第7期(2016年度)AITC成果発表会
クラウド・テクノロジー活用部会

音の機械学習：話者識別

2017年9月19日

富士通株式会社　松井　唯
市川　研二

※この発表は個人の見解であり、所属する組織の公式見解ではありません

amazon echo

https://smashop.jp/blog/3261


What can I help
you with?

様々な形で音声認識技術が実用化

connpass
produced by Be PROUD

ダッシュボード　カテゴリー一覧　新着イベン

キーワード
自然言語

開催日
2017/08/15 ～ 2017/09/14

都道府県

イベント検索

登録したイベントが検索に

テキストマイニング決定版
見える化エンジン

膨大なテキスト情報から
分析が可能。1200社の導
／見える化エンジン

検索結果 (26件)

connpass

produced by Be PROUD

ダッシュボード　カテゴリー覧　新着イベン

キーワード

画像認識

開催日

2017/08/15 ～ 2017/09/14

都道府県

イベント検索

登録したイベントが検索に

テキストマイニング決定版

見える化エンジン

膨大なテキスト情報から
分析が可能。1200社の導
／見える化エンジン

検索結果 (25件)

connpass
produced by Be PROUD

ダッシュボード　カテゴリー覧　新着イベン

キーワード
音声認識

開催日
2017/08/15 ～ 2017/09/14

都道府県

イベント検索

登録したイベントが検索に

か、会社のソフトが
不正コピーだった
にゃんて・・・

BSA

不正コピーソフトを使用中の
BSAにお知らせ下さい。有力情報に
報奨金を差し上げます（2013年 実
*報奨金の提供には一定の条件があり
詳細はHPをご確認下さい。

