

MRI脳画像のVBMデータを用いた機械学習による男女識別

1210386 安岡 薫 【知能情報学研究室】

1 はじめに

近年、機械学習を用いて脳構造画像から、性別や年齢などの個人属性を推定する研究が進められており、文献[1]では畳み込みニューラルネットワーク(CNN)を用いて脳構造画像から97%の精度で男女を識別できたとされている。しかし、CNNは識別の理由が明らかでないことが多く、医学への応用は説明可能性の観点での課題がある。そこで本研究では、MRI脳画像から計測した各脳部位の容積を対象に、様々な分類モデルで機械学習をすることで、男女識別の分析を行う。また識別に関連する脳部位を特定する。

2 実験

2.1 データセット

本研究では学習データとして脳ドック検診者の男性1728名、女性1272名のMRI脳画像からVBM (voxel-based morphometry)を用いて計測した部位ごとの値、計160個の特徴量を用いた。またテストデータとして学習データとは異なる男性420名、女性330名から学習データと同様の特徴量を用いた。

2.2 前処理

実験の前処理として、計測した被験者一人あたり160個の特徴量は、各人のICV(頭蓋内容積)により除算する。これにより、各人の脳の大きさにより補正する。更にデータセットに対してスケーリングを行う。ここで言うスケーリングは正規化(1)と、標準化(2)を用いる。これらは以下の式により実現する。

$$x_{\text{new}}^i = \frac{x^i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

$$x_{\text{new}}^i = \frac{x^i - \mu}{\sigma} \quad (2)$$

(x^i : 元の値, 被験者 $i \in \{1, 2, \dots, 3000\}$)

μ : 平均値, σ : 標準偏差)

本研究では上記の方法によりスケーリングした状態と、スケーリングをしない状態の3条件で実験を行う。

2.3 実験手順

本実験では、準備したデータセットを用いて様々な分類モデルで機械学習を行う。その際に設定したハイパーパラメータによる影響、識別精度を比較する。各分類モデルの識別率を向上させるためにハイパーパラメータを設定する。ハイパーパラメータは、グリッドサーチにより探索した値で機械学習を行った結果、最も精度が良いものを選択する。SVM(Support Vector Machine)のハイパーパラメータの探索範囲は文献[2]を参考にしてい

表1 識別精度(%)

前処理	clf1	clf2	clf3	clf4	clf5	clf6	clf7	clf8	avg.
なし	87.47	88.13	87.87	87.20	87.33	87.87	82.80	71.87	85.07
正規化	88.40	88.13	87.07	88.13	87.47	89.47	82.67	79.73	86.38
標準化	88.13	88.00	87.73	87.87	87.47	90.27	82.80	80.13	86.55

clf1:ロジスティック回帰(正則化なし), clf2:ロジスティック回帰(L1正則化),
clf3:ロジスティック回帰(L2正則化), clf4:ロジスティック回帰(ElasticNet),
clf5:SVM(linear), clf6:SVM(rbf), clf7:Random Forest, clf8:k近傍法

表2 回帰係数

説明変数	x_{137}	x_{87}		x_{160}
係数	1.103	-0.7520		0.0000

(x_{137} :第三脳室, x_{87} :左鳥距皮質, x_{160} :左島皮質)

る。ハイパーパラメータを設定し、機械学習を行った各分類モデルを用いてテストデータに対して識別を行い、それぞれの結果を比較する。そしてロジスティック回帰のそれぞれの説明変数に対する回帰係数から識別に影響を与える特徴量を求める。

2.4 実験結果

各分類モデルのテストデータに対する男女識別の結果を表1に示す。標準化を行った場合のSVM(rbf)が90.27%と最も識別率が高く、前処理の条件ごとの8つの分類モデルのテスト結果の平均値においても、標準化を行った場合が86.55%と最も高い結果となった。またロジスティック回帰における説明変数の回帰係数から、第三脳室や左鳥距皮質などが、男女識別に影響を与えている傾向が見られた。ここではその一例として前処理(標準化)した場合のロジスティック回帰(L1正則化)の説明変数の回帰係数を、絶対値の降順で並べたものの一部を表2に示す。

3 おわりに

本研究ではMRI脳画像からVBMを用いて計測した容積を対象に機械学習を行い男女識別を行った。その結果、標準化したデータセットに対しSVM(rbf)で機械学習をし、男女識別を行った場合が最も高い識別率となった。またロジスティック回帰の説明変数の回帰係数から男女識別に影響を与えた部位を特定することができた。

参考文献

- [1] S.Yoshida, et al., Prediction of human characteristics from brain structural MRI using CNN, ISFT 2019.
- [2] Chih-Wei Hsu, et al., A Practical Guide to Support Vector Classification, National Taiwan University, 2003.