

语音性别识别

1. 定义

1.1 项目概览

本项目是基于 kaggle 上的经过信号处理后的音频数据提取出部分特征，通过建立模型，进行性别的识别，是一个二分类问题。

以分类为目标，对集成学习进行研究，使用的模型是随机森林^[1]和 XGBoost^[2]。随机森林和 XGBoost 都是集成学习的模型，集成学习通过结合多个弱分类器通过某种策略结合在一起，组成一个强分类器。

集成学习分 bagging 和 boosting。Bagging 基于 bootstraping 自主采样方法，重复多次有放回地随机抽出样本进行训练单个模型，再根据这些模型的预测结果经过类似投票的方法得到分类结果。Boosting 则是先从初始训练集训练处一个基学习器，后续的训练样本基于前一次的训练结果进行调整，迭代多次，直到基学习器数目达到设定值，最终将这些基学习器进行加权结合。

项目数据集来自 kaggle，数据集包含 1584 条被标记为男的数据，1584 条被标记为女的数据。每条数据包含关于音频信号的特征（如音频信号的频率、频谱等的均值、峰值等）。

1.2 问题说明

数据集中已经把音频文件做了信号处理，抽取出特定的特征，并且都标识好男女标签，输入特征已经提供，接下来要做的是针对数据分布特征进行数据的预处理，在分类的工作中，需要依据这些特征来进行模型训练。

1.3 指标

问题是个二分类问题，正反例比例为 1 : 1，数据是平衡的，此处采用准确率来作为度量模型能力的指标。

关于准确率的计算是向模型输入测试集的标签，得出对应的预测值，与分测试集中的标签逐个对比，预测成功的数量占测试集的比例，计算公式为：

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

2. 分析

2.1 数据研究

数据集包含 3168 个样本，其中 50% 为男性，50% 为女性，数据集中包含以下特征：

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

对数据进行初步数据的检查：

```
In [26]: df.info() #可见数据无缺失情况

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3168 entries, 0 to 3167
Data columns (total 21 columns):
meanfreq    3168 non-null float64
sd           3168 non-null float64
median       3168 non-null float64
Q25          3168 non-null float64
Q75          3168 non-null float64
IQR          3168 non-null float64
skew         3168 non-null float64
kurt         3168 non-null float64
sp.ent       3168 non-null float64
sfm          3168 non-null float64
mode         3168 non-null float64
centroid     3168 non-null float64
meanfun      3168 non-null float64
minfun       3168 non-null float64
maxfun       3168 non-null float64
meandom      3168 non-null float64
mindom       3168 non-null float64
maxdom       3168 non-null float64
dfrange      3168 non-null float64
modindx      3168 non-null float64
label        3168 non-null object
dtypes: float64(20), object(1)
memory usage: 519.8+ KB
```

数据个特征并无缺失值，无需做缺失值处理。

除了 label 是字符串外，其他特征都是浮点型数值。

| | df.describe() | | | | | | | | | | | | |
|-------|---------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------|--|
| | meanfreq | sd | median | Q25 | Q75 | IQR | skew | kurt | sp.ent | sfm | mode | c | |
| count | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168.000000 | 3168. | |
| mean | 0.180907 | 0.057126 | 0.185621 | 0.140456 | 0.224765 | 0.084309 | 3.140168 | 36.568461 | 0.895127 | 0.408216 | 0.165282 | 0. | |
| std | 0.029918 | 0.016652 | 0.036360 | 0.048680 | 0.023639 | 0.042783 | 4.240529 | 134.928661 | 0.044980 | 0.177521 | 0.077203 | 0. | |
| min | 0.039363 | 0.018363 | 0.010975 | 0.000229 | 0.042946 | 0.014558 | 0.141735 | 2.068455 | 0.738651 | 0.036876 | 0.000000 | 0. | |
| 25% | 0.163662 | 0.041954 | 0.169593 | 0.111087 | 0.208747 | 0.042560 | 1.649569 | 5.669547 | 0.861811 | 0.258041 | 0.118016 | 0. | |
| 50% | 0.184483 | 0.059155 | 0.190032 | 0.140286 | 0.225684 | 0.094280 | 2.197101 | 8.318463 | 0.901767 | 0.396335 | 0.186599 | 0. | |
| 75% | 0.199146 | 0.067020 | 0.210618 | 0.175939 | 0.243660 | 0.114175 | 2.931694 | 13.648905 | 0.928713 | 0.533676 | 0.221104 | 0. | |
| max | 0.251124 | 0.115273 | 0.261224 | 0.247347 | 0.273469 | 0.252225 | 34.725453 | 1309.612887 | 0.981997 | 0.842936 | 0.280000 | 0. | |

