过程可以总结为如下步骤：

（1）随机产生训练集，根据每个训练集生成对应的决策树，确定每棵树的深度，每个节点使用到的特征数量。

（2）将子节点使用的特征数量作为分类特征集，并使用分裂法对特征向量的节点进行分裂。

（3）每棵树都不剪枝，使其完整长成。

（4）输入测试样本集，使用每棵决策树对其进行分类测试，将其结果进行保存。

（5）采用投票法确定类别，将类别中票数最多的一类作为测试样本的类别。

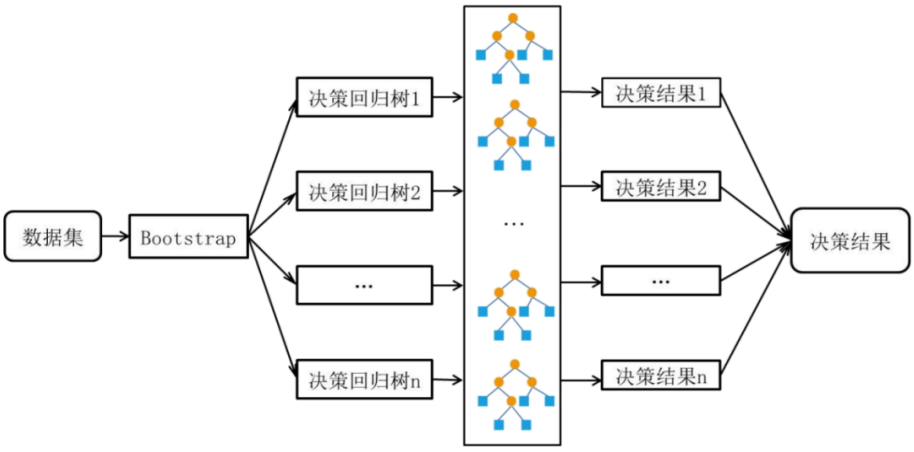

图 1　随机森林的训练过程

本文通过具体的行为类别，构建随机森林模型筛选出 10 个特征信息。

（二）第二部分：核密度估计比较不同类别下特征差异

非参数核密度估计[5]可完全基于数据驱动实现可靠性指标的概率密度估计，是一种从数据样本本身出发，研究数据分布特征的方法。本文采用非参数核密度估计分别对随机森林筛选出的 10 个特征信息，对其在静态、动态、静态姿势转换的三大类行为类别进行比较，绘制非参数核密度估计图，以直观表示筛选出的指标在不同行为类别下的特征差异。

令第j个属性特征x_j的T个样本为$x_{1_j}, x_{2_j}, \ldots, x_{T_j}$，则基于非参数核密度理论可得不同特征$x_j$的概率密度函数$f_k(x_j)$为：

$$f_k(x_j) = \frac{1}{Th}\sum_{i=1}^{n} K\left(\frac{x_j - x_{ij}}{h}\right) \tag{1}$$

式中h为带宽；T为样本数；x_{ij}为第j个属性特征的第i个样本值；K为核函数。

为保证被估计概率密度函数的连续性，核函数$K(x_j)$应为对称平滑非负函数，需满足以下约束：

$$\begin{cases} \int K(x_i)\mathrm{d}x_i = 1 \\ \int x_{ij}K(x_i)\mathrm{d}x_i = 0 \\ \int x_{ij}^2 K(x_i)\mathrm{d}x_i = a > 0 \end{cases} \tag{2}$$

式中a为正常数。

由于满足公式（2）条件的核函数选择多样，且不同的核函数对非参数密度估计的精确性影响不大，因此本文选择高斯函数作为概率密度估计的核函数：

$$K(x_i) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_i^2}{2}\right) \tag{3}$$

由公式（1）（3），则概率密度函数的非参数和密度估计变为：

$$f_k(k_j) = \frac{1}{\sqrt{2\pi}Th}\sum_{i=1}^{n}\exp\left[-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right] \tag{4}$$

据此，绘制关于选出的十种特征的非参数核密度估计图。

（三）第三部分：混淆矩阵与对比检验评价模型

上述模型已通过随机森林对训练集筛选出 10 个重要的观测特征信息。本部分通过测试集中的数据，对模型的准确性进行分析检验，判断模型是否存在欠拟合或过拟合现象。绘制混淆矩阵直观展示模型的结果，并比较特征提取前后模型精度的差异，以此证明特征提取的合理性。

混淆矩阵是一种常用的分类模型评估方法，它把预测结果按照真实类别的不同进行组合，并以矩阵的形式展现出来。其主要由以下四个指标组成：

真阳性（True Positive，TP）：正确预测为正样本的数量；

假阳性（False Positive，FP）：错误预测为正样本的数量；

真阴性（True Negative，TN）：正确预测为负样本的数量；

假阴性（False Negative，FN）：错误预测为负样本的数量；

混淆矩阵通过这四个指标，可以直观评估分类模型的性能表现，比如模型的准确性、召回率、精度等。

5.1.2 模型的求解

（一）第一部分：随机森林筛选重要观测特征的求解

通过随机森林模型，筛选出以下 10 个重要特征。

表 1　随机森林筛选出的 10 个重要特征

编号	原始变量名称	变量对应物理含义
1	tGravityAcc-max()-X	不同时间下，重力加速度信号的 X 轴分量的最大值
2	tGravityAcc-mean()-Y	不同时间下，重力加速度信号 Y 轴分量的平均值
3	tGravityAcc-energy()-Y	不同时间下，重力加速度信号 Y 轴分量的能量
4	tGravityAcc-arCoeff()-Y,1	不同时间下，重力加速度信号 Y 轴分量的自回归系数
5	tBodyAcc-mean()-X	不同时间下，身体加速度信号 X 轴分量的平均值
6	tBodyAcc-max()-X	不同时间下，身体加速度信号 X 轴分量的最大值
7	tBodyAcc-mean()-Y	不同时间下，身体加速度信号的 Y 轴分量的平均值
8	tBodyAcc-correlation()-X,Y	不同时间下，身体加速度信号的 X、Y 轴分量的相关性
9	fBodyAccMag-std()	频域上，身体加速度信号的幅度的标准差
10	tBodyGyro-correlation()-Y,Z	不同时间下，角速度信号 Y 轴和 Z 轴分量的相关性

手机的坐标系统相对固定，与手机运动方向无关。绘制如下智能手机的传感器记录的三轴信息与手机所处平面的关系，以更好地展示 XYZ 轴与各组物理量之间的关系。

查阅相关文献后[1]，总结出传感器的具体作用如下：

线性加速度传感器得到用户的身体加速度信号，可通过此信号判断出：实验者通过自身运动产生的线性加速度引起的惯性力。在实验中可评估智能手机运动轨迹的平动速度变化情况。重力加速度传感器可以将重力在手机的所处平面上分解，判断手机的水平面，从而计算

出手机屏幕和水平面的交角。在实验中可评估智能手机自身的摆动程度。陀螺仪测量出的角速度信号，可以判断实验者提供角动量对手机产生的扭力。在实验中可评估智能手机运动轨迹的转动情况。

