数学学院案例分析报告

案例名称：手机用户行为分析

报告负责人：魏丹怡（202032164）

完成时间：2021.05.02

目　录

一、绪论.....1
1、研究背景.....1
2、研究意义.....1
二、数据来源和基本情况.....1
1、数据来源.....1
2、基本情况.....1
三、数据预处理.....2
1、三维.....2
2、二维.....3
四、用户行为分类.....3
1、形式化和分类器数学推导.....3
2、计算机仿真过程.....5
3、评估方法.....5
4、全部特征分类结果.....5
5、PCA 降维特征分类结果.....8
六、结果对比.....10
七、实践总结.....10

一、 绪论

1、研究背景

传统人体活动识别方法可分为两大类：计算机视觉与可穿戴设备检测。根据监控生成的图片或视频进行计算机视觉判别分析，此种方法采集数据量大，容易受到光照、场景以及检测的个体差异等因素影响，难以得到稳定的识别精度。另一方面，随着微型传感器技术的发展，出现了体积小、功耗低、穿戴便捷的惯性传感器，借助加速度计、陀螺仪等可穿戴传感器,可以在不泄露隐私的情况下采集用户的活动信息,协助完成生活辅助、健康检测等工作。

2、研究意义

基于传感器数据的 HAR 能够有效地避免基于视频数据动作识别带来的隐私泄露等问题，且计算复杂度比视频数据低，使得基于传感器的 HAR 研究更加具有现实意义。目前，AR 的研究应用包括健康监控、智慧家庭、工业环境和运动员监测等。例如：老人通过佩戴智能手环检测是否有走路姿态异常；工人通过佩戴传感器来记录和规范操作动作。因此，准确地识别和记录人体的活动姿态，能够为人们提供更为精确的服务。对多形态多位置的复杂传感器数据,如何利用传感器数据特点提取具有良好判别力的特征以提高基于传感器数据的的 HAR 准确率，是具有研究价值和现实意义的研究问题。

二、 数据来源和基本情况

1、数据来源

此数据集为手机用户的人体活动识别。研究对象为 30 名年龄在 19 到 48 岁之间的志愿者。每个人都被要求在佩戴三星 Galaxy S II 智能手机的同时执行六种活动，分别是起立、坐下、躺下、行走、上楼和下楼。每个受试者都进行了两次实验：在第一次实验中，智能手机被固定在腰带的左侧,而在第二次实验中,智能手机是由用户自己按照自己的喜好放置的，每次执行活动之间有 5 秒的间隔。

2、基本情况

数据集是关于六种手机用户的人体活动的数据，包含 6 个类别。其中训练集有 7352 个样本，测试集有 2947 个样本。每个样本有 561 个特征值，这些特征值中包含均值，相关性，信号幅度区域（SMA）和自回归系数，不同频带的能量，频率偏度和向量之间的角度等。

人体活动的类别为：行走（WALKING）、上楼（WALKING_UPSTAIRS）、下楼（WALKING _DOWNSTAIRS），坐下（SITTING）、起立（STANDING）以及躺下（LAYING）分别由数字 0、1、2、3、4、5 表示，其中 0 表示行走，1 表示上楼，2 表示下楼，3 表示坐下、4 表示起立、5 表示躺下。

三、数据预处理

由于本次实验数据中每个样本对应的特征数为 561 个，数量较多，因此，我采用 PCA 算法对数据进行降维处理。在后续实验中，将使用两组不同特征（一组为全部特征，一组为 PCA 降维后特征）进行分类并对比其分类结果。

