语音信号时域分类实验

郜炫齐 刘贵涛 刘少腾 潘昊璇 吴思源

Contents

1. 特征提取
2. 分类方法
   - 二分类
   - 多分类
3. 结果分析
4. 参考资料

特征提取

数据收集与划分

5 位同学每个人录了 20 组数据，每条语音数据为 2 秒，有效部分为 1 秒

预加重与加窗函数

预加重为了增加高频分辨率。使用 $H(z) = 1 - 0.95z^{-1}$ 的一阶 FIR 高通滤波器进行预加重

加窗分帧实际上是为了对分帧信号做一个平滑处理，等效于对信号做一个滤波。帧率设置为每秒 64 帧和 128 帧，方便后面基 2FFT 运算。本实验中加了矩形窗和 Hanning 窗：

矩形窗

$$w(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & others \end{cases}$$

Hanning 窗

$$w(n) = \begin{cases} 0.5\ 1 - \cos\frac{2\pi n}{N-1} & 0 \le n \le N-1 \\ 0 & others \end{cases}$$

音频处理

采用双门限方法剪辑音频数据，取出有效部分
帧率设置为每秒 64 帧和 128 帧[1]

MFCC 的阶数为 14（1∼14 个都提取），滤波器为 20 25 组

[1]方便进行基 2 快速傅里叶变换

预处理得到的一维信号

分帧加窗之后得到的是二维信号，先将其转化为一维信号，方便特征提取

依据：语音信号短时平稳，认为一帧之内的信号特性近似相同。

对于时域信号，可以选择的有

- 能量
- 短时平均过零率
- 每帧的平均值

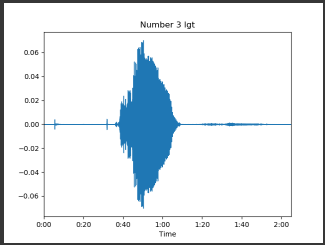
对于频域信号，可以选择的有

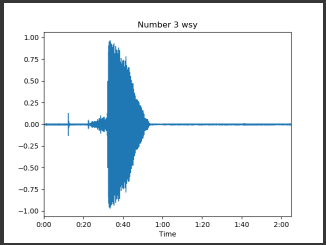
- MFCC

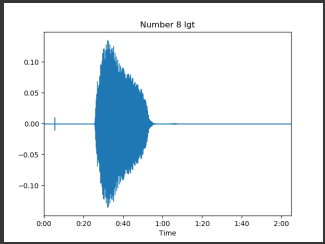
提取序列的总体特征

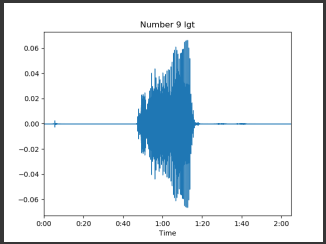
由于本问题中的语音信号较为简单、特征明显。
同一语音信号在时域或长或短、或快或慢，但是其表达的内容是相同的
故考虑使用其序列总体特征进行分类。

特征分析


Number 3 lgt
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0:00
0:20
0:40
1:00
1:20
1:40
2:00
Time


Number 3 wsy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0:00
0:20
0:40
1:00
1:20
1:40
2:00
Time


Number 8 lgt
0.10
0.05
0.00
−0.05
−0.10
0:00
0:20
0:40
1:00
1:20
1:40
2:00
Time


Number 9 lgt
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0:00
0:20
0:40
1:00
1:20
1:40
2:00
Time