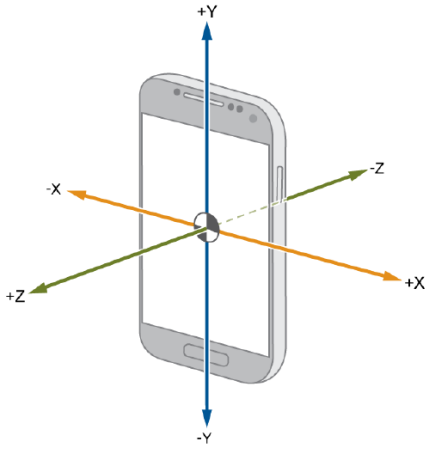

图 2　智能手机的传感器示意图

对 10 个重要特征的解释：

上述指标有 4 个与重力加速度信号相关、5 个与身体加速度信号相关，仅有 1 个与角速度相关。联系实际情况，做实验所述动作对手机在空间中的旋转较小，而对手机自身的旋转与手机的二维运动情况影响较大。而具体的分类，如选取的最大值、平均值、自回归系数等，则是最能代表该物理量在不同行为类别下特征差异的指标。

参考智能手机的坐标系统图，推测由于智能手机的固定位置，导致手机所处平面与身体平面平行，实验过程中 Z 轴方向近似不变，故身体加速度与重力加速度指标在 Z 轴方向上的变化情况小，在对指标的初步筛选中被剔除。

综上所述，将筛选出的 10 个重要特征联系实验的背景，初步说明本文筛选出的特征具有合理性。

（二）第二部分：核密度估计比较不同类别下特征差异的求解

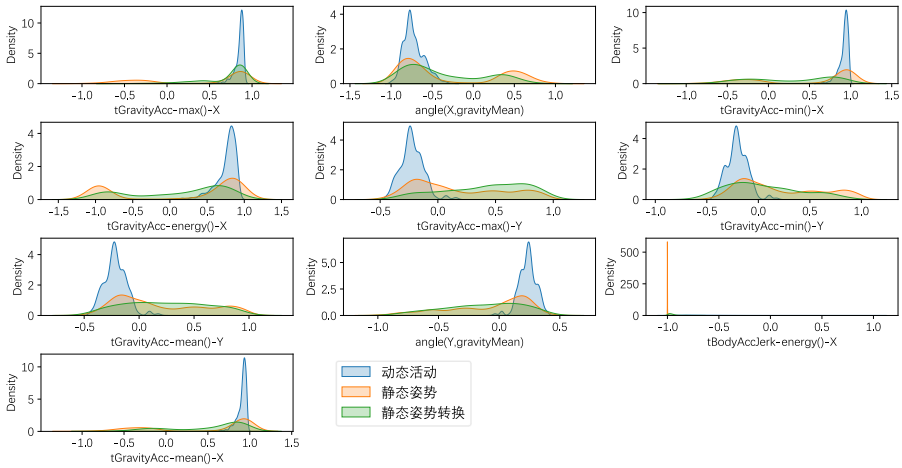

图 3　核密度估计情况

本文对 12 种人体行为，归纳为 3 个大类，即动态活动、静态姿势、静态姿势转换。对于上文选取的 10 个重要特征绘制核密度估计图，定性直观比较重要特征在不同类别下的特

征差异。本文对核密度估计情况的展示如图 3：

由图可比较得出，以上 10 种指标的核密度估计曲线在形状或峰值处均具有较大差异，故可以认为在不同类别下特征的差异较为明显。

（三）第三部分：混淆矩阵与对比检验评价模型的求解

对筛选后的 10 个变量与未筛选的 561 个变量，分别使用随机森林算法进行分类，通过测试集的精度判别进一步比较选取特征的合理性。通过随机计算得出变量筛选前后训练集的混淆矩阵图：

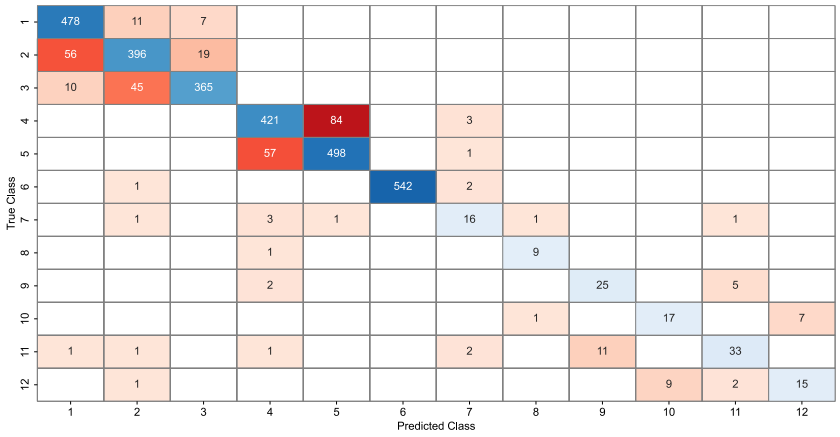

图 4　未进行变量筛选前的混淆矩阵

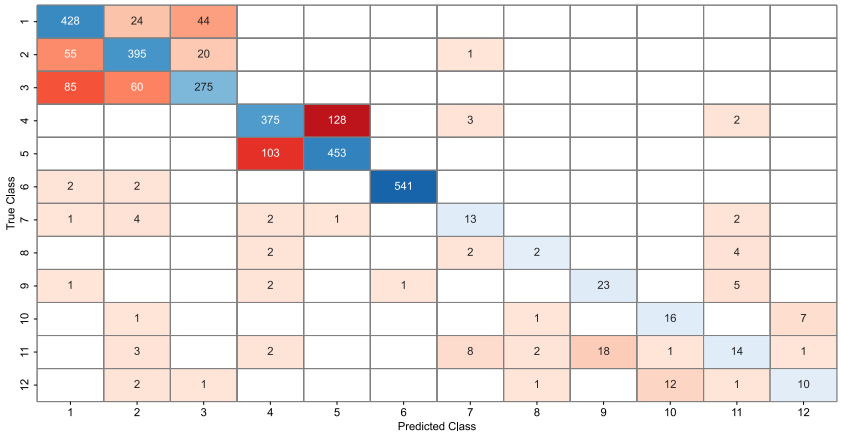

图 5　筛选变量后的混淆矩阵

由上图可以看出变量筛选前后的混淆矩阵基本类似，模型训练的具体信息如下：

表 2　训练结果

是否进行变量筛选	预测准确度（%）	训练用时（秒）
未进行变量筛选	89.03	35.46
变量筛选后	80.49	1.48

根据模型训练的参数可以看出，变量筛选前后预测准确率基本相当，但由于变量数目显著降低，模型训练用时有着明显下降，筛选后的变量在实际的人体行为识别研究中有着更好的应用。

5.2 问题二 基于 SVM 模型对 6 种行为类别的分类研究

问题二包括两部分，第一部分为构建对动态、静态姿势共 6 种具体行为的人体行为识别模型。本文采用支持向量机（SVM）模型对 6 种行为类别首先做动态、静态姿势的二分类，然后再对动态、静态活动中具体行为活动做进一步鉴别分类。分布进行分类。第二部分为评估特征选择、参数调优等对识别结果的影响。本文对三次 SVM 分类分别进行参数调优，并将未筛选的 561 个指标和筛选后的指标进行模型分类精度比较，证明特征选择的意义。

5.2.1 模型的建立

（一）第一部分：SVM 对 6 种行为活动的分类

支持向量机[6]（Support Vector Machine，SVM）属于监督式学习，在统计分类和回归分析中应用较广。SVM 算法的主要优点为：解决高维特征的分类问题和回归问题很有效；仅仅使用一部分支持向量来做超平面的决策，无需依赖全部数据。有大量的核函数可以使用，从而可以很灵活的来解决各种非线性的分类回归问题。样本量不是海量数据的时候,分类准确率高，泛化能力强。查阅相关文献后，发现类似的人体行为活动分类大多采用 SVM 模型分类。为本文的模型选择提供了理论依据。[1,2,3]

