# 基于智能手机传感器的用户行为识别

任谦 陈晓东 高臻 刘羽霄 李知赫 张若飞

## 1 项目进展

#### 1.1 数据预处理

本周数据预处理部分工作主要为：1）WISDM数据集的滑窗采样；2）增加tsfresh提取指定特征的方法；3）增加tsfresh增加自定义特征方法。在tsfresh框架基础上，提供了尽可能多地方便数据处理，调用及分析的API接口，并对之前遗留的部分数据处理程序问题进行了完善。

#### 1.2 特征提取及选择

本周特征提取及选择工作分为两部分：1）基于tsfresh库已有特征的分析，主要为时域特征，结合文献调研结果，在频域特征提取部分进行了适当补充；2）通过XGBoost算法进行特征重要性排序进而特征选择，并实现分类，效果相比于SFFS特征选择器有显著提升。

**1.2.1 频域特征补充**

本周主要完成了频域特征在tsfresh库上的实现，以及程序对本地tsfresh库的调用。

之前程序调用的为tsfresh库，没有办法修改，无法实现自定制特征（即增加新的特征）。在修改本地tsfresh库的初始化和调用后，保证可以在根目录下对本地tsfresh库调用，并且自定制特征，便于后续特征工程工作的进行。

关于频域特征部分，本周新补充实现了9个频域特征，分别为：

 fft\_dc

 fft\_mean,

 fft\_var

 fft\_std

 fft\_kurt

 fft\_shape\_mean

 fft\_shape\_std

 fft\_shape\_skew,

 fft\_shape\_kurt

增加特征后对结果稍有改善，但不明显。下周将继续补充fft\_max等频域特征，以及组合特征，观察对实验结果影响情况。

**1.2.2 XGBoost模型**

本周更换了XGBoost模型，相比于随机森林分类性能更好且自带特征重要性排序的功能，稳定性优于SFFS方法。以下分析都是基于TD数据集。由于训练数据和测试数据是随机划分的，而特征选择又是与数据集相关的，因此为了降低随机性的影响，我们对数据集随机划分20次并相应的训练20个XBGoost模型，每次都选择重要性排在前100的特征，最后统计出现次数超过10次的特征。同时出于尽量减少特征数量的原则，将只出现1次或2次的特征（因为有6个物理特征，因此某个统计特征如果非常重要将出现6次）且去除后结果有提升的特征去除，最后选出了如下13大类统计特征：

#### 自回归系数

#### 自相关系数

#### 高于均值的最长子序列

#### 低于均值的最长子序列

#### ql分位数与qh分位数间的平均绝对连续变化值

#### 时间翻转不对称统计量

#### 分位数

#### 最大值

#### 最小值

#### 中位数

#### 求和

#### 序列中n近邻峰值的个数

#### 估计交叉功率谱密度

#### 这13个特征的选择立竿见影，不仅将特征提取（7290条样本）时间从3h降低到10min，且准确率也比用全部特征提高0.5%~1%。最后经过简单的参数调整，平均准确率达到了94.2%，相比于公司的模型Myna的准确率91.7%提高了2.5%。有了这些特征之后我们也将其运用到UCI数据集上，现在初步尝试的准确率也达到了94%，至少说明这几个特征的选择还是有一定的泛化能力的，不过UCI还需要进一步分析。

#### 最后在人工提取特征方面，本周我们尝试加入速度特征，即加速度的积分（累加和），但是由于采样的不连续和噪声干扰，这种做法会使误差累积最后出现荒谬的结果，因此现在的做法是每个滑窗开始时都进行一次归零，类似于惯性制导中的矫正。最后反映在准确率上时还是能提高0.5%左右的准确率，说明这个特征或许还是有意义的，需要进一步分析。

#### 1.3 深度学习LSTM

结合文献调研结果，本周深度学习的工作方向由原先的CNN模型改用LSTM模型。针对UCI数据集，对网上的LSTM 的训练代码和文档：https://github.com/guillaume-chevalier/LSTM-Human-Activity-Recognition在沙箱端进行实现，但是预测精度并不理想。因此自行用keras搭建网络进行实现。考虑到沙箱有限的计算资源，初步网络是包含了两层LSTM网络、一层全链接Dense层的神经网络，并且为防止过拟合，在两层LSTM网络中均设计了Dropout。在沙箱端进行实现，在UCI数据集上的训练精度达到98%左右，最低能够保证97.5%以上（下图中的epoch是第30-40个epoch）。