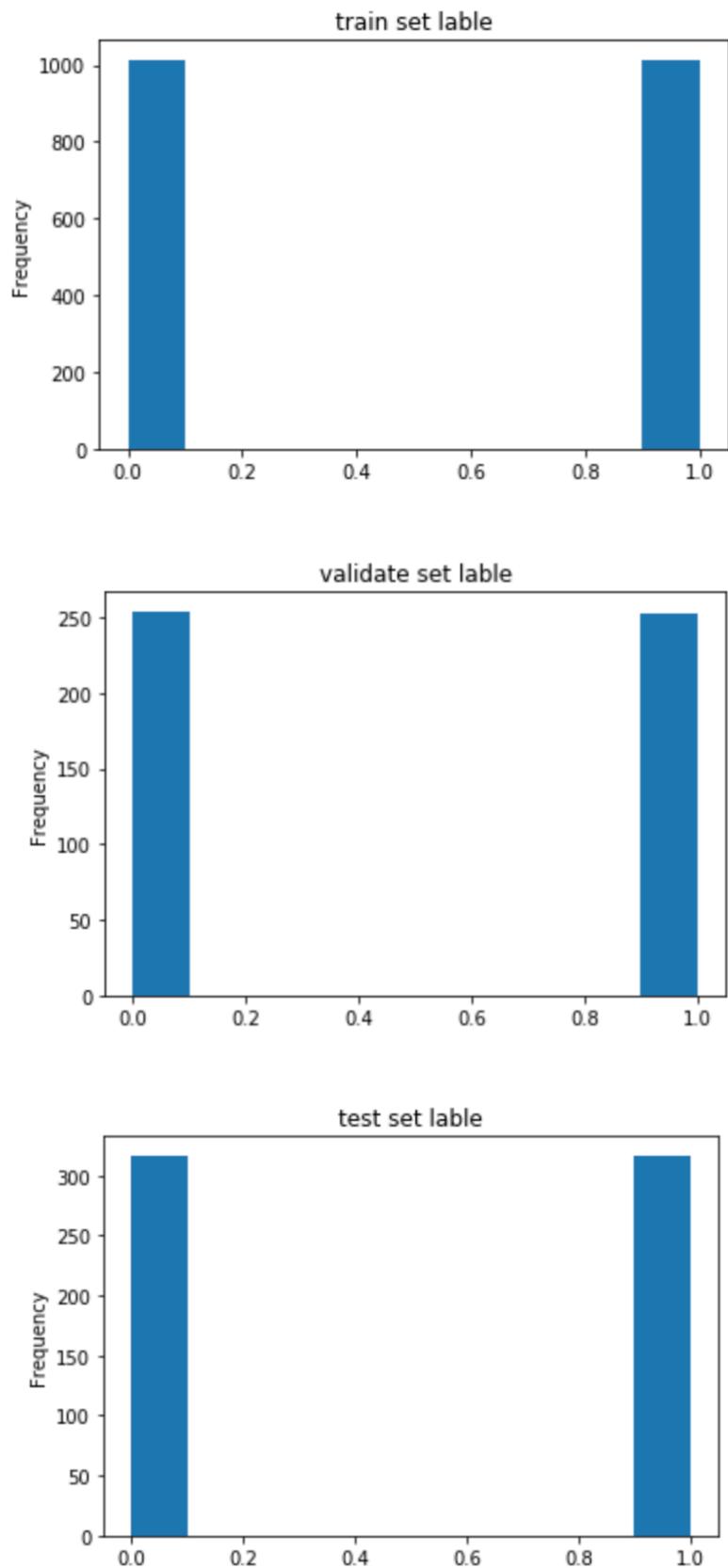
数据描述中，skew、kurt 特征的极大值超出 3σ ，存在异常值，需要处理。

数据集只提供 3168 条数据，总体的男女比例是 1 : 1，所以需要对提供的训练数据切分出训练集、验证集和测试集。训练集和验证集是用于训练模型的，测试集用于评估模型。

