项目总结

## 一、项目概述

人机识别项目，旨在通过传感器设备数据，识别出是人操作，还是机器操作。

在人机识别项目中，我全程项目的各个环节，包括数据采集、数据预处理、特征工程，数据建模、模型校验，算法改进。

项目完成情况：在测试集上做二分类，precision和recall都在96%以上。

|  |  |  |
| --- | --- | --- |
|  | precision | recall |
| 预期成果 | 90%+ | 90%+ |
| 实际成果 | 96%+ | 96%+ |

## 二、项目进度和关键成果

项目资源14人日，各阶段关键成果如下：

|  | 项目资源(人日) | 完成情况 | 关键成果 |
| --- | --- | --- | --- |
| 数据采集 | 0.5 | 已完成 | 手机设备安装sdk，采集数据。 |
| 数据清洗 | 2 | 已完成 | 将文件数据转成需要的格式，去除噪声数据，做归一化、取log等。 |
| 数据采样加滑动窗口 | 1 | 已完成 | 将一批样本，用滑动的方式，分成多段数据，各批次做TopN样本截断。 |
| 按批次切分数据集 | 0.5 | 已完成 | 拆分训练集和测试集，按批次编号来切分。 |
| 特征工程 | 3 | 已完成 | 从原始的25个特征，衍生得到950个特征。 |
| 特征转换加滑动窗口 | 1 | 已完成 | 使转换后的样本行数变多。 |
| 特征选择 | 0.5 | 已完成 | 根据特征的信息增益来提取特征，得到188个优质特征。 |
| 模型选择 | 0.5 | 已完成 | 尝试使用不同模型，对比发现LASSO在测试集上表现优异。 |
| 模型泛化 | 1 | 已完成 | 使用其他机型数据，打压泛化较差的特征的权重。 |
| 模型调优 | 1 | 已完成 | 调整模型参数，加大None样本的权重。将模型结果离线保存。 |
| 批量预测算法 | 2.5 | 已完成 | 写了两套加权算法。 |
| 项目总结报告 | 0.5 | 已完成 |  |

## 三、特征工程

处理数据得到EMA、DIF、DEA、MACD等技术指标，针对数据区间段特点，做出如下特征转换：

feature\_all = list()

feature\_all.append(self.time\_mean()) # 均值

feature\_all.append(self.time\_var()) # 方差

feature\_all.append(self.time\_std()) # 标准差

feature\_all.append(self.time\_mode()) # 众数

feature\_all.append(self.time\_max()) # 最大值

feature\_all.append(self.time\_min()) # 最小值

feature\_all.append(self.time\_over\_zero()) # 过零点个数。 共线性

feature\_all.append(self.time\_range()) # 最大值与最小值之差

feature\_all.append(self.fft\_dc()) # 直流分量。 Ian。

feature\_all.append(self.fft\_shape\_mean()) # 图形的均值、方差、标准差、偏度、峰度

feature\_all.append(self.fft\_shape\_std() \*\* 2)

feature\_all.append(self.fft\_shape\_std())

feature\_all.append(self.fft\_shape\_skew())

feature\_all.append(self.fft\_shape\_kurt())

feature\_all.append(self.fft\_mean()) # 幅度的均值、方差、标准差、偏度、峰度

feature\_all.append(self.fft\_var())

feature\_all.append(self.fft\_std())

feature\_all.append(self.fft\_skew())

feature\_all.append(self.fft\_kurt())

## 四、算法

人机识别项目使用LASSO算法和分段预测加权算法，训练阶段使用提取出的特征来训练模型，预测阶段也要使用这些提取的特征来预测。

使用LASSO做二分类，准确率和召回率都达到90%以上。使用分段预测加权算法，对LASSO预测的结果投票，准确率和召回率都达到96%以上。

LASSO是Least Absolute Shrinkage And Selection Operator的编写，译作最小绝对值收敛和选择算子。LASSO的Cost函数是：



优化的目标是使Cost函数最小，Cost函数对w做L1惩罚。

LASSO算法是使用L1正则的回归分析方法。使用LASSO算法的好处是，可以在参数估计的同时实现变量的选择，较好的解决回归分析中的多重共线性问题，并且能够很好的解释结果。当协变量共线时，LASSO系数估计值不一定唯一。

使用LASSO算法时，对训练样本要求比较高，要对样本做好预处理，防止个别特征拟合过慢甚至不能拟合的情况。还要对数据的异常值做处理，否则可能过拟合或预测值大大超出目标label的范围。

人机识别项目算法亮点：

* LASSO是线性模型，可解释性强，计算效率高，使用矩阵相乘即可得到预测结果。根据奥卡姆剃刀原理，在所有可能选择的模型中，能够很好地解释已知数据并且十分简单的才是最好的模型。
* 计算时加大None样本的权重，打压泛化能力差的特征的权重，提升预测None样本的准确率。
* 将一份数据，使用滑动窗口生成多段数据，分别对多段数据做预测，各段的预测结果加权投票得到最终结果。

## 五、项目思考

**项目中的不足：**

1. 缺少行为识别类项目的最佳实践积累，缺少过程资产。在项目进行过程中，尝试很多失败方案。
2. 测试数据机型较少，对数据理解不深入也导致前期模型效果超高，而泛化结果很差。后期已调整这部分的代码，做到模型的泛化能力较强。

**待优化内容：**

1. 做滑动窗口时，得到最优窗口大小和步长大小。
2. 在保证预测精确度的前提下，尽可能缩小样本行数，综合提高预测速度。
3. 针对不同机型做出个性化模型。

**进一步建议：**

目前预测过程分为以下三大步，建议将第1步的计算，在sdk中实现，来降低网络数据传输。

1. 将一组传感器文件转成标准数据格式（csv格式或json格式），并做数据预处理（比如保留前100行数据）。
2. 对数据做特征工程和特征提取。
3. 使用保存好的模型进行预测。