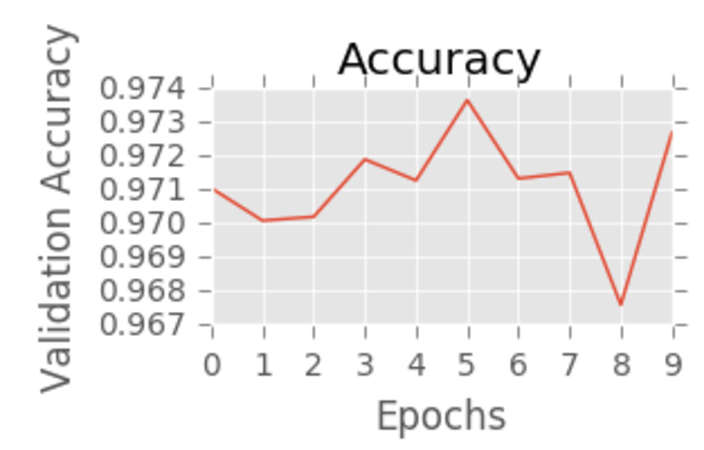


图1 UCI数据集使用LSTM网络的精度

#### 1.4 适应性分析

适应性分析主要考虑不同时间窗口长度和采样频率组合，对数据分类的影响，主要聚焦于三个方面：1）对整体分类准确率的影响，如何确定有效的时间窗口与采样频率；2）不同组合情况下利用SFFS进行特征选择的结果，对于理解加速度传感器进行活动类型分类的作用机制，评估特征的相对重要性；3）不同时间窗口长度及采样频率对于各个具体活动类别分类精度的影响，是否存在差异，进而对易分类类别降低采样频率，在保证分类准确率的情况下，降低手机端耗能。

因按照SFFS特征选择方法的思路，特征提取和特征选择的工作不具有重复性，且限于沙箱运行环境速度问题，本周工作仅以WISDM为例，该数据及初始采样频率20Hz，依据最少类别数据量进行平衡数据抽取，共取得270000条数据进行后续实验。具体实验设置:

* 时间窗口长度4种：2s,3s,4s,5s
* 采样频率：20HZ,10HZ,5HZ
* 分类器：随机森林，决策树，KNN，SVM

**1.4.1 整体准确率**

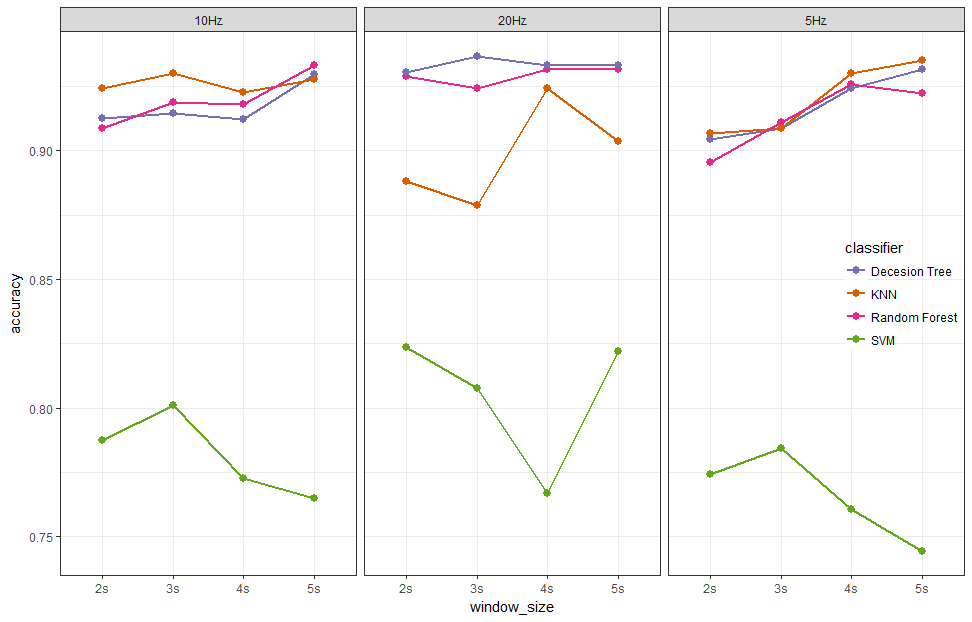


图2 时间窗口，采样频率组合下不同分类器的精度

## 如图2所示，整体分类准确率差别不大（SVM除外），整体为92%-94%左右。具体而言，分类器RF,DT,KNN的分类精度及稳定性方面均明显优于SVM， RF和DT效果差不多，考虑到特征是经过SFFS选择后的特征，各次实验特征集合均不超过10个，可以理解RF,DT在本例效果相近的缘由。

## 此外可以观察到，采样频率为10HZ,5HZ时，随着时间窗口尺寸增加，分类准确率越高，当时间窗口尺寸为4s或者5s时，分类精度基本稳定；另一方面，采样频率越高，准确率相对高且稳定，但是在本数据集情况下，各个频率间分类精度差别不大，即便是降采样到5HZ，准确率依然维持在90%以上。导致上述结果可能是数据集本身的原因，如数据采集过于理想化，后续将在TD数据集上进行类似适应性分析。

**1.4.2 特征选择结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 2s | 3s | 4s | 5s |
| 5HZ | u'1.0\_\_minimum'  u'1.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_6\_\_w\_10'  u'4.0\_\_quantile\_\_q\_0.8' u'0.0\_\_quantile\_\_q\_0.6' u'2.0\_\_quantile\_\_q\_0.8'  u'1.0\_\_sum\_values'  u'1.0\_\_agg\_linear\_trend\_\_f\_agg\_"min"\_\_chunk\_len\_10\_\_attr\_"intercept"' | u'3.0\_\_quantile\_\_q\_0.8' u'1.0\_\_mean' u'2.0\_\_quantile\_\_q\_0.2'  u'2.0\_\_autocorrelation\_\_lag\_1' u'1.0\_\_index\_mass\_quantile\_\_q\_0.9'  u'3.0\_\_quantile\_\_q\_0.9'  u'4.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_11\_\_w\_10'  u'3.0\_\_symmetry\_looking\_\_r\_0.25' | u'3.0\_\_quantile\_\_q\_0.9'  u'1.0\_\_agg\_linear\_trend\_\_f\_agg\_"mean"\_\_chunk\_len\_50\_\_attr\_"intercept"'  u'2.0\_\_quantile\_\_q\_0.2'  u'0.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_1\_\_w\_20' | u'3.0\_\_quantile\_\_q\_0.8' u'1.0\_\_median'  u'2.0\_\_agg\_linear\_trend\_\_f\_agg\_"min"\_\_chunk\_len\_50\_\_attr\_"intercept"'  u'2.0\_\_mean\_autocorrelation'  u'1.0\_\_percentage\_of\_reoccurring\_values\_to\_all\_values' |
| 10HZ | u'1.0\_\_minimum'  u'1.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_6\_\_w\_10'  u'4.0\_\_quantile\_\_q\_0.8' u'0.0\_\_quantile\_\_q\_0.6' u'2.0\_\_quantile\_\_q\_0.8'  u'1.0\_\_sum\_values'  u'1.0\_\_agg\_linear\_trend\_\_f\_agg\_"min"\_\_chunk\_len\_10\_\_attr\_"intercept"' | u'4.0\_\_quantile\_\_q\_0.9'  u'1.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_12\_\_w\_20'  u'0.0\_\_quantile\_\_q\_0.9' u'1.0\_\_linear\_trend\_\_attr\_"stderr"'  u'2.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_1\_\_w\_2' | u'3.0\_\_quantile\_\_q\_0.9'  u'1.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_13\_\_w\_20'  u'0.0\_\_quantile\_\_q\_0.7'  u'2.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_9\_\_w\_10' | u'3.0\_\_quantile\_\_q\_0.8' u'1.0\_\_mean' u'0.0\_\_quantile\_\_q\_0.6'  u'1.0\_\_maximum' u'0.0\_\_friedrich\_coefficients\_\_m\_3\_\_r\_30\_\_coeff\_2' |
| 20HZ | u'1.0\_\_minimum' u'1.0\_\_sum\_values' u'0.0\_\_quantile\_\_q\_0.8'  u'2.0\_\_quantile\_\_q\_0.8' u'2.0\_\_autocorrelation\_\_lag\_1'  u'0.0\_\_autocorrelation\_\_lag\_2' | u'1.0\_\_minimum' u'1.0\_\_quantile\_\_q\_0.6' u'0.0\_\_quantile\_\_q\_0.8'  u'2.0\_\_quantile\_\_q\_0.9' u'1.0\_\_mean\_abs\_change\_quantiles\_\_qh\_1.0\_\_ql\_0.0'  u'0.0\_\_maximum' | u'3.0\_\_quantile\_\_q\_0.8'  u'1.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_9\_\_w\_10'  u'0.0\_\_quantile\_\_q\_0.4'  u'1.0\_\_agg\_linear\_trend\_\_f\_agg\_"min"\_\_chunk\_len\_10\_\_attr\_"intercept"'  u'2.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_12\_\_w\_10'  u'1.0\_\_mean\_abs\_change' u'3.0\_\_symmetry\_looking\_\_r\_0.55' | u'4.0\_\_quantile\_\_q\_0.9'  u'1.0\_\_cwt\_coefficients\_\_widths\_(2, 5, 10, 20)\_\_coeff\_11\_\_w\_20'  u'0.0\_\_quantile\_\_q\_0.6' u'1.0\_\_maximum'  u'0.0\_\_linear\_trend\_\_attr\_"intercept"' |