検索結果（　件）

自然言語≒画像処理≒音声認識

26件　25件　9件

スクナイ

音の現状

野良エンジニアが少ない

なぜ

音を扱うのが難しい

なぜ

1. ファイル入出力や特徴量抽出が難しい

→ライブラリを使えば簡単：昨年の報告

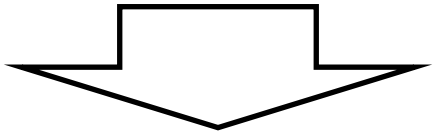
2. 手法が分からない：今年の報告

本日取り扱う技術

話者識別

アプローチ

手法が分からない



手法を調べて実装する

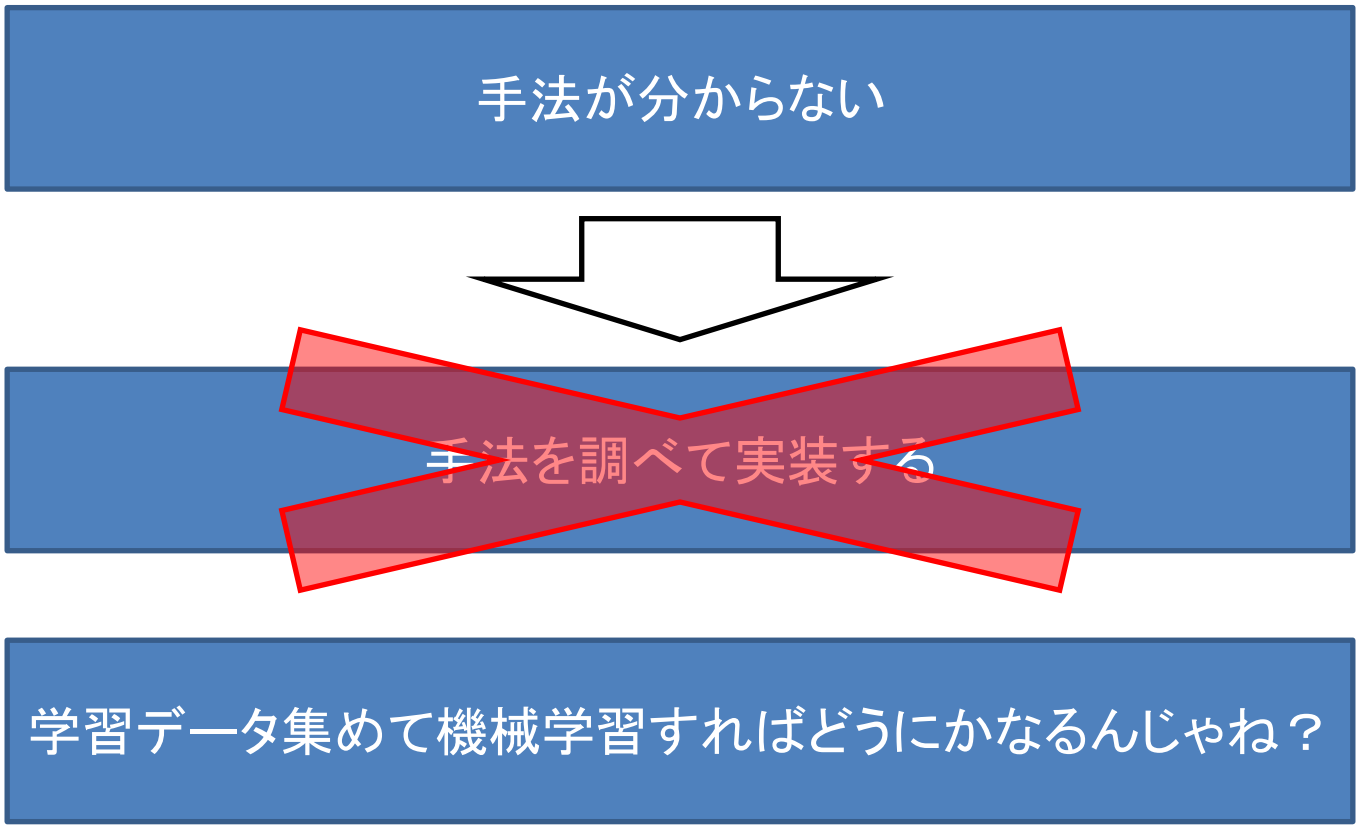
話者識別で一般的な手法

- i-vector
  1. 発話の音響特徴量を混合ガウス分布でモデル化
  2. 平均ベクトルを結合しGMMスーパーベクトルを算出
  3. 個人性を分離できる空間に射影して因子分析
  4. ここで得られる写像がi-vector

わからん・・・

詳しくはこちらをご覧ください
小川 哲司, 塩田 さやか, “i-vectorを用いた話者認識,” 日本音響学会誌, Vol. 70, No. 6, pp. 332-339, 2013.

今回のアプローチ



粗が目立つかもしれませんがご了承ください

やったこと

- 音声を集める
- 特徴量をたくさん出す
- 特徴量を選ぶ
- 分類する

やったこと

- 音声を集める
- 特徴量をたくさん出す
- 特徴量を選ぶ
- 分類する

音声を集める

- 学習用データ
  - 内容：ATR音素バランス文
  - 量：5~10秒程度×25文×10名
    - クラウド・テクノロジー活用部会参加者6名の音声を収録
    - 話速バリエーション型音声データベース（SRV-DB）を利用　4名分 http://www.it.ice.uec.ac.jp/SRV-DB/

やったこと

- 音声を集める
- 特徴量をたくさん出す
- 特徴量を選ぶ
- 分類する

特徴量をたくさん出す

- OpenSMILEを用いて特徴量抽出
  - OpenSMILE：ミュンヘン工科大学が開発している特徴量抽出のためのフリーソフト
  - 感情認識の分野でよく用いられている

- 様々な特徴量セットが用意されているが、どれがいいかわからなかったのでデフォルトの「The INTERSPEECH 2009 Emotion Challenge feature set」を使用（感情認識用）