1、三维

采取 PCA 算法提取特征，设置主成分个数为 3，得到其贡献率如下图所示：

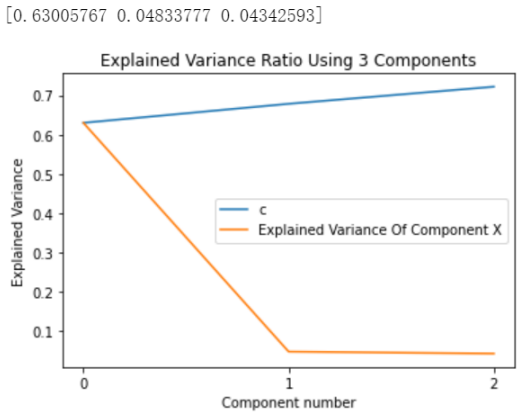

图 1 主成分贡献率

可以看出，所有主成分的累计贡献率为 72%，其中第一主成分的贡献率最大。绘制散点图矩阵分析三个主成分之间的关系：

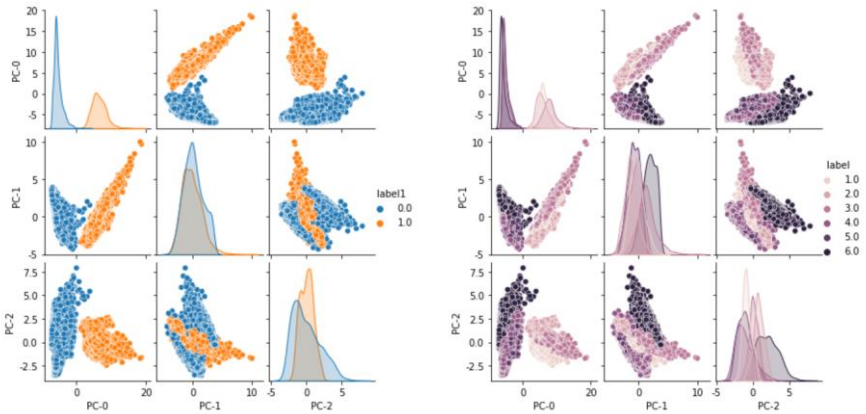

图 2 主成分之间关系

左图为二分类情况下样本点的分布，两个类别各包括 3 种运动状态，其中“0 类”包括坐下、起立、躺下，“1 类”包括走路、上楼、下楼。右图为六分类情况下样本点的分布，即每种运动状态各为一类。

可以看出，在将主成分一、主成分二作为特征，或将主成分一、主成分三作为特征时，不同类别样本之间存在较明显的分类边界，因此，我合理假设仅使用两个主成分，即可较好进行后续分类研究。

