机器学习纳米学位

声音性别识别

Fayren udacity

April 10th, 2018

Proposal

本文旨在介绍使用机器学习的方法实现男性和女生声音的自动识别。

第一部分背景介绍中，首先介绍了声音性别识别工作的背景知识，包括传统的处理思路和机器学习思路的异同以及优缺点。

第二部分问题描述 中，详细的介绍的声音性别识别问题，阐述了面临的问题是什么，传统思路的难点在哪里以及本文采用机器学习方案的优势。

第三部分数据输入中主要介绍了数据的预处理方式。

第四部分解决方案陈述中，主要介绍了项目的总体设计方案和整体流程。

第五部分基准模型中，陈述了本文拟采用的算法模型以及其参数设置。

第六部分评估衡量中，主要陈述了模型优劣评估的方案。

背景介绍

声纹性别识别是一个复杂的工作，作为人类本身可以很容易的听出来男生或者女生，但是机器如何能分辨男性声音和女性声音呢？

一方面，在机器学习技术风行之前，大多采用硬编码的方式来做声音性别识别，也就是在编写程序时，事先分别定义好男生女生的声音特点，再用写好的程序去匹配声音特点，这种传统的编写规则程序不仅区分准确率很难保证，规则编写更是非常复杂。

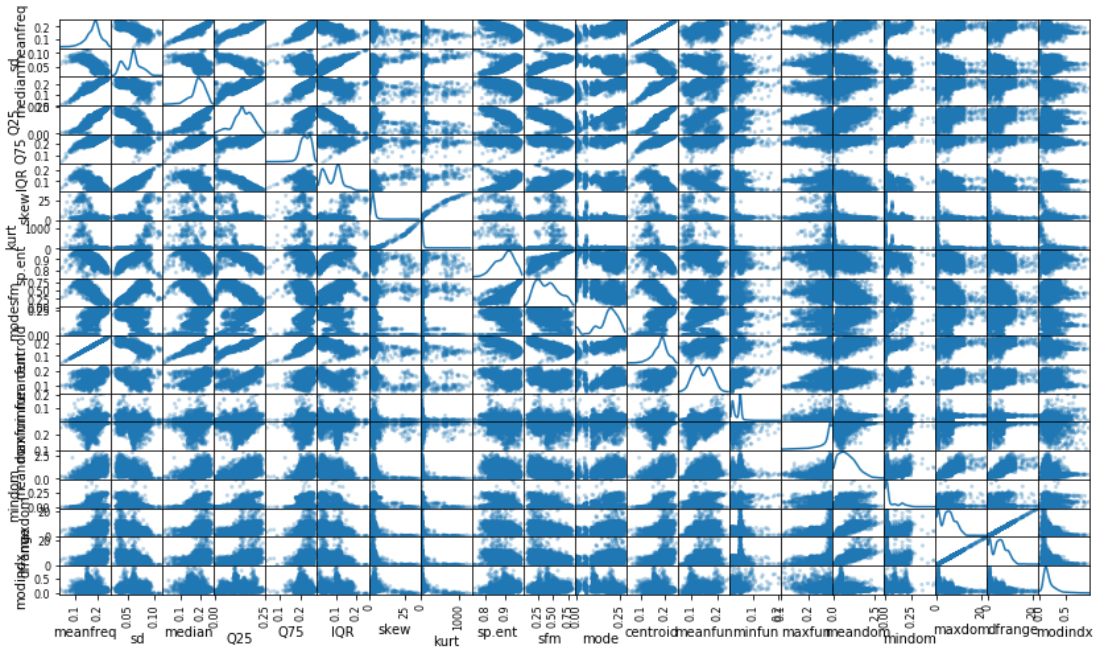
另一方面，随着机器学习技术的发展，使用机器学习方法判断一段音频是男性发出还是女性发出变得容易了许多。通过对声音数据的学习，程序能在大量的数据中统计学习得到女性男性的声音特点。再在习得的经验基础上再来做声音识别，这样不仅大大提高了识别的正确率，同时还不用编写复杂的规则，程序也拥有更灵活适应数据的特点。