如图2.2所示为一个线性可分的最优分类界面，实心圆点和空心圆点分别表示两类样本，实线为分类线，虚线上的样本即为支持向量。使用$x_1, x_2, \ldots, x_n$代表这些数据点，每个x所代表的数据都有m维，用y代表其属于的类别，w是分类界面的法向量，b代表常数，则分类界面可以表示为式（5）的形式：

$$y = w^T x + b \tag{5}$$

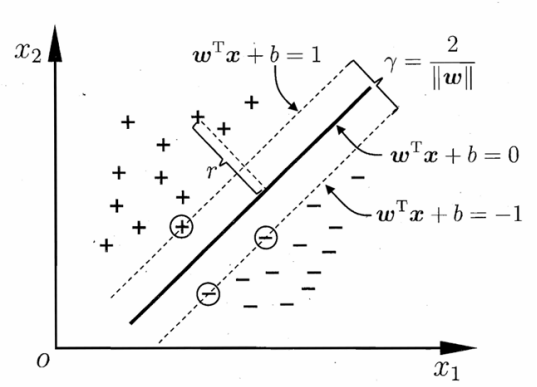

图 6　线性可分的最优分类界面图

分类器的支持向量为令$w^T x + b = 1$及$w^T x + b = -1$的数据点，所要寻找最优分类界面即寻找最优的w与b。

间隔d是两条支持向量所在虚线之间的距离，表示为式（6）：

$$d = \frac{w}{\|w\|}(x_1 - x_2) = \frac{2}{\|w\|} \tag{6}$$

因此可以得出最优超平面便是求得最大间隔d，最大间隔d就是求得最小$\|w\|$，也即求最小的$\|w\|^2$，即将最优超平面的求取抽象为式（7）所示函数：

$$\begin{aligned} &\min \frac{\|w\|^2}{2} \\ &\text{s.t. } y_i(w^T x + b) - 1 \geq 0 \end{aligned} \tag{7}$$

定义拉格朗日函数如式（8）所示：

$$L(w,b,a)=\frac{\|w\|^2}{2}-\sum_{i=1}^{N}\alpha_i[y_i(wx_i+b)-1] \tag{8}$$

其中α_i为拉格朗日乘子，对w和b求偏导并置零：

$$\frac{\partial L}{\partial w}=0\Longrightarrow w=\sum_{i=0}^{N}\alpha_i y_i x_i \tag{9}$$

$$\text{s.t. }\frac{\partial L}{\partial b}=0\Longrightarrow\sum_{i=1}^{N}\alpha_i y_i x_i$$

整理得到式（10）：

$$L(w,b,a)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \tag{10}$$

式（10）中$\alpha_i\alpha_j$和$y_i y_j$均是常数，而$x_i^T x_j$是线性核核函数的计算公式。

在样例线性可分的基础上进一步拓展，当样例线性不可分时，则可尝试使用核函数来将特征映射到高维。然而，映射也无法保证完全可分。因此需要将模型进行调整，以保证在不可分的情况下，也能够尽可能地找到分隔超平面。此时应允许一些点游离并在模型中违背限制条件（函数间隔大于 1）。新模型如式（11）所示：

$$\min_{r,w,b}\frac{\|w\|^2}{2}+Q\sum_{i=1}^{m}\xi_i \tag{11}$$

$$\text{s.t.}\begin{cases}y_i(w^Tx+b)\geq 1-\xi_i, i=1,2,\dots,m\\ \xi_i\geq 0, i=1,2,\dots,m\end{cases}$$

引入非负参数ξ_i（也称为松弛变量）后，就允许某些样本点的函数间隔小于 1，即在最大间隔区间里，函数间隔允许为负数，即样本点在对方的区域中。而放松限制条件后，需要调整目标函数，以对离群点进行惩罚，$Q\sum_{i=1}^{m}\xi_i$的表示离群点越多，目标函数值越大，而要求的是尽可能小的目标函数值，以此实现对离群点的惩罚。其中Q是离群点的权重，Q越大表明离群点对目标函数影响越大。因此，目标函数控制了离群点的数目和程度，使大部分样本点仍然遵守限制条件。模型修改后，拉格朗日公式也进行修改，结果如式（12）所示：

$$L(w,b,\xi,\alpha,r)=\frac{\|w\|^2}{2}+Q\sum_{i=1}^{m}\xi_i-Q\sum_{i=1}^{N}\alpha_i[y_i(w^Tx_i+b)-1]-\sum_{i=1}^{m}r_i\xi_i \tag{12}$$

式（12）中α_i和r_i都是拉格朗日乘子，因此有：

$$\max_{\alpha}H(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \tag{13}$$

$$\text{s.t.}\begin{cases}0\leq\alpha_i\leq Q, i=1,2,\dots,m\\ \sum_{i=0}^{N}\alpha_i y_i=0\end{cases}$$

$$\begin{cases}\alpha_i=0\Longrightarrow y_i(w^Tx+b)\leq 1\\ \alpha_i=Q\Longrightarrow y_i(w^Tx+b)\leq 1\\ 0<\alpha_i<Q\Longrightarrow y_i(w^Tx+b)=1\end{cases} \tag{14}$$

式（14）表明在两条间隔线外的样本点之前的系数为 0，离群样本点前面的系数为Q，而支持向量（也即在超平面两边的最大间隔线上）的样本点之前的系数在$(0,Q)$上。某些在最大间隔线上的样本点也不是支持向量，相反也可能是离群点。

通过式（14）可以求出变量α_i，进而w和b也可以求出，因此得到完整的分类公式，将测试数据输入求解的模型中，可通过运算结果判断测试数据属于哪个类别。

已有文献中指出：对六种动作采用 SVM 模型进行多分类，存在对静态动作分类精度不足的缺陷。本文考虑实际情况，对动态与静态动作准确预测的应用价值，要远大于对具体动作预测的价值。因此本文先对动作的动态、静态进行判别分析，然后再区分具体动作类型。一定程度上，弥补了已有文献的不足。[2]

（二）第二部分：参数寻优以及特征选择的意义

本文对 SVM 的常用核函数进行参数寻优，得出适用于上述三个分类器的最佳分类方法。并参考问题一，通过混淆矩阵，直观评估分类模型的性能表现，比如模型的准确性、召回率、精度等。

5.2.2 模型的求解

（一）第一部分：SVM 对 6 种行为活动的分类求解

（1）利用 SVM 模型做动静态判别的二分类

在 Python 中输入训练集，训练模型后，在测试集中基于 SVM 模型对动态、静态姿势判别的二分类模型，混淆矩阵中的 TRUE 表示结果为动态的人体行为，绘制可视化的混淆矩阵图如下：

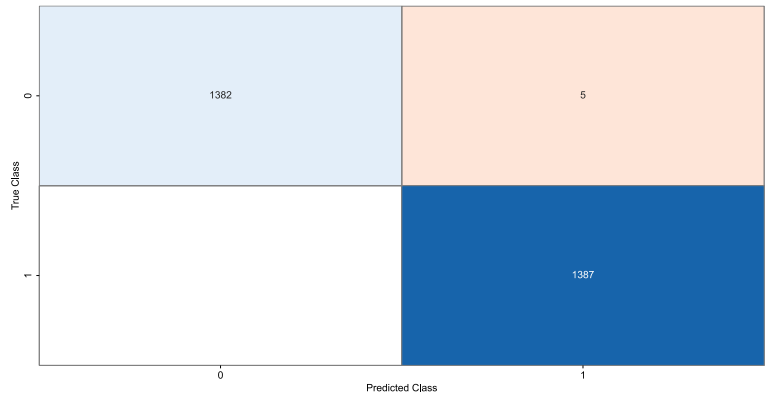

图 7　动静态判别的混淆矩阵图

根据测试结果，SVM 模型的精度达到了 99.82%。在错误预测中，仅有 5 个动态的人体行为的错误预测。从错误类型看，此模型对静态模型的识别准确性更高。从总体精度看，进一步验证了上述特征选择的合理性。以及此分类器在实际应用中能够对人体行为中静态、动态姿势的判别发挥较好的作用。

