深層学習による著者の属性推定

M2021SS001 安部沙桜里
指導教員：松田眞一

1　はじめに

先行研究にあたる財津・金 [11, 12] では，ブログの文章を対象に，ランダムフォレスト法（Random Forest）およびサポートベクターマシン（Support Vector Machine)を用いて文章の書き方の特徴を調べることにより，性別および年齢層の推定を検証している．

本研究では，渡邊・松田 [10] において，テキストデータに対して著者個人を推定する手法として高い判別精度が示された深層学習が，性別および年齢層という著者の属性を推定する場合における有効性を持つかどうか検証する．

2　深層学習 (DeepLearning)

深層学習とは，機械学習の技法の 1 つである．機械学習は，機械の計算力や高速処理能力を基に，大量の訓練データを使って学習する．これにより複雑な問題を YES か NO で答えられるパターン認識の問題に置き換え規則性を見つけ出し，モデルを構築する．その学習したモデルを用いて，問題となる未知のデータに対して適切だと考えられるパターンを予測することで解を得ることができる．この機械学習の技法に，ニューラルネットワークというものがある．これは，人間の脳にある神経細胞（ニューロン）の繋がりを数理モデル化したもので，データを与える入力層，結果を出す出力層，それらの間の層である中間層で構成されている．ニューラルネットワークには，パラメータ学習が進まなくなる勾配消失問題などの技術的な問題によって，十分な学習ができず，良い精度が出ないという欠点があった．しかし，トロント大学の G.Hinton 教授らがオートエンコーダ（自己符号化器）の深層化に成功し，研究が進んだ．これにより，中間層を増やした多層ニューラルネットワークによる学習のモデルの推定が実現可能となり，この多層構造のニューラルネットワークを基に強化学習などの技法を取り入れたものが，深層学習と呼ばれるようになった．深層学習は，従来のニューラルネットワークよりも得られるモデルの予測精度を格段に向上させることができるようになった．画像認識や音声認識，自然言語処理を始めとしてさまざまな分野で活用されている．(北 [2]，巣籠 [8] 参照）

3　データについて

3.1　用いるデータ

本研究では，先行研究を参考に「アメーバブログ」[1] と「にほんブログ村」[6] から抽出した文章を対象に分析を行った．ブログは 2 つの性別（男性，女性）× 6 つの年齢層（10 代から 60 代）の計 12 グループにおいて各 20 名，計 240 名分を選定した．具体的には，例えば 20 代男性の場合,「アメーバブログ」については,「芸能人・有名人ブログ」のうち，所属事務所のサイト等で性別と生年月日が判明している者，または「一般人ブログ」のうち，ブログ内で性別と年齢が判断できる者のブログを選定し，「にほんブログ村」については,「その他日記・20 代男性日記」を対象に，ブログ内で性別と年齢が判別できる者のブログを選定した．また，書き手の年齢層を揃えるために，記事は 2018 年から 2022 年までに投稿されたものとし，記事投稿時点の年齢を基にグループを振り分けた．

3.2　データの処理

選定したブログから抽出した記事の文章には，解析には不要な情報が含まれているため，以下のように文章のクリーニング作業を行った．

1. 文字化けを起こす絵文字や環境依存文字などを同じ意味になる文字に置き換え，または削除
2. 顔文字，URL など不要なものを削除.
3. 一部の記事における，文章の前後に挿入された全角空白を削除.