特徴量をたくさん出す

- 抽出結果：384次元

name	pcm_RMSenergy_sma_max	pcm_RMSenergy_sma_min	pcm_RMSenergy_sma_range	pcm_RMSenergy_sma_maxPos	pcm_RMSenergy_sma_minPos	pcm_RMSenergy_sma_amean	pcm_RMSenergy_sma_linregc1	pcm_RMSenergy_sma_linre
'ATR_AM00_0800_000.wav'	2.38E-02	2.91E-05	2.38E-02	140	8	4.73E-03	-9.03E-06	
'ATR_AM00_0800_001.wav'	2.08E-02	2.99E-05	2.08E-02	54	314	4.91E-03	-1.09E-05	
'ATR_AM00_0800_002.wav'	1.57E-02	3.07E-05	1.57E-02	83	4	4.24E-03	-5.94E-06	
'ATR_AM00_0800_003.wav'	1.59E-02	0	1.59E-02	251	0	3.86E-03	2.05E-06	
'ATR_AM00_0800_004.wav'	1.49E-02	3.33E-05	1.49E-02	300	452	4.04E-03	-3.69E-06	
'ATR_AM00_0800_005.wav'	1.35E-02	2.94E-05	1.35E-02	144	2	3.76E-03	-5.58E-06	
'ATR_AM00_0800_006.wav'	1.40E-02	2.62E-05	1.40E-02	466	7	4.00E-03	-4.16E-06	
'ATR_AM00_0800_007.wav'	1.00E-02	2.34E-05	1.00E-02	92	4	3.41E-03	-4.00E-06	
'ATR_AM00_0800_008.wav'	2.82E-02	2.99E-05	2.81E-02	332	254	4.46E-03	-3.30E-07	
'ATR_AM00_0800_009.wav'	2.87E-02	2.86E-05	2.87E-02	56	9	4.74E-03	-9.26E-06	
'ATR_AM00_0800_010.wav'	1.45E-02	2.56E-05	1.45E-02	244	10	4.60E-03	-1.95E-06	
'ATR_AM00_0800_011.wav'	2.08E-02	2.53E-05	2.08E-02	85	505	4.35E-03	-1.86E-06	
'ATR_AM00_0800_012.wav'	1.76E-02	2.55E-05	1.75E-02	34	492	3.93E-03	-1.05E-06	

特徴量をたくさん出す

接頭	説明
pcm_fftMag_mfcc_sma.[N]	MFCC[N]次元目
pcm_RMSenergy_sma	パワー
pcm_zcr_sma	波形のゼロ交差率
F0_sma	基本周波数
voice_Prob_sma	声である確率

×

中間	説明
[無し]	通常
de_	時間微分値

×

接尾	説明
max, min	最大値,最小値
amean	平均
maxPos, minPos	最大値, 最小値のある位置
range	値の範囲 (最大値と最小値の幅)
linregc1, linregc2, linregerrQ	線形近似の勾配度, オフセット, 二乗誤差
stddev	標準偏差
skewness	歪度
kurtosis	尖度

=

384次元

MFCC:音声の周波数軸での概形を表現する特徴量

特徴量をたくさん出す

接頭	説明
pcm_fftMag_mfcc_sma.[N]	MFCC[N]次元目
pcm_RMSenergy_sma	パワー
pcm_zcr_sma	波形のゼロ交差率
F0_sma	基本周波数
voice_Prob._sma	声である確率

接尾	説明
max, min	最大値,最小値
amean	平均
maxPos, minPos	最大値, 最小値のある位置
range	値の範囲 (最大値-最小値の幅)
linregc1, linregc2	線形近似の勾配度, オフセット,二乗誤差
stddev	標準偏差
skewness	歪度
kurtosis	尖度

