深層学習を用いた文章の書き手の同定

渡邊 翔* 松田 眞一†
E-Mail: matsu@nanzan-u.ac.jp

三品・松田 [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いたが，2 値判別では MART 法，多値判別では RandomForest 法がそれぞれ良い結果を出していた．本論文では近年画像認識などで話題である深層学習に着目し，三品・松田 [10] において特に高い判別精度を出した品詞の n-gram 分布と読点前の文字の分布を変数として用いて各分類法の精度比較を行う．検証の結果，2 値判別，多値判別ともに標本サイズが十分に大きい時には深層学習が大幅に有効であることが分かったが，標本サイズの減少により判別精度が下がりやすい結果になった．しかし，深層学習は標本サイズが小さい場合でもモデルに合ったパラメータをチューニングすることで十分な精度を出すことができ，様々な文章に対して適応できることが検証できた．

1 はじめに

テキストを計量的に扱い，それに基づいて著者不明の文章の書き手の同定をしたり，著者の真贋を判定したりする計量文献学は，欧米で 100 年以上の歴史がある．日本では，日本語の文章が英文と違って分かち書き（単語と単語の間にスペースを入れる書き方）をしないことやひらがな，カタカナ，漢字など文字の種類が多いことから研究が遅れていた．そういった中で村上征勝氏は手作業で整理したデータを元に多くの結果を残してきた.（村上 [11] 参照）その後，計算機の発達に伴い，形態素解析技術が進歩し，文章を自動的に単語に分け品詞の分類が行えるようになった．

本論文の直接的な先行研究は三品・松田 [10] であるが，それは金・村上 [9] からの派生である．金・村上 [9] は形態素解析を取り入れた上で，当時判別手法として主流となってきた RandomForest 法を小説，日記，作文の文章の書き手の同定問題に組み入れて既存の方法と比較を行い，RandomForest 法が優れていることを示した．三品・松田 [10] では金・村上 [9] の日記や作文と同様のデータが入手できなかったため，同じ設定で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いる．分類法としては MART 法を追加し，結果として多値判別では RandomForest 法が相変わらず優れているものの 2 値判別では MART 法の方が優れていることを突き止めた．

本論文で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．本論文においても三品・松田 [10] と同じ著者の小説，ブログを用いてモデルの検証をし，三品・松田 [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．

*南山大学大学院理工学研究科システム数理専攻
†南山大学理工学部システム数理学科

2 著者推定とは

著者推定とは，著者の書き方の特徴に基づいて書き手を推定することである．最近ではインターネットの普及により電子化された書籍の入手が容易となり，コンピュータが自然言語を扱えるようになったことで文章を自動で計測し数値化できるようになった．著者の特徴を数値として抽出することで与えられた文章データから著者推定ができるようになった．特に機械学習を用いた方法が主流であり，自動スパムメール判定にも応用されている．(石田 [4] 参照)

3 分類法

本論文では深層学習のみを新たに計算し，三品・松田 [10] において比較検証された MART 法，RandomForest 法，Bagging 法，AdaBoost 法の結果と直接比較する．深層学習以外の分類法の詳細については三品・松田 [10] を参照のこと．

3.1 深層学習 (DeepLearning) について

ニューラルネットワーク（多層パーセプトロン）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があった．中間層のもっと多いニューラルネットワークはモデルの推定が困難であるとされていたが，その理論が進み実現可能となった．それがトロント大学の G.Hinton 氏によって提案されたディープ・オートエンコーダという手法である．この手法と従来からあるバックプロパゲーション（誤差逆伝播法）を組み合わせて学習を行うことで中間層が 2 層以上でも推定を可能にしたとされている．さらに，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，この多層構造のニューラルネットワークが深層学習と呼ばれるようになった．深層学習は中間層が多い分，より複雑な構造の判別が可能となり，従来のニューラルネットワークより格段に良い性能を示す例が相次いだ．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 [12] 参照)

