机器学习毕业项目报告

项目名称： 语音识别性别

姓名：王川

开题时间：2018-05-16

1. 定义

1.1 项目概述

该项目内容来源于 Kory Becker 发布于 kaggle 的语音性别识别(Gender Recognition by Voice)挑战，其目的是使用机器学习的方法判断一段音频信号是男性还是女性。

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。

1.2 问题陈述

本项目要解决的问题是:判断一段语音的说话者是男性还是女性，这是一个有监督的二分类问题。声音的信息已进行了特征提取，所以，实际上是根据声音的特征统计量数据集来判断是说话者男性还是女性,。

为了解决该问题，采用了机器学习的方法，具体过程是:首先，对声音的数据集进 行探索和处理，得到适用于机器学习模型训练的数据。然后，将数据集划分为训练集、 验证集和测试集，这些数据集将会适用在模型训练、验证和测试的不同阶段。之后，完成对模型的训练及验证，并完成数据集特征的取舍和模型性能的提升。最后，完成对模型的测试，得到模型对语音性别的预测准确率。

最终会获得在该数据集下较高的预测精度，以及有效的机器学习模型。获得的机器 学习模型能够使用语音特征信息判断说话者的性别，并达到较高的预测准确率。

1.3 评价指标

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:

accuracy (base) = A(正确识别的性别的数量) / A+B(总数量)

使用机器学习模型的预测准确率至少要优于基准测试的正确率。即：

accuracy (machine learning) >= accuracy (base)

2. 分析

2.1 数据研究

2.1.1 数据集来源及特征

输入的数据集来源于[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，分析频率范围为0hz-280hz。

数据集中的特征见下图表1。

表1 语音识别数据集特征

特征	描述
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

从表1可以看出，数据集中主要是声音样本频率的相关的特征，主要要主频(平均值、最小值、最大值、范围)，基频(最小值、最大值、平均值)，频率分布(偏度、峰度)，信息量度量参数(频谱熵、频谱平坦度)以及其他基本统计量(平均数、四分位数、中位数、众数)等。当然还有表示男性或女性的类别标志标量。

2.1.2 数据集样本描述

在数据集中抽取了100#，200#和3000#行的两个样本量，样本各特征数据见表2。其中，100#为男性样本，200#和3000#为女性样本。

根据[4]可知，男声的基音频率分布范围约为0~200Hz，女声的基音频率分布范围为200~500Hz。从表中可以看到男性样本平均基频(meanfun)为0.119977kHz，女性样本平均基频为 0.188126kHz和0.172260kHz，确实是在前述范围之内，且女性平均基频约为男性平均基频的两倍。

表2 抽取的3个样本

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm
0	0.119977	0.081471	0.109892	0.060215	0.192688	0.132473	22.400866	560.122789	0.854878	0.505371
1	0.188126	0.082495	0.218128	0.165719	0.247730	0.082010	2.384851	11.942546	0.933937	0.592861
2	0.172260	0.061118	0.167037	0.143519	0.215185	0.071667	2.226554	8.718907	0.934892	0.568876

mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0.000000	0.119977	0.081201	0.015795	0.262295	0.007812	0.007812	0.007812	0.000000	0.000000	male
0.000000	0.188126	0.181134	0.016649	0.266667	0.297917	0.007812	0.687500	0.679688	0.287356	female
0.058704	0.172260	0.151881	0.069444	0.277778	0.930176	0.029297	1.459961	1.430664	0.477816	female

meanfreq和centroid的值在两个样本内完全一致。在女性样本中的这个量大于男性样本中的这个量，这说明女性样本的声音要比男性样本的声音要响亮。因为女性的声音音调一般较高，而男性一般声音音量较大，综合看，在这两个样本中频率高的女性样本占了上风。

IQR、kurt、sd 均反映的是声音总体的频率变化范围的离散程度。kurt 的定义为四阶样本中心距除以样本方差的平方。所以，男性样本的 kurt 要小于女性样本，而女性男性样本 sd(标准差:方差开方)的值要高于男性是合理的。sd 值大，则数据分散程度明显，进而 IQR 值大。这些量说明了男性样本的声音频率分布要更离散，即变化范围更广。声音频率的变化范围因个体及语音内容而异。日常生活中，男性在声调变化方面似乎更有优势(比如男性更容易模仿女性的声音，反之却不那么容易)，所以这两个样本的声音频率变化范围差异似乎和这个观察契合。

从 skew 可以看出，女性样本频率分布的右偏程度稍高。

sp.ent和sfm度量的是语音的信息量，因语音内容而异，只要在0~1之间均为正常。在只有一个基频的情况下，主频一般为基频的倍数。在有多个基频的具体语音信息中，较难有这种关系，只能粗略看出在两个样本中meandom较meanfun大。其余特征，例如基频的最值、主频的最值、四分位量等有很大的随机性，和说话者的发音习惯、语音内容均有关系。

2.1.3 数据集特征统计量

数据集特征统计量见表3:

表3 数据集特征统计量

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode
count	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000
mean	0.180907	0.057126	0.185621	0.140456	0.224765	0.084309	3.140168	36.568461	0.895127	0.408216	0.165282
std	0.029918	0.016652	0.036360	0.048680	0.023639	0.042783	4.240529	134.928661	0.044980	0.177521	0.077203
min	0.039363	0.018363	0.010975	0.000229	0.042946	0.014558	0.141735	2.068455	0.738651	0.036876	0.000000
25%	0.163662	0.041954	0.169593	0.111087	0.208747	0.042560	1.649569	5.669547	0.861811	0.258041	0.118016
50%	0.184838	0.059155	0.190032	0.140286	0.225684	0.094280	2.197101	8.318463	0.901767	0.396335	0.186599
75%	0.199146	0.067020	0.210618	0.175939	0.243660	0.114175	2.931694	13.648905	0.928713	0.533676	0.221104
max	0.251124	0.115273	0.261224	0.247347	0.273469	0.252225	34.725453	1309.612887	0.981997	0.842936	0.280000

centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000
0.180907	0.142807	0.036802	0.258842	0.829211	0.052647	5.047277	4.994630	0.173752
0.029918	0.032304	0.019220	0.030077	0.525205	0.063299	3.521157	3.520039	0.119454
0.039363	0.055565	0.009775	0.103093	0.007812	0.004883	0.007812	0.000000	0.000000
0.163662	0.116998	0.018223	0.253968	0.419828	0.007812	2.070312	2.044922	0.099766
0.184838	0.140519	0.046110	0.271186	0.765795	0.023438	4.992188	4.945312	0.139357
0.199146	0.169581	0.047904	0.277457	1.177166	0.070312	7.007812	6.992188	0.209183
0.251124	0.237636	0.204082	0.279114	2.957682	0.458984	21.867188	21.843750	0.932374