クリーニングを終えたテキストデータを基に，1 つのブログにつき 1 テキスト，総数 240 個のテキストデータを作成した．このテキストデータは，1000 文字を基準とし，1000 文字以降の意味が通じる文の末尾を区切りとした．

3.3　形態素解析

日本語による文章を統計的に解析するためには，文章情報を単語の出現頻度や品詞の情報といった解析可能な数値データへと変換する必要がある．しかし，英語を含む多くの言語は文章内の単語の区切りがあらかじめ明白であるのに対して，日本語の文章は，まず文節を得なければならない．これに関しては，コンピュータによって判別を行う，自然言語処理技術の形態素解析を用いた．形態素とは言語学で用いられる専門用語であり，言語における意味を持つ最小の単位とされている．文を形態素の単位まで分割することを形態素解析という．

本研究の形態素解析には，形態素解析フリーソフト「MeCab」[4] が実装されている，多言語テキストマイニングツール「MTMineR」[5] を使用した．

3.4　変数

本研究では，財津・金 [11, 12] を参考に，性別および年齢層それぞれの判別に有効である文体的特徴をテキストデータごとに算出し，変数とした．

3.4.1　性別推定の文体的特徴

性別の推定に用いた変数は，以下の 2 通りである．

性別推定の文体的特徴 A

- 漢字，平仮名，片仮名の全文字種における相対度数
- 名詞の全品詞種における相対度数

- 動詞（自立語），動詞（非自立語），形容詞（自立語），接続助詞，助詞（連体化），感動詞の度数
- 接続助詞「し」の度数
- 助動詞「なかっ」の度数
- 読点「、」の度数
- 小書き文字「っ」，小書き文字「ゃ」の度数
- 文字「私」，文字「僕」の度数

性別推定の文体的特徴 B　性別推定の文体的特徴 A のうち，動詞（自立語），動詞（非自立語），形容詞（自立語），接続助詞，助詞（連体化），感動詞の度数を相対度数にした組み合わせである．

3.4.2　年齢層推定の文体的特徴

年齢層の推定に用いた変数は，以下の 2 通りである．

年齢層推定の文体的特徴 A

- 名詞（一般），名詞（接尾-助数詞），名詞（接尾-助動詞語幹）の度数
- 係助詞「は」＋読点「、」の度数
- 助動詞「です」＋接続助詞「けど」の度数
- 副詞「ずっと」の度数
- 品詞の bigram の，名詞－名詞，名詞 (数) －名詞 (接尾-助数詞)，記号－名詞，助詞（連体化）－名詞（一般），助動詞－記号，助動詞－形容詞，副詞－副詞の度数

年齢層推定の文体的特徴 B　年齢層推定の文体的特徴 A のうち，名詞（一般），名詞（接尾-助数詞），名詞（接尾-助動詞語幹）の度数，品詞の bigram の，名詞－名詞，名詞 (数) －名詞 (接尾-助数詞)，記号－名詞，助詞（連体化）－名詞（一般），助動詞－記号，助動詞－形容詞，副詞－副詞の度数を相対度数にした組み合わせである．

3.4.3　n-gram

n-gram とは単語，または品詞が n 個繋がった組み合わせのことをいう．$n = 1$ の場合 unigram，$n = 2$ の場合 bigram，$n = 3$ の場合 trigram という．

4　検証

4.1　検証方法

本研究では交差検証として財津・金 [5,6] と同様，Leave One Out 法を用いて性別および年齢層の推定精度の検証を行った．Leave One Out 法では，データ全体から 1 つのデータのみをテストデータとして取り出し，残りを学習データに用いる．後に，テストデータと学習データを 1 個ずつ入れ替えて繰り返し，すべてのデータがテストデータになるように検証を行うものである．

4.2　評価指標

性別推定における評価には，男性と女性を対象とした 2 値分類を行った．年齢層推定には 6 つの年齢層を一度にまとめて分類する多クラス分類と，6 つの年齢層を 2 分割したものを対象とした 2 値分類を行った．6 つの年齢層を 2 分割した分割区分は以下の 5 通りである．

1. 「10 代」vs.「20 代から 60 代」
2. 「10 代から 20 代」vs.「30 代から 60 代」
3. 「10 代から 30 代」vs.「40 代から 60 代」
4. 「10 代から 40 代」vs.「50 代から 60 代」
5. 「10 代から 50 代」vs.「60 代」