2、二维

仍采取 PCA 算法提取特征，设置主成分个数为 2，得到其贡献率如下图所示：

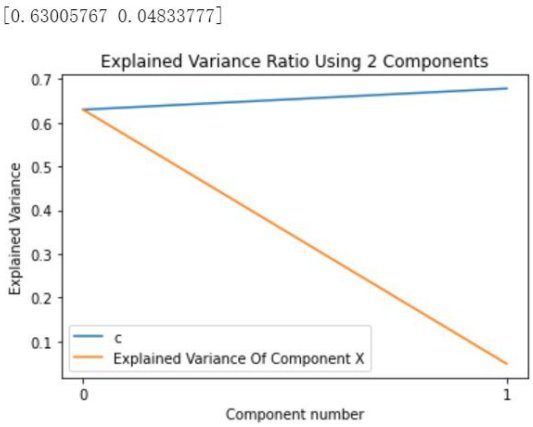

图 3 主成分贡献率

可以看出，所有主成分的累计贡献率为 68%。为了探索使用主成分是否可以对样本进行较好分类，绘制训练集样本散点图如下：

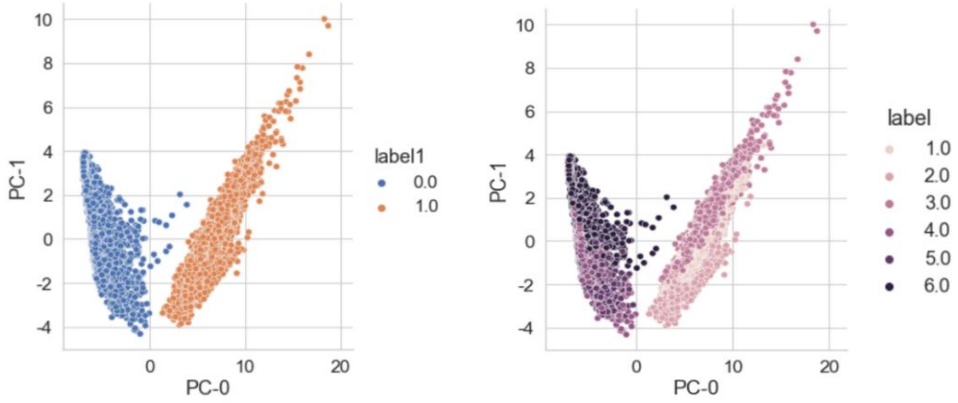

图 4 散点图分布

不难看出，无论将样本分为几类，在使用两个主成分作为特征时，各类样本之间都存在分类边界，但六分类的分类边界不够清晰。因此，使用两个主成分可以进行后续分类实验，但将两大类细分为六类时，分类效果也许不够理想。

四、用户行为分类

1、形式化和分类器数学推导

（1）SVM 算法

运用支持向量机实现一个多分类问题，我使用一对多的方法，即训练时依次把某个类别的样本归为一类,其他剩余的样本归为另一类，这样 k 个类别的样本就构造出了 k 个 SVM。

在此案例中，要将手机用户行为分为六类（记为0、1、2、3、4、5），于是抽取训练集的时候，分别抽取：

（a）0所对应的向量作为正集，1、2、3、4、5所对应的向量作为负集；

（b）1所对应的向量作为正集，0、2、3、4、5所对应的向量作为负集；

（c）2所对应的向量作为正集，0、1、3、4、5所对应的向量作为负集；

（d）3所对应的向量作为正集，0、1、2、4、5所对应的向量作为负集；

（e）4所对应的向量作为正集，0、1、2、3、5所对应的向量作为负集；

（f）5所对应的向量作为正集，0、1、2、3、4所对应的向量作为负集；

使用这六个训练集分别进行训练，得到六个训练结果。在测试的时候，把对应的测试向量分别利用这六个训练结果文件进行测试。最后每个测试都有一个结果，最终的结果便是这六个值中最大的一个作为分类结果。

将手机用户行为的属性以坐标形式表示，建立以下支持向量机模型：

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$subject\ to\ y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0。i = 1, \cdots, n$$

其中参数C代表在线性不可分的情况下，对分类错误的惩罚程度。ζ_i是对于第i样本点的分类损失，如果分类正确则是0，如果分类有所偏差则对应一个线性的值，ζ_i的求和是总误差，优化的目标当这个值越小越好，越小代表对训练集的分类越精准。目标函数中另一项的最小化的优化方向则是使间隔大小最大。

（2）Logistic Regression算法

逻辑回归是监督学习中的一种，它根据大量带有分类标签的特征变量来训练优化模型，在根据模型来预测只有特征变量的分类标签。在手机用户行为分类案例中，我们通过许多带有分类标签的特征变量数据来训练预测信用状况类别的模型。为实现预测分类问题，我们使用了Logistic模型及Sigmoid函数：

$$log_i(Z) = \frac{1}{1 + e^{-Z}}$$

Sigmoid函数有一些特点，比如当$z = 0$是$log_i(\mathrm{z}) = 0$。5当$z < 0$时，$0 < log_i(\mathrm{z}) < 0$。5；当$z > 0$时，0。$5 < log_i(\mathrm{z}) < 1$。所以$log_i(\mathrm{z})$函数的取值范围为$(0,1)$。

