2022年度 言語解析論 レポート課題

氏名:
学生番号:

2022年7月20日

1 概要

本レポートでは，著者推定問題をSVMを用いて解く手法の概要と結果を報告する．

2 手法

判別器の構成は図1に示す通りである．

`mkvector.py` では，1列目に正解ラベルが，2列目に文が書かれたCSVファイルを入力し，MeCabによって形態素に分け，各形態素にIDを割り振った辞書を作る．この辞書に形態素列を再び通し，文を，IDiの形態素が現れるかどうかを第i成分としたBag-of-wordsベクトルとして，疎ベクトルの形式で正解ラベルとともに `.feature` ファイルに出力する．

`svm_train.py` では，`train.feature` に出力された学習データのBag-of-wordsベクトルをNumPyのベクトルに変換し，ラベルとともにscikit-learnのSVMに与えて学習する．学習したモデルをpickleを用いて `svm_model.dat` にdumpする．

`svm_test_analy.py` では，`svm_model.dat` をロードし，`test.feature` に出力されたテストデータのBag-of-wordsベクトルをNumPyのベクトルに変換し，SVMで推定する．推定結果とacculacyを出力する．

3 出力

出力結果の抜粋を以下に示す．

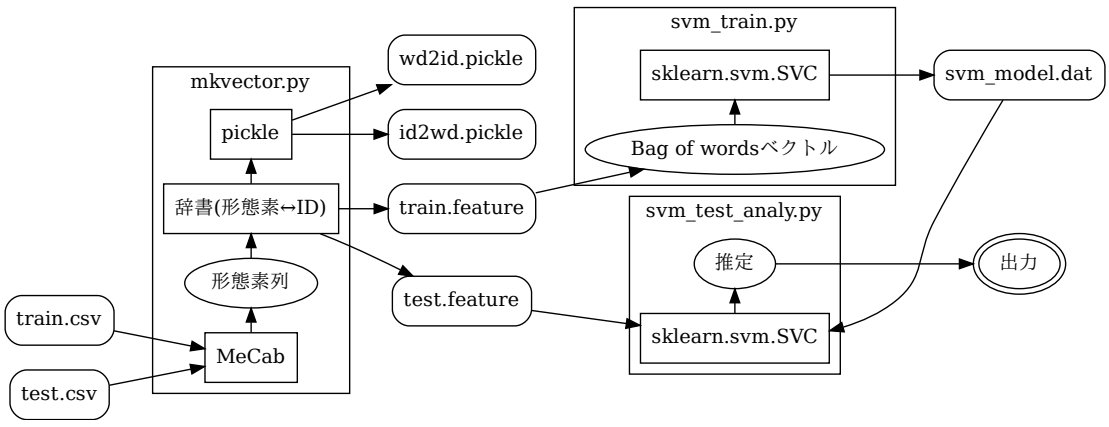

図1　判別器の構成

```
num_axis= 3010
shape (175, 3010)
correct, estimated
0        0       a, 婆さんはどこからとり出したか、眼をつぶった妙子の顔の先へ、一挺のナイフを突きつけました。
1        1       e, 私は仕方がないので母親に貰ったお小遣いをふんぱつして、人力車に乗りました。
2        2       m, 小川君は好奇心が起って溜まらなくなった。
*******（中略）*******
2        2       m, 翌朝深淵の家へは医者が来たり、警部や巡査が来たりして、非常に雑※（「二点しんにょう＋鰊のつくり」、第 4 水準 2-89-93）した。
1        1       e, 数年以前から、いつもあんな苦し相な顔をして居ります。
0        0       a, 五
0        1       a, そうしてこれが出来なければ、勿論二度とお父さんの所へも、帰れなくなるのに違いありません。
acccuracy= 0.903
```

学習データに 3009 種類の形態素が含まれたことから，それ以外の形態素を未知のものとして 1 つの次元で表現し，特徴ベクトルは 3010 次元のベクトルとなっている．

テストデータは 175 個の 3010 次元ベクトルが集まった行列として SVM に入力され，その推定結果が 2 列目に，正解ラベルが 1 列目に出力されている．推定の結果，acculacy は 0.903 であった．