从数据集特征的统计量可以看出，特征大概可以分为 4 类:

a) 数值范围在 0~0.3 之间表征语音频率范围类(mean/sd/median/Q25/Q75/IQR/mode/centroid/meanfun/minfun/maxfun);

b) 数值范围在0~1之间表征信息携带类(sp.ent/sfm/modindx);

c) 数值范围在超过1的表征语音分布特性类(skew/kurt);

d) 主频类(meandom/mindom/maxdom/dfrange)。主频通常为基频的数倍，数值范围跨度较宽。

2.1.4 数据集异常说明

关于数据集异常的说明有一下几点:

a) 数据集未见缺失(NA)数据，不需要补全;

b) 数据集有两个样本完全相同，需要去重;

c) 数据集从几个不同的方面描述语音信息，度量衡并不一致，所以数据集需要进行归一化处理。数据集中可能的离群数据在归一化后再进行判断。

d) meanfreq和centroid表征同一特征，可以删除一个;

e) 表征男女类别的标志信息需要进行编码;

2.2 数据集可视化

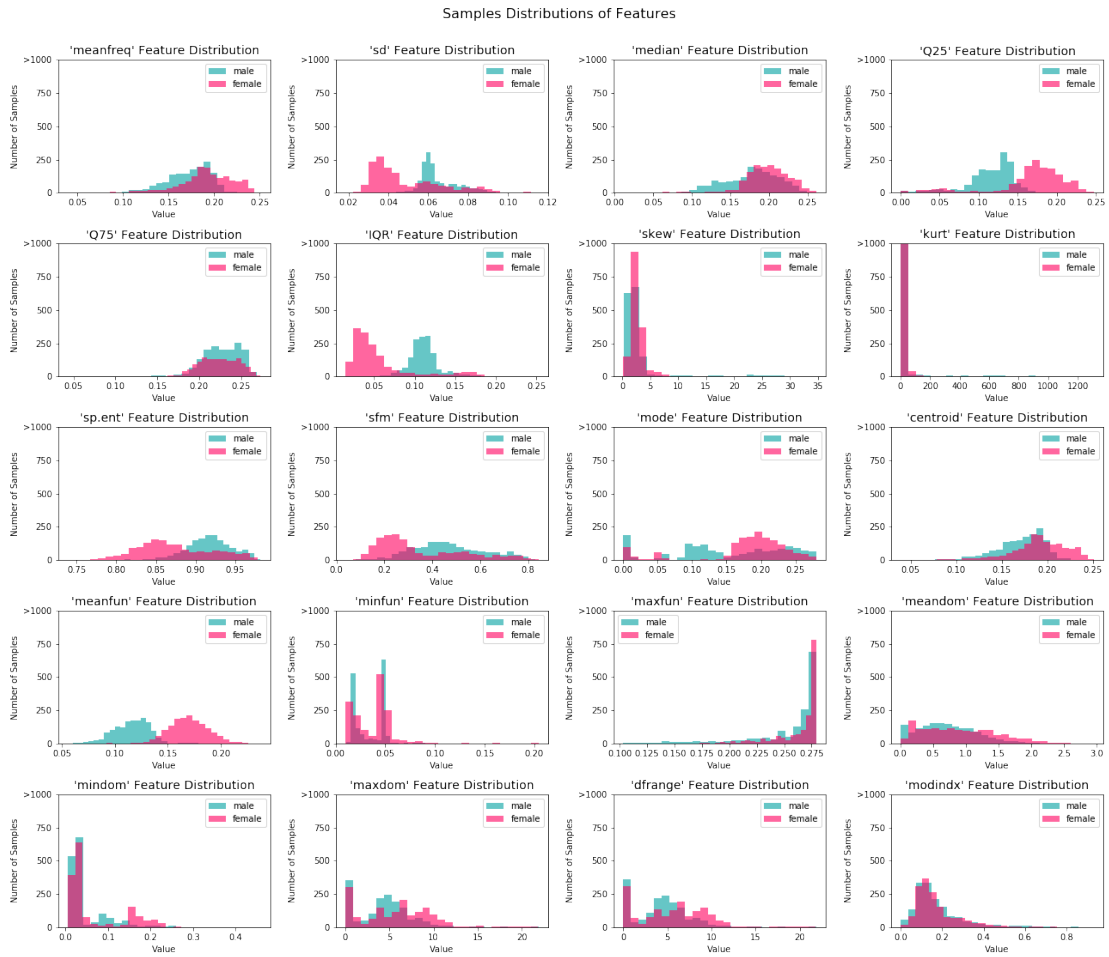
为了更加直观的看出男女样本在不同特征上的频数分布情况，将数据集特征的样本分布绘制成柱状图，如下图。

根据男女样本的频数分布范围是否存在差异，将特征分类两类:

a) 存在明显差异类，如sd/Q25/IQR/sp.ent/sfm/meanfun;

b) 较近似类，如除a)中所列特征;

modindx/meandom/maxfun 分布范围几乎重合。skew/kurt存在严重的右偏分布，后续需要取对数处理。



2.3 算法与方法

本项目采用的是机器学习算法对语音的性别特征进行识别。项目中涉及数据集预处理、特征选择和机器学习算法实现等三方面的技术，下面分别进行说明。

2.3.1 数据集预处理

数据集预处理部分包括特征的初步筛选、数据的归一化处理、异常值的去除以及性别类别的编码等。

(1) 初步筛选特征

删除不必要的特征的好处在于，可以减少后续机器学习模型的训练样本和训练时间。过多的特征容易产生复杂的模型，会降低模型的泛化能力。根据前文对数据集特征的分 析，因meanfreq与centriod 表相同特性，可以将meanfreq特征删除;数据集可视化中男 女性样本分布范围相近的特征可以考虑删除，但这种方法并不可以量化 “相近” 的标准，基于保守考虑，先不做其余特征删除处理。