4.3　2 値分類

ある特定のクラスを C とした時，各データに対する予測結果は以下の表 1 の分割表で示される

表 1　分類結果の分割表

		分類結果	
		C である	C でない
データ	C に属する	a	c
	C に属さない	b	d

$$\text{正解率}: \frac{a+d}{a+b+c+d} \tag{1}$$

$$\text{精度}: P = \frac{a}{a+b} \tag{2}$$

$$\text{再現率}: R = \frac{a}{a+c} \tag{3}$$

$$F\text{ 値}: F = \frac{2 \times P \times R}{P+R} \tag{4}$$

正解率とは，すべてのデータの中で正しく推定したかを示す指標であり式 (1) で求めるとする．精度 P は，C と推定して実際に C であった割合を示し，再現率 R は，C を C と正しく推定した割合であり，それぞれ式 (2)，式 (3) で求めるとする．なお，精度と再現率はトレードオフの関係にある．このため，精度と再現率の両者を折衷した評価指標として F 値がある．本研究では，式 (4) より算出した．(高村・奥村 [9] 参照)

4.4　多クラス分類

3 つ以上のクラスの中から 1 つを選択する多クラス分類の場合は，一つの多値分類問題として評価した．各データに対して，正しいクラスを予測できれば正解，そうでない場合は不正解とし，式 (5) で示される正解率を評価指標とした．(高村・奥村 [9] 参照)

$$\text{正解率}: \frac{\text{(正解した評価事例)}}{\text{(評価事例数)}} \tag{5}$$

4.5　平均二乗誤差 (Mean Squared Error)

平均二乗誤差 (Mean Squared Error) が小さいほど，誤差が少ないモデルであると言える．本研究では，実際の値を y_i，予測値を $\hat{y}_i$，データの総数を n とした時，以下の式 (6) より算出した．

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (6)$$

平均二乗誤差の算出には，真値と予測値の差を次のように定めた．予測値が真値と同じ場合は 0，真値より 1 つ下または 1 つ上の分類として判別された場合は 1，真値より 2 つ下または 2 つ上の分類として判別された場合は 2 とする．

5 分析結果

深層学習を行うにあたり，統計ソフト R の 'h2o' パッケージ [7] にて実装した．学習回数は 10000 回とし，活性化関数は Rectifier とし，その他の各種パラメータはデフォルトのままで検証を行った．また，分類法内で使われる乱数に伴い評価が異なることがあるため，Leave One Out 法によるモデルの学習と評価の実験を，5 回繰り返した評価指標の算術平均を最終的な評価とした．

5.1 性別推定

3.4.1 節にて示した性別推定の文体的特徴 A と性別推定の文体的特徴 B，それぞれの変数 17 項目を基に，240 個のデータに対して推定精度を検証した．2 値分類による結果（正解率，精度，再現率，F 値）は，それぞれ表 2，表 3 である．

結果より，性別推定の文体的特徴 B の方が推定精度の高いモデルが構築できたと考えられる．ただし，先行研究の財津・金 [11] において，ランダムフォレスト法による検証では，文体的特徴 A と同様の文体的特徴を変数とする正解率が 0.84，F 値に関しては男性で 0.863，女性で 0.857 が得られている．サポートベクターマシンによる検証では，正解率が 0.71，F 値に関しては男性で 0.695，女性で 0.724 が得られているため，いずれの手法と比べても推定精度が低い結果となった．

表 2 性別推定の推定精度 (文体的特徴 A)

分類	精度	再現率	F 値
男性	0.664	0.595	0.627
女性	0.633	0.698	0.664
正解率	0.647		

表 3 性別推定の推定精度 (文体的特徴 B)

分類	精度	再現率	F 値
男性	0.655	0.630	0.642
女性	0.644	0.668	0.656
正解率	0.649		

5.2 年齢層推定

5.2.1 年齢層推定（分割なし）

3.4.2 節にて示した年齢層推定の文体的特徴 A と年齢層推定の文体的特徴 B，それぞれの変数 13 項目を基に，240 個のデータに対して推定精度を検証した．多クラス分類による結果は年齢層推定の文体的特徴 A では，正解率が 0.226，MSE が 2.145 となった．年齢層推定の文体的特徴 B では，正解率が 0.222，MSE が 2.098 となった．正解率のみを見ると年齢層推定の文体的特徴 A の方が推定精度が高いと言えるが，データがどれだけ離れた年代として推定されるかを考えた MSE においては，文体的特徴 B の方が推定精度が勝る結果になった．

