机器学习毕业项目开题报告

项目名称： 语音识别性别

姓名：王川

开题时间：2017-05-10

1.项目背景

语音识别技术（Speech recognition）是指将人类语音中的词汇内容转换为计算机可读的输入，例如按键、二进制编码或者字符序列。语音识别技术应用于各种场合，语已经进入工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品等各个领域。而性别语音识别则是语音识别中的一个重要分支，性别辨识可应用于需验证男女性别的场合，例如根据语音识别性别并与身份证号码中性别位号码相互验证，

可以对二者的一致性进行确认。其次，说话人的性别判定的准确性能够影响到语音识别系统的性能，如果性别判定错误，那么系统将会按照错误的模型进行匹配，最后识别得到的结果就不会很理想。再而如果人机交互中，如果机器能够正确识别出来人的性别，就可以针对人的性别才去更加智能化的交互。

因此，该项目根据才音频信号中已提取出特征的数据作为输入数据，采用不同的机器学算法对语音的性别特征进行识别，分析不同特征对区性别的影响程度，对不同的机器学算法在改分类问题上的表现进行讨论。

2.问题描述

该项目使用机器学习的方法判断一段音频信号是男性还是女性，数据已经从音频信号中提取出特征，因此我们需要根据声音的特征来判断是男性还是女性。一般而言，男声的基音频率分布范围为 0~ 200H z，女声的基音频率分布范围为 200~ 500H z。因此,准确而可靠地估计基音周期对于说

话人性别识别非常重要。

3.输入数据

输入的数据集来源于kaggle【3】，这个数据集包含3168个样本，其中50%为男性，50%为女性，并且语音样本已经通过声学分析已经使用R语言脚本处理过提取了特征。数据集中的特征见下图。

- meanfreq: 频率平均值 (in kHz)
- sd: 频率标准差
- median: 频率中位数 (in kHz)
- Q25: 频率第一四分位数 (in kHz)
- Q75: 频率第三四分位数 (in kHz)
- IQR: 频率四分位数间距 (in kHz)
- skew: 频谱偏度
- kurt: 频谱峰度
- sp.ent: 频谱熵
- sfm: 频谱平坦度
- mode: 频率众数
- centroid: 频谱质心
- peakf: 峰值频率
- meanfun: 平均基音频率
- minfun: 最小基音频率
- maxfun: 最大基音频率
- meandom: 平均主频
- mindom: 最小主频
- maxdom: 最大主频
- dfrange: 主频范围
- modindx: 累积相邻两帧绝对基频频差除以频率范围
- label: 男性或者女性

在使用数据集之前会对数据集的基本特征进行分析，对可能存在的异常数据点进行删除处理，分析特征之前的关联和差异性，采用特征方法进行探索，确定与性别判断相关性高的特征。

输入数据按比例随机分为训练集和测试集，测试集中又会按一定比例随机分成部分数据作为测试模型的验证集。训练，验证，测试，都将分别使用不同的数据集。

4.解决办法

该项目采用机器学的方法对数据集进行分类，在数据集的清理，探索部分将经过特征工程，主成分分析等进行处理；而对数据集进行分类的机器学方法可以采用决策树，神经网络，随机森林，支持向量机，逻辑回归，XGBoost 中的一个或则几个，机器学习模型好坏的评价指标是对性别预测的准确率。

5.基准模型

根据【4】中可知：男声的基音频率分布范围为 0~ 200Hz,

女声的基音频率分布范围为 200~ 500Hz，为了确定机器学习算法与基于非人工智能的方法相比能获得更好的结果，可以采用基准模型用于测量初始准确度。由此，可以根据音频率来确定一个基准模型，大于 200 的为女性，小于 200 的为男性。可以用与机器学习模型一样的衡量指标-预测准确率来确定该基准模型的准确率。

