语音性别识别

机器学习纳米学位毕业项目

陈晓杰
2018年9月

I. 定义

项目概览

随着人工智能算法的发展，大家都希望从各处获取、提取和处理数据，以获得更多有价值的信息。语音数据的提取和处理则是其中一个重要的方向。近年来出现了许多语音识别的应用方案，现已影响着我们的日常生活，如iPhone的siri语音助手、各类语音输入法、声纹锁等等。

说话人识别[1]是一项利用声音特征来识别人的语音识别技术。该项技术的历史可以追溯到四十年前，并在当时已经发现了不同个体间的语音特征的区别。在本项目中，需要解决的是根据语音特征来识别性别，是说话人识别的一种特例，因为最终的只需要分类成男性或者女性。

在机器学习的领域中，分类是它的一项主要的应用。在此项目中，数据集是已经从音频信号中提取出的特征和对应的分类标识组成的，因此可以利用监督学习[2]的模型解决此问题，例如逻辑回归[3]、决策树[4]、随机森林[5]、SVM[6]、神经网络[7]、GBDT[8]和XGBoost[9]等算法。

问题说明

该项目解决的是一个音频分类的问题，使用机器学习的方，判断一段音频信号是由男性还是女性发出的。该项目使用的数据中，每一条数据都由若干个音频特征和一个分类标识组成，所以本质上该项目解决的是一个监督学习的分类问题，并且是一个典型的二分类问题。

指标

对于该项目的二分类问题，输出的标签是不同的性别，在识别性别问题的角度上看，正确地识别出两种性别同等重要。因此，适用的评估标准是分类准确率(Accuracy)。

- 准确率(Accuracy)
  准确率可以直观地体现出分类器的性能，准确率越接近1，性能越好。准确率的定义是正确分类出来的样本数占样本总数目的比率：

$$Accuracy = \frac{\sum_{i=1}^{n} I\left(y_i = \hat{y}_i\right)}{n}$$

y_i 表示第 i 个样本的标签，$\hat{y}_i$ 表示 i 个样本的预测值，如果 y_i 和 $\hat{y}_i$ 相等，$I=1$，否则 $I=0$。

II. 分析

数据研究

本项目中的数据集来自KORY BECKER在16年6月的语音性别识别项目[10]，共有3168个样本。样本数据是由音频文件解析的，这些音频文件来自男性和女性发言者。通过运用R语言的seewave和tuneR的包对语音样本进行了预处理[11]，分析频率范围为0hz-280hz（人类声音范围）。

如表1所示，该数据集有21个特征，分别是 “meanfreq”, “sd”, “median”, “Q25”, “Q75”, “IQR”, “skew”, “kurt”,”sp.ent”, “sfm”, “mode”, “centroid”, “meanfun”, “minfun”, “maxfun”, “meandom”, “mindom”, “maxdom”, “dfrange”, “modindx” 和 “label”。前20列作为输入特征，都是连续型的数值，最后一列lebel是性别的标识，为 “male” 或 “female”。

Classification	Feature names
Input	meanfreq, sd, median, Q25, Q75, IQR, skew, kurt, sp.ent, sfm, mode, centroid, meanfun, minfun, maxfun, meandom, mindom, maxdom, dfrange, modindx
Output	label

表1. 数据集特征

每个输入特征代表音频信号在频域内的分布情况，都是浮点型数值，详细信息如下[10]：

- meanfreq : 频率平均值
- sd : 频率标准差
- median : 频率中位数
- Q25 : 频率第一四分位数
- Q75 : 频率第三四分位数
- IQR : 频率四分位数间距
- skew : 频谱偏度[12]，体现音频信号频谱的对称性
- kurt : 频谱峰度[13]，体现音频信号频谱的峰部的尖度
- sp.ent : 频谱熵
- sfm : 频谱平坦度[14]
- mode : 频率众数
- centroid : 频谱质心[15]
- 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

输出用 “male” 或 “female” 分别表示男性和女性：

- label : 性别标识，两种分类(“male” 和 “female”)

表2展示的是数据集中前5个样本的特征详情，输入的特征都是浮点型的数值，通过用pandas工具统计没有发现数据缺失。

	Sample 0	Sample 1	Sample 2	Sample 3	Sample 4
meanfreq	0.059781	0.0660087	0.0773155	0.151228	0.13512
sd	0.0642413	0.06731	0.0838294	0.0721106	0.0791461
median	0.0320269	0.0402287	0.0367185	0.158011	0.124656
Q25	0.0150715	0.0194139	0.00870106	0.0965817	0.0787202
Q75	0.0901934	0.0926662	0.131908	0.207955	0.206045
IQR	0.075122	0.0732523	0.123207	0.111374	0.127325
skew	12.8635	22.4233	30.7572	1.23283	1.10117
kurt	274.403	634.614	1024.93	4.1773	4.33371
sp.ent	0.893369	0.892193	0.846389	0.963322	0.971955
sfm	0.491918	0.513724	0.478905	0.727232	0.783568
mode	0.0	0.0	0.0	0.0838782	0.104261
centroid	0.059781	0.0660087	0.0773155	0.151228	0.13512
meanfun	0.0842791	0.107937	0.0987063	0.0889648	0.106398
minfun	0.0157017	0.0158259	0.0156556	0.0177976	0.0169312
maxfun	0.275862	0.25	0.271186	0.25	0.266667
meandom	0.0078125	0.00901442	0.00799006	0.201497	0.712812
mindom	0.0078125	0.0078125	0.0078125	0.0078125	0.0078125
maxdom	0.0078125	0.0546875	0.015625	0.5625	5.48438
dfrange	0.0	0.046875	0.0078125	0.554688	5.47656
modindx	0.0	0.0526316	0.0465116	0.247119	0.208274
label	male	male	male	male	male