时间序列特征

对于每一类型的特征序列，提取其序列总体特征

1. 平均值
2. 最大值
3. 标准差
4. 上升部分比例

数据收集与划分

将采集到的每类 100 个数据打乱之后划分训练集和测试集

训练集	验证集	测试集
850	50	100

分类方法
二分类

1. 决策树
2. 支持向量机
3. Boosting 方法

多分类结果的 Confusion Matrix

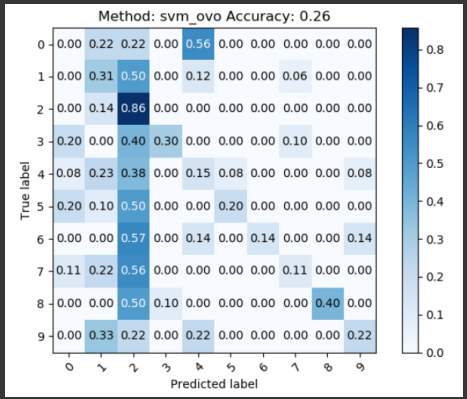


图: SVM 多分类

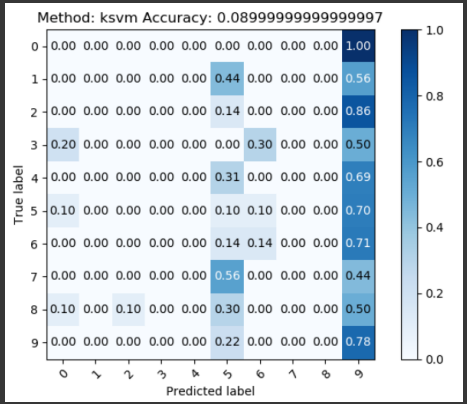


图: 线性 SVM 多分类

多分类结果的 Confusion Matrix

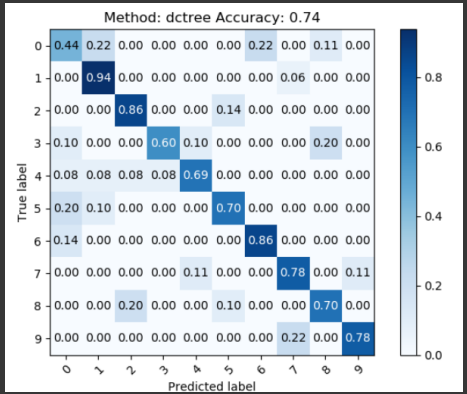


图: 决策树多分类

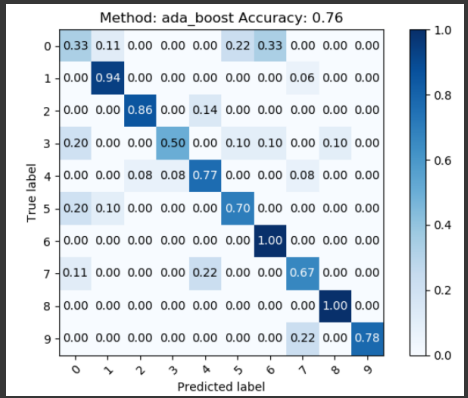


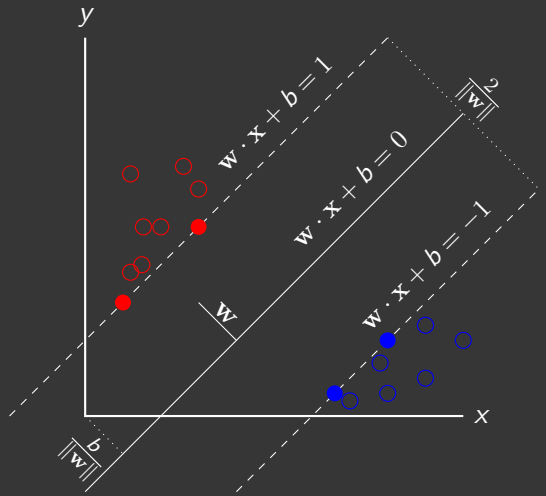
图: AdaBoosting 方法

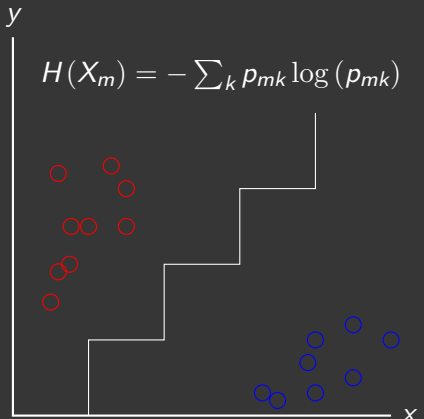
2

[2] 20 个分类器，采用 ECOC 方法

决策边界

支持向量机与决策树





方法比较

SVM 利用线性超平面分类，对于线性可分的数据泛化能力很好。但是需要较多的数据才能收敛。

决策树 使用线性超平面的组合进行分类，拟合能力很强，容易收敛，但是也容易过拟合。

实验结果：在训练集上 SVM 类方法比较难收敛

现有的多分类方法

- One - vs - Rest $\Longrightarrow$ 正负样本不均衡的问题
- One - vs - One $\Longrightarrow$ 计算量可能过大，分类器过多
- Error-Correcting Output Codes