（2）利用 SVM 模型对动态和静态的进一步判别：

对训练集的动态和静态数据进行筛选，在 Python 中分别根据数据对其进行训练，在测试集中预测结果如下：

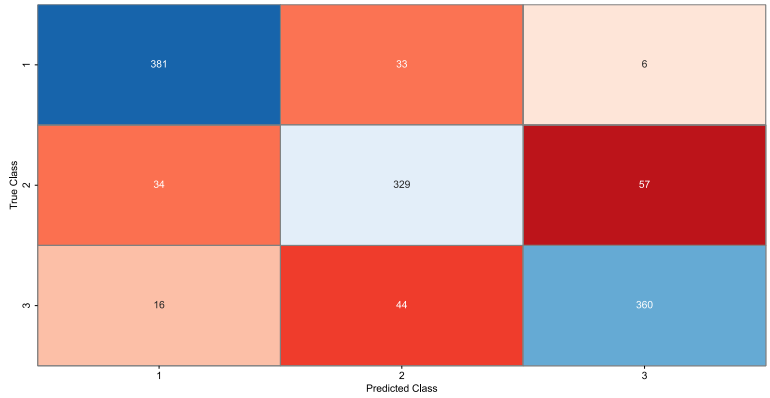

图 8　对动态数据的测试结果

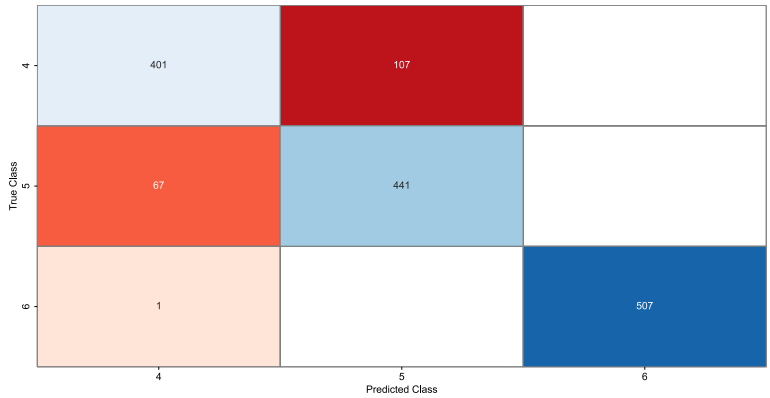

图 9　对静态数据的测试结果

根据测试结果，对动态数据的预测精度为 84.92%，对静态数据的预测精度为 88.76%。在对静态数据的预测中，对“躺着”人体行为的预测结果准确度高，而对站着和坐着预测结果稍差。而在动态数据的预测中，对“下楼”的人体行为识别准确度较高，“上楼”和“步行”的预测结果稍差。

（二）第二部分：参数寻优以及特征选择的意义

（1）对 SVM 分类器进行参数优化

根据已有的核函数，对 SVM 分类器进行参数寻优，以寻求最优结果。对二分类和两次三分类的 SVM 模型结果进行多次参数调优，得到如下表的最优分类结果：

表 3　参数优化后最优分类结果

分类内容	精度（%）	最优核函数与算法
对动态与静态的二分类	99.63	高斯核、SMO
对动态行为的详细分类	88.18	线性核、SMO
对静态行为的详细分类	89.62	高斯核、L1QP

（2）特征选择对 SVM 分类结果的影响

对已筛选的 10 个观测特征信息和未筛选的 561 个观测特征信息，在分类精度和分类所需时间两项指标中比较其差异。结果如下表：

表 4　特征选择后分类结果

分类内容	是否进行变量选择	分类精度（%）	分类所需时间
对动态与静态的二分类	未进行变量选择	99.87	8.40s
	进行变量选择	99.51	<1s
对动态行为的详细分类	未进行变量选择	98.09	7.98s
	进行变量选择	87.44	<1s
对静态行为的详细分类	未进行变量选择	95.68	7.35s
	进行变量选择	88.20	<1s

由上表可以看出，筛选变量后，分类精度比筛选变量前略有降低，但分类所需时间明显减少。筛选变量后在可接受范围内，略微降低了分类器的精度，但模型的复杂程度大幅下降，因此特征选择在实际的应用中有着更好地运用价值。

本文先对动作的动态、静态进行判别分析，然后再区分具体动作类型。已有文献则将对六种动作采用 SVM 多分类。本文提供的方法对静态动作的分类效果优于原文献。一定程度上，弥补了已有文献的不足。[2]

5.3 问题三 引入静态姿势转变的分类模型

问题三包含两个部分，第一部分为：对 12 中行为类别重构人体行为识别模型。本文通过 k 近邻法（KNN）重新构建基于特征信息的人体行为识别模型，并进行参数调优，指标筛选前后的分类精度比较，以及通过混淆矩阵，给出较高误判率的人体行为类别，并通过数据本身和人体运动的两个角度给出其原因。第二部分需改进上述模型，本文从观测特征的重新选取出发，通过随机森林筛选观测特征信息，并构建 3 组二分类模型，并通过 ROC 受试者工作曲线分析评价构建的模型。

5.3.1 模型的建立

（一）第一部分：KNN 重构人体行为识别模型

上文中构建了适合二分类的 SVM 模型，并用改进后的 SVM 模型完成了三分类。以此分两步完成了对 6 个行为类别的判定。考虑到在实际人体行为中，静态姿势转换与动态姿势所产生的动作类似，且容易对二者中的具体类别进行误判。故针对本问题，本文对 12 种行为分类，通过同一模型进行判别，提升模型的应用性。相比于问题二中的二分类与三分类，SVM 多分类的缺陷在于处理大规模多分类问题时，需要构建大量的分类器，对计算资源要求较高。同时，对于非线性分类问题，SVM 需要将输入样本进行映射，构建高维特征空间，计算复杂度也较高。因此认为 SVM 模型不适合 12 种行为类别的分类。本文使用适合多分类的 k 近邻法（KNN）重新构建基于特征信息的人体行为识别模型。

本文使用 k 近邻算法进行分类，原因如下：

（1）样本数量较大：本题给出的训练数据集包含了 7767 个样本，这样的样本数量足够支持 k 近邻算法的分类任务。

（2）样本分布较为均匀：本题给出的数据集中，不同类别的样本数量相对均衡，不存在明显的样本不平衡问题。

（3）k 近邻算法适用于非线性分类问题：人体行为识别问题是一个非线性分类问题，k 近邻算法可以通过计算样本之间的距离来确定每个测试样本的类别，适用于这种非线性分类问题。且文献也采用过 k 近邻算法进行多分类且效果较好。[1]

Step1：KNN 算法介绍

k 近邻（K-Nearest Neighbor，KNN）分类算法是从训练样本中找到和待分类样本最接近的 k 条记录，然后依据其类别来判定待分类样本的类别。

在算法流程中，每输入一个待分类样本，则计算这个未知样本与所有训练样本的相似性，一般采用n维欧氏距离d，如式（15）所示：

$$d = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2} \tag{15}$$

$x_1, x_2, \ldots, x_n$和$x_1', x_2', \ldots, x_n'$分别是未知样本和训练样本的n个特征向量的值，然后选取其中k个最小距离，并且计算这k个训练样本中属于每个类别的样本个数。最后可将待分类样本归类为与其相距最近的k个训练样本中最多数的一类。