表2. 样本数据

	count	mean	std	min	25%	50%	75%	max
meanfreq	3168	0.180907	0.029918	0.039363	0.163662	0.184838	0.199146	0.251124
sd	3168	0.057126	0.016652	0.018363	0.041954	0.059155	0.067020	0.115273
median	3168	0.185621	0.036360	0.010975	0.169593	0.190032	0.210618	0.261224
Q25	3168	0.140456	0.048680	0.000229	0.111087	0.140286	0.175939	0.247347
Q75	3168	0.224765	0.023639	0.042946	0.208747	0.225684	0.243660	0.273469
IQR	3168	0.084309	0.042783	0.014558	0.042560	0.094280	0.114175	0.252225
skew	3168	3.140168	4.240529	0.141735	1.649569	2.197101	2.931694	34.72545
kurt	3168	36.56846	134.9286	2.068455	5.669547	8.318463	13.64890	1309.612
sp.ent	3168	0.895127	0.044980	0.738651	0.861811	0.901767	0.928713	0.981997
sfm	3168	0.408216	0.177521	0.036876	0.258041	0.396335	0.533676	0.842936
mode	3168	0.165282	0.077203	0.000000	0.118016	0.186599	0.221104	0.280000
centroid	3168	0.180907	0.029918	0.039363	0.163662	0.184838	0.199146	0.251124
meanfun	3168	0.142807	0.032304	0.055565	0.116998	0.140519	0.169581	0.237636
minfun	3168	0.036802	0.019220	0.009775	0.018223	0.046110	0.047904	0.204082
maxfun	3168	0.258842	0.030077	0.103093	0.253968	0.271186	0.277457	0.279114
meandom	3168	0.829211	0.525205	0.007812	0.419828	0.765795	1.177166	2.957682
mindom	3168	0.052647	0.063299	0.004883	0.007812	0.023438	0.070312	0.458984
maxdom	3168	5.047277	3.521157	0.007812	2.070312	4.992188	7.007812	21.86718
dfrange	3168	4.994630	3.520039	0.000000	2.044922	4.945312	6.992188	21.84375
modindx	3168	0.173752	0.119454	0.000000	0.099766	0.139357	0.209183	0.932374

表3. 各列数据统计详情

从表3中的各列数据的统计特性可以看出，”skew”, “kurt”, “maxdom” 和 “dfrange” 的值比其他特征的数量级大，如果用到一些对特征数值大小敏感的模型(如KNN、SVM和LR等)，则需要对其进行归一化处理。