(2) 数据集归一化处理

由 2.1.4 节，度量衡不同导致样本数据的范围分布差异。在数据上面施加一个缩放 并不会改变数据分布的形式，且规一化保证了每一个特征在使用监督学习器的时候能够被平等的对待。对于skew/kurt存在严重偏态分布的变量，则需要先进行对数转换。将数据转换成对数，这样非常大或小的值不会对学习算法产生负面的影响,并且使用对数变换显著降低了由于异常值所造成的数据范围异常。

进行对数转换时有一个问题:对于本身数值范围分布在0~1的数据要不要进行对数处理?对这些数据一般采用先加1，再取对数的处理方式。在数值范围分布较小(0~1)时，直接取对数再归一化，或直接归一化其实对数据集的分布方式改变较小。故本文对这种分布的特征采用直接归一化的方式。对于skew/kurt 则采用先对数再归一化的方式。

(3) 异常值的去除对数据进行归一化处理后，可根据数据点的离群情况来判断的异常样本。离群点的存在会对模型的训练造成干扰。异常值被定义为小于Q1-1.5IQR或大于Q3+1.5IQR的值。实际去除异常点的过程中，考虑到特征之间的关联性，将同时有几个特征数值判断为异常值的样本进行去除，这样能增加判断样本异常的置信度。总的数据集样本较少，应减少数据量的损失。

(4) 性别标签的编码性别类为标签值，为了便于算法模型使用，将male编为0，将female编为1。

2.3.2 采用的机器学习模型

从[8]推荐的几种算法以及开题报告导师给出的建议，选择了随机森林、xgboost、和逻辑回归作为本项目完成的算法。

(1) 随机森林

随机森林是一种集成学习算法。它的特点是:具有较高的精度;不容易发生过拟合;能够将弱学习器进行联合，获得较好的预测效果;数据不平衡会导致分类导致分类精度下降，训练比较耗费时间。随机森林因其基学习器的多样性，使最终集成后的泛化能力较强。

本项目的特征较多，为随机森林提供了较大的特征选择范围，这将可能增加进一步提升随机森林的泛化能力。

(2) xgboost,

xgboost全称是 “eXtreme Gradient Boosting” ，是在GBDT的基础上对boosting算法进行的改进，内部决策树使用的是回归树。回归树的分裂结点对于平方损失函数，拟合的就是残差；对于一般损失函数（梯度下降），拟合的就是残差的近似值，分裂结点划分时枚举所有特征的值，选取划分点。

在有监督学习中，我们通常会构造一个目标函数和一个预测函数，使用训练样本对目标函数最小化学习到相关的参数，然后用预测函数和训练样本得到的参数来对未知的样本进行分类的标注或者数值的预测。一般目标函数是如下形

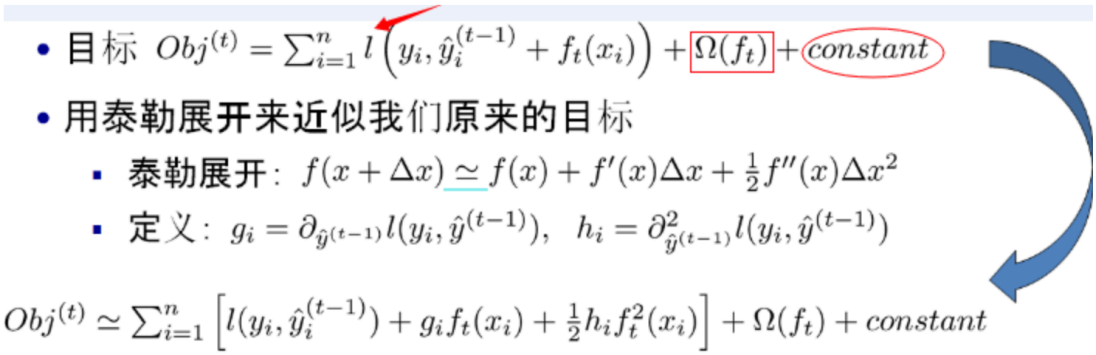
式的，我们通过对目标函数最小化，求解模型参数Θ。预测函数、损失函数、正则化因子在不同模型下是各不相同的。

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta)$$

Training Loss measures how well model fit on training data

Regularization, measures complexity of model

xgboost 与 gbdt 比较大的不同就是目标函数的定义，如下图：



红色箭头指向的 l 即为损失函数；红色方框为正则项，包括 L1、L2；红色圆圈为常数项。xgboost 利用泰勒展开三项，做一个近似，我们可以很清晰地看到，最终的目标函数只依赖于每个数据点的在误差函数上的一阶导数和二阶导数。

xgboost的loss由两部分构成，前者优化经验误差，后者是控制泛化误差：

xgboost的目标函数如下：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Training loss　　Complexity of the Trees

相比较于传统的GBDT，xgboost正则更加的细化了，包括传统的L2正则以及叶子数目的正则项：

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Number of leaves　　L2 norm of leaf scores

优势：

1） 传统 GBDT 以 CART 作为基分类器，xgboost 还支持线性分类器，这个时候 xgboost 相当于带 L1 和 L2 正则化项的逻辑斯蒂回归（分类问题）或者线性回归（回归问题）。

2）传统 GBDT 在优化时只用到一阶导数信息，xgboost 则对代价函数进行了二阶泰勒展开，同时用到了一阶和二阶导数。顺便提一下，xgboost 工具支持自定义代价函数，只要函数可一阶和二阶求导。

3）xgboost 在代价函数里加入了正则项，用于控制模型的复杂度。正则项里包含了树的叶子节点个数、每个叶子节点上输出的 score 的 L2 模的平方和。从 Bias-variance trade off 角度来讲，正则项降低了模型的 variance，使学习出来的模型更加简单，防止过拟合，这也是 xgboost 优于传统 GBDT 的一个特性。