4 文章データについて

4.1 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ため，それを手作業で行う困難さがある．そこで，コンピュータによる分かち書き，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自動で単語の

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技術の研究である．本論文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)

4.2　用いる文章データ

小説データとして青空文庫 [2] 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から昭和初期の文豪の作品で，三品・松田 [10] で用いられているものと同じ作品，著者 10 人各 20 編の作品‡を用いて各分類法の評価を行う．用いる小説リストを表 1 に示す．ブログデータも三品・松田 [10] と同様に著者 5 人各 10 編の作品を用いる．

表 1: 用いる小説リスト

著者名	作品名
芥川龍之介	或阿呆の一生，或日の大石内蔵助，偸盗，玄鶴山房，歯車，母，春，芋粥，地獄変，彼，奇怪な再会，煙管，毛利先生，おぎん，お律と子等と，路上，将軍，少年，杜子春，保吉の手帳から
太宰治	愛と美について，兄たち，老ハイデルベルヒ，美少女，小さいアルバム，地球図，千代女，断崖の錯覚，男女同権，誰，誰も知らぬ，富嶽百景，服装に就いて，玩具，逆行，八十八夜，恥，花吹雪，春の盗賊，皮膚と心
泉鏡花	売色鴨南蛮，鷭狩，絵本の春，縁結び，伯爵の釵，半島の一奇抄，遺稿，化鳥，木の子説法，小春の狐，高野聖，国貞えがく，草迷宮，眉かくしの霊，女客，婦系図，怨霊借用，七宝の柱，歌行燈，薬草取
菊池寛	仇討禁止令，仇討三態，芥川の事ども，青木の出京，M侯爵と写真師，勲章を貰う話，身投げ救助業，三浦右衛門の最後，無名作家の日記，恩を返す話，恩讐の彼方に，大島ができる話，乱世，船医の立場，勝負事，俊寛，出世，忠直卿行状記，若杉裁判長，ゼラール中尉
森鴎外	阿部一族，普請中，二人の友，雁，護持院原の敵討，百物語，じいさんばあさん，かのように，寒山拾得，カズイスチカ，妄想，鶏，最後の一句，堺事件，杯,，山椒大夫，青年，高瀬舟，ヰタ・セクスアリス
夏目漱石	坊ちゃん，虞美人草 1，虞美人草 2，彼岸過迄 1，彼岸過迄 2，一夜，薤露行，行人 1，行人 2，琴のそら音，草枕，倫敦塔，幻影の盾，門，三四郎，趣味の遺伝，それから，吾輩は猫である 1，吾輩は猫である 2，吾輩は猫である 3
岡本綺堂	穴，白猿伝・其他 (唐)，半鐘の怪，石灯籠，異妖変，影を踏まれた女，勘平の死，箕輪心中，お化け師匠，お文の魂，青蛙堂鬼談，宣室志 (唐)，心中浪華の春雨，捜神後記 (六朝)，捜神記 (六朝)，鳥辺山心中，寄席と芝居と，湯屋の二階，酉陽雑俎 (唐)，ゆず湯
佐々木味津三	青眉の女，足のある幽霊，血染めの手形，達磨を好く遊女，毒色のくちびる，笛の秘密，へび使い小町，袈裟切り太夫，曲芸三人娘，京人形大尽，卍のいれずみ，明月一夜騒動，身代わり花嫁，耳のない浪人，村正騒動，生首の進物，七化け役者，南蛮幽霊，なぞの八卦見，千柿の鍔
島崎藤村	嵐，ある女の生涯，朝飯，分配，千曲川のスケッチ，鳥帽子山麓の牧場，船，岩石の間，家 (上)，伊香保土産，旧主人，芽生，桃の雫，並木，伸び支度，幼き日，三人，刺繍，食堂，藁草履
海野十三	暗号音盤事件，暗号の役割，あの世から便りをする話，ある宇宙塵の秘密，英本土上陸作戦前夜，骸骨館，生きている腸，鞄らしくない鞄，化学者と夜店商人，鍵から抜け出した女，快星ガン，海底都市，火薬船，鬼仏洞事件，奇賊悲願，奇賊は支払う，恐ろしき通夜，宇宙尖兵，宇宙戦隊，宇宙の迷子

