机器学习纳米学位

毕业项目

程铭 - 2017, 12, 19

开题报告

项目背景

利用音频的形式对性别进行识别目前存在一定的应用价值，比如在一些需要分辨男女身份的场合，或是可以将性别识别的结果用作身份验证的特征之一，在客服中根据性别自动接入相应的客服人员等。传统的判别方法大多是基于音频信号上的一些特性，如以男声的基音频率普遍较女声低来进行分类，分类方法相对单一，准确率低，因此有待进一步的改进。

参考文献：

小波的提升方法在基音提取中的应用[J]. 彭辉,宁飞,孔宇. 山东大学学报(理学版). 2003(01)

根据语音分形维和基音周期的说话人性别识别研究[J]. 王振华. 生物医学工程学杂志 2008(04)

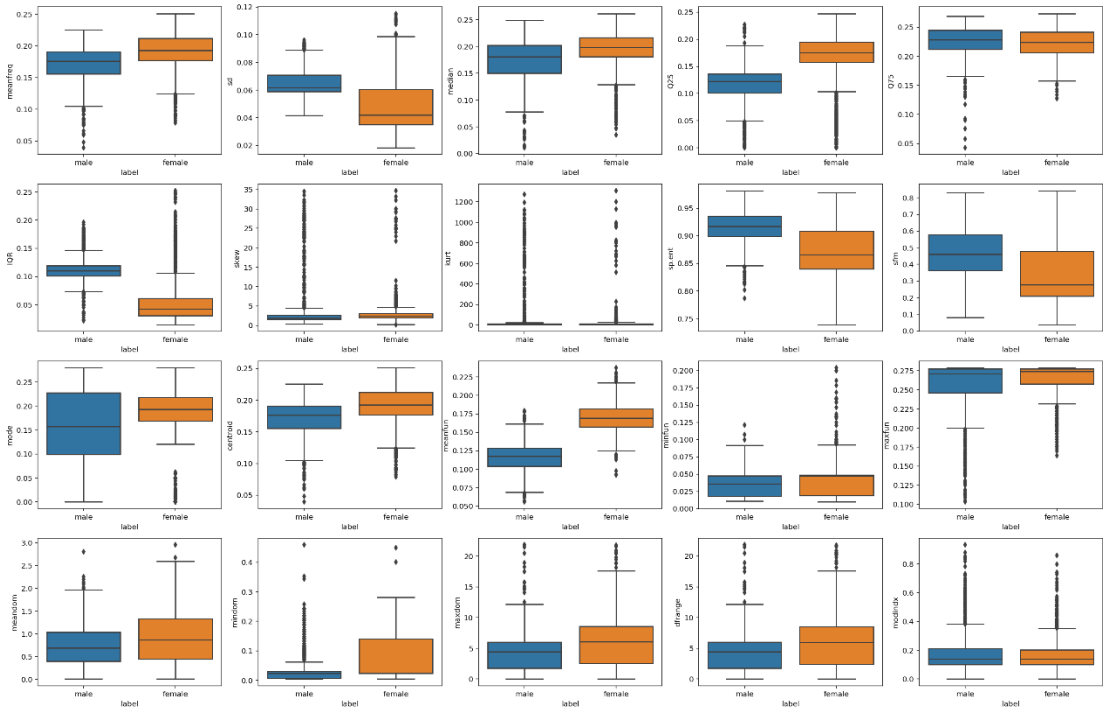
问题描述

该问题是利用音频信号处理作为相关的 domain knowledge，将原始音频信号用来提取部分信息作为可用特征的二分类问题。

数据或输入

数据集由 Kaggle(https://www.kaggle.com/primaryobjects/voicegender)获得，其中一共包含了 3168 个样本，每个样本对应了 20 个经音频信号处理后提取的参数作为特征，及其相应的标签。特征包含频率平均值，频率标准差，频率中位数，频率第一四分位数，频率第三四分位数，频率四分位数间距，频谱偏度，频谱峰度，频谱熵，频谱平坦度，频率众数，频谱质心，峰值频率，平均基音频率，最小基音频率，最大基音频率，平均主频，最小最大主频，主频范围及累积相邻两帧绝对基频频差除以频率范围。