4）Shrinkage（缩减），相当于学习速率（xgboost 中的 eta）。xgboost 在进行完一次迭代后，会将叶子节点的权重乘上该系数，主要是为了削弱每棵树的影响，让后面有更大的学习空间。实际应用中，一般把 eta 设置得小一点，然后迭代次数设置得大一点。（补充：传统 GBDT 的实现也有学习速率）

5）列抽样（column subsampling）。xgboost 借鉴了随机森林的做法，支持列抽样，不仅能降低过拟合，还能减少计算，这也是 xgboost 异于传统 gbdt 的一个特性。对缺失值的处理。对于特征的值有缺失的样本，xgboost 可以自动学习出它的分裂方向。

6）xgboost 工具支持并行。boosting 不是一种串行的结构吗?怎么并行的？注意 xgboost 的并行不是 tree 粒度的并行，xgboost 也是一次迭代完才能进行下一次迭代的（第 t 次迭代的代价函数里包含了前面 t-1 次迭代的预测值）。xgboost 的并行是在特征粒度上的。我们知道，决策树的学习最耗时的一个步骤就是对特征的值进行排序（因为要确定最佳分割点），xgboost 在训练之前，预先对数据进行了排序，然后保存为 block 结构，后面的迭代中重复地使用这个结构，大大减小计算量。这个 block 结构也使得并行成为了可能，在进行节点的分裂时，需要计算每个特征的增益，最终选增益最大的那个特征去做分裂，那么各个特征的增益计算就可以开多线程进行。可并行的近似直方图算法。树节点在进行分裂时，我们需要计算每个特征的每个分割点对应的增益，即用贪心法枚举所有可能的分割点。当数据无法一次载入内存或者在分布式情况下，贪心算法效率就会变得很低，所以 xgboost 还提出了一种可并行的近似直方图算法，用于高效地生成候选的分割点。

(3) 支持向量机

该模型的特点是:可以解决高维问题;解决小样本下的机器学习问题;泛化能力较强;样本很多时，效率不是很高;对缺失数据较敏感。本项目中数据集较小，而涉及的特征较多，SVM比较适合解决高维问题，所以选择SVM进行尝试。

(4) 逻辑回归

该模型的特点是:模型容易使用和解释;可适用于类别和连续性的自变量;预测结果介于是0和1之间的概率，可以根据需求对预测概率低的结果做删除处理。选择逻辑回归的原因在于它的训练速度快、训练代价小，且较适用于二分类问题。且本项目的数据集较小，采用该模型也是适宜的。

2.3.3 特征选择

基于之前的陈述，对特征的处理仅做了删除meanfreq的处理。其余特征的选择算法在下述讨论。

常见的特征选择方法有3种:过滤式选择、包裹式选择和嵌入式选择。

过滤式方法先对数据集进行特征选择，然后再训练学习器，特征选择过程与后续学习器无关。过滤式选择的缺点在于没有考虑到特征之间的关联作用，可能把有用的关联特征误踢掉。

包裹式方法则是将最终使用的学习器的性能作为特征子集的评价准则，为学习器“量身定制”特征子集。从最终学习器的性能来看，包裹式特征选择比过滤式特征选择更好，但由于在特征选择过程中要进行多次训练学习器，包裹式特征的选择的计算开销通常要比过滤式特征选择大得多。典型的包裹型算法为“递归特征删除算法”。

嵌入式特征选择方法将特征选择过程与机器学习训练过程融为一体，两者在同一个优化过程中完成，即在学习器训练过程中自动地进行了特征选择。这与前两中选择方法中特征选择过程和学习器训练过程的区分进行有一定的区别。

本项目中选择包裹式方法。具体的说，是采用了 “递归特征删除算法” 。选择该方法的原因是，本项目数据集较小，进行多次的训练也不会花很长时间;另外，该算法较易理解，结果易于可视化和分析。

2.4 基准模型

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[6]。

根据领域知识[1]，可以选择基音频率0.2kHz指标来构建一个基准模型。即meanfun大于0.2kHz直接判定为女性，meanfun小于0.2kHz判定为男性。采用与机器学习模型一样的衡量指标，即预测准确率来确定该基准模型的准确率。

accuracy(base)=0.53

3. 方法

3.1 数据预处理

除重复元素和meanfreq特征，对skew/kurt做对数处理，异常样本判断并且删除删除异常数据，最终剩余样本 3111 个。并且对数据集进行了归一化处理，然后对性别标签进行了编码。

3.2 实施

3.2.1 特征筛选

本节将采用基于交叉验证的“递归特征删除算法”来进行特征的选择。该算法的支持 来自于 sklearn.feature_selection.RFECV。用了4中机器学习算法，

a. 随机森林模型 RandomForestClassifier(random_state=0)

b. xgboost XGBClassifier()

c. 支持向量机模型 SVC(kernel="linear", random_state=0)

d. 逻辑回归模型 LogisticRegression(random_state=0)

对3.1节预处理完成的数据集进行拆分，从数据集随机选取中20%作为测试集，然后从剩下的80%数据中，随机抽取20%的数据作为验证集，剩下的数据为训练集。采用sklearn.feature_selection.RFECV计算不同学习器情况下，特征的重要程度；其中RFECV采用的参(step=1,cv=StratifiedKFold(2),scoring='accuracy')，下图为各种特征模型排序。

a. RandomForestClassifier：

	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
0	1	2	1	8	1	3	7	1	1	1	1	1	6	10	4	9	1	11	5

b. XGBClassifier

	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
0	1	1	1	1	1	5	7	2	1	1	4	1	1	8	1	6	1	3	1

c. SVC

	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
0	4	10	6	2	1	3	7	1	1	14	9	1	1	15	13	11	8	12	5

d. LogisticRegression

	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
0	1	6	1	1	1	4	1	1	1	2	1	1	1	9	10	5	7	8	3

