文章の書き手の同定における深層学習の適応に関する研究

M2016SS011 渡邊翔
指導教員：松田眞一

1　はじめに

三品 [10] において小説やブログの文章の書き手の同定における分類法の精度比較を行っており，RandomForest 法や MART 法などを用いて文章の書き方の特徴を調べることにより予測している．そこで私は近年画像認識などで話題である深層学習に着目し，分類法に深層学習を追加して各分類法の精度比較を行い深層学習の有効性を検証していく．本研究においても同じ著者の小説，ブログを用いてモデルの検証をし，三品 [10] において特に高い判別精度を示した品詞の n-gram 分布と読点前の文字の分布を比較対象とする．

2　著者推定とは

文章データをそのまま解析するのは不可能であるので，単語の出現頻度や長さなどを用いて数値データに変換し文字の分布から著者の特徴を抽出しコンピュータで自動判別させることにより著者の推定ができるようになる．これらの研究は自動スパムメール判定などにも応用できる．(石田 [4] 参照)

3　分類手法

以下に精度比較に用いた分類手法を示す．深層学習以外の詳細については三品 [10] を参照のこと．

3.1　深層学習 (DeepLearning)

ニューラルネットワークとは人間の脳の神経の働きを真似たもので，入力層，中間層，出力層の 3 つの層からなるモデルでデータを学習し出力結果を出していた．しかしニューラルネットワークは変数が莫大に多いと良い精度を出しにくく，少なすぎる変数においても良い精度が出ないという欠点もあった．コンピュータの計算能力の向上や繰り返し計算による強化学習など学習アルゴリズムの改善により，より複雑なニューラルネットワークの実現ができるようになった．それが深層学習と呼ばれるようになった．深層学習は中間層の数を増やすことで多層構造のニューラルネットワークを再現した機械学習法である．多層化したことによりニューラルネットワークでは対応できなかった非線形な場合にも対応できるのが大きな利点である．各層には複数のユニットが存在し，入力層と出力層のユニット数は入力された変数，出力されるものの個数に相当する．中間層の層の数，ユニット数，学習回数などの決定は解析者の主観に委ねられるもので多くは用いるデータ依存になる．(岡谷 [11] 参照)

3.2　MART 法

Boosting 法の一種であり，ある分割規則に従い樹木を成長させていき作成した樹木の細かい枝の刈り込みを行い弱分類器とする．その弱分類器を逐次損失関数の傾きにより重みをつけて合成していくことで強分類器とする手法である．

3.3　RandomForest 法

アンサンブル学習法の一種であり，ランダムに抽出した変数から多数の決定木を弱分類器として作成する．その複数の決定木から分類問題では多数決をとることにより強分類器とする手法である．

3.4　Bagging 法

アンサンブル学習法の一種であり，ブートストラップと呼ばれる復元抽出法で作成される複数の学習データを用いて弱分類器を作成し多数決をとることにより強分類器とする手法である．

3.5　AdaBoost 法

複数の弱分類器を作成しそれらを合成することによって強分類器とする Boosting 法の一種であり，逐次学習データの調整をしながら弱分類器を作成し，誤り率の高い分類器に重みをつけて合成していくことで強分類器とする手法である．

4　文章データについて

4.1　用いる文章データ

小説データとして青空文庫 [1] や電子図書館 [3] からダウンロードできる著作権の切れている明治の文豪 (芥川龍之介，太宰治，泉鏡花，菊池寛，森鴎外，夏目漱石，岡本綺堂，佐々木味津三，島崎藤村，海野十三) の作品で，三品 [10] で用いられているものと同じ 20 編の作品を用いて各分類器の評価を行う．ブログデータも三品 [10] と同様に著者 5 人各 10 編の作品を用いる．