例如，将特征按箱子形图进行可视化，左侧蓝色为 ‘male’， 右侧黄色为 ‘female’，其中比较明显的存在男女分布不同的特征有 meanfun, Q25, IQR，其余特征在 male 和 female 上箱形图的位置高低也有一定的区分度，因此这些特征可以用来有效地帮助解决分类问题。



解决方法描述

在本毕业项目中，需要先利用提取出的如频率均值、频率标准差、频谱偏度、频谱峰度等音频信号处理的参数作为 20 个特征，再采用机器学习中监督学习的方法训练分类模型。尝试不同的模型，根据测试集的准确率来选择相对最优，最适合这个问题的模型，再进行进一步的参数调优，最后以在准确率上得到一个很好的评分。

评估标准

在开始训练前将数据分为训练集，验证集和测试集三部分；在该身份识别中，选用准确率(Accuracy=预测正确的总数/测试数据总数)来衡量模型的表现；

利用准确率作为评估标准来进行网格搜索，找出最优分类器参数 ;最后在测试集上进行预测，和在训练集上的准确率对比是否存在过拟合/欠拟合，计算准确率是否达到 98%以上，最终做出分类器模型是否合格的标准。

基准模型

现有的基准模型有支持向量机，逻辑回归，高斯贝叶斯，随机森林，神经网络等一系列模型可供选择，根据该数据集的样本较少，特征复杂，初步选用支持向量机，逻辑回归及随机森林做简单测试。经过简单的尝试，其中在本案例下随机森林获得了相对较好的结果，因此选定随机森林最为基准模型，进行下一步的调试。

随机森林是一种基于集成学习的方法，训练时采用 booststrap 的取样方式，并利用了弱分类器的思想，可以很好的避免过拟合的情况发生，模型的泛化能力强；能根据训练来自动学习