‡太宰治に関しては三品・松田 [10] で 21 作品となっている．本論文では 20 作品にするため「二十世紀旗手」を外した．

4.3 文章のクリーニング

青空文庫 [2] 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も存在する．そのような不要な情報などを削除することを文章のクリーニング作業という．解析者によってクリーニングの基準が異なるため，三品・松田 [10] と同様に手順を明示する必要がある．本論文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. 漢文や英文を含んだ文を削除する．
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[1] 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．

4.4 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・松田 [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．本論文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たものである．本論文では $n = 2$，すなわち bigram で表されるものを扱い，品詞情報に焦点を当てて集計し，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 2 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 2: bigram の相対頻度表

bigram	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．なお，ある一定の出現頻度未満の文字についてはその他の項目としてまとめた．

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.（三品・松田 [10] 参照）

5.2.1　2 値判別の評価指標

ある対象の著者 $i(i = 1, 2, \cdots, n)$ とその他の著者にそれぞれ A_i，B_i とラベルをつけたグループを G_i としたときの 2 値判別の書き手の同定結果を集計した分割表は表 3 で表される．

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，本論文では三品・松田 [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

表 3: 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

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．

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

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

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

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た値を 2 値判別における F 値と呼び，総合的な指標とする．

5.2.2 多値判別の評価指標

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 値判別に分割し直してから，5.2.1 節と同様の手順でそれぞれの著者の再現率と正解率を求める．求めたそれぞれの著者の再現率と正解率を以下のように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．

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

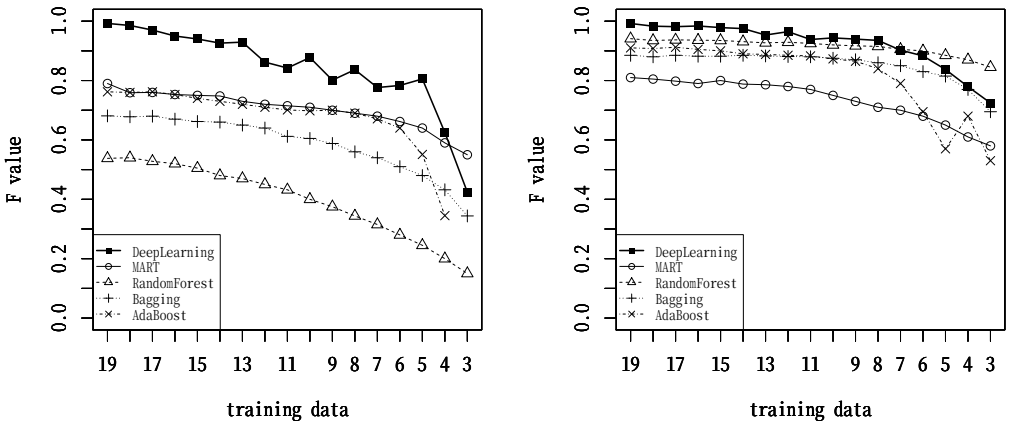
6 深層学習検証結果

深層学習は統計ソフト R の ‘h2o’ パッケージ [13] 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．正解率と F

値は似たような傾向が見られたためここでは F 値のみを示す．深層学習以外の分類法の値は三品・松田 [10] および三品の修士論文において検証結果の値が掲載されていなかったため，目視で確認して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・松田 [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，それ未満の文字についてはその他の項目としてまとめて集計した．

6.1　小説 n-gram 分布

品詞の n-gram 分布の項目数を集計したところ，小説データでは 158 項目となった．小説データについて 2 値判別と多値判別を行い，各分類法との比較グラフを図 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．