探索性可视化

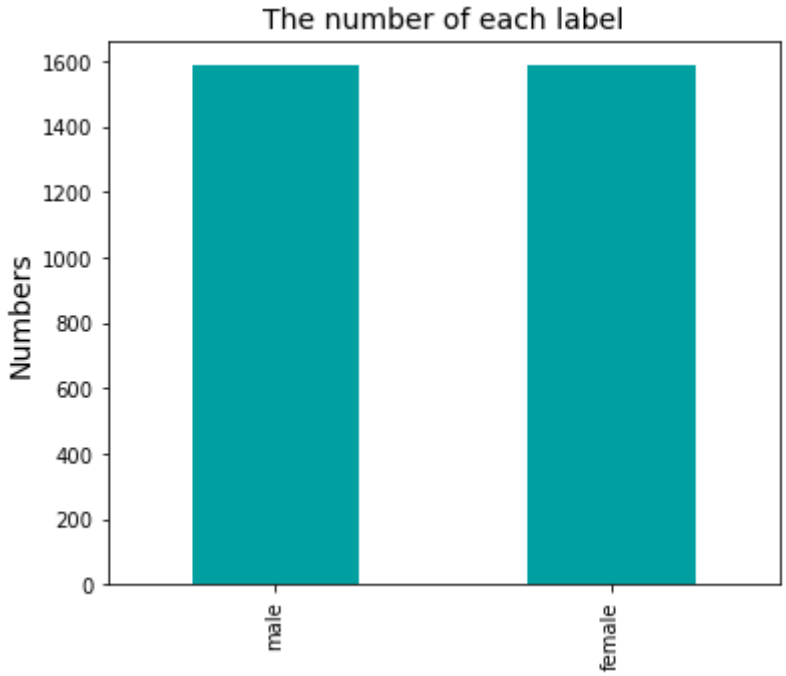

图1. 各标签的样本数

图1展示了数据集3168个样本中，标签为 “male” 和 “female” 的数量。经统计，男性和女性的样本数均为1584，各占总数据的50%。

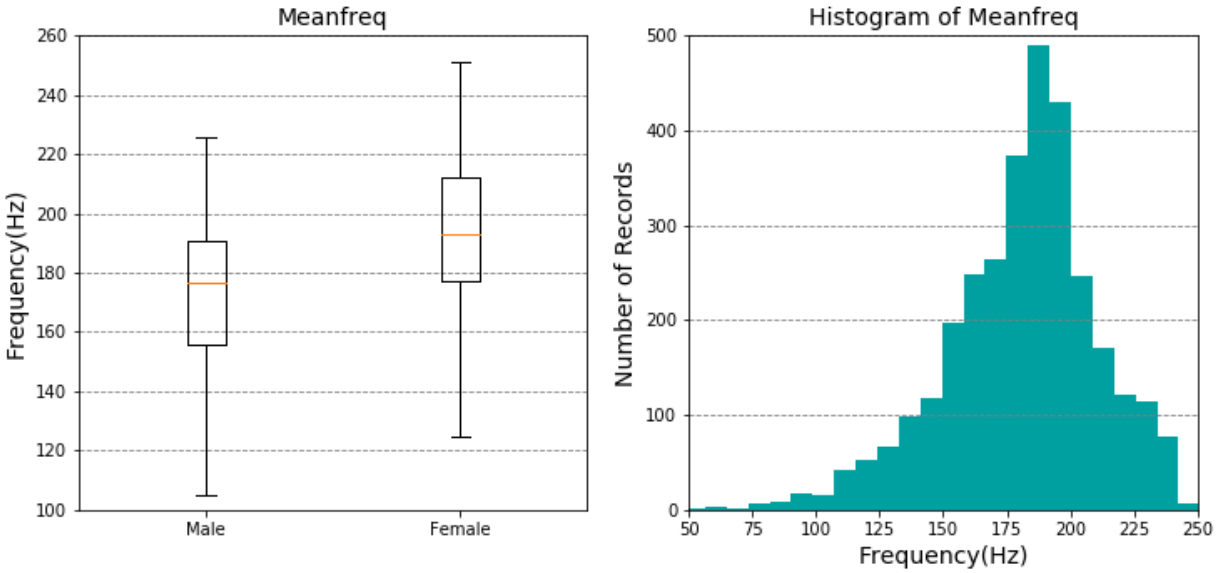

图2. 频率平均值分布

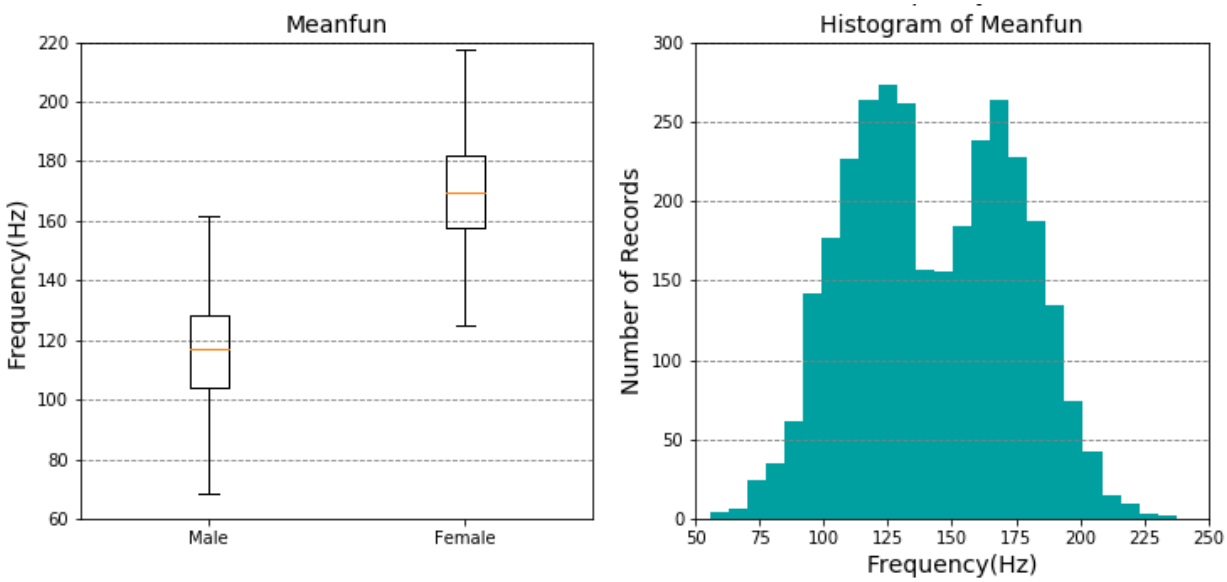

图3. 平均基音频率分布

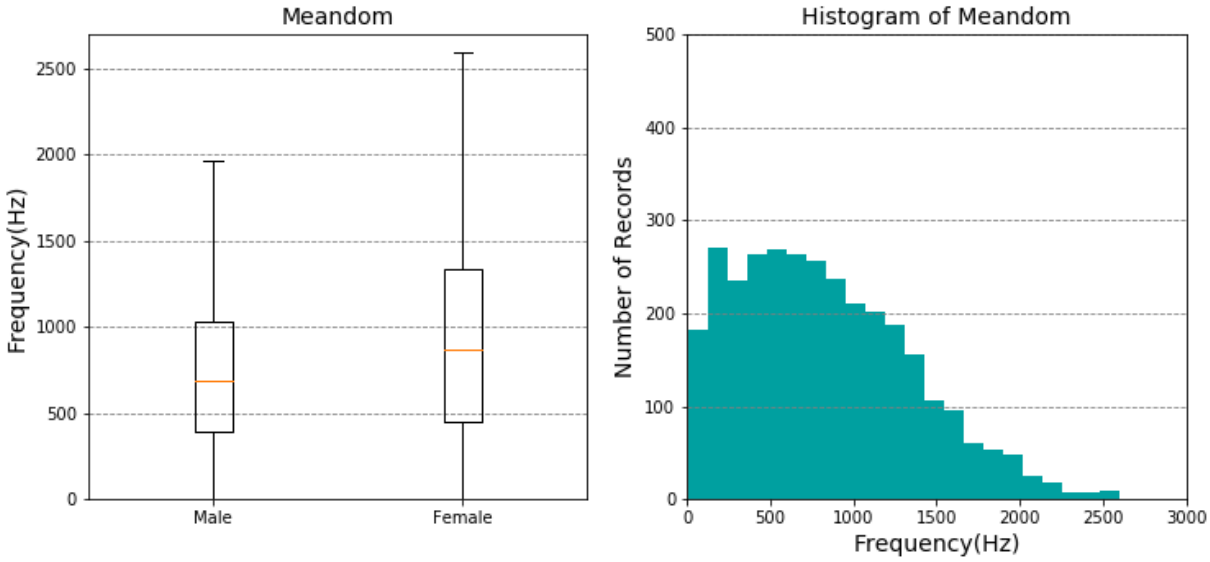

图4. 平均主频分布

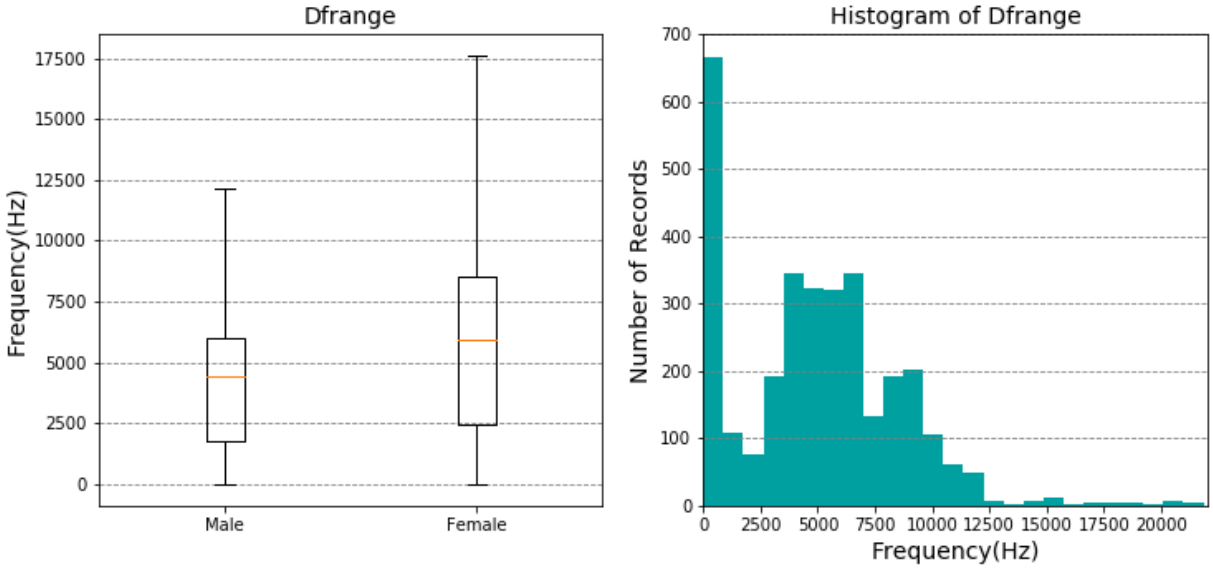

图5. 主频范围分布

图2、图3、图4和图5分别展示了不同性别的语音样本的频率平均值、平均基音频率、平均主频和主频范围的分布，绘制箱线图时去除了异常值的显示。这四幅图直观地展示出男性和女性语音的频率相关特征的分布情况，可以看到女性的语音频率相关特征的数值大于男性的，与我们的日常认知（女性讲话的声音频率比男性的高）相符。
此外，在图3平均基音频分布中，从箱线图可以看出男女的分布边界在这四个特征中最为清晰，同时从柱状图的双峰的分布，也可以明显看出不同性别分布的分割点，约在150Hz。所以，平均主频将是样本分类的关键特征。

算法与方法