5.2.2 年齢層推定（2 分割）

3.4.2 節にて示した年齢層推定の文体的特徴 A と年齢層推定の文体的特徴 B，それぞれの変数 13 項目を基に，240 個のデータに対して推定精度を検証した．2 値分類による結果のうち，紙面の都合上，文体的特徴 A・文体的特徴 B 共にもっとも正解率の高かった分割区分 1.「10 代」vs.「20 代から 60 代」と，正解率の低かった区分の一つとして分割区分 3.「10 代から 30 代」vs.「40 代から 60 代」の結果のみ示す．

文体的特徴 A のグループでは,「分割区分 1.」にて 0.803 のもっとも高い推定精度が得られた．ただし，精度および再現率においては，データ数の少ない 10 代では F 値は 0.305 と低い結果になった．文体的特徴 B のグループでの「分割区分 1. の精度」は，0.796 であり，文体的特徴 A のグループよりも低い精度となっている．

表 4 「10 代」vs.「20 代− 60 代」文体的特徴 A

分類	精度	再現率	F 値
10 代	0.369	0.260	0.305
20 代-60 代	0.860	0.911	0.885
正解率	0.803		

表 5 「10 代− 30 代」vs.「40 代− 60 代」文体的特徴 A

分類	精度	再現率	F 値
10 代− 30 代	0.588	0.642	0.614
40 代− 60 代	0.606	0.550	0.576
正解率	0.596		

表 6 「10 代」vs.「20 代− 60 代」文体的特徴 B

分類	精度	再現率	F 値
10 代	0.347	0.255	0.294
20 代− 60 代	0.858	0.904	0.881
正解率	0.796		

表 7 「10 代− 30 代」vs.「40 代− 60 代」文体的特徴 B

分類	精度	再現率	F 値
10 代− 30 代	0.574	0.623	0.597
40 代− 60 代	0.588	0.537	0.561
正解率	0.580		

6 要素の追加

より精度の高いモデルを構築するため，3.4.1 節および 3.4.2 節以外の除外された文体的特徴を変数として追加することによって，推定精度が向上するか検証した．基本の変数をいずれも文体的特徴 B とし，単語および品詞の n-gram(unigram，bigram，trigram) の相対度数，読点前・

句点前の文字の相対度数，文の長さの度数をそれぞれ変数として追加した．

6.1 性別推定

性別推定における文体的特徴を変数として追加した時の推定精度のうち，正解率が上がった要素についての結果を，表 8 に示す．特に品詞の bigram を変数として追加した場合は正解率 0.700，男性 F 値 0.633，女性 F 値 0.746 となり，先行研究におけるサポートベクターマシンによる検証による正解率 0.710，男性 F 値 0.695，女性 F 値 0.724 に近しいまたは高い結果となった．

表 8 性別推定の文体的特徴 B ＋追加文体的特徴

追加した要素	性別	精度	再現率	F 値	正解率
単語の trigram	男性	0.705	0.570	0.630	0.666
	女性	0.639	0.762	0.695	
品詞の unigram	男性	0.775	0.545	0.640	0.693
	女性	0.649	0.842	0.733	
品詞の bigram	男性	0.814	0.518	0.633	0.700
	女性	0.647	0.882	0.746	
読点前の文字	男性	0.671	0.615	0.642	0.657
	女性	0.645	0.698	0.670	
句点前の文字	男性	0.721	0.623	0.668	0.691
	女性	0.668	0.758	0.710	