3.2.2 数据训练

根据上面的4中模型，在不进行特征删除的情况下，进行模型训练，查看结果：

a. RandomForestClassifier

	0	1	2
acc_train	0.926667	0.960000	1.000000
acc_val	0.947791	0.955823	0.973896
pred_time	0.005134	0.010436	0.005619
train_time	0.059978	0.111715	0.147623

b. XGBClassifier

	0	1	2
acc_train	0.926667	0.970000	1.000000
acc_val	0.951807	0.969880	0.979920
pred_time	0.000746	0.001212	0.002752
train_time	0.012327	0.040465	0.601335

c. SVC

	0	1	2
acc_train	0.883333	0.976667	0.980000
acc_val	0.845382	0.967871	0.975904
pred_time	0.001426	0.008375	0.011717
train_time	0.005971	0.001498	0.044124

d. LogisticRegression

	0	1	2
acc_train	0.906667	0.950000	0.970000
acc_val	0.881526	0.951807	0.969880
pred_time	0.000273	0.000264	0.000601
train_time	0.000679	0.001360	0.013367

3.2.3 特征删除

根据上图需要将数据集中对学习模型重要程度较低的特征进行删除，即需要删除所有值大于1的特征，整理了一下如下：

a. RandomForestClassifier

median, Q75, skew, kurt, minfun, maxfun, meandom, mindom, dfrange, modindx

b. XGBClassifier

Skew,kurt,sp.ent,centroid,maxfun,mindom,dfrange

c. SVC

sd, median, Q25, Q75, skew, kurt, mode, centroid, maxfun, meandom, mindom, maxdom, dfrange, modindx

d. LogisticRegression

median, skew, mode, maxfun, meandom, mindom, maxdom, dfrange, modindx

3.2.4 模型的训练过程

删除特征后获得的不同学习模型对应的含有不同特征的数据集。对这些数据集进行和前述中相同的拆分工作，即将数据集按一定比例拆分为训练集、验证集和测试集。拆分时使用相同的随机因子，以保持不同应用于不同算法的数据集样本一致。

将数据集应用于 上面4 种学习模型的训练。获得的在训练集和验证集上的准确率(accuracy)以及训练和预测所花费的时间如下图。

a. RandomForestClassifier

	0	1	2
acc_train	0.936667	0.973333	1.000000
acc_val	0.963855	0.967871	0.975904
pred_time	0.004282	0.008083	0.040591
train_time	0.047669	0.053222	0.106707

b. XGBClassifier

	0	1	2
acc_train	0.926667	0.970000	1.000000
acc_val	0.951807	0.971888	0.979920
pred_time	0.001025	0.001487	0.006444
train_time	0.011136	0.031632	0.300668

c. SVC

	0	1	2
acc_train	0.843333	0.976667	0.970000
acc_val	0.823293	0.965863	0.973896
pred_time	0.001897	0.003477	0.003335
train_time	0.002661	0.001177	0.015097

d. LogisticRegression

	0	1	2
acc_train	0.906667	0.953333	0.973333
acc_val	0.891566	0.959839	0.969880
pred_time	0.000166	0.000160	0.000786
train_time	0.000529	0.000836	0.012776

4个学习模型对于100%的训练数据的序号是2，在验证集上的预测率都超过0.95。xgboost模型所需的训练时间要较其他模型更长，但总体来说，由于数据集较小，所需的时间均较短。SVC和 LogisticRegression 模型在训练集和验证集上的表现相近。而 RandomForestClassifier和XGBClassifier模型在训练集上准确率是1，而在验证集上分别为0.975904和 0.979920，这有可能是发生了过拟合。

特征删除前与特征删除后的数据对比：

	删除前		删除后	
RandomForestClassifier	acc_train	1.000000	acc_train	1.000000
	acc_val	0.973896	acc_val	0.975904
XGBClassifier	acc_train	1.000000	acc_train	1.000000
	acc_val	0.979920	acc_val	0.979920
SVC	acc_train	0.980000	acc_train	0.970000
	acc_val	0.975904	acc_val	0.973896
LogisticRegression	acc_train	0.970000	acc_train	0.973333
	acc_val	0.969880	acc_val	0.969880

由图可知，总体来说删除前和删除后变化不是很大，XGBClassifier 基本没有发生变化，SVC 预测率怎变低了，而 RandomForestClassifier 和 LogisticRegression 稍微有所提升，说明特征删选的时候对于不同的模型可能存在着误删除。

3.3 模型优化

这里仅对RandomForestClassifier和XGBClassifier进行模型优化，选择GridSearchCV进行网格优化。具体步骤如下：

a. 确定RandomForestClassifier需要寻优的参数为min_samples_split和n_estimators，即最小可拆分样本数和树的个数，默认为 2 和 10，寻优空间定为{'n_estimators' : range(4,14,2), 'min_samples_split' : range(2,7,1)}。

b. 确定XGBClassifier需要寻优的参数为max_depth、n_estimators、learning_rate，参照[10]，设置可优化参数：{'max_depth':range(2, 7), 'n_estimators':range(100, 1100, 200), 'learning_rate':[0.05, 0.1, 0.25, 0.5, 1.0]}

c. 利用 GridSearchCV 方法来构建网格搜索模型，并使用默认参数。采用交叉验证打分的方式获得优选模型。

优化后RandomForestClassifier的最优参数及结果是：

```
RandomForestClassifier
Unoptimized model
Accuracy score on validation data: 0.9759

Optimized Model
Final accuracy score on the validation data: 0.9799
Parameter setting that gave the best results on the hold out data:
```

	min_samples_split	n_estimators
0	5	12

优化后XGBClassifier的最优参数及结果是：

```
XGBClassifier
Unoptimized model
Accuracy score on validation data: 0.9799

Optimized Model
Final accuracy score on the validation data: 0.9880
Parameter setting that gave the best results on the hold out data:
```

	learning_rate	max_depth	n_estimators
0	1.0	2	300

4. 结果

4.1 模型的评估与验证

将训练集随机拆分出 1%，10%，100%(100%就是原训练数据集)的数据集训练 4 种 学习模型，然后对测试集的数据进行预测，SVC 和 LogisticRegression采用未优化的默认参数，RandomForestClassifier 和 XGBClassifier采用调优后的参数，获得的准确率如下：

表4 各学习模型(对应 3 个不同大小的训练集)的测试准确率

	1%	10%	100%
RandomForestClassifier	0.945425	0.963082	0.982343
XGBClassifier	0.942215	0.966292	0.980738
SVC	0.820225	0.966292	0.971108
LogisticRegression	0.879615	0.948636	0.966292