问题描述

区分声音的性别很明显是一个分类问题。首先在区分男性声音和女性声音之前我们得先了解如何去表达声音，而且是一种可量化表达声音的方案。本文将根据以下特征来研究如何区分男性声音和女性声音：

- 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

本文在研究声纹特征时发现一般声纹特征分布如下：



如上图所示，在各个特征空间离，特征呈现正态分布的特点。

在背景介绍中，我们知道这里同样可以使用传统的编写规则的方式实现声音性别的区分，例如一般女生的声音音调比男生高，也就是频率会很高，对主频范围设置一个值，大于某个值的认为是女生。当然这么简单的规则肯定无法收获很高的正确率。虽然这里可以通过复杂一点的规则提高区分的正确率，但是编写规则将会是异常麻烦的。

所以，机器学习方法在这里就表现出它得天独厚的优势。通过对样本的分析，让程序习得男生女生的声音各个特征的一般分布，从而分析得到是男生还是女生的声音，避免了规则的编写，同时提高了程序的灵活性。

数据输入

本文实验中用到的数据集包含 3168 个样本，其中 50%为男性，50%为女性。样本数据中包含问题描述中提到的各个声音特征以及该声音对应的性别。

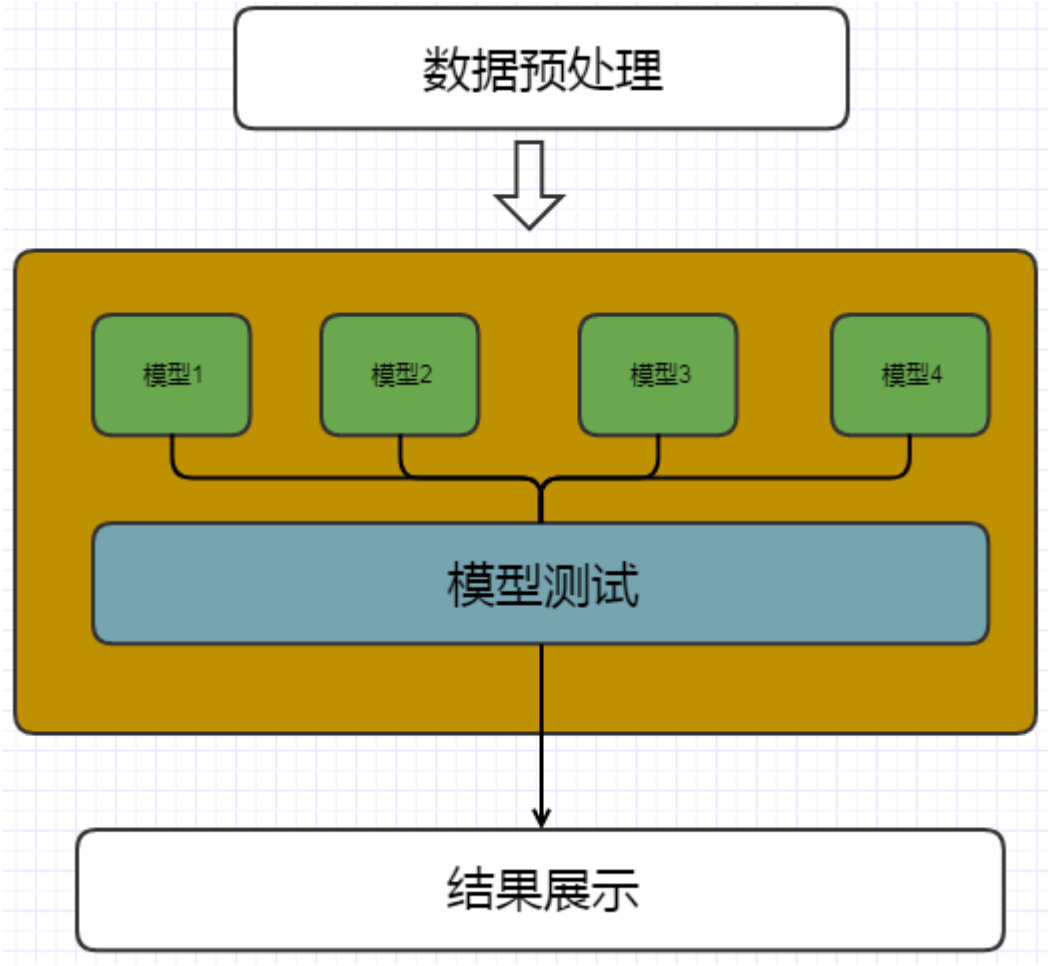
由于本文采用机器学习的方法来对声音性别进行识别，在训练之前的数据处理时，主要做数据做了一下操作：