2.2 探索可视化

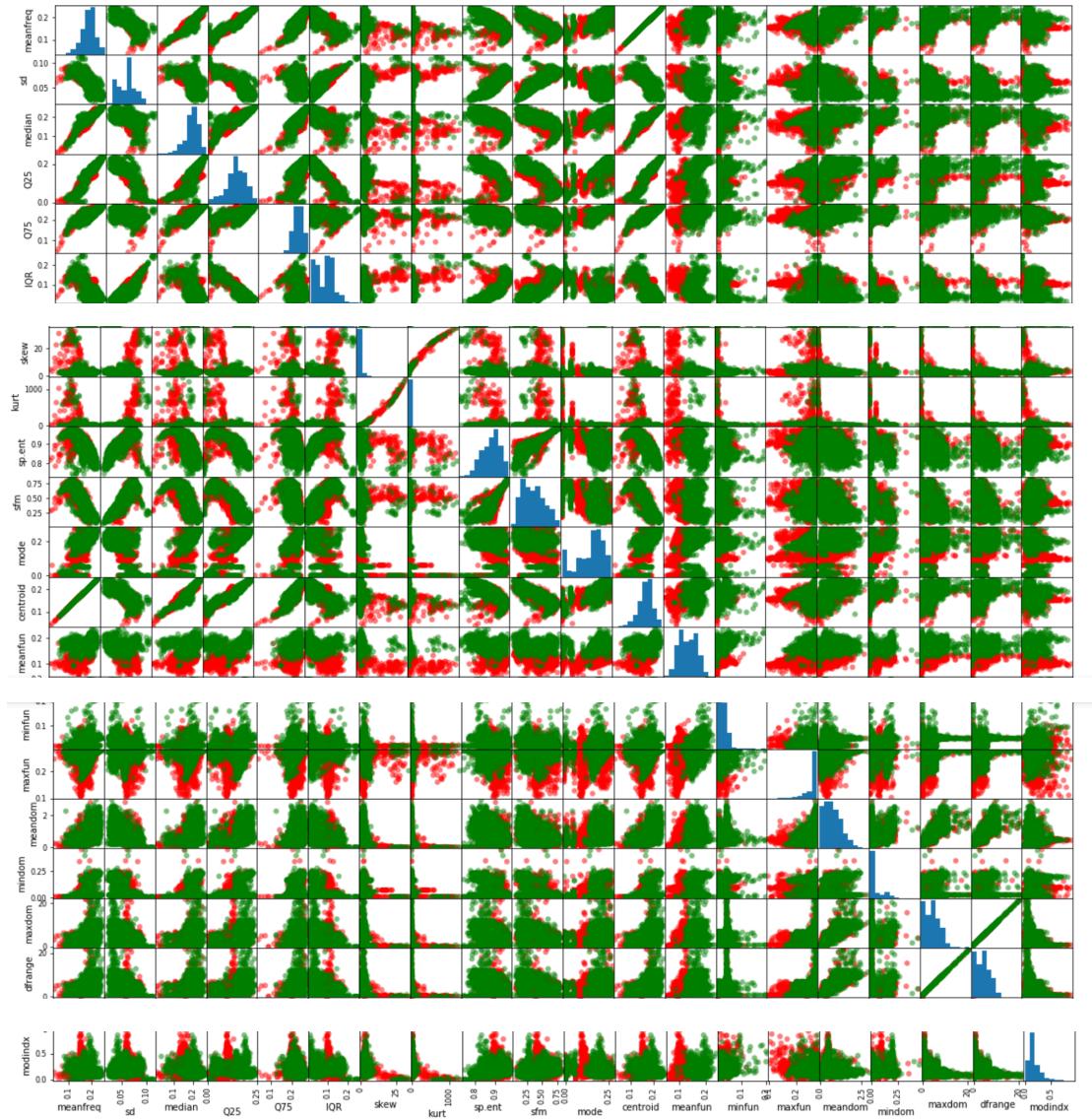
对所有样本进行切分，现已 8 : 2 的比例切分出训练集和测试集，这对切分出的训练集进一步以 8 : 2 的比例切分出真正的训练集和验证集。