Step2：KNN 算法的参数调优

KNN 算法主要调整的参数包含调整k值和调整距离的计算方法。一般距离的计算采用欧式距离，欧氏距离是最常见的距离度量，衡量的是多维空间中各个点之间的绝对距离。在实际调参中，一般调整k值，即代表考虑的邻居个数，k一般低于训练样本数的平方根。考虑上述要求，本文对k值选择在 2~88 内调整，以寻求一个最合适的分类结果。

（二）第二部分：随机森林改进分类模型

Step1：随机森林模型

考虑到问题二已对动态姿势与静态姿势，以及具体的静态姿势的判别有较高的精度。本问题可用适合用于变量选择的随机森林模型，将六个具体的静态姿势转换拆解为，由站着、坐着、躺着发起的三组二分类。重新在 561 个特征值中进行筛选，构建模型。

Step2：ROC 曲线

接收者操作特征或称 ROC 曲线是一种对于灵敏度进行描述的功能图像。ROC 曲线可以通过描述真阳性率（TPR）和假阳性率（FPR）来实现。它能够很好的描述分类器对于不均衡分布的样本的分类性能。

下面对 ROC 曲线的说明。

对于一个分类器的分类结果，一般有以下四种情况：

1. 真阳性（TP）：判断为 1，实际上也为 1。
2. 伪阳性（FP）：判断为 1，实际上为 0。
3. 真阴性（TN）：判断为 0，实际上也为 0。
4. 伪阴性（FN）：判断为 0，实际上为 1。

TPR：在所有实际为阳性的样本中，被正确地判断为阳性的比率。

FPR：在所有实际为阴性的样本中，被错误地判断为阳性的比率。

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

ROC 空间将伪阳性率（FPR）定义为x轴，真阳性率（TPR）定义为y轴。

本文通过 ROC 曲线评估上述模型的准确性。

5.3.2 模型的求解

（一）第一部分：KNN 重构人体行为识别模型的求解

Step1：特征选择对 KNN 分类结果的影响

在k值为默认值 1 的条件下，对已筛选的 10 个观测特征信息和未筛选的 561 个观测特征信息，在分类精度和分类所需时间两项指标中比较其差异。结果如下表：

表 5　特征选择对 KNN 分类结果的影响

是否进行变量选择	分类精度（%）	分类所需时间
未进行变量选择	82.23	27.57s
进行变量选择	86.68	3.04s

由上表可以看出，筛选变量后，分类精度比筛选变量前有所提高，且分类所需时间明显减少。筛选变量后不仅提高了分类器的精度，而且大大降低了模型的时间复杂度，因此本题下情境下，多分类前先进行特征选择有利于精度的提升。且本题不同于第二问的是，本题采用的是 KNN 算法进行多分类，核心在于输入样本的种类取决于与其相距最近的k个训练样本中最多数的一类。若特征过多会使模型错判与输入样本最近的样本致使分类结果错误。

Step2：k 值的选取对 KNN 分类结果的影响

改变超参数k，观察对已筛选的 10 个观测特征信息和未筛选的 561 个观测特征信息，在分类精度上的差异。结果如下：

表 6　k值的选取对 KNN 分类结果的影响

k 值	未进行选择分类精度（%）	进行选择分类精度（%）
1	82.23	86.68
10	83.10	88.68
20	83.19	88.98

观察上表可知：

随着k值增大，是否进行特征选择的分类精度都有所提高，其中k值由 1 升至 10 时精度增加较多，推测此时因为模型变得更加稳定，不容易受到噪声或异常值的影响及样本分布的影响。当k值由 10 升至 20 时，变化幅度较小，推测是因为当k值增大到一定程度时，样本之间的距离已经不再是主要的影响因素，因此准确率不再随k值的增大而发生显著变化。

不论k值如何变化，进行特征选择后的精度均高于未进行特征选择后的精度。推测是因为数据集中存在影响进行行为类别判断的噪声或异常值，进行特征选择后显著提高了模型分

类精度。

Step3：通过混淆矩阵分析误判率

在 MATLAB 中输入训练集，训练模型后，在测试集中基于 k 近邻多分类模型对十二种姿势判别的多分类模型，混淆矩阵中的横纵坐标分别对于十二种人体行为，绘制可视化的混淆矩阵图如下：

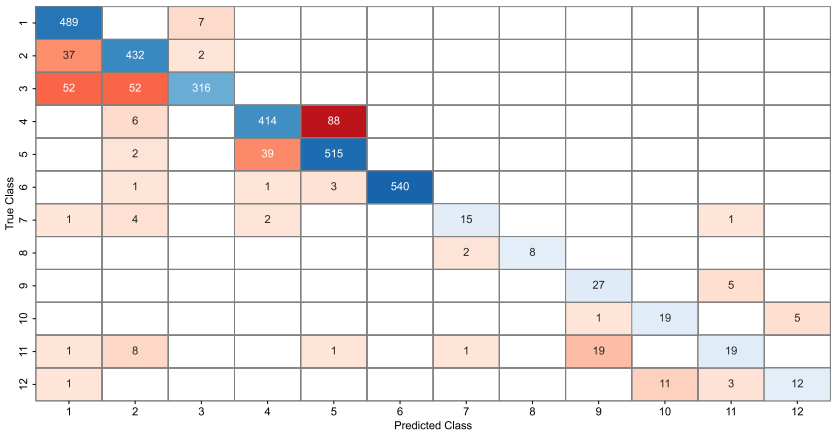

图 10　多分类混淆矩阵

由混淆矩阵分析可知，模型对于躺着的姿态判别最为准确，准确率达 100%，符合第二问结论。且模型对于 1~6 号无姿势转换的行为判别率较高，对于 7~12 号静态姿势转换的行为判别率显著降低。有相关文献证明，姿势转换的行为判别确实是人类行为识别任务的一大难点[1]。

在静态姿势转换的行为判别率中最低的三个行为是：站着转换为躺着(11)、站着转换为坐着(7)、躺着转换为坐着(10)。准确率分别为 34.61%、43.38%、48.16%。

本文分别从人体力学和数据处理角度层面分析原因。

1、人体力学：在动作的力学特征中，坐、站、躺都是静态姿势，姿势的力学特征相似，判别较为困难；在身体的重心变化中，不同的动作和姿势会导致身体重心的变化，但由于坐、站、躺三种姿势皆为静态姿势，所以进行姿态转换时重心变化不大。所以判别较为困难。

2、在数据处理方面，本题提供的针对静态姿势转化的样本数量不足，模型无法充分学习到它们的特征，从而导致分类效果不佳。

3、前文所筛选的变量中，特征提取不充分。这三个行为的特征可能与其他行为相似度较高，难以区分。因此，需要更加细致地分析这些行为的特征，提取更加有效的特征。

（二）第二部分：随机森林对分类模型的改进

Step1：通过随机森林筛选变量

上文推测模型误判率较高的原因可能是由于前文所筛选的变量，特征提取不够充分，导致模型难以对静态姿势转换做具体的区分。

由于问题二中的 SVM 模型已对具体的静态姿势有着较好的区分，本部分选择基于适合用变量选择的随机森林模型，将六个具体的静态姿势转换做更细致的分类，具体做法为：将 6 个静态姿势转变拆解为：由站着、坐着、躺着发起的三组二分类。本着在模型精度变化不大的前提下，尽可能减小选择变量的数量，以达到更好的使用价值。在上述准则的要求下，