随机森林[16]是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习（Ensemble Learning）方法。随机森林是一种很灵活实用的方法，它有如下几个特点：

- 在当前所有算法中，具有极好的准确率
- 能够有效地运行在大数据集上
- 能够处理具有高维特征的输入样本，而且不需要降维

- 能够评估各个特征在分类问题上的重要性
- 在生成过程中，能够获取到内部生成误差的一种无偏估计
- 对于缺省值问题也能够获得很好得结果
- ……

基于随机森林算法的广泛适应和良好表现，本项目采用随机森林的模型对数据集进行学习和分类，识别语音的性别。项目的工作由以下几部分组成：

1. 数据准备：从kaggel[11]下载数据集，探索数据并将其统计性质可视化。
2. 数据预处理：数据集标签'label'的数值化、分离特征和标签、对极大值极小值分布差值过大的特征施加对数转换、全体特征归一化、数据混洗和切分。
3. 创建分类器：采用sklearn的随机森林模型RandomForestClassifier创建一个分类器。
4. 训练分类器：用训练集对分类器进行训练。
5. 参数调整：用一种贪心的坐标下降法[17]进行超参调整，调节的参数和范围分别为：n_estimators(1-80，步进5)、max_features(3-8，步进1)、min_samples_split(2-8，步进1)、criterion(gini和entropy)。
6. 将参数调优后的RF分类器对测试集的数据进行语音性别识别，计算其准确率，对其预测效果进行分析。