1. 原数据将 label 从其中分离，分别得到训练数据和 label。
2. 将源数据去除事先随机预留了 20%的数据作为测试集，用来衡量模型的性能。
3. 其次剩下的 80%数据中的 20%数据用来做训练时的验证集，用来找出模型的最佳效果。剩下的 80%数据作为训练数据

解决方案陈述

由于问题的复杂性，难以编写规则来直接分类声音，本文使用机器学习的方法来实现声音性别的分类。

项目的总体训练流程如下：



主要包含以下几个步骤：

1. 数据预处理，切割得到测试集，训练集和验证集。
2. 分别使用多个模型训练来拟合数据。
3. 评测各个模型的性能，对于每个模型选出最佳性能下的参数。
4. 展示最终结果。

基准模型

如前两章介绍，本文拟准备使用机器学习的方法来识别声音的性别。最终选择了如下四种模型做对比试验。其中，随机森林，支持向量机，LR 回归为传统的机器学习模型，最后本文拟采用 DNN 神经网络试验。

- 随机森林 (Random Forest)
- 支撑向量机 (SVM)
- Logistic 回归（LogisticRegression）
- 深度神经网络（DNN）

其中 LR 和支持向量机可以直接使用，随机森林模型采用 GridSearch 方法找到最合适的决策树数量。DNN 神经网络拟采用如下的结构：

Layer (type)	Output Shape	Param #
the_input (InputLayer)	(None, 20)	0
dense1 (Dense)	(None, 128)	2688
activation_21 (Activation)	(None, 128)	0
dropout_19 (Dropout)	(None, 128)	0
dense2 (Dense)	(None, 256)	33024
activation_22 (Activation)	(None, 256)	0
dropout_20 (Dropout)	(None, 256)	0
dense3 (Dense)	(None, 256)	65792
activation_23 (Activation)	(None, 256)	0
dropout_21 (Dropout)	(None, 256)	0
dense4 (Dense)	(None, 128)	32896
activation_24 (Activation)	(None, 128)	0
dropout_22 (Dropout)	(None, 128)	0
dense5 (Dense)	(None, 128)	16512
activation_25 (Activation)	(None, 128)	0
dropout_23 (Dropout)	(None, 128)	0
dense-final (Dense)	(None, 1)	129
sigmoid (Activation)	(None, 1)	0