如上表所示，利用SFFS对各次实验进行特征选择结果：

整体而言quantile(红色标注), cwt\_coefficients\_\_widths(蓝色标注)连续小波变换系数出现频次最多。前者quantile在12次实验中均有出现，且单次实验出现频次常多于2次，cwt\_coefficients\_\_widths连续小波变换系数出现8次。该结果说明分位数特征的重要性，也可以一定程度上解释降采样频率后分类精度变化较小，因降采样对序列分位数影响较小。

其次agg\_linear\_trend聚成块后的线性最小二乘回归, linear\_trend对时间序列做最小二乘回归, minimum, sum\_values出现频率较多，均为2-4次，说明这些特征具有一般重要意义。

**1.4.3 混淆矩阵-类别间差异**

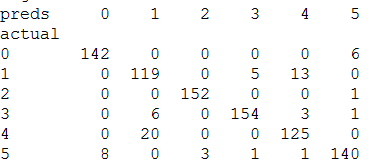


图3 预测及实际类别混淆矩阵（时间窗口3s,采样频率20Hz）

各次实验的混淆矩阵相似，此处以时间窗口长度3s，采样频率20HZ实验为例，影响精度主要原因在于1,4类别的错分，即Sitting及LyingDown，其次是0,5类别错分，即为Walking，Stairs，其余类别的precesion,recall,F1-score等指标都相对理想（高于0.95）。各类别间分类结果受时间窗口及采样频率影响较弱，受数据集本身采集方式影响较大。

后续实验考虑在TD数据集进行探索，以期得到可以依据活动类别设定有效采样频率与时间窗口的方法。

## 2 项目工作计划

未来一周的工作包括如下三部分：

1. 特征提取及选择：继续补充频域特征；尝试采用分层等方法提高分类精度，理解XGBoost的分类机制，进而探讨提高精度的有效手段。
2. 深度学习：将继续进行LSTM的数据可视化工作，对LSTM网络进行结构调整，参数调整；针对文献中出现较为频繁的CNN网络进行实现
3. 适应性分析: 在TD数据（采样频率100HZ）探讨不同时间窗口尺寸和采样频率对分类的影响，同时考虑分类计算耗时，综合评估有效时间窗口及采样频率；在相关实验基础上继续挖掘不同类别间差异，及不同窗口和采样组合及相应特征的关系。