可以看出，XGBClassifier和 RandomForestClassifier 使用 1%的训练集训练的 结果已经有了较高的准确率，后续使用了 10%和 100%的训练集训练的结果仅仅提升了 几个百分点。这说明，这两种学习模型对数据集的改变敏感性较小。

而对 SVC 和 LogisticRegression 两种学习模型而言，使用 1%的数据集仅仅获得了 85%左右的准确率，这显然是不够好的。当使用 10%的训练集进行训练后，准确率上升 到了 95%左右，即使使用 100%的训练集进行训练，准确率仅提升 1~2 个百分点，这说 明这两种学习器对数据量的有一定的需求，但是，当数据量达到一定程度后(如本项目中的10%的训练集)，也开始变得对数据集的改变敏感性较低。

综上，这几种模型在数据量达到一定量后的稳健性较好。XGBClassifier和RandomForestClassifier相比于SVC和LogisticRegression模型训练效果更容易趋于稳定。

4.2 与基准模型进行比较

根据上表4可知，对于100%的训练集本项目采用的4个机器学习模型均获得了超过0.95的准确率，其中随机森林模型的准确率达到了0.982343。而根据2.3节可知 accuracy(base) = 0.53，可知本项目所采用的学习模型所获得的准确率均超过基准模型。

本项目所采用的4个模型的准确率均达到0.95以上，算是一个比较好的结果，基本上解决了根据语音特征确定说话者性别这一问题。

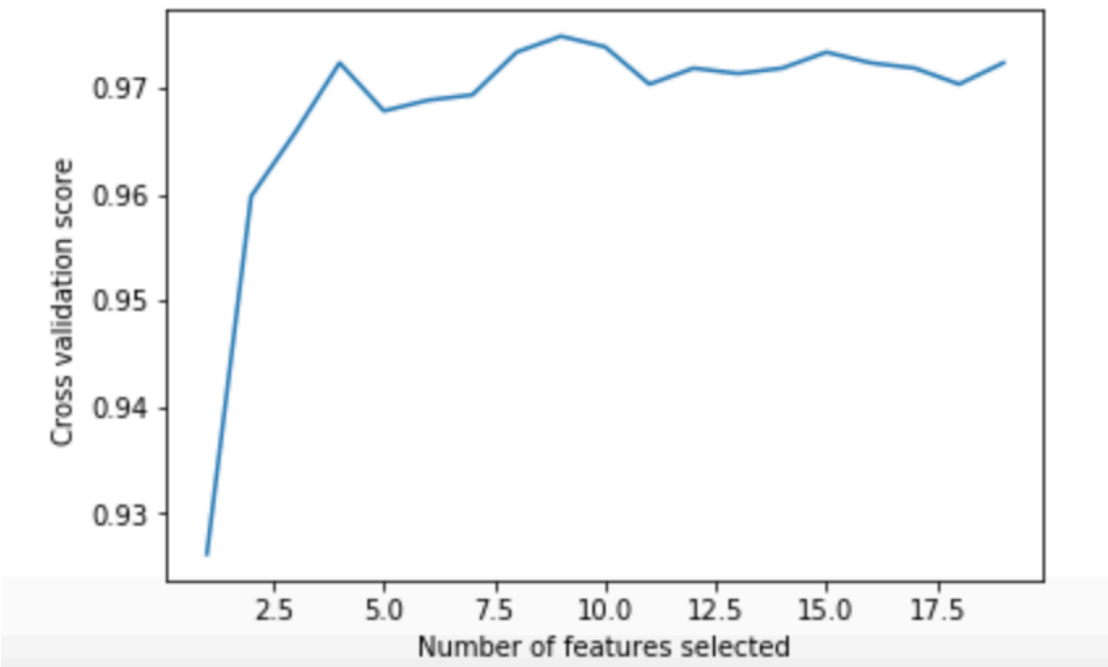