到不同特征的权重占比，对特征选择的要求不高；训练速度快，且适用本数据集情况下的高噪音的情况。

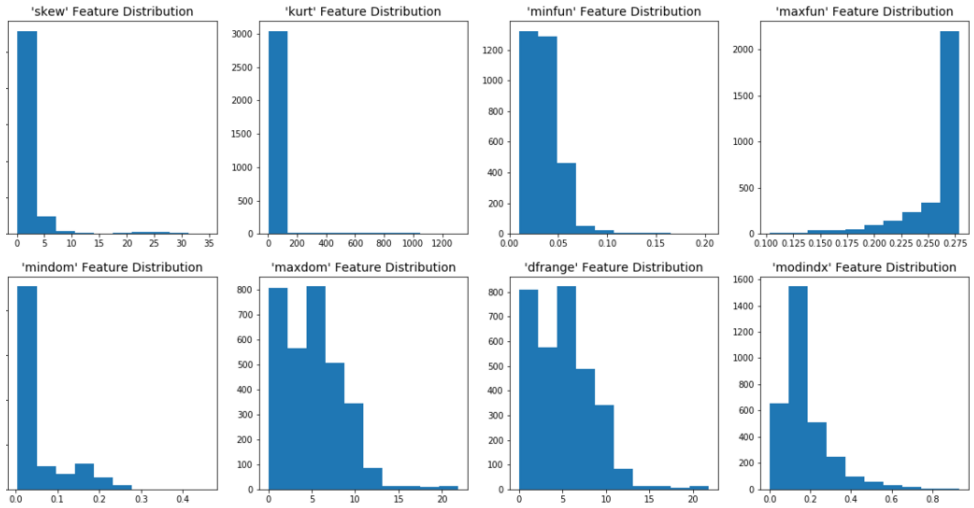
项目设计

数据预处理

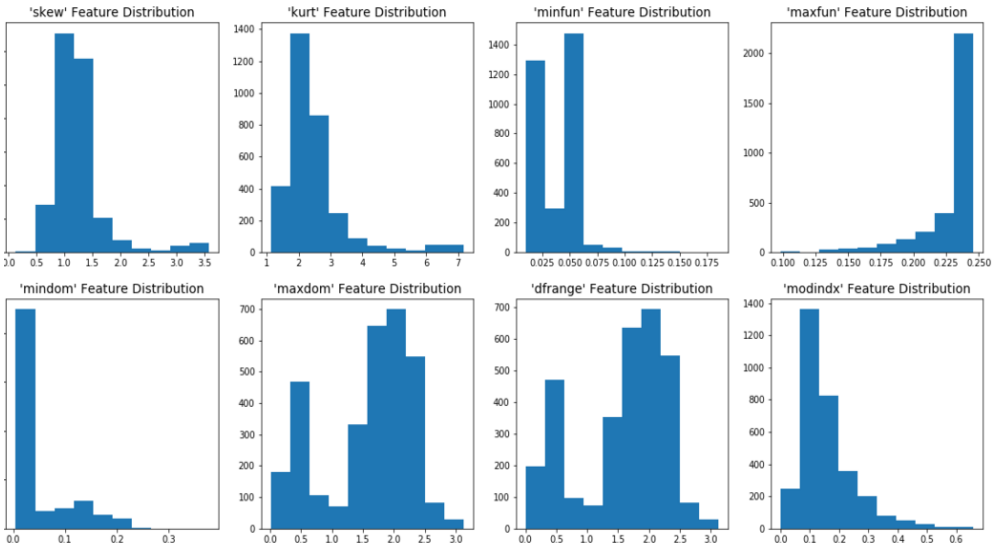
1、 利用 pandas 框架导入 csv 格式的数据，观察数据特征，一共包含多少个数据点，每个数据点多少个特征；

2、 将 label 一栏单独从原数据集中剥离，作为标签使用，同时对原数据格式 ‘male’ / ‘female’ 独热编码为 1/0；

3、 对不同特征的分布进行可视化，发现下列特征存在倾斜分布，如下图所示：

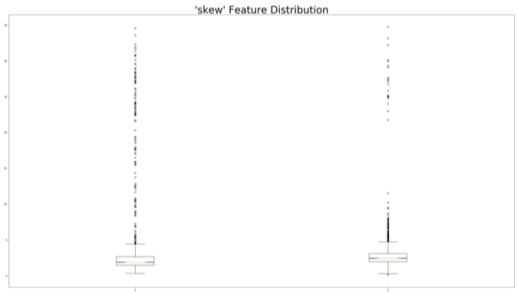


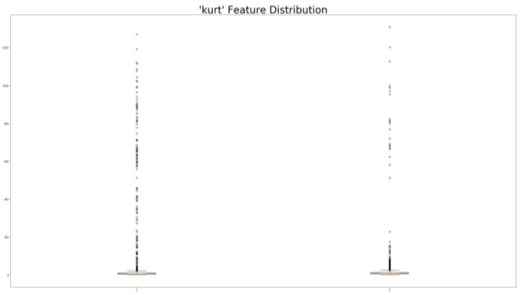
因此对特征'skew', 'kurt', 'minfun', 'maxfun', 'mindom', 'maxdom', 'dfrange', 'modindx'进行非线性变换（log），可以观察到其倾斜分布有了一定程度的改善：



4、对所有特征进行归一化（I 利用 MinMaxScaler），以避免因数值大小的问题引起的权重不均衡；

5、 经观察及尝试， 将特征 ‘skew’， ‘kurt’ 进行箱形图可视化，每张图的左侧为 male，右侧为 female，如下图所示：





经过观察，可以发现 male 和 female 的不同类别，在该两个特征上的分布并没有区别，且都偏侧化严重，经过后期的模型简单验证，发现确实删除这两项特征可以提高模型的表现，因此在此处将特征 ‘skew’ 及 ‘kurt’ 删除，不再使用；

6、后对所有余下特征进行归一化（利用 MinMaxScaler），以避免因数值大小引起的权重不均衡；

7、 将原数据集分为训练集和测试集(20%)，再将训练集分出小部分作为验证集(20%)，数据预处理部分完成。

<u>建立模型</u>

1、 利用 sklearn 库导入随机森林模型，进行初步尝试；

2、 建立 GridSearchCV 网格搜索，对随机森林模型进行调参，找到最优模型，并通过在测试集上的 Accuracy 和 F-score 来衡量最终表现。