6.2 年齢層推定（分割なし）

分割なしの年齢層推定においても，文体的特徴を変数として追加した時の推定精度のうち，正解率が上がった要素についての結果を，表 9 にて示す．特に単語の unigram を変数として追加した場合は，正解率が 0.298 となり文体的特徴 B のみの精度よりも約 7%上がり，MSE も 2.00 を下回る結果となった．

表 9 年齢層推定の文体的特徴 B ＋追加の文体的特徴

追加した要素	正解率	MSE
単語の unigram	0.298	1.806
単語の bigram	0.227	2.067
品詞の bigram	0.286	2.025
読点前の文字	0.236	2.050
文の長さ	0.224	1.997

7 パラメータチューニング

深層学習はさまざまなモデルパラメータを調節することで，より高い精度のモデルを得ることができる．紙面の都合上，ドロップアウトによる結果のみを示す．ドロップアウトとは学習用データに対してユニットを一定の割合で無効化することで過学習を防ぐ方法である．これを有効にしデフォルトの 0.5 に値を設定して検証した．性別推定の正解率は 0.613 と推定精度が低下したが，分割なしの年齢層の正解率は 0.318，MSE は 1.798 となり推定精度が向上した．

8 まとめ

性別推定について，先行研究である財津・金 [11] ではランダムフォレスト法にて 80%以上の推定精度が得られたが，今回の研究ではそれに劣る結果となった．

分割なしの年齢層推定は，ピンポイントに推定することは困難であると考えられるが，MSE の結果から 2 区分以内に分類されやすいことが分かる．また，特定の年齢層において隣同士の年齢層区分に分類された数を追加した割合では，20 代のデータに対して，10 代〜30 代に分類されたデータの割合は，文体的特徴 B での検証では 0.570，文体的特徴 B+単語の unigram での検証では 0.735 と大幅に高くなった．

一方，他の文体的特徴を変数として追加することやパラメータチューニングによって，性別推定および年齢層推定の精度は向上した．これは深層学習が変数からデータの背景にあるルールやパターンを学習するため，検討する変数の候補を広くした方がモデルの推定精度の向上に繋がるためと考えられる．

9 おわりに

本研究では，性別および年齢層の推定において，深層学習の有効性について検証した．先行研究よりも深層学習による検証は精度の低い結果となったが，より多くの変数を検討することやパラメータチューニングにより精度の向上がみられた．今回用いた手法とは異なる深層学習の分類手法を用いることでより精度の高いモデルを構築することができる可能性もある．

参考文献

[1] アメーバブログ：https://ameblo.jp/ (2022/5 閲覧)

[2] 北栄輔 :『R で学ぶデータサイエンス　—データマイニングの基礎から深層学習まで—』，オーム社，2018.

[3] 黒橋禎夫・柴田知秀 :『自然言語処理概論』，サイエンス社，2016.

[4] MeCab：http://taku910.github.io/mecab/ (2022/5 閲覧)

[5] MTMineR : https://mjin.doshisha.ac.jp/MTMineR/mt.html (2022/8 閲覧)

[6] にほんブログ村：https://blogmura.com/ (2022/5 閲覧)

[7] Package‘h2o’：https://cran.r-project.org/web/packages/h2o/h2o.pdf (2022/5 閲覧)

[8] 巣籠悠輔 :『Deep Learning Java プログラミング　深層学習の理論と実装』，インプレス，2016.

[9] 高村大也・奥村学（監修）:『言語処理のための機械学習入門』，コロナ社，2012.

[10] 渡邊翔・松田眞一 :『深層学習を用いた文章の書き手の同定』，南山大学紀要『アカデミア』理工学編，18，1-13，2018.

[11] 財津亘・金明哲 :『ランダムフォレストによる著者の性別推定　—犯罪者プロファイリング実現に向けた検討—』，情報知識学会誌，27[3]，261-274，2017.

[12] 財津亘・金明哲 :『機械学習を用いた著者の年齢層推定　—犯罪者プロファイリング実現に向けて—』，情報知識学会誌，59[2]，57-65，2018.