对于训练集、验证集、测试集三者的标签频数绘图：



可见切分出的三个数据集男女比例都是均衡的。

查看两两特征的分类散点图：



可见 meanfun,maxfun 两个特征的正反例区分度比较高，初步推测分类时这两个特征提供比较大的贡献。

2.3 算法与方法

针对语音识别的训练集可知，输入数据是一组关于已经处理过的音频信号数据，输出是男女的类别。音频信号数据是特征空间，标签是输出空间。

用于分类的算法有很多，如逻辑回归、支持向量机、决策树等。

逻辑回归是在特征空间中求解多个超平面来是错误分类尽可能少，他要求数据是线性可分或者几部的线性可分，scatter_matrix 可见没有某一个特征显示出明显的线性可分，所以逻辑回归不合适。

支持向量机是在特征空间内求解分类超平面，使得分类超平面在分类边界离正反例距离尽可能远。支持向量机可以通过使用核技巧，把数据映射到高维度，然后再高维度里求解分类超平面，通过核技巧，支持向量机可以解决非线性问题。

支持向量机的常用核函数有线性核函数、径向基核函数和 sigmoid 核函数。

线性核函数：

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$$

径向基核函数/高斯核函数：

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

Sigmoid 核函数：

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta \mathbf{x}_i^T \mathbf{x}_j + \theta)$$

Svm 的优化目标公式为^[6]：

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \ell_{0/1}(y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1)$$

决策树则是依据信息论，通过信息增益或者信息增益率来对信息不断的进行分叉，这样可以把同一类的样本尽可能地归到同一类。

集成学习是使用多个弱分类器通过一定策略组合起来，作为一个新的强分类器，从而达到分类作用。理论依据为基分类器的错误率通过 Hoeffding 不等式展开，可以得出分类器数据的增大，集成错误率将指数级下降^[5]。

集成学习从个体的分类器的依赖关系强弱，分 bagging 和 boosting：

Bagging 通过重复多次有放回地随机抽出样本进行训练单个模型，再根据这些模型的预测结果经过类似投票的方法得到分类结果。这种方法的各个分类器依赖关系弱，代表是随机森林。

Boosting 则是先从初始训练集训练处一个基学习器，后续的训练样本基于前一次的训练结果进行调整，迭代多次，直到基学习器数目达到设定值，最终将这些基学习器进行加权结合。这种方法使得之前学习器做错的训练样本可以在后续得到更大的关注。此处使用 XGBoost。

随机森林是 XGBoost 都是以决策树为基学习器，随机森林是通过随机样本的放回、随机选择特征，构建决策树，然后投票得出分类结果；但是 XGBoost 的每个分类器都是根据已训练的分类器的性能进行训练，譬如对第一个学习器训练之后，增大错误样本的权重，同时减少正确样本的权重，再利用第二个学习器对于第一个样本进行学习。

GDBT 和 XGBoost 都是利用损失函数的负梯度来拟合每轮损失函数，但是 XGBoost 是支持并行，基于特征上的并行。

Xgboost 是 GB 算法的实现，xgboost 在目标函数中有经验误差和泛化误差，xgboost 目标函数：

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

Training loss **Complexity of the Trees**

其中泛化误差为：

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

Number of leaves **L2 norm of leaf scores**

泛化误差与叶子节点数和叶子节点的值有关。

2.4 基准测试

由于数据中男女各占 50%，假设设置所有的分类为其中一种（譬如都分类为男性），则会有 50% 的准确率，所以基准模型可以设为基准准确率为 50%。相关资料中显示 svm^[3] 和高斯混合模型^[4] 的准确率可高达 98.7%^[3] 和 99.62%^[4]，这里设置基准为 98%。

