语音性别识别项目
开题报告

Max Liu 刘子铭*

2018年8月16日

目录

*电子邮件:simonliu245@gmail.com

1 领域背景

语音性别识别还仅仅是语音识别领域的一个小分支，且技术上已经成熟，人们已经开始关注更精细的问题[1]。自20世纪70年代人们创立隐马尔可夫（HMM）模型以来，语音识别领域就开始逐步发展。其核心技术包括隐马尔可夫模型（HMM），混合高斯模型（GMM），MFCC，LM等等。现如今，在语音识别领域效果最好的当属深度学习技术[2]，其错误率在以往最好的系统GMM-HMM框架的基础上相对还要下降30%或更多，这一成就也使得语音识别技术有了更强的实用性。

我选择该项目有两个主要动机，其一是各平台广泛使用语音识别功能，如Siri，Cortana等等交互应用炙手可热的的趋势就可以预见语音识别技术还会有更广泛的天地；选择该项目的另一原因是一个偶然，前不久我偶然读了一篇论文[3]，它描述了一种可以提取有实际意义的重要特征的算法SMuRFS。我想现在大多数算法已经可以保证语音性别识别的准确性，但提取的特征是否有声学上的实际意义呢，这就是我的思考所在。

可惜的是可惜我还没有读懂上述论文的R代码，因此我还不能做具体实现，不过后续我会做一些尝试，这里我只叙述我的一些具体想法。上述论文描述了传统的特征提取算法的一大缺陷是可能只能提取出统计学上的重要特征而非与实际问题密切相关的特征，为此它提出了名为SMuRFS的算法，目的是识别有着实际意义的重要特征。如果我们能根据该算法提取音频特征，就有可能得到比性别识别更好的结果，比如情绪识别等等。

2 问题陈述

该语音性别识别项目的基本任务是准确判断一段音频信号属于男性还是女性[4]。

很显然，这是一个二分类问题，属于监督学习的范畴，且是一个可解问题。我的基本任务是训练出这样一个模型，对给出的音频信号，它能够得出准确的性别。另外，该问题实用性较强的重要原因是音频信号是可通过R脚本量化处理的[4]，且可以通过训练机器学习模型得到的预测器即时得到给定音频的预测结果。

3 数据集及输入

该项目训练用的数据集来源于Kaggle相关项目[4]，是已经经过R脚本量化后的数据集，包括特征和标签值。根据Kaggle上的描述，这些特征是seewave和tuneR这两个R软件包从原始音频信号中提取出的声学特性。一些提取的特征部分如下所示，这些声学特性的具体含义详见资料[4]。

表 1: 部分特征及其含义

meanfreq	频率平均值(in kHz)
sd	频率标准差
...	...
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

当然，把数据作为输入之前是有预处理步骤的。这些特征只是原始的输入数据，预处理的步骤我会在后续的notebook中详细给出分析过程，这里只说明思路。实际上我已经观察了数据集，在初步观察时我发现了三个问题，分别对应的处理措施是独热编码、打乱数据的顺序以及剔除缺失的数据，我将在notebook中详细解释我的思考过程。除了这些，数据并没有出现太大的波动，因此我会考虑是否对数据集进行标准化，考虑的依据是分别训练出的模型的性能。

另外，在得到音频信号的归属之前，先要训练预测器。数据集应当被随机分割为训练集验证集和测试集，随机是为了保证训练样本不会轻易出现同性语音占比太多；多分出验证集是因为不能使用测试集进行调参，测试集是最终的测评集合。我计划训练多个监督模型并按一定标准选出最佳分类器，用于最终的输出预测。

实际操作识别一段音频信号的时候可能需要手动运行R脚本处理音频信号并将得到的输出作为输入向量，已经训练好的分类器会根据输入向量返回相应的结果。

4 结果陈述

在上一部分中我们已经介绍了问题的解决方案。我们可以训练不同的监督模型并对预测表现进行比较，比较的依据是一些评价指标，这将在第6部分介绍。得出各个模型的结果后我们也会把结果与Kaggle上已有的结果进行对比，最后选择出最佳分类器。