对随机森林模型算法调优，筛选出每组二分类的最优观测特征信息为 6 个，三组二分类变量筛选结果如下：

表 7　三组二分类变量筛选结果

转变前动作	编号	原始变量名称	变量对应物理含义
站着	1	tGravityAcc-max()-Y	不同时间重力加速度信号 Y 轴分量的最大值
	2	tGravityAcc-mean()-Y	不同时间重力加速度信号 Y 轴分量的平均值
	3	tBodyAcc-entropy()-Y	不同时间身体加速度信号 Y 轴分量的信号复杂度
	4	tBodyAcc-max()-Z	不同时间身体加速度信号 Z 轴分量的最大值
	5	tBodyGyro-entropy()-X	不同时间身体陀螺仪信号 X 轴分量的信号复杂度
	6	tBodyGyro-energy()-X	不同时间身体陀螺仪信号 X 轴分量的能量
坐着	1	tBodyAcc-mean()-Y	不同时间身体加速度信号 Y 轴分量的平均值
	2	tBodyAcc-arCoeff()-X,4	不同时间身体加速度信号 X 轴分量的自回归系数
	3	tBodyAcc-mean()-X	不同时间身体加速度信号 X 轴分量的平均值
	4	tBodyAcc-min()-Y	不同时间身体加速度信号 Y 轴分量的最小值
	5	tBodyAcc-max()-Y	不同时间身体加速度信号 Y 轴分量的最大值
	6	tGravityAcc-correlation()-X,Y	不同时间重力加速度信号 XY 轴分量之间的相关性
躺着	1	tBodyGyroJerkMag-max()	不同时间角速度变化率的最大值
	2	tBodyGyroMag-arCoeff()3	不同时间角速度的大小的自回归系数
	3	tBodyGyroMag-min()	不同时间角速度大小的最小值
	4	tBodyAcc-mean()-Z	不同时间身体加速度信号 Z 轴分量的平均值
	5	tBodyAcc-mean()-Y	不同时间身体加速度信号 Y 轴分量的平均值
	6	tBodyAccMag-arCoeff()3	不同时间加速度的大小的自回归系数

注：转变前的动作，即由该动作发起。如转变前的动作为站着，即包含站着转换为坐着(7)和站着转换为躺着(11)。

根据上表，随机森林变量筛选时，对不同的发起动作筛选出的变量不同，证明用均一的

指标对 12 个动作的判别具有不合理性质。结合问题一中绘制的智能手机的传感器示意图，对模型选择的变量作出如下解释。

在站着转换为坐着和站着转换为躺着的区分中，随机森林模型选取出了：重力加速度、身体加速度与角速度三类物理量。联系人体的实际动作，在站着转换为坐着中，人体的旋转换为扭动情况较小，同时人体的动作幅度较小；而在站着转换为躺着中，人体会产生较大的旋转扭动，动作幅度更大。介于两者的动作区别，直观分析智能手机传感器的重力加速度、身体加速度与角速度会有较为明显的差异。

在坐着转换为站着和坐着转换为躺着的区分中，随机森林模型则主要聚焦于身体加速度信号的不同。联系人体的实际动作，二者的动作幅度相当，但二者的身体加速度方向存在明显不同。介于两者的动作区别，直观分析智能手机传感器的身体加速度值会有明显差异。

在躺着转换为站着和躺着转换为坐着的区分中，随机森林模型选取出了：角速度、身体加速度与总体加速度三类物理量。联系人体的实际动作，二者在身体摆动、与动作幅度均有较大差异。但对手机屏幕的旋转均有较大程度的旋转，故此时针对重力加速度会因此失效。介于两者的动作区别，直观分析智能手机传感器的角速度、身体加速度与总体加速度会有较为明显的差异。

在不同物理量下的具体的分类，如选取的最大值、信号复杂度、自回归系数等，则是最能代表该物理量在不同行为类别下特征差异的指标。

Step2：随机森林模型的预测结果

对于三类动作分别绘制如下的混淆矩阵图，直观展示分类结果。

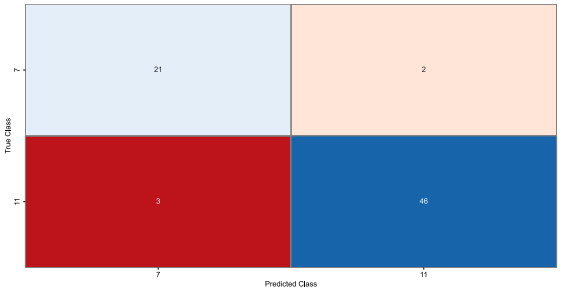

图 11　由站着转出的混淆矩阵

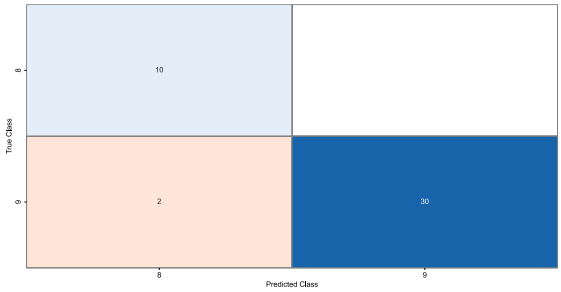

图 12　由坐着转出的混淆矩阵

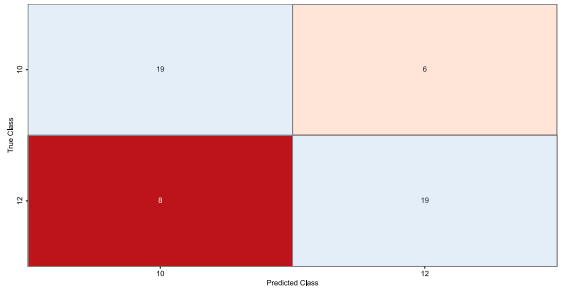

图 13　由躺着转出的混淆矩阵

由混淆矩阵可以看出，除由躺着发起的动作外，其余动作的分类的结果相比于 KNN 的分类已有了很大程度的改善。对于由站着发起的动作和由坐着发起的动作，其预测准确率均在 90%以上。对由躺着发起的动作准确率尚可，接近 70%。

对由躺着发起的动作较之前两者仍不够准确，推测是因为躺着和坐着、站着这两个姿势

的特征差异较小，身体的重心、肌肉的活动模式等都比较相似，因此难以区分。此外，躺着转换为坐着和站着的过程中，体需要克服重力的作用，需要进行较大的肌肉活动，这也增加了分类的难度。不同年龄段的志愿者由躺姿转变为坐姿和站姿时往往行为差异巨大，且数据集中为这两种姿态转换提供的数据量较小，因此分类效果不好。

Step3：使用 ROC 曲线进一步评价模型

绘制出模型的 ROC 曲线，曲线结果如下：

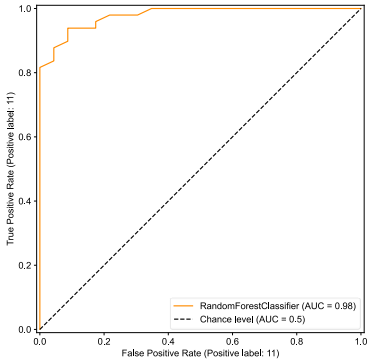

图 14　由站着转出的 ROC 曲线

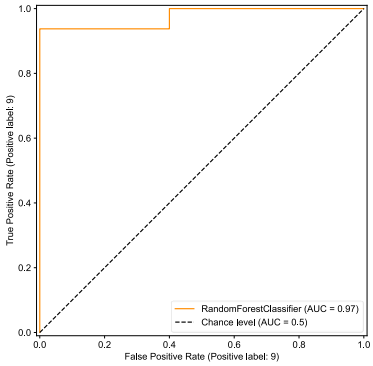

图 15　由坐着转出的 ROC 曲线

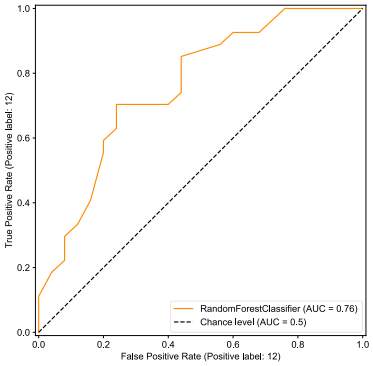

图 16　由躺着转出的 ROC 曲线

观察 ROC 曲线发现，由站着转出的动作和由坐着转出的动作的 ROC 曲线整体偏于左上部分，说明模型的真正率相对较高，模型能够较好地检测出正样本。而且，临界点靠近左上角也表明模型在分类阈值上的表现比较优秀，即在不同的分类阈值下，模型都能够保持较高的真正率和较低的假正率。这都说明该随机森林模型分类模型在分类由站着和坐着转出的动作时有较高的准确性和鲁棒性。