3. 方法

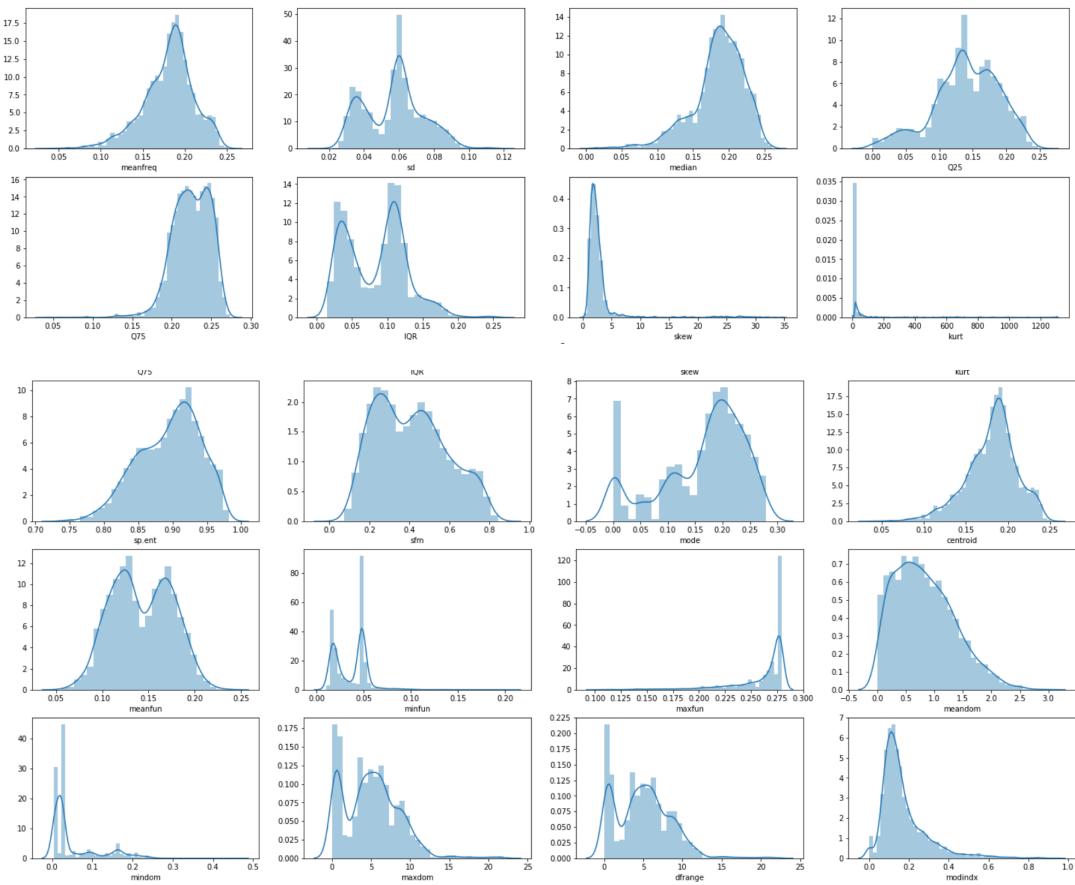
3.1 数据预处理

数据集已经提供了 3168 条数据，进行非空检查，可看到无缺数情况，不用进行缺失值处理。

数据的标签 label 列中值为英文的 male 和 female，此处需转为数值 0 和 1。

使用 pandas.DataFrame.skew() 观察数据倾斜情况，可得到四个特征的分布发生了比较大的倾斜，需要进行处理。分别为特征 skew(4.933314), kurt(5.872586), maxfun(-2.238535), modindx(2.064335) 四个特征存在比较大的偏斜。

查看各个特征的频数直方图，可见：



可见部分特征存在严重偏斜状况，集合以上的 skre() 函数，对 skew, kurt, maxfun, modindx 进行 log 转换。

3.2 实施

分别创建默认参数的随机森林和 XGBoost 模型，使用训练集进行训练，使用验证集和训练集得出准确率结果：

◦

| 模型 | 验证集上准确率 | 测试集准确率 |
|---------|---------|--------|
| 随机森林 | 96.25% | 97.63% |
| XGBoost | 97.04% | 97.63% |
| SVC | 73.96% | 76.81% |

3.3 改进

使用 GridSearchCV, 分别对各模型进行调参，：

| 模型 | 候选参数 | 最优参数 |
|---------|---|---|
| 随机森林 | 'n_estimators' : range(10,100,5), 'max_depth':range(3,200,5) | { 'max_depth': 13, 'n_estimators': 95 } |
| XGBoost | n_estimators 通过 early_stopping_rounds 确定； | {} |