4.2　文章のクリーニング

青空文庫からダウンロードしたテキストファイルにはルビ (ふりがな) やタイトルのような文章の解析には不要な情報を削除したりする文章のクリーニング作業が必要となる．解析者によってクリーニングの基準が異なるため，本研究では以下のような作業を手作業で行った．

1. ルビやタイトルのような不要なものを削除する．
2. コンピュータで表示されない文字などを同じ意味になる単語に置き換え，解析ソフトの辞書に登録する．

3. 地の文以外の単独で現れる会話文を削除する．
4. 漢文や英文などを削除する．
5. 全角空白を半角空白に置換する．

1 のルビの削除には web[2] から入手でき，Windows のコマンドプロンプト上で実行することができる自動ルビ削除プログラム「delruby.exe」を用いた．

4.3 形態素解析

日本語の文章を解析するためには単語の出現頻度や品詞情報などを集計した数値データに変換する必要がある．しかし日本語の文章は英文のように単語の分かち書きがなされていないので，分かち書きを手作業で行うことが大変困難である．そこでコンピュータによる自然言語処理技術である形態素解析を用いる．形態素解析とは文章を単語の最小単位に分割し，必要に応じて単語の品詞情報も追加する方法である．本研究では形態素解析フリーソフトの MeCab を統計ソフト R 上で実行することができる RMeCab を用いる．(石田 [4] 参照)

4.4 変数について

金 [5, 6, 7, 8] において，単語の長さの分布，品詞の n-gram 分布などが書き手の特徴を表していると示されており，三品 [10] でもそれらの変数に着目し，書き手の同定における各分類法の精度評価を行っている．本研究ではそれらの変数の中で特に高い判別精度が示された品詞の n-gram 分布と読点前の文字の分布を扱う変数とする．著者や作品によってはサイズ (文字数) が異なるので，各変数については文章データそれぞれの相対頻度を用いることとする．

- n-gram 分布
  n-gram とは文字，単語，品詞情報が n 個繋がった形で表されることである．本研究では $n = 2$，すなわち bigram で表されるものを扱い，品詞情報に焦点を置いて集計をし，今回扱う変数とする．
  【例】
  $n = 2$，bigram の出現頻度を総数で割った相対頻度を以下の文を例にして表 1 に示す．
  例文：彼は急いでコンビニに向かった。
  彼 [名詞] は [助詞] 急い [動詞] で [助詞] コンビニ [名詞] に [助詞] 向かっ [動詞] た [助動詞]。[記号]

表 1 bigram の相対頻度表

n-gram	相対頻度
[助詞-動詞]	0.25
[助詞-名詞]	0.125
[助動詞-記号]	0.125
[動詞-助詞]	0.125
[動詞-助動詞]	0.125
[名詞-助詞]	0.25

- 読点前の文字の分布
  読点前の文字を集計し，それぞれの文字を相対頻度で表したものを今回扱う変数とする．また，ある一定の出現頻度以下の文字についてはその他の項目にまとめた．

5 モデル検証

5.1 検証方法

小説データでは著者 10 人各 20 編の作品のデータセットからサンプリングをすることにより学習データとテストデータに分割し，分類法の評価を行う．ブログデータでは著者 5 人各 10 編のデータセットで同様に分類法の評価を行う．学習データの標本サイズ (作品数) の違いによる判別精度を見るために，金・村上 [9] にならって標本サイズを S としたとき，学習データとして各著者から $(S-1, S-2, \cdots, 3)$ 個ずつランダムサンプリングし，それ以外のデータをテストデータとする．以上のように分類法内で使われる乱数やランダムサンプリングにより評価が異なることがあるので，モデルの学習と評価の実験を 100 回繰り返した評価指標の平均値を分類法の精度とする．

5.2 評価指標

判別種類はある 1 人の著者とその他の著者を対象とし，2 分類を行う 2 値判別と複数の著者を対象とし，複数の著者を 1 度にまとめて分類を行う多値判別である．同定結果の評価指標は正解率と F 値で表すとする．