由躺着转出的动作的 ROC 曲线高于对角线但有一定波动，则表明模型的预测能力高于随机猜测，但仍然存在一定程度的误分类。但 ROC 曲线的高低并不一定完全反映出分类模型的优劣，因为 ROC 曲线只关注真正率和假正率这两个指标，而忽略了其他重要的指标，例如准确率和查全率等。

5.4 问题四 佩戴位置发生变化时的改进

问题四的第一部分为讨论佩戴位置发生改变时模型是否依然使用。第二部分为讨论选择的特征有效性是否发生变化。第三部分为是否可以扩充其他的特征提高分类准确率。

5.4.1 对位置发生改变后模型适用性的讨论

查阅文献后，发现以下指标对位置较为敏感，即位置发生微小变动后指标的值会发生变化。具体包括：垂直和水平加速度的最大值、陀螺仪（三轴）的最大值、三轴陀螺仪变化率的最大值、水平加速度的标准差、重力加速度和身体加速度之间角度的平均值、水平加速度范围。对比问题一中已筛选出的 10 个重要特征，比较后发现前文所构建的模型，包含了上述指标中的垂直和水平加速度的最大值(tGravityAcc-max()-X)和水平加速度范围(tBodyAcc-max()-X、tBodyAcc-mean()-X、tBodyAcc-mean()-Y)。可以推断前文所构建的模型对位置的要求具有敏感性，故上述所构建的模型及其选出的关键特征不适用。

为了考虑智能手机的位置，之前的大多数研究都采用了固定的位置，比如裤子口袋，腰部等。尽管智能手机用户可能会把它拿在手里或其他地方。这个设置限制了用户的自由。但手机位置的改变常会导致动作的显著错误分类。文献指出这个问题可以分为两种解决方案。一种是找到一些对位置不敏感的特征，以此进行分类；另一种是识别位置，并在具体部位构建出不同的模型。本文对两种方法都予以实施，以确保结果的准确性和全面性。

5.4.2 用不敏感特征判别的决策树模型

（一）模型的建立

Step1：建模思路

查找文献后发现以下 9 类物理量对位置较为不敏感。[2]具体如下表：

表 8　9 类对位置不敏感的物理量

物理量名称	在数据集中的体现
垂直加速度范围	tBodyAcc-max()-Z、tBodyAcc-min()-Z
垂直加速度变化率最大值	tBodyAccJerk-max()-Z
用 FFT 计算的垂直加速度时间	fBodyAccMag-meanFreq()
合成加速度能量	fBodyAccMag-energy()
垂直加速度过零率	tBodyAcc-min()-Z、tBodyAcc-max()-Z、tBodyAcc-mean()-Z
合成加速度平均值	fBodyAcc-mean()-X fBodyAcc-mean()-Y、fBodyAcc-mean()-Z
合成加速度方差	fBodyAcc-var()-X、fBodyAcc-var()-Y、fBodyAcc-var()-Z
水平加速度平均值	fBodyAcc-mean()-X、fBodyAcc-mean()-Y
合成加速度范围	tBodyAcc-max()-X、tBodyAcc-min()-X

注：上表在文献中总结而得，数据集中的某个具体指标可能在多个物理量中有体现。且本数据集不包含合成速度方差中所述的 3 个具体物理量。

本文通过在训练集中对以上 9 类物理量的筛选构建决策树模型，并通过测试集对模型的结果进行分析。

Step2：决策树模型

本问采取决策树进行建模来讨论选择的特征有效性是否发生变化，原因如下：

（1）决策树算法可根据特征的重要性进行特征选择，从而确定哪些特征对于分类任务最为重要。这可以帮助我们更加准确地确定哪些特征对于不同位置下的行为识别具有不同的影响。

（2）决策树算法可以直观地展示特征间的关系，帮助我们更好地理解不同特征之间的作用和影响。让我们更加深入地分析数据，深度剖析特征有效性。

（3）决策树算法具有较高的可解释性和可视化性，可以更加直观地理解模型的分类过程和结果。从而让我们更好地理解模型的优缺点，进行模型优化和改进。

决策树是一种树形结构，可用图形来形象化地表示。而在机器学习模型中，决策树是一种预测模型，表示对象属性与对象值之间的一种映射关系。在决策树结构中，每个内部节点对应于一个属性，并且具有对应于该属性的每个可能值的分枝，每个叶节点表示预测的一种类别。

决策树分类算法中具有代表性的是迭代二叉树 3 代（Iterative Dichotomiser 3，ID3）及其改进 C4.5 算法，以及分类与回归树（Classification And Regression Tree，CART）算法。其中，ID3 由 RossQuinlan 首先提出，以信息熵和信息增益为特征选择的评估标准。首先计算数据集中的每个特征的信息熵或者信息增益，每次选择具有最小的信息熵或最大的信息增益的特征作为分支标准，此过程一直递归到无剩余特征为止。

设M为训练的数据集，其中M有C个类别，并且假设样本集合M中第c类样本所占的比例为$p_c(c = 1,2,\dots,C)$，那么集合M的信息熵定义如式所示：

$$Info(M) = -\sum_{c=1}^{C} p_c \log_2 p_c \tag{18}$$

如果根据特征X有n个不同取值的离散特征，则使用特征 X 将数据集 M 划分为 n 个分枝后的信息熵，定义如式（19）所示：

$$Info_X(M) = \sum_{i=1}^{n} \left[\left(\frac{|M_i|}{|M|} \right) Info(M_i) \right] \tag{19}$$

通过式（18）和（19），特征X的信息增益的定义如式（20）所示：

$$Gain(M,X) = Info(M) - Info_X(M) \tag{20}$$

C4.5 算法是基于 ID3 算法的改进，与 ID3 不同，其采用信息增益率代替信息增益来选择特征，解决了用信息增益选择特征时可选特征不足的问题。C4.5 算法产生的分类规则便于理解，通常其准确率较高。但是在构造决策树的过程中，需要对数据集进行多次扫描与排序，因而算法效率较低。

C4.5 算法使用增益率的定义如式（21）所示：

$$Gain(M,X) = \frac{Gain(M,X)}{IV(X)} \tag{21}$$

其中IV称作分支标准X的固有值，作为分支标准的特征属性，其可取值越多则IV值越大，其定义如式（22）所示：

$$IV(X) = -\sum_{i=1}^{n} \frac{|M_i|}{|M|} \log_2 \frac{|M_i|}{|M|} \tag{22}$$

CART 分类算法由 Leo Breiman 等人首先提出，该算法可用于创建分类树与回归树，本文只讨论 CART 创建决策树的问题。

CART 算法假设决策树是一棵二叉树，属于二分递归分割方法。在构造决策树的过 程中，每个特征节点都对样本数据进行二分，即对非叶子特征节点生成两个分支。每次 对决策树内部特征节点进行决策时只能有“是”或“否”两个取值，即使一个特征的取 值有多个，在划分时也只能分割成“是”与“否”二者之一。

