语音性别识别

机器学习纳米学位开题报告

陈晓杰
2018年8月

1. 问题描述

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[1]的分类问题。

2. 项目背景

说话人识别[2]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。

3. 数据或输入

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]。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）。

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	...	centroid	me
0	0.059781	0.064241	0.032027	0.015071	0.090193	0.075122	12.863462	274.402906	0.893369	0.491918	...	0.059781	0.0
1	0.066009	0.067310	0.040229	0.019414	0.092666	0.073252	22.423285	634.613855	0.892193	0.513724	...	0.066009	0.1
2	0.077316	0.083829	0.036718	0.008701	0.131908	0.123207	30.757155	1024.927705	0.846389	0.478905	...	0.077316	0.0

3 rows × 21 columns

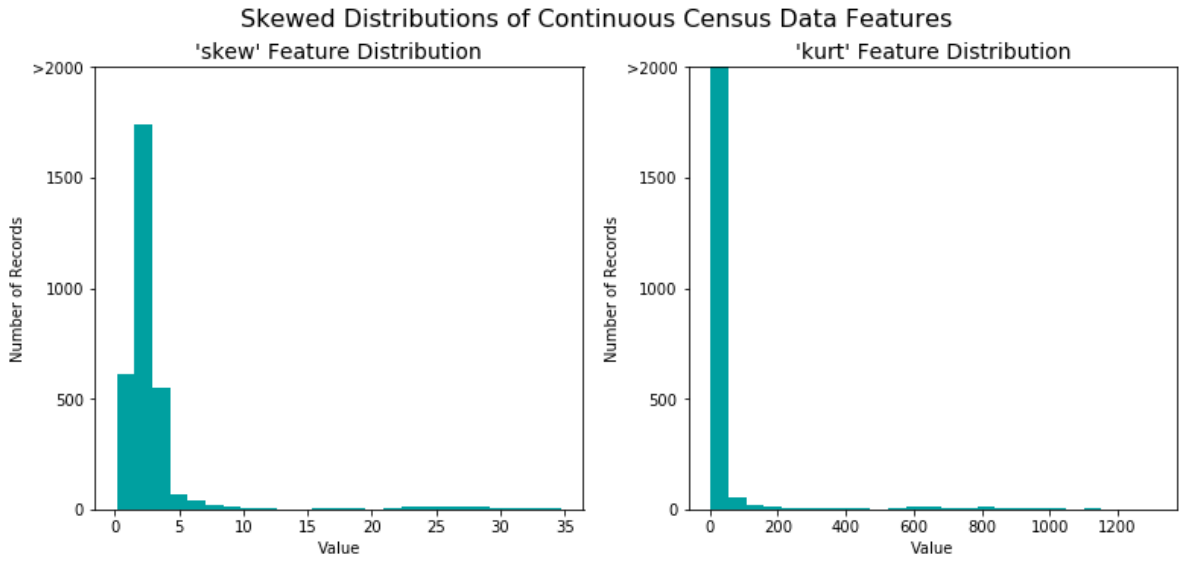
```
数据集中的总样本数： 3168
男性语音样本数： 1584
男性语音样本数： 1584
男性样本和女性样本的占比分别为： 50.0%， 50.0%
```

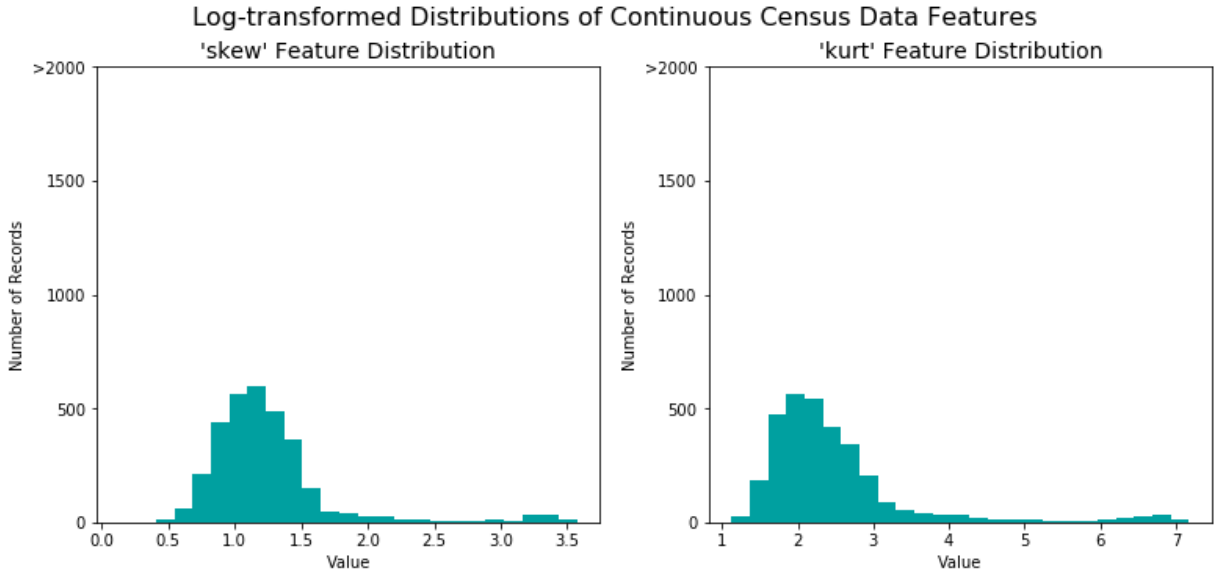
图中展示了项目数据集中的前三个样本。样本总数是3186个，每个样本有21个数据，包括20个特征和1个标签。数据集中的最后一列'label'将是需要预测的列，表示这个样本对应的音频的性别为男