5.2.1 2 値判別

ある対象の著者 $i(i = 1, 2, \cdots, n)$ とその他の著者をそれぞれ A_i，B_i とラベルをつけたグループを G_i とした時の 2 値判別の書き手の同定結果を集計した分割表は表 2 で表される．

表 2 2 値判別同定結果の分割表

G_i		分類結果	
		A_i	B_i
データ	A_i	a_i	c_i
	B_i	b_i	d_i

$$\text{再現率}：R_i = \frac{a_i}{a_i + c_i} \tag{1}$$

$$\text{精度 1}：P_i = \frac{a_i}{a_i + b_i} \tag{2}$$

$$\text{精度 2(正解率)}：P_i = \frac{a_i + d_i}{a_i + b_i + c_i + d_i} \tag{3}$$

A_i と判別されるべきものの内どれだけ正しく判別されたかを表す指標を再現率 R_i とし，式 (1) で求めるとする．また A_i と判別されたものの内どれだけ正しく判別されたかを表す指標を精度 P_i とする．本来ならば式 (2) で求められるが，本研究では三品 [10] と同様に少ない標本サイ

ズでの精度比較を行い，直接結果を比較するために式 (3) を今回扱う精度とし，正解率と呼ぶ．式 (2) の分母である $a_i + b_i$ が同定結果によっては 0 となりうることもあるからである．

再現率と正解率はどちらか一方が上がれば他方は下がるトレードオフの関係になっているので，再現率と正解率の調和平均である以下の式で求めたものを F 値とする．

$$F_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{4}$$

F 値は複数の値が出るので，式 (4) で求めた著者 n 人分の値を平均して出して値を 2 値判別における F 値とする．

5.2.2 多値判別

多値判別の場合，多値判別の同定結果の分割表から各著者 $i(i = 1, 2, \cdots, n)$ とその他の著者の 2 値判別に分割してから，それぞれの書き手の再現率と正解率を求める．それぞれの著者の再現率と正解率を平均して出した値を多値判別における再現率 $\hat{R}$，正解率 $\hat{P}$ とし，以下の式で求めるとする．

$$\text{再現率} : \hat{R} = \frac{1}{n} \sum_{i=1}^{n} \frac{a_i}{a_i + c_i} \tag{5}$$

$$\text{正解率} : \hat{P} = \frac{1}{n} \sum_{i=1}^{n} \frac{a_i + d_i}{a_i + b_i + c_i + d_i} \tag{6}$$

2 値判別と同様に次式で F 値を定義する．

$$F = \frac{2 \times \hat{P} \times \hat{R}}{\hat{P} + \hat{R}} \tag{7}$$

6 深層学習検証結果

深層学習は統計ソフト R の 'h2o' パッケージ [12] を用いて実装する．学習回数を 10000 回とし，その他のパラメータは R の関数のデフォルトのままで検証を行った．紙面の都合上 F 値のみ示し，各標本サイズにおけるグラフも一部抜粋したグラフのみを示す．深層学習以外の分類法の値は三品 [10] において検証結果の値が掲載されていなかったため，目視で確認し大まかな値を計測しグラフを再現した．また，ブログデータでは用いる記事や著者に違いがあり直接比較することが困難であるため深層学習の結果のみを示す．読点前の文字の分布は三品 [10] と同様に文字の出現頻度を小説データは 50 以上，ブログデータでは 3 以上を基準とし，それ以下の文字についてはその他の項目にまとめて集計した．

6.1 小説データ

n-gram 分布 (158 項目) の 2 値判別，多値判別を行い，多値判別の各分類法による比較グラフを図 1 に示す．2 値判別においては標本サイズが 3 の場合のみ深層学習の有効性は示されなかった．多値判別においては標本サイズが十分大きい場合には非常に高い判別精度を出していたが標本サイズが小さくなるにつれて精度の減少が目立つ結果になった．しかし標本サイズ 19 のように学習データが十分にある場合は最大 0.992 と非常に高い判別精度となった．

