

开题报告

项目背景

俗语云“未见其人，先闻其实”，说明每个人的声音会像指纹那样，是有固定的“声纹”，而即使不认识的人，也能通过声音判别出是男是女，说明是性别的声纹是有固定的模式判别，而此项目则是让机器能辩雌雄。

问题描述

性别语音识别是一个分类问题。通过对音频信号提取出有可能关联的特征，训练出识别模型。

输入数据

数据集包含 3168 个样本。

```
In [23]: df['label'].value_counts()

Out[23]: female      1584
         male        1584
         Name: label, dtype: int64
```

男性女性各占一半。

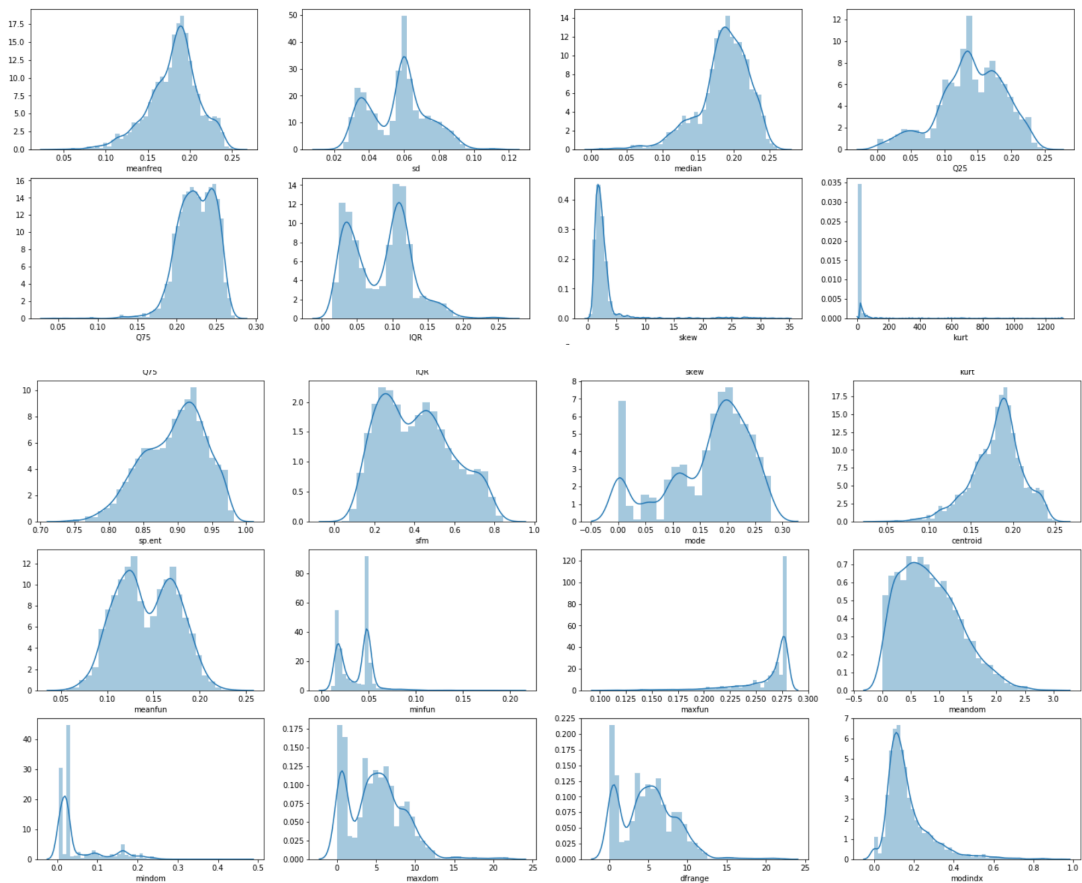
```
df.info() #可见数据无缺失情况
```

```
<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
```

数据不存在缺失情况，不需要缺失值填值处理。

```
In [89]: df.skew()
```

```
Out[89]: meanfreq    -0.617495
          sd          0.136916
          median     -1.012785
          Q25        -0.490877
          Q75        -0.900311
          IQR         0.295432
          skew        4.933314
          kurt        5.872586
          sp.ent      -0.430934
          sfm         0.339958
          mode        -0.837236
          centroid    -0.617495
          meanfun     0.039141
          minfun      1.878004
          maxfun      -2.238535
          meandom     0.611022
          mindom      1.661114
          maxdom      0.726189
          dfrange     0.728261
          modindx     2.064335
          dtype: float64
```



结合以上的直方图和 skew() 看，特征 skew, kurt, maxfun, modindx 四个特征存在比较大的偏斜，对以上特征进行 log 转换。
另外，把 label 编码成数字值。

解决办法

问题为二分类问题，本次分别使用随机森林^[1]、XGBoost^[2]模型。选择原因是出于集成学习不会发生过拟合效果。

基准模型

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%。

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%。

模型调优采用 GridSearch 来对特征中的某一项或多项进行调优，如随机森林，设置 max_depth 的候选值进行网格搜索。

评估指标

分类问题，男女的类别标签比例是 1:1，数据集是平衡的，可采用准确率进行评价。

设计大纲

数据预处理：结合数据探索，对数据偏斜度高（>2）的特征进行对数转换。另外，由于标签值是 male 或 female，需要对标签值转成数值型。

使用 sk-learn 的 train_test_split 将数据以 8:2 比例切分出训练集和测试集。

模型搭建：使用 sk-learn 中 RandomForestClassifier 和 XGBClassifier。

模型训练：随机森林设置决策树的数目 n_estimators，决策树的最大深度 max_depth 等参数；XGBoost 则使用 tree booster, 树的最大深度 max_depth

模型调参：利用 sk-learn 的 GridSearchCV 对以上参数进行调参。训练集以 8:2 切分出训练集和验证集。以训练集进行拟合，以验证集和预测值计算出准确率。

模型评估：使用 `sklearn.metrics` 中的 `accuracy_score` 进行评估。使用测试集计算出准确率。

可视化：对随机森林和 XGBoost 的调参过程的准确率，以直方图形式展示出来，展示出不同模型、参数间的差距。

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] XGB00ST 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