Total params: 151,041
Trainable params: 151,041
Non-trainable params: 0

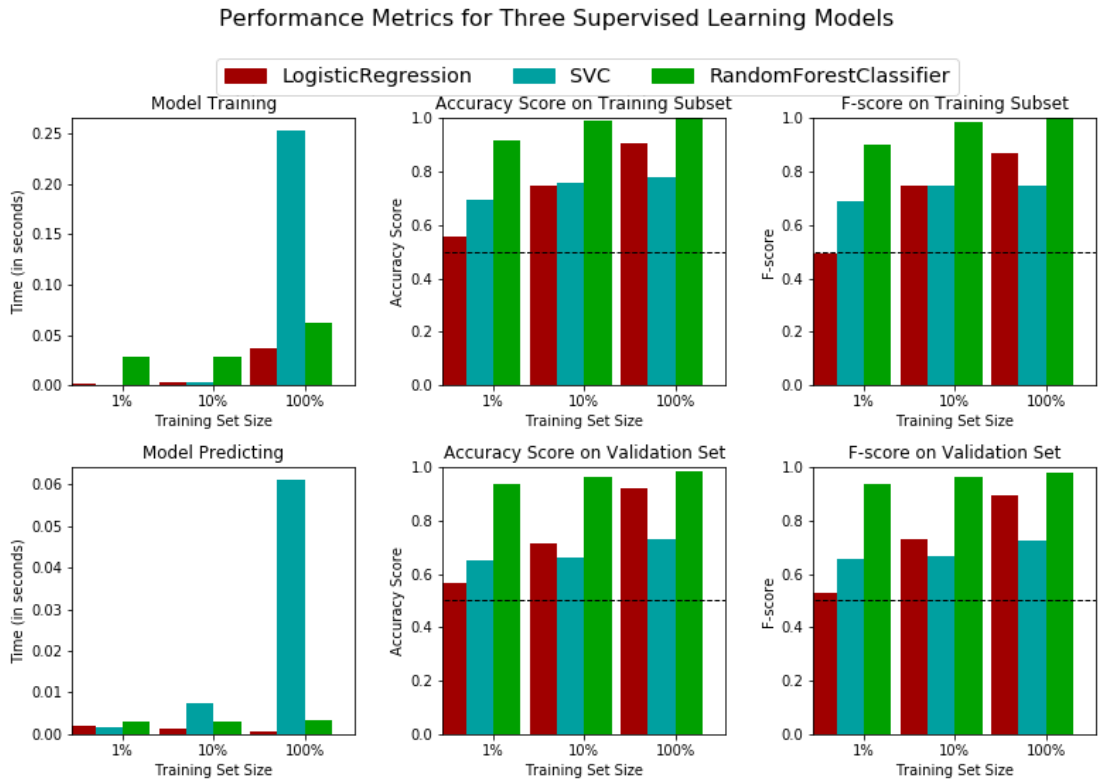
评估衡量

本文主要涉及四种模型，评估上主要以训练集正确率和验证集正确率来评估模型性能。评估公式：

$$accuracy = T/(T+F)$$

其中 T 表示判断正确的样本数量，F 表示判断错误的样本数量。

最后输出如下图所示的正确率图：



项目设计

总结本文的实现方案，首先明确本文要解决的问题是实现一个能区分声音的性别的程序。其次，本文分析了实验中所获取到的数据，阐述了声音文件的特征抽取，这些特征对于性别的识别至关重要。紧接着是机器学习模型的选型，本文在实现上拟采用四种模型进行实验，分别是逻辑回归模型，SVM 支持向量机模型，随机森林模型和 DNN 深度神经网络模型。这些模型在已经获取到声音文件特征的情况下可以直接开始训练。最后，需要对本文所选择的模型优劣进行评估，针对前三种机器学习模型，使用传统的测试集正确率，验证集正确率和 Fscore 进行建成。针对 DNN 神经网络的训练，拟准备添加正确率曲线和 loss 曲线观察模型的收敛。

参考文献

[1] Thomas S. Data-driven Neural Network Based Feature Front-ends for Automatic Speech Recognition[Ph.D. dissertation], Johns Hopkins University, Baltimore, USA, 2012.

[2] Wu Wei-Lan, Cai Meng, Tian Yao, Yang Xiao-Hao, Chen Zhen-Feng, Liu Jia, Xia Shan-Hong. Bottleneck features and subspace Gaussian mixture models for low-resource speech recognition. Journal of University of Chinese Academy of Sciences, 2015, 32 (1):97-102(吴蔚澜，蔡猛，田垚，杨晓昊，陈振锋，刘加，夏善红. 低数据资源条件下基于 Bottleneck 特征与 SGMM 模型的语音识别系统. 中国科学院大学学报, 2015, 32 (1):97-102)

[3] Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors. Computer Science, 2012, 3 (4):212-223