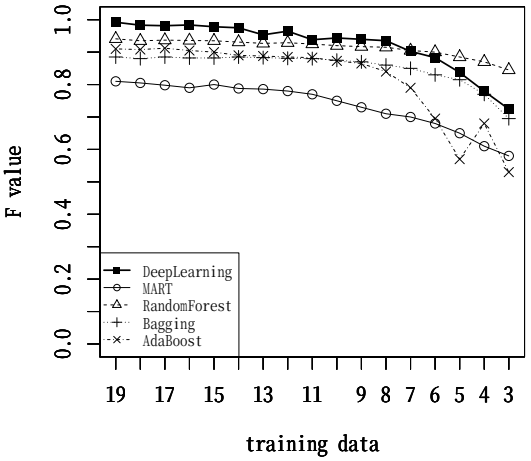

図 1 小説 n-gram 分布 F 値のグラフ (多値判別)

次に読点前の文字の分布 (33 項目) の 2 値判別，多値判別を行い，2 値判別の各分類手法との比較グラフを図 2 に示す．2 値判別，多値判別ともにすべての標本サイズにおいて深層学習の有効性が示された．2 値判別では MART 法が最も判別精度が高かったがその結果を大きく上回る結果となった．標本サイズ 19 で最大 0.998 と非常に高い判別精度となった．

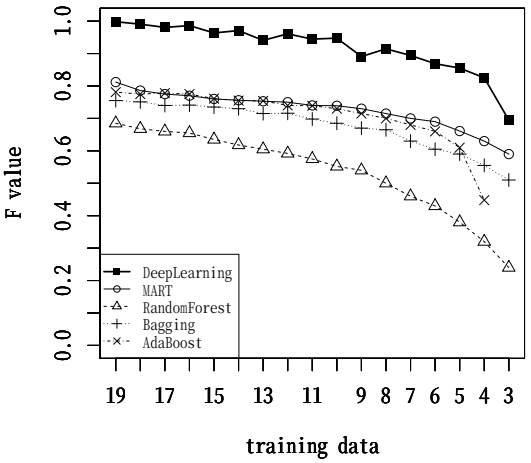

図 2 小説読点前の文字の分布 F 値のグラフ (2 値判別)

6.2 ブログデータ

ブログデータについても同様の変数 (n-gram 分布 113 項目，読点前の文字の分布 36 項目) を用いて 2 値判別，多値判別を行い，多値判別の深層学習の結果を表 3 に示す．両方の変数において同様の傾向が見られ，2 値判別ではすべての標本サイズにおいて深層学習の有効性が示され，多値判別においては標本サイズが小さい場合には三品 [10] の

RandomForest 法の方が高い有効性を示し深層学習の有効性は示されなかった．

表 3　ブログ n-gram 分布の正解率と F 値の値 (多値判別)

標本サイズ	9	8	7	6	5	4	3
F 値	0.972	0.960	0.957	0.899	0.873	0.793	0.800

7　パラメータチューニング

深層学習などの機械学習法には様々なパラメータが存在し，モデルに合わせたパラメータを調整 (チューニング) することでより高い精度を出したりコストの低いモデルを構築することが可能である．今回は学習回数，中間層の数，ユニットの数，ドロップアウトに焦点を当てて検証をし，紙面の都合上小説 n-gram 分布のみの結果を示す．

学習回数を 1000，100，10 と減らして検証をしたところ 1000 回と 100 回の間に精度の差が見られたので 1000 回程度で十分に精度を出すことができると考えた．以下の検証は時間コストを少なくするため学習回数を 1000 回に設定してチューニングを行った．