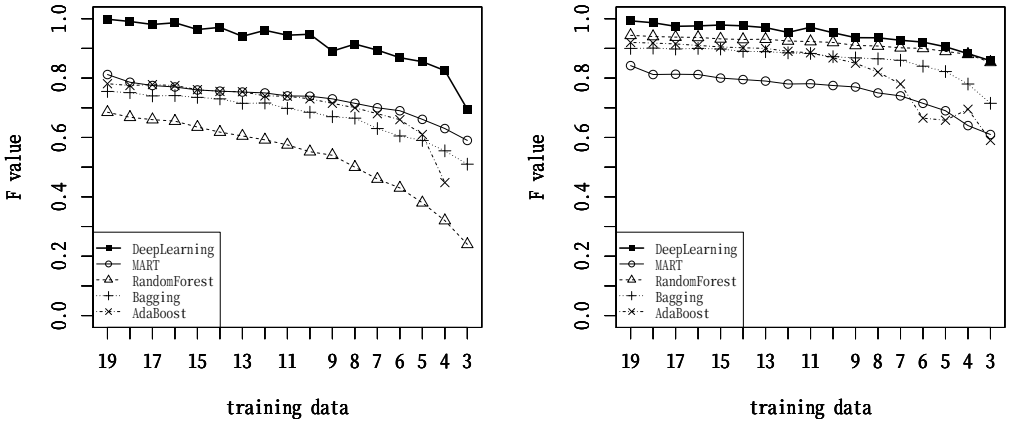
2 値判別　　多値判別

図 1: 小説 n-gram 分布の F 値のグラフ

6.2　小説読点前の文字の分布

三品・松田 [10] と同様に文字の出現頻度 50 以上を基準とし，それ未満の文字についてはその他の項目にまとめて集計したところ小説データは 33 項目となった．小説データについて 2 値判別，多値判別を行い，各分類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．



2 値判別　　多値判別

図 2: 小説読点前の文字の分布の F 値のグラフ

6.3 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，n-gram 分布の多値判別の深層学習の結果のみを図 3 に示す．

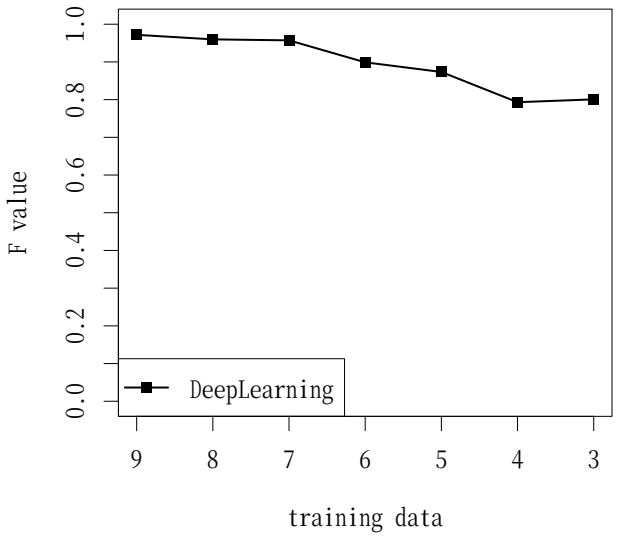


図 3: ブログ n-gram 分布 F 値のグラフ (多値判別)

両方の変数において小説と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・松田 [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．しかし，標本サイズ 9 で最大 0.972 と非常に高い判別精度となった．

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，調整 (チューニング) をすることでより精度を高めたりコストの低いモデルを構築することが可能である．今回深層学習として実装した R の ‘h2o’ パッケージにも様々なパラメータが存在するので岡谷 [12] を参考に重要だと思われるパラメータを中心にチューニングし，より良いモデルが構築することができるのかを検証していく．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小説データに対して検証する．検証結果の値の比較は F 値を用いることにし，標本サイズは 19，11，3 とする．ここでは，n-gram 分布の多値判別の結果のみを示す．