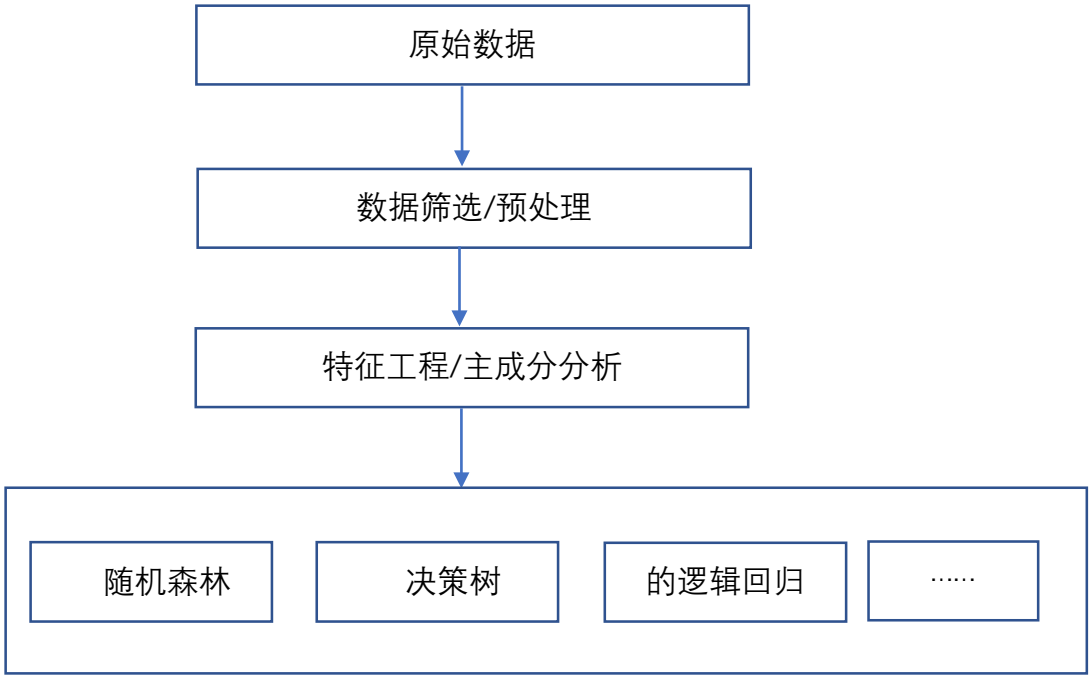
6.评估指标

由于男女的分类只有两种分类，所以不管是把声音识别为男性还是女性，都不存在着任何偏向性，所以问题本身就没有查全率和查准率的问题。对模型识别男性还是女性的好坏指标应该采用识别的准确率来衡量，及准确识别的测试样本数量与所有测试样本的数量之比：

accauracy= A(正确识别的性别的数量) / A+B（总数量）

7.设计大纲

如下图:



项目为以下几个步骤进行:

1. 原始数据的预处理，包括数据的清洗和部分数据的可视化以及数据的编码。
2. 对数据的特征进行分析，确定哪些特征与性别相关，以及不相关的特征与性别相关特征的关系。
3. 根据特征，对数据进行训练，验证，测试数据集的拆分，选择而一种或则多种机器学模型进行训练，预测，获得预测精度。
4. 将获得的预测精度与基准模型的预测精度进行比较，讨论和分析不同机器学习模型的结果。

8.参考文献

1. http://www.ctiforum.com/factory/list/www.delta.com/delta10_1105.htm 语音技术——性别辨识和语者验证

2. http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/ 一个使用 R 语言分

析数据集的例子
3. https://www.kaggle.com/primaryobjects/voicegender 语音下载材料
4. https://wenku.baidu.com/view/9010c72fb4daa58da0114a36.html 基于高斯混合模型的语音性别识别
5. http://www.doc88.com/p-4961318524195.html 语音性别识别报告
6. http://cea.ceaj.org/CN/article/downloadArticleFile.do?attachType=PDF&id=22878 基于性别识别的分类 CHMM 语音识别