中間	説明
[無し]	通常
de_	時間微分値

MFCC:音声の周波数軸での概形を表現する特徴量

384次元

どれがいいかは
機械学習が選んでくれ

やったこと

- 音声を集める
- 特徴量をたくさん出す
- 特徴量を選ぶ
- 分類する

特徴量を選ぶ

- SVM (Support Vector Machine)
  - 教師あり学習を用いるパターン認識モデルの一つ
  - 現在知られている手法の中でも認識性能が優れた学習モデルの一つ

https://ja.wikipedia.org/wiki/**サポートベクターマシン** より

- 遺伝的アルゴリズム
  - 進化的アルゴリズムの一つ
  - 解の候補を遺伝子に見立て、近似解を探索

https://ja.wikipe**dia.org/wiki/遺伝的アルゴリズム** より

特徴量を選ぶ

- 全特徴量 (384次元) で学習してみた
  - 10名分のデータ

```
> # 正解している行数
> nrow(y[y$result == y$ans,])
[1] 70
> # 全データ数
> nrow(y)
[1] 125
> # 精度
> acc
[1] 0.56
```

いらない特徴量がある?

→ 精度: 56%

次元数が大きいため計算量が大きい

特徴量を選ぶ

- 4名分のデータを使用
  - 話速バリエーション型音声データベースから
- 遺伝子型 : 各特徴量を学習に使用するかしないかを表すbit列

- 各bitを一定確率で反転した個体を30個作成
  - SVMで学習
  - モデルを評価
    - 正答数
    - 特徴量の次元数によるペナルティ
- 評価値が高い個体1個を次世代とする

1000回実行

- 特徴量: 384次元 → 11次元

特徴量を選ぶ

- 4名分のデータを使用
  - 話速バリエーション型音声データベースから
- 遺伝子型 : 各特徴量を学習に使用するかしないかを表すbit列

- 各bitを一定確率で反転した個体を30個作成
  - SVMで学習
  - モデルを評価
    - 正答数
    - 特徴量の次元数によるペナルティ
- 評価値が高い個体1個を次世代とする

1000回
実行

- 特徴量: 384次元 → 11次元

選んだ特徴量

選択された特徴量	
pcm_fftMag_mfcc_sma.2._linregerrQ	MFCC2次元目の線形近似の二乗誤差
pcm_fftMag_mfcc_sma.3._linregerrQ	MFCC3次元目の線形近似の二乗誤差
pcm_fftMag_mfcc_sma.4._min	MFCC4次元目の最小値
pcm_fftMag_mfcc_sma.9._amean	MFCC9次元目の線形近似の平均
pcm_fftMag_mfcc_sma.11._amean	MFCC11次元目の線形近似の平均
pcm_fftMag_mfcc_sma_de.8._kurtosis	MFCC8次元目時間微分の尖度
pcm_fftMag_mfcc_sma_de.2._range	MFCC2次元目時間微分の幅
pcm_fftMag_mfcc_sma_de.3._linregerrQ	MFCC3次元目時間微分の線形近似の二乗誤差
pcm_fftMag_mfcc_sma_de.5._minPos	MFCC5次元目時間微分の最小値の位置
pcm_fftMag_mfcc_sma_de.8._linregc2	MFCC8次元目時間微分のオフセット
pcm_zcr_sma_max	ゼロ交差の最大値

ソースコード

```
library(“GA”)
library(“kernlab”)

#カレントディレクトリ変更
setwd(‘/Users/me/Documents/170919_AITC_washa’)
```

インポート他

```
#データをロード
#Web上の音声から抽出した特徴量
x <- read.table(“webdata.csv”, header=T, sep=“;”)

#nameカラムを話者名にする
x$name <- substring(x$name, 5, 8)

#データの分割
rowdata<-nrow(x)
random_ids<-sample(rowdata, rowdata*0.5)

#トレーニングデータ
training<-x[random_ids, ]

#テストデータ
predicting<-x[-random_ids, ]
```