| | | |
|-----|---|---|
| | <pre>'max_depth': range(3, 10, 2), 'min_child_weight': range(1, 10, 2) 'gamma': [i / 10.0 for i in range(0, 5)] 'learning_rate': [i *0.01 for i in range(0, 11)] 'reg_alpha': [1e-5, 1e-2, 0.1, 1, 100, 1000] 'reg_lambda': [1e-5, 1e-2, 0.1, 1, 100, 1000] 'subsample': [i *0.1 for i in range(6, 10)], 'colsample_bytree': [i *0.1 for i in range(6, 10)]</pre> | <pre>'colsample_bytree': 0.7, 'gamma': 0.1, 'learning_rate': 0.04, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 160, 'reg_alpha': 1e-05, 'reg_lambda': 1e-05, 'subsample': 0.9}</pre> |
| SVM | <pre>{'kernel': ['rbf'], 'gamma': [1e-3, 1e-4], 'C': [0.1, 1, 10, 50, 100, 1000]}, {'kernel': ['linear'], 'C': [0.1, 1, 10, 50, 100, 1000]}</pre> | <pre>{'C': 1000, 'kernel': 'linear'}</pre> |

调参后的模型，准确率为：

| 模型 | 测试集准确率 |
|---------|--------|
| 随机森林 | 98.11% |
| XGBoost | 98.11% |
| SVC | 97.95% |

模型融合^[7]：

融合以上三个模型，在测试机上得出的准确率是 98.11%

4. 结果

经过使用网格搜索调参的随机森林和 XGoost 在测试集上的准确率已经接近 98% 的准确率，其中随机森林的最优参数是

```
{'max_depth': 13, 'n_estimators': 95},
```

XGBoost 的最优参数是

```
{
'colsample_bytree': 0.7,
'gamma': 0.1,
'learning_rate': 0.04,
'max_depth': 3,
'min_child_weight': 1,
```

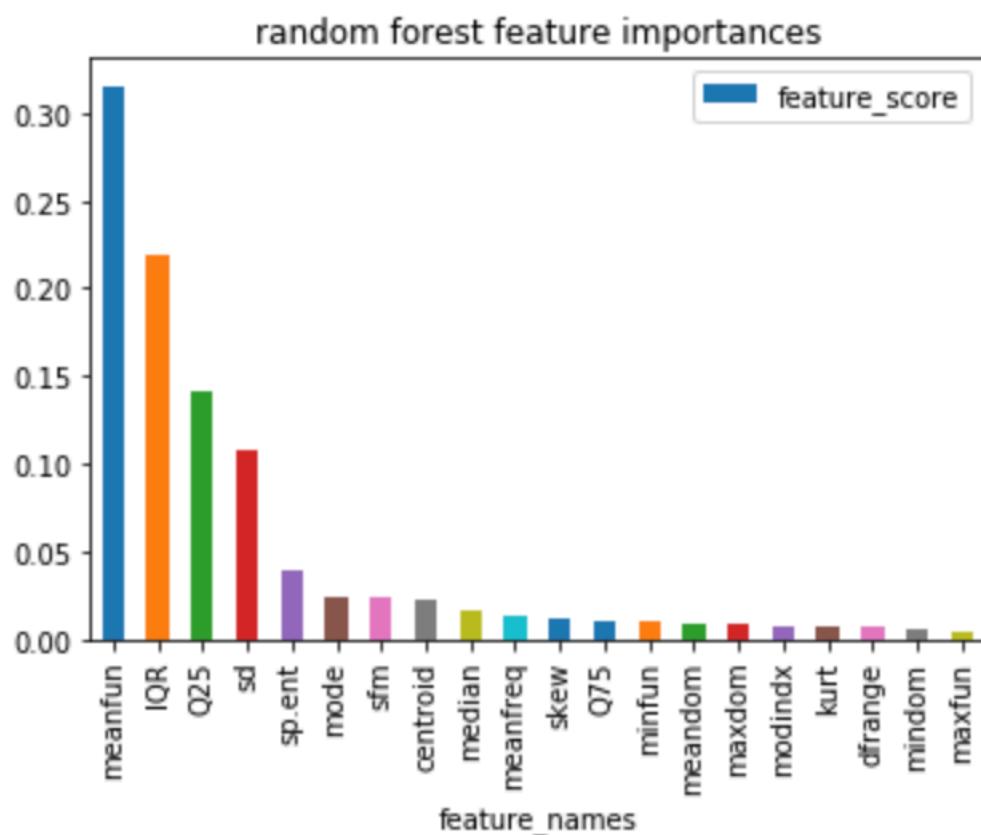
```
'n_estimators': 160, 'reg_alpha': 1e-05,  
'reg_lambda': 1e-05,  
'subsample': 0.9  
}  
,  
SVC 的最优参数是{'c': 1000, 'kernel': 'linear'}
```