CART 需要构造尽可能大的决策树，需要用测试集的数据对已生成的决策树进行减枝，然后选择其中的最优子树。CART 算法使用基尼指数作为分枝标准，基尼指数是度量数据不纯度的一种方法，为介于 0 到 1 的数。基尼指数值越小，说明训练样本集的纯 净度越高，反之则说明训练样本集的纯净度越低，样本的类别越复杂。基尼指数反映了从训练数据集中随机抽取两个子样本类别不同的概率。基尼指数的定义如式（23）所示：

$$Gini(M) = 1 - \sum_{i=1}^{m} p_i^2 \tag{23}$$

式（23）中M表示一个训练样本的数据集，m表示M中的类别C的个数，p_i表示M中任意一个子样本的记录属于C中第i个类别的概率。如果所有的记录都属于同一类，那么$p_1 = 1$，$Gini(D) = 0$，此时样本不纯度最低。

在 CART 算法中使用基尼指数构造二叉决策树时，将计算每个特征所有取值的基尼指数，然后计算特征分裂后的基尼指数。分裂后的基尼指数定义如式（24）所示：

$$Gini_s(M) = \sum_{i=1}^{2} \left[\frac{|M_i|}{|M|} Gini(M_i) \right] \tag{24}$$

Step3：模型的求解

对决策树模型所得结果绘制混淆矩阵图：

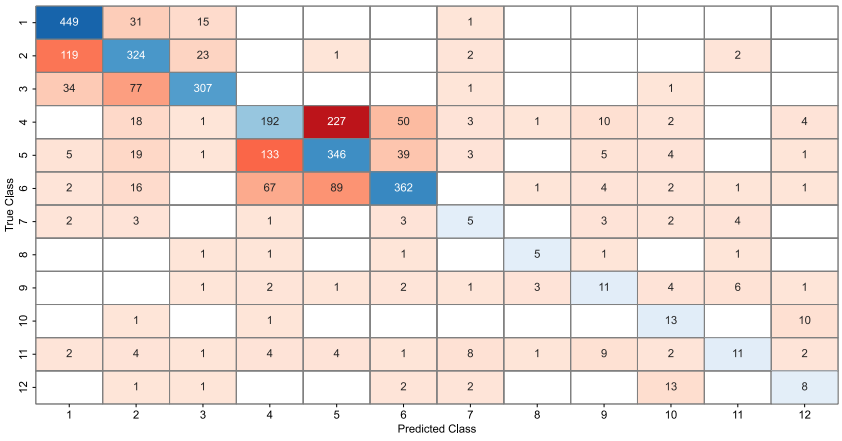

图 17　决策树模型的混淆矩阵图

根据决策树的混淆矩阵发现，根据对位置不敏感信息对人体行为识别预测的准确率为64.29%，识别精度尚可。但模型可以对人体动作属于静态或动态作出显著判断，印证了前文首先对动态或静态采取判断，再细分其具体行为的合理性。模型的不足之处在于对静态和动态中具体行为的划分，以及对静态动作转换的具体划分精度不足。

5.4.3 对位置识别的实验设计

筛选出对位置不敏感的指标构建模型后，发现模型仍存在较多不足。为进一步优化对人体行为识别的研究，本文提出对位置识别的实验设计。精确识别智能手机所处人体位置后，针对不同位置构建出不同模型，以提高位置发生改变后，对人体行为识别的精确性。

针对智能手机佩戴的位置发生改变，本文考虑新增以下实验，提升在不同部位人体行为识别研究的判别准确性。

表 9　新增实验组合设计

编号	佩戴位置	新增参数	编号	佩戴位置	新增参数	编号	佩戴位置	新增参数
1	手	——	5	手	震动频率	9	手	触摸压力
2	腰部	——	6	腰部	震动频率	10	腰部	触摸压力
3	腿部	——	7	腿	震动频率	11	腿	触摸压力
4	胸部	——	8	胸部	震动频率	12	胸部	触摸压力

实验组合设计理由如下：

（1）本问题的目标是考虑智能手机佩戴位置发生改变后，如何提高人体行为识别研究的准确率。考虑到日常生活中，智能常见手机佩戴的位置。可归纳为：拿在手上、上衣口袋（腰部）、裤子口袋（腿部）、西服口袋（胸部）。以此明确了本文对实验佩戴位置的设计。

（2）针对新增参数的设计，前文的分析证明特征的有效性会发生显著变化。查阅相关文献的前人研究，并结合生活实际。智能手机佩戴在不同部位，人体活动中使智能手机受到的震动频率和触摸压力会显著不同。可通过这些参数来具体判别智能手机的佩戴位置，再针对不同部位的数据构建不同的模型。

（3）实验设计中，1~4 组未增加对新增物理量的测量，作为对照组，与 5~12 的实验组相互对比，可探究新增其他物理量的测量对预测结果的影响。

（4）设计震动频率和触摸压力的两组实验组，可以判断哪一种对不同部位识别的准确度更高。两组实验组可以进行交互验证。

六、模型的评价、改进与推广

6.1 模型的优点

1、在降维过程中，用机器学习的方法进行变量选择，增强了模型的可解释性。

2、本文针对不同的问题提出了不同的机器学习算法，提升了模型优度，增加了模型的可解释性。

3、在已有文献的基础上进行研究，所提出的分类方法，在一定程度上弥补了文献中对静态人体行为识别精度较低的缺陷。

4、从数据本身和人体动作的两个角度，给出了人体行为识别误判率高的原因，解释更为全面。

5、提出了对位置识别的实验设计，可根据实验进一步深入完成人体行为识别研究。

6.2 模型的缺点

1、利用已有数据，对由躺转出的行为识别表现欠优。

6.3 模型的改进

1、进行实验，获得智能手机在不同佩戴位置下的一手数据。在实际应用背景下，对用户的人体行为识别研究进一步展开。

2、溯源题中所给数据，即找到该项实验的原始数据，进一步解释模型。

3、进一步增加样本数量，尝试更多机器学习算法，如神经网络等。寻求最佳的分类模型。

6.4 模型的推广

1、依据题目中所给信息，建立了人体行为识别研究的分类模型，对各种状态下的人体行为识别提供了的参考依据，在人工智能领域有一定的参考价值。

2、本文对模型的合理性与敏感性的验证方法，可用于其他数学问题和模型检验。

七、参考文献

[1]Anguita D, Ghio A, Oneto L, et al. A public domain dataset for human activity recognition using smartphones[C]//Esann. 2013, 3: 3.

[2]Yang R, Wang B. PACP: A position-independent activity recognition method using smartphone sensors[J]. Information, 2016, 7(4): 72.

[3]Ling Chen,Xiaoze Liu,Liangying Peng,Menghan Wu. Deep learning based multimodal complex human activity recognition using wearable devices[J]. Applied Intelligence,2020(prepublish).

[4]曾津,周建军.高维数据变量选择方法综述[J].数理统计与管理 2017.36(04).678-692.DOI 10.13860/j.cnki.sltj.20170329-001.

[5]王晓东,田俊.聚类分析分类结果合理性考核方法[J].数学的实践与认识,2008(20):110-113.

[6]程平,何昱衡,辜榕容.基于支持向量机机器学习算法的项目人员绩效评价研究——基于 A 风景园林规划研究院规划设计类项目[J].中国管理会计,2020(01):32-43.