5 基准模型

我们的基准模型[5]是天真预测器，默认将所有音频识别为男性（male），它的准确率应该在50%附近。按概率说，基准模型与随意猜测的预测准确性相近。因此，天真预测器理应是合理的基准模型。

经过训练的机器学习模型理应做到比基准模型的表现更优才能说明训练有了良好效果。当然，既然我们提到了Kaggle，我们试图将准确性提高到Kaggle上已经提到的表现，即使不能，也要与之接近。

表 2: Accuracy on Kaggle

Baseline (always predict male)	50% / 50%
Logistic Regression	97% / 98%
CART	96% / 97%
Random Forest	100% / 98%
SVM	100% / 99%
XGBoost	100% / 99%

6 评估指标

我们的评估指标可以使用分类模型的经典指标，准确性，精确率，召回率和F1 measure[6]，这些都可以作为量化模型性能的工具。其中F1 measure是召回率和精确率的调和平均，既有平均价值，也会对召回率和精确率接近的模型给予更高的分数，这样的模型往往更加实用。

不过这里我们使用的主要评估指标还是准确性（Accuracy），与Kaggle上的相同，其它参数只是帮助我们对模型有更客观的认识。其计算公式如下所示：

$$\begin{aligned} Accuracy &= \frac{\#True\ Positive + \#True\ Negative}{\#True\ Positive + \#True\ Negative + \#False\ Positive + \#False\ Negative} \\ &= \frac{\#Rightly\ Predictions}{\#All\ Samples}. \end{aligned}$$

7 项目设计

综合以上6个部分我们就大致可以得出项目思路了。

首先便是数据的预处理步骤，我会先导入全部数据集，输出几条数据以观察数据集的结构，根据需要对数据集进行独热编码以方便处理。再观察是否有缺失数据，如果有，考虑去除缺失数据。又根据第3部分中的描述，我会考虑做数据的打乱顺序。接着可以做适当的可视化观察，简单地筛选特征并观察是否有二分类的情况，为模型的训练提供一个先验。这里我打算使用的有两种方式，一种是利用pandas中的corr()函数求取与label标签相关性最强的几个特征，另一种是用sklearn中随机森林分类器的feature_importances_函数来得到特征重要性。恰好，它们二者可以相互验证。

其次便是模型的训练步骤，把已经处理好的数据集适当分割为训练集，验证集和测试集三部分。然后初始化各分类器，比如K近邻，支持向量机，决策树等等并训练它们，最后再测试集上获取评估参数，并比较不同模型的性能，最后得到一个最佳分类器作为问题的解决方案。当然，根据Kory Becker[5]的描述，我们可以使用融合的模型，但是由于篇幅所限，这里就不多赘述了。

如果可能，后续我会考虑之前提到的SMuRFS，即序列特征多元随机森林选择算法，根据我的推测，它可能直接得出与性别直接相关的特征，那样就可以把该问题拓展，比如给定其它音频的标签，就可以选择出对应相关的重要声学特征，获得其它理论成果。

8 参考文献

【1】腾讯AI Lab副主任俞栋：语音识别领域四大前沿问题亟待研究[EB/OL].

http://tech.qq.com/a/20170528/017297.htm.

【2】俞栋, 邓力. 解析深度学习-语音识别实践[M]. 北京:电子工业出版社, 2016.

【3】Mayer J, Rahman R, Ghosh S, et al. Sequential Feature Selection and Inference using Multivariate Random Forests[J]. Bioinformatics, 2017, 34(8).

【4】Kory, Becker. Gender Recognition by Voice—Kaggle[EB/OL].

https://www.kaggle.com/primaryobjects/voicegender.

【5】Kory, Becker. Identifying the Gender of a Voice using Machine Learning[EB/OL].

http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/.

【6】范淼, 李超. Python机器学习及实践－－从零开始通往Kaggle竞赛之路[M]. 北京:清华大学出版社, 2016. 41-43