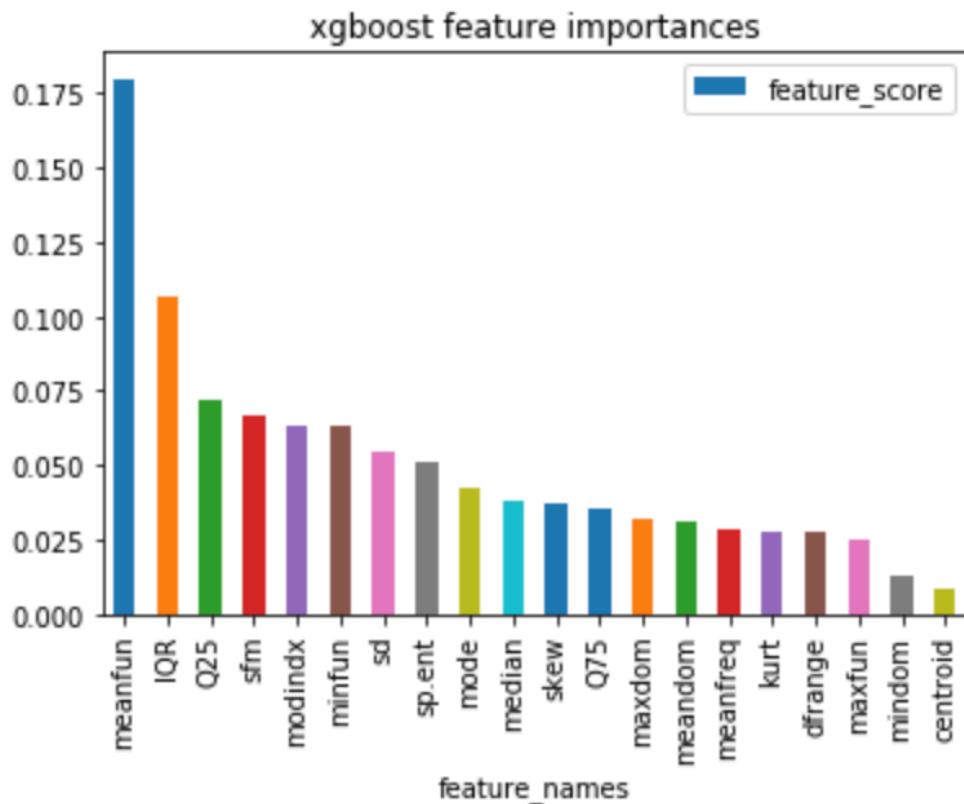
。
max_depth 对应的是单棵树的最大深度，随机森林和 XGBoost 都是用决策树作为弱分类器的选型，虽然说深度加大可以加强决策树的拟合能力，但是容易出现过拟合情况，集成学的基分类器是一种弱分类器，所以比较低的深度是比较合理的。

5. 结论

5.1 总结

分别对两个最优参数的模型的特征重要度进行作图：





可见两者都把 meanfun 占最高的贡献度。前三的特征贡献度排名一致。

本次试验使用了集成学习的两个方面的模型，来眼球两者准确率，两者都能达成设定目标。

5.2 改进

由于数据集是从 kaggle 上获取，所以后续改进可以尝试着自己采集真实生活中的音频来进行信号处理，提取出对应的信号。

而对于本次实验室基于对集成学习的研究，针对准确率的提升没有进行其他模型的探讨，譬如 SVM，后续的改进可以采用其他的分类模型进行相应的调参与本次试验对比，得到更优的模型。

参考文献

- [1] Ho, Tin Kam (1995). [Random Decision Forests](#) (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.
- [2] XGBOOST github 地址：<https://github.com/dmlc/xgboost>
- [3] 肖汉光,何为.基于 MFCC 和 SVM 的说话人性别识别[J].重庆大学学报（自然科学版）,2009,7:770-774
- [4] 张超琼,苗夺谦,岳晓冬.基于高斯混合模型的语音性别识别[J].计算机应用,2008,z2:360-362,365
- [5] 周志华.机器学习[M].北京:清华大学出版社,2015.172-173
- [6] 周志华.机器学习[M].北京:清华大学出版社,2015.129-130
- [7] 模型融合：<https://mlwave.com/kaggle-ensembling-guide/>