5. 结论

5.1 自由形态的可视化

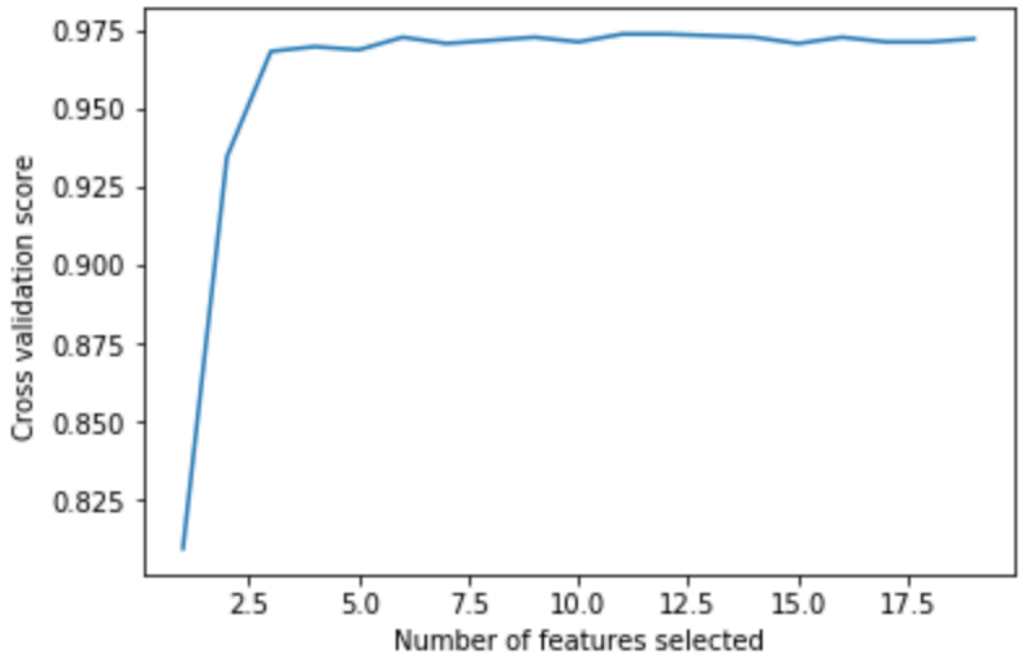
5.1.1 特征的筛选

使用上面4种不同学习模型时RFECV得分与特征选择个数之间的关系曲线如下图：

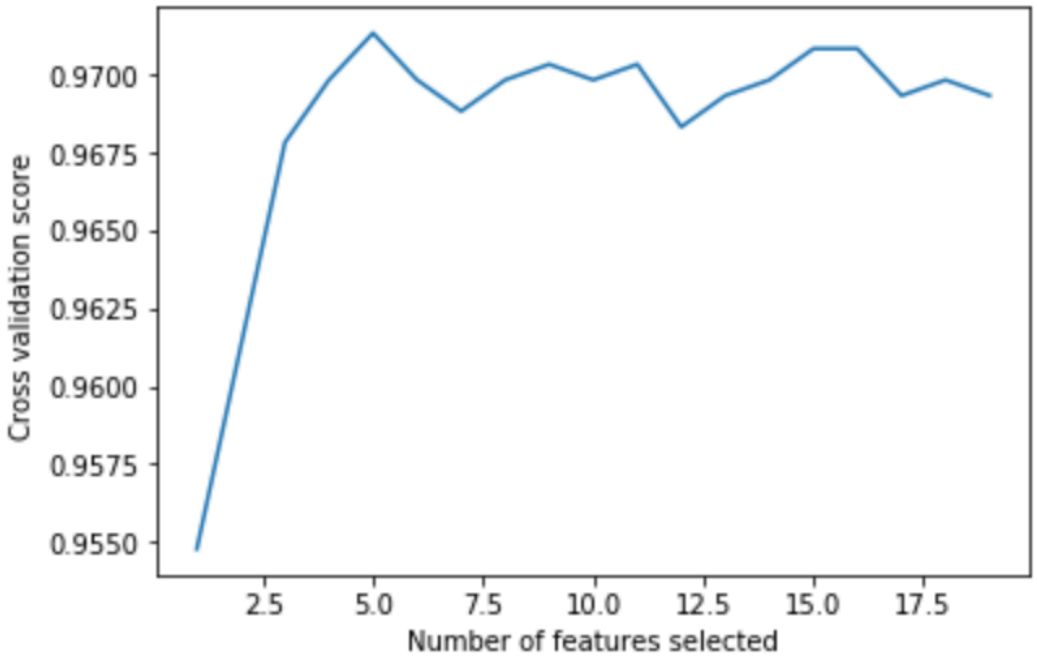
a. RandomForestClassifier



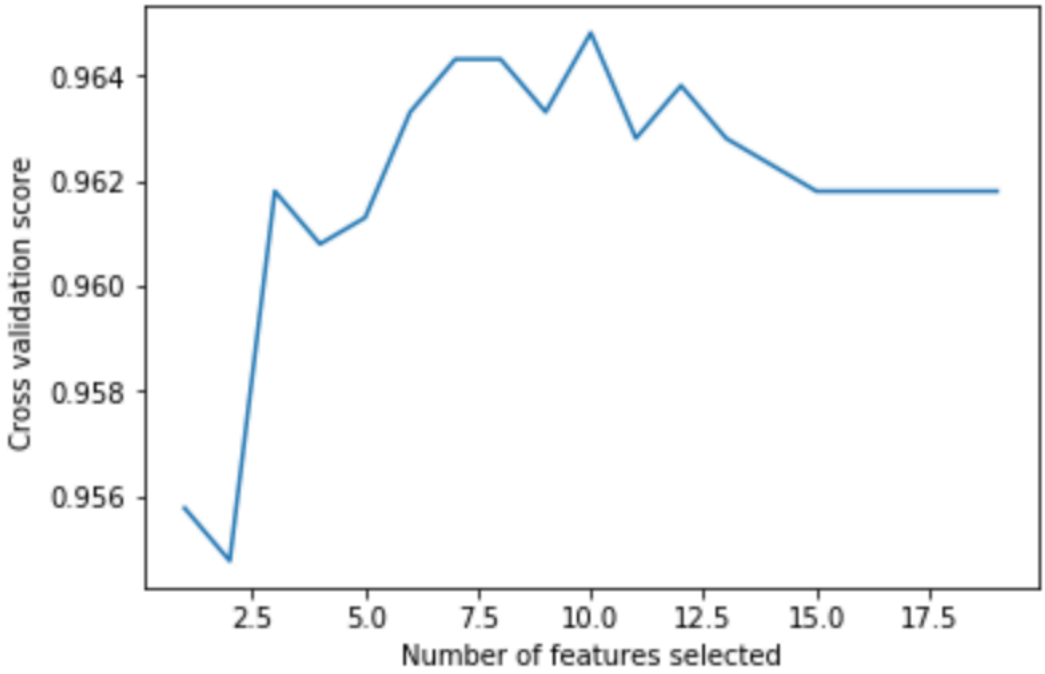
b. XGBClassifier



c. SVC



d. LogisticRegression



一般情况下，特征越多，模型越复杂，精度越高，模型泛化能力越差;特征越少，模型越简洁，精度越低，模型泛化能力越强;交叉验证打分正好可以评判特征数量变化情况下模型的优劣，来确定最佳的特征个数。

从图以看出，随着特征选择数量的增加，RFECV得分会先增大，然后再趋于水平或下降，这说明特征的选择并非越多越好，而是有一优选个数。其中RandomForestClassifier模型最优特征个数为9，XGBClassifier模型最优特征个数为12，SVC模型最优特征个数为5，LogisticRegression模型最优特征个数为10。

采用sklearn.feature_selection.RFECV 计算不同学习器情况下，特征的重要程度；其中 RFECV 采用的参数是(step=1, cv=StratifiedKFold(2),scoring='accuracy')