其中回归的基本方程为：

$$z = w_0 + \sum_i^N w_i x_i$$

我们可以把$log_i(\mathrm{z})$的函数值看成类别为1的概率预测值，当$log_i(\mathrm{z}) < 0$。5时，我们预

测的分类为 0；当$log_i(z) >= 0$。5时，我们预测的分类为 1。对于行为分类这种六分类问题，我们只需将五个二分类的逻辑回归组合即可实现。

2、计算机仿真过程

（1）SVM

在 jupyter notebook 中调用 svc 构建基于支持向量机的手机用户行为分类模型。

（2）Logistic Regression

在 jupyter notebook 中调用 sklearn。linear_model 包构建基于逻辑回归的手机用户行为分类模型。

3、评估方法

（1）混淆矩阵

混淆矩阵向我们展式了查准率(准确率)与查全率(召回率)：

查准率(P) = $\frac{TP}{TP+FP}$ ，即在被判别为正类别的样本中，确实为正类别的比例是多少；

查全率(R) = $\frac{TP}{TP+FN}$ ，即在所有正类别样本中，被正确判别为正类别的比例是多少。

（2）ROC 曲线

模型训练完成之后，每个样本都会获得对应的两个概率值，一个是样本为正样本的概率，一个是样本为负样本的概率。把每个样本为正样本的概率取出来，进行排序，然后选定一个阈值，将大于这个阈值的样本判定为正样本，小于阈值的样本判定为负样本，可以得到两个值，一个是真正率，一个是假正率：

真正率（TPR）= $\frac{TP}{TP+FN}$ ，即模型判定为正样本且实际为正样本的样本数与所有的正样本数之；

假正率（FPR）= $\frac{FP}{TN+FP}$ ，即模型判定为正样本实际为负样本的样本数与所有的负样本数之比。

我们每选定一个阈值，就能得到一对真正率和假正率，由于判定为正样本的概率值区间为[0,1]，那么阈值必然在这个区间内选择，因此在此区间内不停地选择不同的阈值，重复这个过程，就能得到一系列的真正率和假正率，以这两个序列作为横纵坐标，即可得到 ROC 曲线了。而 ROC 曲线下方的面积，即为 AUC 值。

4、全部特征分类结果

（1）SVM

通过 Sklearn 的 SVC 模型对手机用户行为进行分类，分类结果如下：

Accuracy of SVM Classifier: 0.9504580929759077

	precision	recall	f1-score	support
1.0	0.94	0.98	0.96	496
2.0	0.93	0.96	0.94	471
3.0	0.99	0.91	0.95	420
4.0	0.94	0.89	0.91	491
5.0	0.91	0.95	0.93	532
6.0	1.00	1.00	1.00	537
accuracy			0.95	2947
macro avg	0.95	0.95	0.95	2947
weighted avg	0.95	0.95	0.95	2947

可以看出，模型分类的精度达到了 95%。

通过混淆矩阵来观察预测分类和实际分类情况，绘制混淆矩阵的热点图如下：

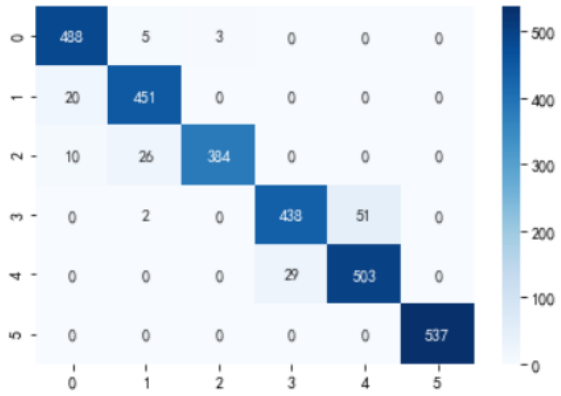

图 5　SVM 混淆矩阵（a）

由上图可看出，我将大部分的用户行为均分类正确，只有少部分人的身体姿态较为相似的行为被分类错误。其中错误最明显的是将坐下误分为起立，错误样本个数为 51。其次，将起立误分为坐下，错误样本个数为 29。总体来看，此模型分类结果良好。