データのロード
分割

ソースコード

```
#遺伝的アルゴリズム用のfitness関数
fitness <- function(filter){
    filter <- c(1, filter)
                # 先頭行(name)は必ず残す
    t <- training[,filter == 1]
    p <- predicting[,filter == 1]
    x_svm<-ksvm(name ~., data=t )

    # テストデータを推論
    result_predict<-predict(x_svm, p)

    # 一致している件数をカウントする
    # できるだけ項目数を少なくするため、使っている項目数をペナルティとして引く
    y <- data.frame(result=result_predict, ans=p$name)
    nrow(y[y$result == y$ans,]) – (length(filter[filter==1])/1000
}
```

学習・評価用の関数

```
# 遺伝的アルゴリズム
GA <- ga(type = “binary”, fitness = fitness, nBits = ncol(x) – 1, popSize = 30, maxiter = 1000, pmutation=0.2)
summary(GA)
plot(GA)
```

遺伝的アルゴリズム

```
# 特徴量の組み合わせを出力
sol <- c(1, GA@solution[1,])
xx <- x[,sol == 1]
# 使用した特徴量名(およびname)を出力
names(xx)
```

結果出力

やったこと

- 音声を集める
- 特徴量をたくさん出す
- 特徴量を選ぶ
- 分類する

分類する

- 10名分のデータを使用
  - クラウド・テクノロジー活用部会参加者6名
  - 話速バリエーション型音声データベース（SRV-DB）4名分
- SVMで学習
- テストデータを分類
- 精度：56% → 98%

ソースコード

```
library(“GA”)
library(“kernlab”)

# カレントディレクトリ変更
setwd(‘/Users/me/Documents/170919_AITC_washa’)
```

インポート他

```
# データをロード
# Web上の音声から抽出した特徴量
x1 <- read.table(“webdata.csv”, header=T, sep=“;”)
# AITC CT部会メンバーの音声から抽出した特徴量
x2 <- read.table("aitcdata.csv", header=T, sep=",")
# 結合
x <- rbind(x1, x2)

# 厳選した特徴量のカラムだけ残す
x <- x[, which(colnames(x) %in%c("name","pcm_fftMag_mfcc_sma.2._linregerrQ","pcm_fftMag_mfcc_sma.3._linregerrQ",
      "pcm_fftMag_mfcc_sma.4._min","pcm_fftMag_mfcc_sma.9._amean","pcm_fftMag_mfcc_sma.11._amean","pcm_zcr_sma_max",
      "pcm_fftMag_mfcc_sma_de.2._range","pcm_fftMag_mfcc_sma_de.3._linregerrQ","pcm_fftMag_mfcc_sma_de.5._minPos",
      "pcm_fftMag_mfcc_sma_de.8._linregc2","pcm_fftMag_mfcc_sma_de.8._kurtosis"))]

# nameカラムを話者名にする
x$name <- substring(x$name, 5, 8)
```

データのロード
分割

ソースコード

```
# データの分割
rowdata<-nrow(x)
random_ids<-sample(rowdata, rowdata*0.5)

#トレーニングデータ
training<-x[random_ids, ]

#テストデータ
predicting<-x[-random_ids, ]
```

データのロード
分割

```
t <- training
p <- predicting

# 学習
x_svm<-ksvm(name ~., data=t )

#テストデータを分類
result_predict<-predict(x_svm, p)
y <- data.frame(result=result_predict, ans=p$name)

# 結果出力
y
```

学習・分類

考察

- 遺伝的アルゴリズムを用いることで人が関与せずに特徴量選択できた
- 有識者が特徴量選択すればもっとうまくいきそう

→結果に関してd有識者の見解を聞きたい

このあたりの特徴は
個人性には寄与
しないのでは？

接尾	説明
max, min	最大値,最小値
amean	平均
maxPos, minPos	最大値, 最小値のある位置
range	値の範囲 (最大値と最小値の幅)

まとめ

- SVMを使用して10人分の声の分類を行い98%の精度で分類できた
- 自動特徴量選択を行った
  - 特徴量を384次元→11次元に絞り込んだ
  - 精度が56%→98%に向上した
- 今後は様々な環境での使用時の挙動を確認したい