此外，将尝试使用模型融合stacking[18]方法，搭建两层模型，第一层采用参数调优后的SVM、GBDT、KNN和决策树模型，第二层用逻辑回归得到最终的预测结果并计算准确率。

基准测试

根据KORY BECKER的项目，随机森林模型在训练集和测试集的准确率分别达到100%和98%[10]。本项目的目标是争取采用随机森林模型，让准确率接近98%，通过模型融合方法，尝试得到更好的准确率。

III. 方法

数据预处理

通过对数据的研究，本项目需要做的预处理如下：

1. 缺失值处理：通过pandas工具查看数据集的信息，没有发现缺失值，所以不需要额外处理。
2. 标签数值化：数据集中的标签 'label' 非数值特征，需要将其数值化，具体做法是将 'male' 和 'female' 分别用1和0替代。
3. 分离特征和标签：从数据集中分离出20列特征和1列标签。
4. 施加对数转换：'skew'和'kurt'两个特征的数值范围比其他特征大很多，且分布极大值和极小值的差距很大，因此对这两个特征施加了对数转换，使得特征数据的分布接近正态。
5. 归一化："skew", "kurt", "maxdom" 和 "dfrange" 的值比其他特征的数量级大，因此需要对全体特征进行归一化处理。
6. 数据混洗和切分：使用sklearn.model_selection.train_test_split对数据集进行随机的训练集和测试集划分。

实施

在此我将分别用随机森林模型和stacking方法实施性别语音分类。

1. 随机森林
   运用一种坐标下降法分别对n_estimators、max_features、min_samples_split和criterion进行参数调节，具体做法是贪心地对参数逐个搜索，由于随机森林的抽样和特征选取本身具有随机性，所以对具体参数的每一个数值都训练10次，然后计算出模型在测试集得到的10次准确率的均值，作为对应的参数数值的得分，最后，对应得分最高的参数数值作为此参数搜索的最优结果。

   - n_estimators(1-80，步进5)

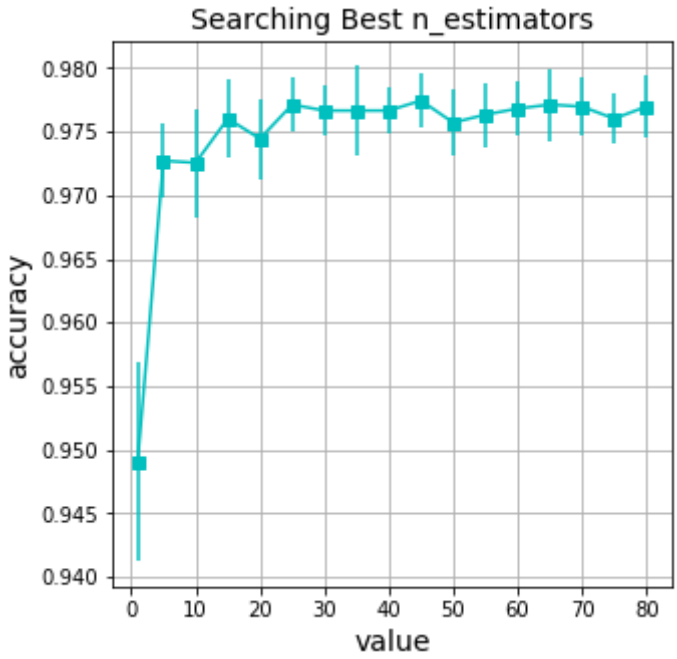

图6. n_estimators搜索

从图6中可以看出，当n_estimators大于20之后准确率达到稳定，在取值为45的时候均值最大，所以n_estimators的搜索结果为45。

   - max_features(3-8，步进1)

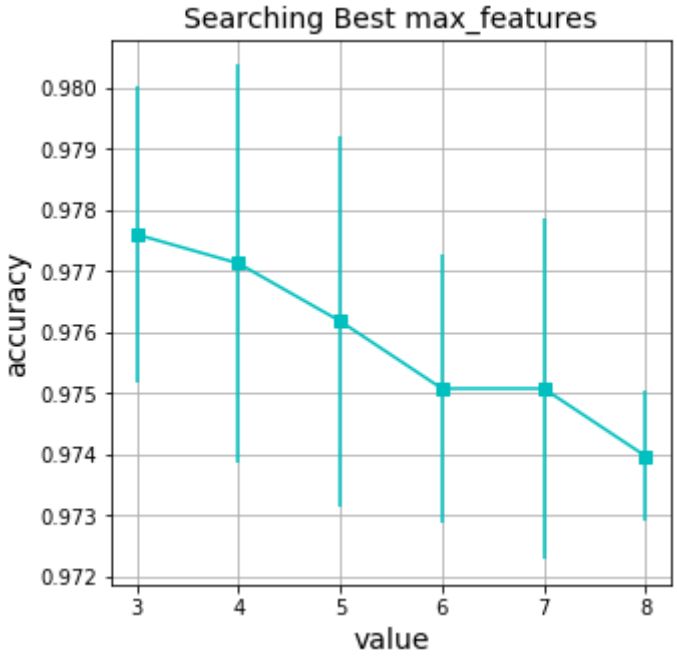

图7. max_features搜索

运用贪心方法，先将n_estimators设为45，再对max_features搜索。从图7中可以看出，max_features在取值范围内准确率呈递减趋势，因此max_features搜索结果为3。