之后，绘制 ROC 曲线对此模型进行评估，如下所示：

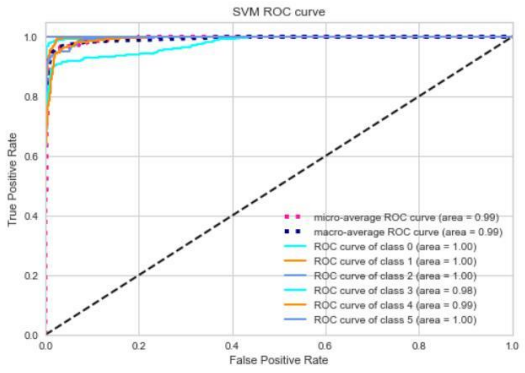

图 6　SVM ROC 曲线（a）

由上图可知，总体 ROC 曲线下方的面积（AUC）为 0.99，各类标签下的 ROC 曲线下方面积分别为 1、1、1、0.98、0.99、1。说明此分类器效果较好。

（2）Logistic Regression

通过 Sklearn 的 LogisticRegression 模型，取 C=1000，solver='lbfgs'，对手机用户

行为进行分类，结果如下：

```
Accuracy of SVM Classifier: 0.9575839837122497
              precision    recall  f1-score   support

         1.0       0.94      0.99      0.97       496
         2.0       0.96      0.94      0.95       471
         3.0       0.99      0.96      0.97       420
         4.0       0.96      0.88      0.92       491
         5.0       0.90      0.97      0.93       532
         6.0       1.00      1.00      1.00       537

    accuracy                           0.96      2947
   macro avg       0.96      0.96      0.96      2947
weighted avg       0.96      0.96      0.96      2947
```

可以看出，模型分类的精度达到了 95.7%。

通过混淆矩阵来观察预测分类和实际分类情况，绘制混淆矩阵的热点图如下：

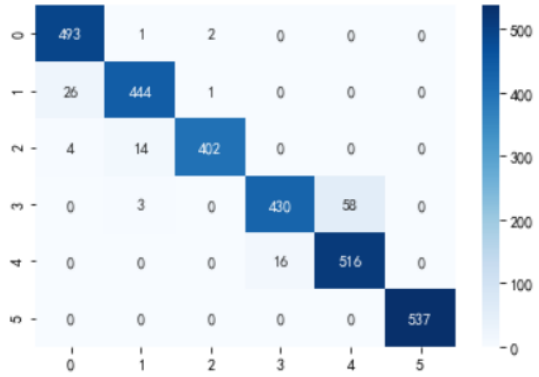

图 7　LR 算法混淆矩阵（a）

由上图可看出，我们将大部分的用户行为均分类正确，只有少部分人的身体姿态较为相似的行为被分类错误。其中错误最明显的是将坐下的行为误分为起立，错误样本个数为 58。其次，将上楼误分为行走，错误样本个数为 26。总体来看，此模型分类结果较为良好。

之后，绘制 ROC 曲线对此模型进行评估，如下所示：

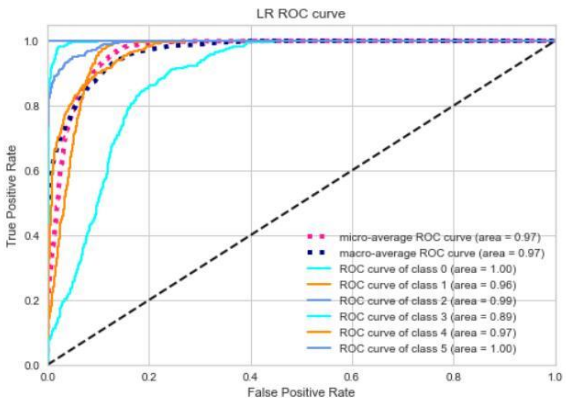

图 8　LR 算法 ROC 曲线（a）

由上图可知，总体 ROC 曲线下方的面积（AUC）为 0.97，各类标签下的 ROC 曲线下方面积分别为 1、0.96、0.99、0.89、0.97、1。说明此分类器效果较好。