ECOC

Error-Correcting Output Codes

将多分类问题转化为多个二分类问题，使用 m 个分类器去解决这 m 个问题，然后将各个分类器的结果编成一组二进制码。通过对比码表来确定最终类别。

优点

有些分类器分错也能得到尽可能正确的结果

虽然不是任意生成的二进制码都有自纠错性的，但是其实 ECOC 码的优劣对分类效果影响不大

多分类器

显然，当分类器个数越多的时候，因为某几个分类器出错导致测试样本错分的概率会下降，结果更好

采用决策树分类器，采样率为 128 帧每秒，仅使用 MFCC 的序列特征进行分类

每增加 20 个分类器测试集准确率增加 1 个百分点

分类器个数较少 (40 以内) 的时候，决策树多分类方式效果会更好。通过调整分类器的个数，可以把准确率从 70 调整到 80 左右。

结果分析

每秒帧数对结果的影响

发现对于每秒 64 帧和每秒 128 帧的情况，每秒 128 帧比每秒 64 帧结果高 3 个点左右

帧数	分类器数	结果
64	40	69
128	40	72

特征聚类

尝试找到时域特征和频域特征在描述特定类别时的相似与不同

从每类任取 10 个数据，然后计算特征空间与其距离最近的 20 个点，之后统计 20 点中正例的数目和正例的百分数

	时域特征聚类		频域特征聚类	
Label 0	50/200	0.25	58/200	0.29
Label 1	91/200	0.455	132/200	0.66
Label 2	89/200	0.445	116/200	0.58
Label 3	59/200	0.295	161/200	0.805
Label 4	55/200	0.275	76/200	0.38
Label 5	87/200	0.435	89/200	0.445
Label 6	73/200	0.365	88/200	0.44
Label 7	159/200	0.795	100/200	0.5
Label 8	94/200	0.47	119/200	0.595
Label 9	152/200	0.76	89/200	0.445

T-SNE

降维算法
利用仿射变换把高维数据点分布映射到概率空间
然后在概率空间寻找高维和低维的相似性
这里调用 sklearn.manifold.TSNE 进行学习

T-SNE 可视化

时域数据的 T-SNE

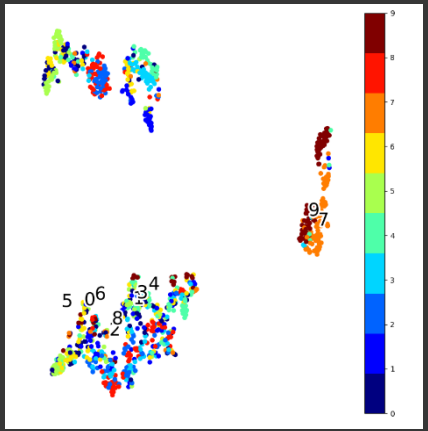


图: 时域 T-SNE 9 和 7 时域特征容易辨别

T-SNE 可视化

频域数据的 T-SNE

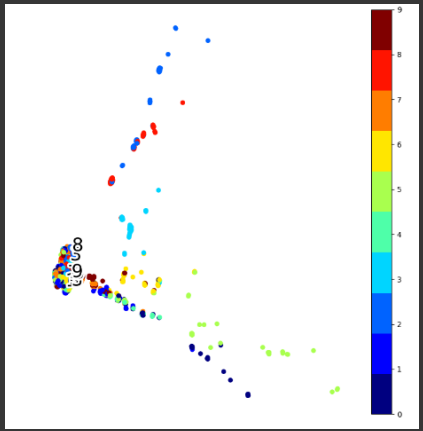


图: 频域 T-SNE 3 容易区分

部分结果比较

特征	每秒帧数	分类器	准确率%
时域特征	128	SVM + ECOC	9
时域特征	128	决策树 + ECOC	55
时域特征	128	SVM ovr	41
时域特征	128	决策树多分类	59
14 阶梅尔倒谱系数 (MFCC-14)	128	SVM ovo	28
14 阶梅尔倒谱系数 (MFCC-14)	128	决策树多分类	70
14 阶梅尔倒谱系数 (MFCC-14)	128	SVM + ECOC	14
14 阶梅尔倒谱系数 (MFCC-14)	128	决策树 + ECOC	78
14 阶梅尔倒谱系数 (MFCC-14)	128	Ada Boost + ECOC	80
14 阶梅尔倒谱系数 (MFCC-14)	128	KNN + ECOC	59

参考工具箱

- 提特征 $\Longrightarrow$ tsfresh [3]
- 机器学习工具箱 $\Longrightarrow$ sci-kit learn [4]
- 语音信号提取 $\Longrightarrow$ librosa [5]
- 数据可视化 $\Longrightarrow$ seaborn [6]

[3] https://tsfresh.readthedocs.io/en/latest/
[4] https://scikit-learn.org/
[5] https://librosa.github.io/librosa/install.html
[6] http://seaborn.pydata.org/

本项目开源地址

https://github.com/edmundwsy/ASR-for-chinese-number