- min_samples_split(2-8，步进1)

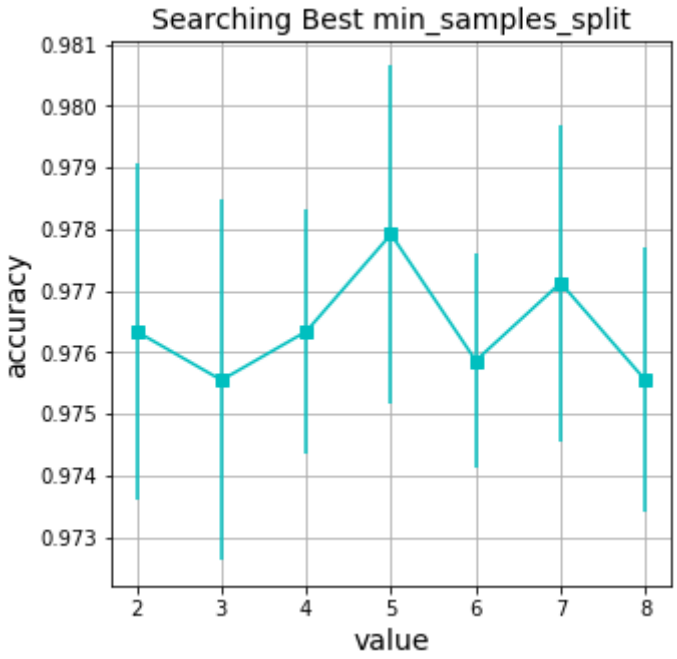

图8. min_samples_split搜索

同上，先将n_estimators设为45，max_features设为3，再对min_samples_split搜索。从图8中可以看出，min_samples_split在取值为5的时候准确率最高。

- criterion(gini和entropy)。
  将n_estimators设为45，max_features设为3，min_samples_split设为5，分别让

criterion取gini和entropy，比较他们的平均准确率，结果分别为0.97735和0.97893，所以criterion取entropy。

2. stacking
   首先分别用GridSearchCV对SVM(C和gamma)、GDBT(n_estimators、max_features和max_depth)、KNN(n_neighbors)和DT(min_samples_split和max_features)进行参数搜索，得到调优后的最佳模型，并将他们作为模型融合的第一层，第二层用训练后的LR的预测作为最终输出，最终获得0.98107的准确率。

改进

在最初随机森林的参数搜索时，在对每个参数值进行得分评估时，用了测试集的数据。为了防止测试集的数据影响到调参，最终将方案进行了改进，采用随机森林的袋外得分替代了之前测试集的得分，最终调优的参数分别为：n_estimators=50、max_features=5、 min_samples_split=4、criterion=’entropy’，在测试集得到的准确率为0.97744。

IV. 结果

模型评估与验证

Models	Random Forest	SVM	GDBT	KNN	DT
Untuned parameters	n_estimators=10 max_features='auto' min_samples_split=2 criterion=gini	C=1.0 gamma='auto'	n_estimators=100 max_features=None max_depth=3	n_neighbors=5	max_features=None min_samples_split=2
Tuned parameters	n_estimators=50 max_features=5 min_samples_split=4 criterion='entropy'	C=5000.0 gamma=0.005	n_estimators=200 max_features=5 max_depth=4	n_neighbors=6	max_features=11 min_samples_split=4

表4. 模型参数调整

Models	Random Forest	SVM	GDBT	KNN	DT
Accuracy(Untuned)	0.97391	0.97003	0.97318	0.98107	0.96056
Accuracy(Tuned)	0.97744	0.97160	0.97949	0.98422	0.96214

表5. 模型准确率

表4中展示的是项目中用到的各模型的参数调整详情，表5展示了各模型参数调整前后的准确率，在各个单独的模型中，经过参数调整后的准确率都有所提升。这五个单独模型的在测试集中的表现排名是：KNN>GDBT>Random Forest>SVM>DT。项目中重点关注的Random Forest在测试集的准确率为0.97744，约为98%，达到了之前给出的基准。

	SVM	GDBT	KNN	DT	LR	Accuracy
Parameters	C=5000.0 gamma=0.005	n_estimators=200 max_features=5 max_depth=4	n_neigh bors=6	max_features=11 min_samples_split= 4	default	0.98107
	C=1.0 gamma='auto'	n_estimators=100 max_features=None max_depth=3	n_neigh bors=5	max_features=None min_samples_split= 2	default	0.98433