5、PCA 降维特征分类结果

（1）SVM

通过 Sklearn 的 SVC 模型对手机用户行为进行分类，分类结果如下：

Accuracy of SVM Classifier: 0.6382762130980658

	precision	recall	f1-score	support
1.0	0.46	0.81	0.59	496
2.0	0.83	0.11	0.20	471
3.0	0.78	0.83	0.81	420
4.0	0.85	0.04	0.09	491
5.0	0.53	0.98	0.69	532
6.0	0.97	0.99	0.98	537
accuracy			0.64	2947
macro avg	0.74	0.63	0.56	2947
weighted avg	0.73	0.64	0.56	2947

可以看出，模型分类的精度只达到了 63.8%，降维后分类效果不是很好。

通过混淆矩阵来观察预测分类和实际分类情况，绘制混淆矩阵的热点图如下：

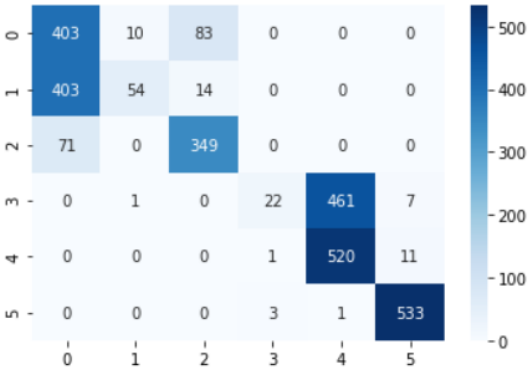

图 9　SVM 混淆矩阵（b）

由上图可看出，将大部分的用户行为均分类正确，只有少部分人的身体姿态较为相似的行为被分类错误。其中错误最明显的是将坐下的行为误分为起立，错误样本个数为 461。其次，将上楼误分为行走，错误样本个数为 403。对于上楼和坐下这两个行为分类结果很差，要多关注这两个行为，进行改正。

之后，绘制 ROC 曲线对此模型进行评估，如下所示：

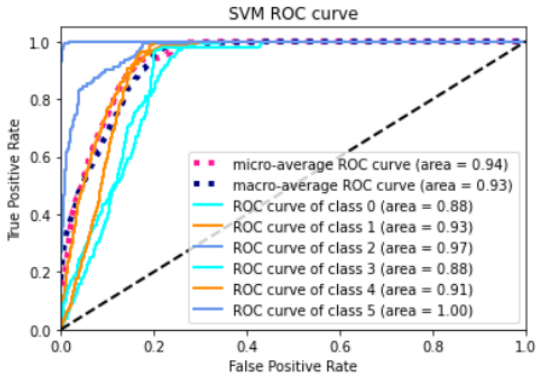

图 10 SVM ROV 曲线（b）

由上图可知，总体 ROC 曲线下方的面积（AUC）为 0.94，各类标签下的 ROC 曲线下方面积分别为 0.88、0.93、0.97、0.88、0.91、1。与使用全部特征分类 ROC 曲线下方面积对比，说明降维后分类效果没有使用全部特征分类的效果好。

（2）Logistic Regression

通过 Sklearn 的 LR 模型对手机用户行为进行分类，分类结果如下：

```
Accuracy of LR Classifier: 0.6257210722768918
              precision    recall  f1-score   support

         1.0       0.45      0.75      0.56       496
         2.0       0.60      0.07      0.13       471
         3.0       0.73      0.90      0.81       420
         4.0       0.53      0.18      0.27       491
         5.0       0.53      0.83      0.65       532
         6.0       0.96      0.99      0.98       537

    accuracy                           0.63      2947
   macro avg       0.63      0.62      0.57      2947
weighted avg       0.63      0.63      0.57      2947
```