性或女性，其他的每一列都是对应音频的具体声音特征。分析数据集中的每一列，没有非数值类型的值，所以无需填补缺失值。

```
# 将数据切分成特征和对应的标签
label_raw = data['label']
features_raw = data.drop('label', axis = 1)
# 将'label'编码成数字值
label_mapping = {'male':1, 'female':0}
label = label_raw.map(label_mapping)
```

在这20个特征里，都是数值的类型，所以无需进行数值的转换；最后一列标签是用'male'和'female'来表示男性和女性，因此需要将该列进行数值化处理，用1表示男性，用0表示女性。





分析每一列的数据时，可以发现'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，因此对其采取了对数转换，以便更好地进行归一化处理。

```
from sklearn.preprocessing import MinMaxScaler
# 初始化一个 scaler，并将它施加到特征上
```

```
scaler = MinMaxScaler()
numerical = features_raw.columns
features_raw[numerical] = scaler.fit_transform(features_raw[numerical])
features = features_raw

# 导入 train_test_split
from sklearn.model_selection import train_test_split
# 将'features'和'label'数据切分成训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(features, label, test
_size = 0.2, random_state = 0,stratify = label)

print("Training set has {} samples.".format(X_train.shape[0]))
print("Testing set has {} samples.".format(X_test.shape[0]))
```

```
Training set has 2534 samples.
Testing set has 634 samples.
```

至此，所有的类别变量都已转化成数值特征并进行了归一化处理，最后对数据集进行混洗和数据切分，最终切分成训练集和测试集。

4. 评估标准

判断一段语音是男性或是女性发出的，这是一个典型的二分类问题。在该项目中，输出的标签是不同的性别，在识别性别问题的角度上看，正确地识别出性别是最重要的，因此，适用的评估标准是分类准确率(Accuracy)和样本预测时间。

- 混淆矩阵(confusion matrix)
  关于二分类问题中，测试数据集的真实值和与预测值，有以下四个关系：
  TN: True Negative，预测值为1，且预测对了
  TP: True Positive，预测值为0，且预测对了
  FN: False Negative，预测值为0，但预测错了
  FP: False Positive，预测值为1，但预测错了

  混淆矩阵的定义如下表所示：

	实际值为 1	**实际值为 0**
预测值为 1	TP	FP
预测值为 0	FN	TN

- 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。根据混淆矩阵，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：

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- 样本预测时间
  样本预测时间可以衡量分类器的对于数据预测的快慢，对于一个测试集，若其样本总数为N，分类器对测试集的预测总时长为T，则样本预测时间的定义如下：

$$t = \frac{T}{N}$$

5. 基准模型

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%[11]。本项目的目标是争取准确率达到98%以上。

6. 项目设计

随机森林[12]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]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[13]进行超参调整，调节的参数和范围分别为：n_estimators(1-80，步进5)、criterion(gini和entropy)、max_features(3-8，步进1)、min_samples_split(2-8，步进1)
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。

---

1. https://zh.wikipedia.org/wiki/監督式學習 ↩
2. https://en.wikipedia.org/wiki/Speaker_recognition ↩
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://www.cnblogs.com/liuyihai/p/8309019.html ↩
13. https://www.zhihu.com/question/48282030/answer/114305326 ↩