表6. stacking准确率

如表6所示，融合模型stacking的准确率超过了98%，在测试集的表现上超过了除KNN以外的单一模型。

理由

经过参数调整后，Random Forest在测试集的准确率已接近了98%的基准，成绩几乎和GDBT持平，比SVM和DT的成绩好。而融合模型stacking的准确率则超过了98%，除KNN以外，取得了比其他单一模型更好的成绩。

V. 结论

自由形态的可视化

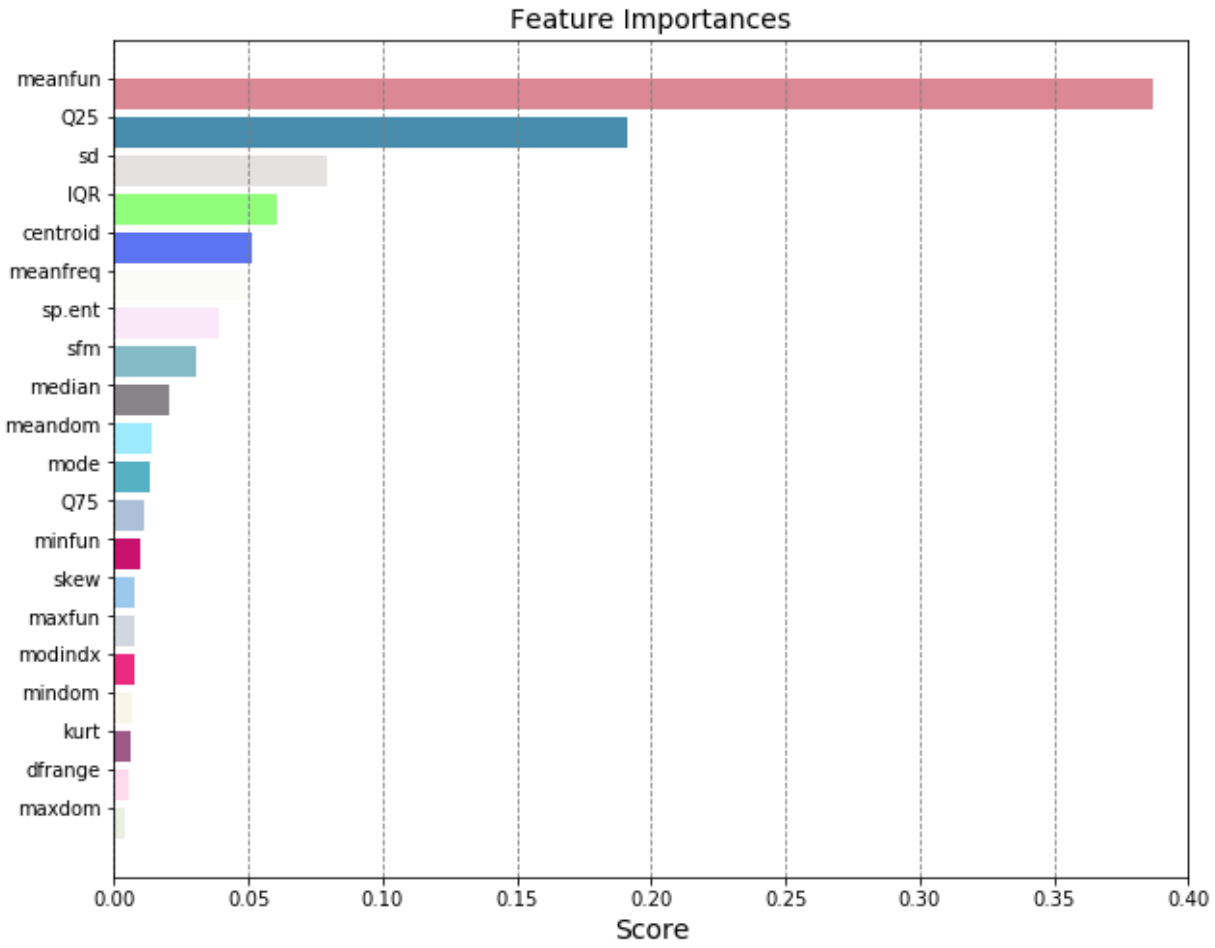

图9. 特征重要性

利用随机森林模型生成的feature_importances_，画出了图9展示的各个输入特征的重要性。从图中可以看出，meanfun平均基音频率的得分最高，而且远高于其他特征的得分，与此前数据研究时粗略分析的特征重要性相吻合，因此它是区分语音是来自男性或女性的最重要的特征。

Features Dimenson	Random Forest	SVM	GDBT	KNN	DT
20	0.97391	0.97003	0.97318	0.98107	0.96056
8(Features Seletion)	0.97160	0.96372	0.97160	0.97476	0.96214

表7. 特征选择的准确率

通过比较图9各特征的得分，选取了 ‘meanfun’, ‘Q25’, ‘sd’, ‘IQR’, ‘centroid’, ‘meanfreq’, ‘sp.ent’ 和 ‘sfm’ 这8个作为输入的特征。表7中展示的是经过特征选择后的各个单一模型的表现，可以看出用前8个特征作为模型输入，在准确率上跟有20个特征的原始数据几乎一致。因此，在性别语音分类这个问题上，这8个特征可以认为已经足够用于取代原样本特征了。