可以看出，模型分类的精度只达到了 62.6%，降维后分类效果不是很好。

通过混淆矩阵来观察预测分类和实际分类情况，绘制混淆矩阵的热点图如下：

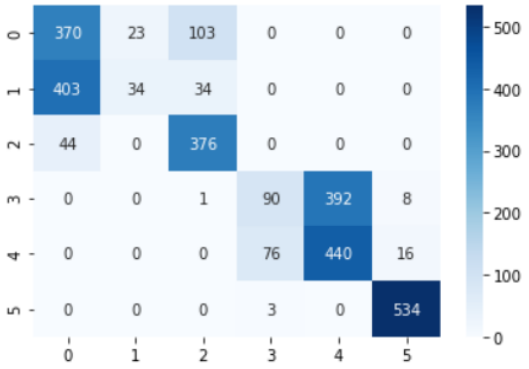

图 11　LR 算法混淆矩阵（b）

由上图可看出，将大部分的用户行为均分类正确，只有少部分人的身体姿态较为相似的行为被分类错误。其中错误最明显的是将上楼的行为误分为行走，错误样本个数为 403。其次，将坐下误分为起立，错误样本个数为 392。对于上楼和坐下这两个行为分类结果很差，要多关注这两个行为，进行改正。

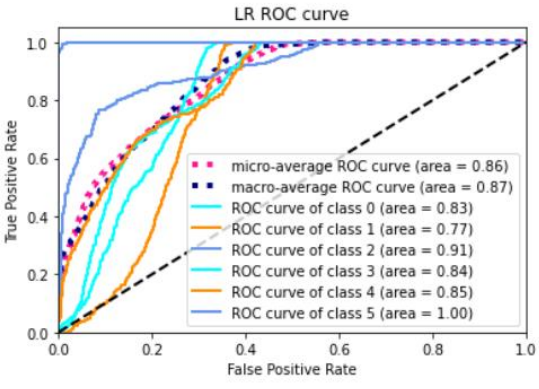

图 12 LR 算法 ROC 曲线（b）

由上图可看出，将大部分的用户行为均分类正确，只有少部分人的身体姿态较为相似的行为被分类错误。其中错误最明显的是将上楼的行为误分为行走，错误样本个数为 403。其次，将坐下误分为起立，错误样本个数为 392。与 svm 模型得到结论一致，对于上楼和坐下这两个行为分类结果很差，要多关注这两个行为，进行改正。

六、结果对比

本文以手机用户行为数据实验数据，采用 SVM、逻辑回归对数据集进行分类。使用 pca 降维后的前两个主成分以及全部数据特征分别构建 SVM、逻辑回归模型进行分类。其中，利用全部特征 SVM 的分类准确率为 95%，逻辑回归的分类准确率是 95.7%，两种分类器准确率均较高且相差不多。使用 PCA 降维后特征构建的模型中，SVM 的分类准确率为 63.8%，逻辑回归的分类准确率是 62.6%，准确率虽然相差不多但整体较差，可见利用全体特征进行分类更好，分类器在本次实验中影响不大。通过混淆矩阵可以看到，不论是利用 PCA 降维后特征还是全体特征，都存在指标误分的情况。其中将上楼误分类为行走以及坐下误分为起立的错位最为集中，可能因为两者运动形态较为相似，在后续研究中可以针对该情况进行特征的优化以及分类器的选取。

七、实践总结

本次实验目的为通过逻辑回归、SVM 以及 PCA 降维算法实现了对于人体活动的数据的各种行为识别。在实验过程中，首先，我对于人体识别有了初步的认识和理解；其次，巩固了 PCA 降维、逻辑回归以及 SVM 算法相关知识点，这个过程不但加深了我对于统计理论的理解，同时在不断的尝试与纠错中，增加了许多编程经验。

我使用全部特征进行分类，分类结果良好，但是经过 pca 降维后，模型的分类准确率降低了 30%左右。由此看来，在此实验中通过特征筛选得到的结果不太理想，有可能是通过 pca 后所提取的特征贡献率较小，不能涵盖所有信息。因此，应该尝试其他的降维方法提取特征，可能会得到较高的分类准确率。