各模型特征重要性排序表：

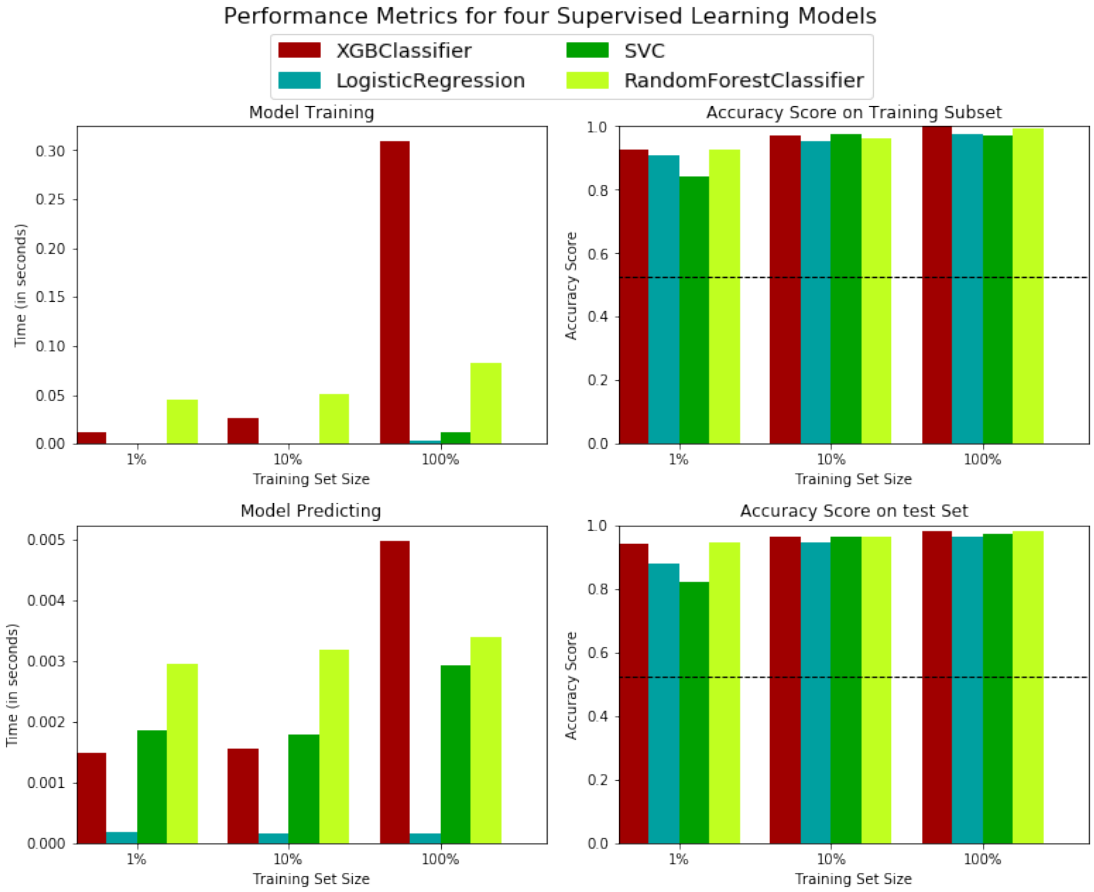
	RandomForestClassifier	**XGBClassifier**	**SVC**	**LogisticRegression**
sd	1	1	4	1
median	2	1	10	6
Q25	1	1	6	1
Q75	8	1	2	1
IQR	1	1	1	1
skew	3	5	3	4
kurt	7	7	7	1
sp.ent	1	2	1	1
sfm	1	1	1	1
mode	1	1	14	2
centroid	1	4	9	1
meanfun	1	1	1	1
minfun	6	1	1	1
maxfun	10	8	15	9

meandom	4	1	13	10
mindom	9	6	11	5
maxdom	1	1	8	7
dfrange	11	3	12	8
modindx	5	1	5	3

对重要程度较低的特征进行删除后保留程度较高的。

5.1.2 模型训练与测试的准确率与时间

上面4种模型在1%，10%，100%训练集上的训练时长和准确率，以及模型在测试集上进行预测的时长和正确率如下图：



其中虚线为基准模型确定的准确率 0.53。

从图中可以看出，几种模型的训练和预测均耗费的时间较短XGBClassifier消费时间相对较多。几种模型在测试集上的准确率略有差异，均超过基准模型，达到0.95左右，其中随机森林模型的准确率最高为0.982343，XGBClassifier次之，支持向量机模型的准确率再次，逻辑回归模型模型最差。

5.2 思考

整个项目是通过语音特征信息作为输入，运用机器学习模型，判断语音说话者的性 别。至少完成了以下几部分的工作:

a) 完成了对数据集的探索，对数据集的特征以及特征的统计量进行了说明;

b) 完成了对数据集的预处理，包括删除重复样本以及样本数据集的归一化;

c) 针对不同机器学习模型对特征进行筛选，使用筛选后的数据集对模型完成训练和验证;

d) 对决随机森林模型和XGBoost进行参数优化;

e) 在测试集上对学习模型进行测试。

特征选择部分算是一个比较困难的地方，因为进行特征选择的方法就比较多。本项目选择了更易理解的包裹式方法，采用的是sklearn.feature_selection，RFECV方法，使用了默认参数。其实，就REFCV方法而言，参数设置不同，对特征选择的结果也不同，也会影响最终模型的训练、预测和验证。综合这一系列影响因素，想要获得一个最优的结果并不容易，本项目选择的是一个相对来说较简单的方式来完成的。

最终四个模型都获得了不错的准确率，均达到0.95以上，算是一个比较好的结果， 基本解决了根据语音特征确定语音说话者性别这一问题。在该问题的表现上，其中随机森林模型的准确率最高为0.982343，XGBoost次之，支持向量机模型的准确率再次，逻辑回归模型模型最差。

5.3 改进

本项目的改进可以从以下几方面考虑:

a) 在特征选择部分，尝试使用嵌入型的特征选择方法。嵌入式的特征选择方法能够选择出更符合学习模型的特征，应该能获得不错的训练和预测效果;

b) 尝试除本项目以外的学习模型，如神经网络、决策树等;

c) 扩大网格搜索模型参数的维度和个数;

本项目中的模型均获得了较高的准确率，采用上述3个改进方式，只是提供了优化的方向。最终获得的模型不一定能获得比目前本项目中模型更好的效果，机器配置限制了优化的速度。

6. 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

7. 优达学城监督学习项目.为 CharityML寻找捐献者.

8. https://github.com/nd009/capstone/tree/master/Gender_Recognition_by_Voice

9. https://blog.csdn.net/a1b2c3d4123456/article/details/52849091

10. https://blog.csdn.net/u013421629/article/details/78642834

11. https://blog.csdn.net/sb19931201/article/details/52557382

12. https://blog.csdn.net/china1000/article/details/51106856

13. https://blog.csdn.net/xiaocong1990/article/details/70230801