中間層の数を初期値の 2 から 1，3〜5 と変化させたところ 3 までは精度は上がったがそれ以上になると精度が下がる傾向がみられ，過学習を起こしてしまう可能性があると考えた．次に各ユニットの数を初期値の 200 から 100，50，25 と減らしていったところ 100 程度まで減らしても精度に大きな影響がないことが分かったが，50 まで減らすと大きく精度が下がる結果となった．

データを学習をする際に一定の確率でユニットを無効化することで過学習を防ぐドロップアウトという方法があり，これを有効にした上でデフォルト値である確率 0.5 に設定をして検証を行った．今まで中間層の数を増やすことにより精度を上げることができたが，ドロップアウトによりそれを大きく上回る結果が得られた．小説 n-gram 分布の標本サイズ 3 の場合で判別精度で 0.859 となった．学習回数 1000 回でのデフォルトの精度 0.741 と比べると約 11% も精度が上がった．このようにドロップアウトはモデルの過学習を抑制する方法として非常に有効であることが分かった．

読点前の文字の分布はチューニングによる精度の大きな変化はみられず，小説データにおいて特別チューニングをしなくても著者の十分な特徴を表すことができる変数であると考えられた．

8　まとめ

三品 [10] の結果では 2 値判別においては MART 法，多値判別において RandomForest 法が最も高い判別精度を出していたが，標本サイズが十分に大きいときにはともに深層学習の有効性が示された．三品 [10] で検証された MART 法と RandomForest 法は標本サイズの違いによる影響は受けにくく正解率や F 値も下がりにくい結果なっているのに対して，今回の深層学習の結果では標本サイズの減少により正解率や F 値の値の変化が大きく見られた．しかし標本サイズが小さい場合でも用いる文章データに合ったモデルにチューニングをすることで十分に精度を出せることも検証できた．従来の研究では判別種類に応じて分類手法を変える必要があったが，十分にデータを収集することが可能であるならば深層学習が 2 値判別でも多値判別でも非常に有効であることが検証できた．これらのことより深層学習は小説やブログをはじめ様々な文章の書き手の同定において有効な分類手法であるといえるだろう．

9　おわりに

本研究では標本サイズが十分に大きい場合に深層学習の有効性が示された．また，標本サイズが小さい場合でもモデルのチューニング次第ではさらに精度を高めることができる可能性も示された．今回行ったチューニング作業はほんの一部でしかない．さらに複雑なパラメータを動かしたり，学習方法を変えることでより良いモデルを構築することができるかもしれない．

参考文献

[1] 青空文庫，http://www.aozora.gr.jp/，2017/9/7 閲覧．

[2] 青空文庫のテキストからルビを削除するには -AOKIDS Home Page，http://www.aokids.jp/others/delruby.html，2017/9/11 閲覧．

[3] 電子図書館，http://www.eonet.ne.jp/~log-inn/，2017/10/12 閲覧．

[4] 石田基広：『R によるテキストマイニング入門』，森北出版，2008.

[5] 金明哲：読点の情報に基づく文献の分類，情報処理学会『全国大会講演論文集』第 46 回人工知能及び認知科学，131-132，1993.

[6] 金明哲：日本語における単語の長さの分布と文章の著者，『社会情報』**5**(2)，13-21，1996.

[7] 金明哲：助詞の分布における書き手の特徴に関する計量分析，『社会情報』**11**(2)，15-23，2002.

[8] 金明哲：分節パターンに基づいた文書の書き手の識別，『行動計量学』**40**(1)，17-28，2013.

[9] 金明哲・村上征勝：ランダムフォレスト法による文章の書き手の同定，『統計数理』**55**(2)，255-268，2007.

[10] 三品光平：『文章の書き手の同定における分類法の精度検証の研究』，南山大学大学院数理情報研究科修士論文，2013.

[11] 岡谷貴之：機械学習プロフェッショナルシリーズ『深層学習』，講談社，2017.

[12] Package‘h2o’，https://cran.r-project.org/web/packages/h2o/h2o.pdf，2017/7/1 閲覧．