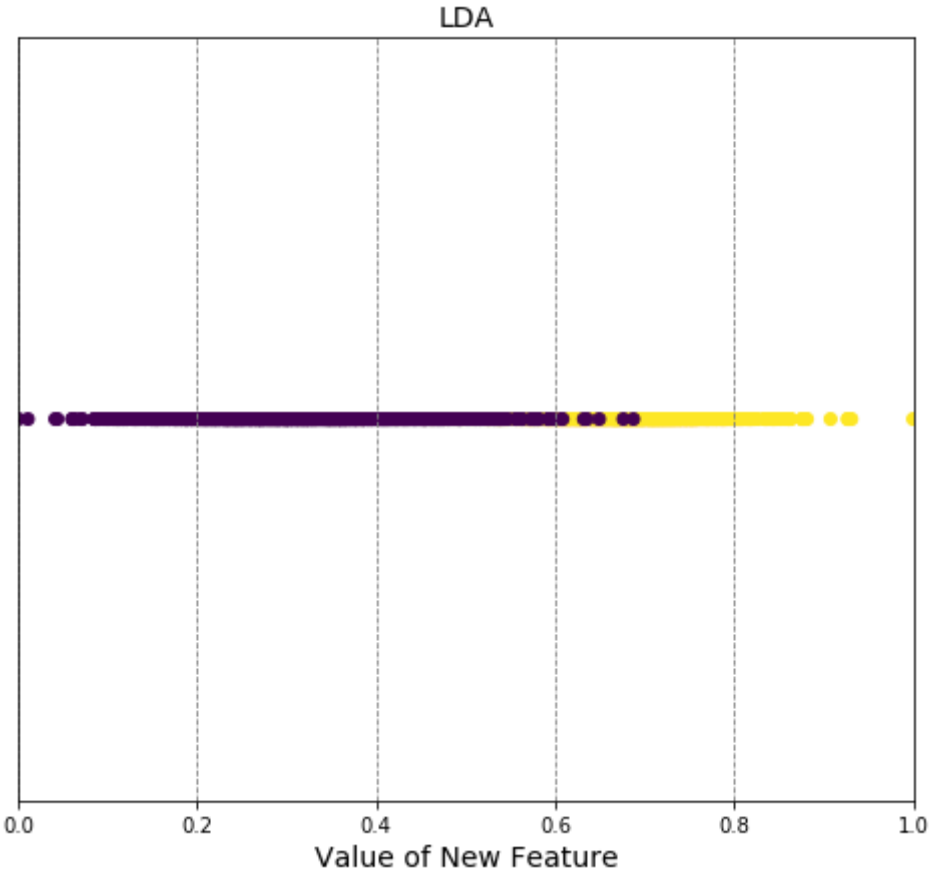

图10. LDA分析

如图10所示，通过用LDA对数据集中的输入特征转换，得到一组一维的输入特征，不同颜色代表着不同的性别，由图中可以看出数据集的分类点在0.6附近。

Features Dimenson	Random Forest	SVM	GDBT	KNN	DT
20	0.97391	0.97003	0.97318	0.98107	0.96056
1(LDA)	0.95583	0.97160	0.95899	0.96529	0.95425

表8. LDA降维后的准确率

表8展示了各个未调参的单独模型在降维数据上的准确率。除了SVM几乎持平，其他模型的表现都略微有所下降。

思考

本项目运用了随机森林和模型融合实现了语音性别分类，经过数据研究、数据预处理、模型训练、模型调参和模型融合，最终得到了合适的模型，并在语音分类上取得了预想的成绩。
在模型融合里有个有趣的现象，模型融合后，精确率接近knn，但却不会超过knn。我认为在融合模型的第一层里，其他模型的准确率没有knn高，因此最终将他“拖累”了。

后续改进

数据特征方面的改进，可以考虑更好的特征选择和特征降维方案，将输入数据的维度降低，加快模型的训练速度，对于参数搜索和svm模型而言，单一模型的训练速度的提高可以大大地提升他们的

效率。模型方面，可以尝试用神经网络模型。因为神经网络善于提取抽象的特征，可能可以得到一个更好的模型。

---

1. https://en.wikipedia.org/wiki/Speaker_recognition ↩
2. https://zh.wikipedia.org/wiki/監督式學習 ↩
3. https://zh.wikipedia.org/wiki/邏輯迴歸 ↩
4. https://zh.wikipedia.org/wiki/决策树 ↩
5. https://zh.wikipedia.org/wiki/随机森林 ↩
6. https://zh.wikipedia.org/wiki/支持向量机 ↩
7. https://zh.wikipedia.org/wiki/人工神经网络 ↩
8. https://en.wikipedia.org/wiki/Gradient_boosting ↩
9. https://en.wikipedia.org/wiki/Xgboost ↩
10. http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/ ↩
11. https://www.kaggle.com/primaryobjects/voicegender ↩
12. https://zh.wikipedia.org/wiki/偏度 ↩
13. https://zh.wikipedia.org/wiki/峰度 ↩
14. https://en.wikipedia.org/wiki/Spectral_flatness ↩
15. https://en.wikipedia.org/wiki/Spectral_centroid ↩
16. https://www.cnblogs.com/liuyihai/p/8309019.html ↩
17. https://www.zhihu.com/question/48282030/answer/114305326 ↩
18. https://blog.csdn.net/MrLevo520/article/details/78161590 ↩