7.1　学習回数

第 6 章では深層学習の学習回数を 10000 回に設定し検証を行ったので，まず学習回数に焦点を当ててどのくらいの回数で同等の結果が得られるのかを検証する．

各標本サイズおける学習回数の違いによる F 値の変化を図 4 に示す．標本サイズが 19，11 の場合には大きな差は見られなかったが，標本サイズ 3 の場合学習回数が 100 回になると少し F 値が下がっていく傾向がみられた．このことより深層学習の学習回数は 1000 回程度行えば同等の結果が得られるのではないかと考えた．学習回数を少なくすることで時間コストも抑えることができるので，本章では学習回数を 1000 回に設定しその他のパラメータを変更して検証をしていく．

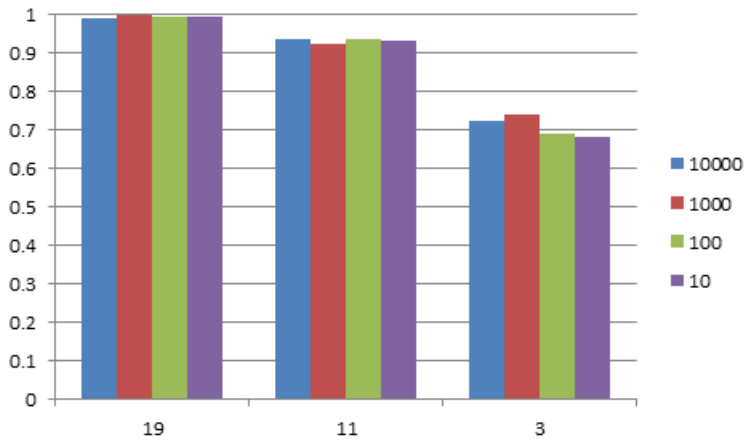


図 4: 学習回数の違いによる F 値の変化 (小説 n-gram 分布)

7.2 中間層とユニット

深層学習は中間層の数と各層のユニットの数を変更することもできる．中間層の数を増やしたりユニットの数を減らしたりすることで各標本サイズにおける傾向を見ていく．‘h2o’パッケージのデフォルトのユニットと中間層の数は 200 × 2 であり，この表記で中間層の数 2，各ユニットの数 200 を表すこととする．

ユニットの数を 200 とし各中間層の数の違いによる F 値の変化を図 5 に示す．学習回数と同様に標本サイズ 3 の場合には F 値の変化がみられた．深層学習の学習アルゴリズムで中間層の数が 1 であるニューラルネットワークを再現してみたが，やはり精度は低くなっており中間層の多層化の重要性が感じられる．しかし，中間層の数が 3 までは F 値が増加しているが，それ以上になると同等の値になるか減少する傾向がみられた．これは中間層の数が増えることによりモデルがより複雑になっていき過学習を起こしやすくなっているのだと考えた．今回の小説データでは中間層の数は 3 程度まで増やすことで十分な精度を出すことができるのではないかと考えた．

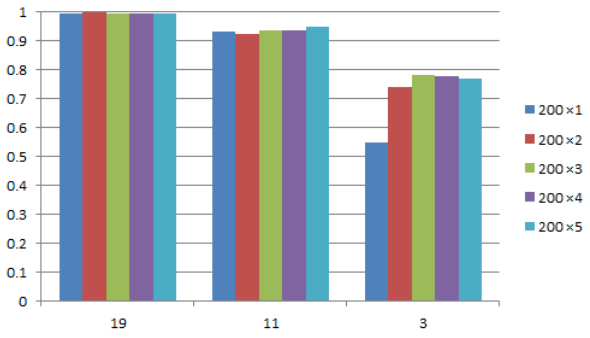


図 5: 中間層の数の違いによる F 値の変化 (小説 n-gram 分布)

次にユニットの数の違いによる F 値の変化を図 6 に示す．ユニットの数を減らしていくにつれて精度は下がっておりユニットの数を 25 にすると急激に下がった結果となった．読点前の文字の分布についても同様に検証をしたが，精度に大きな変化は見られなかった．各変

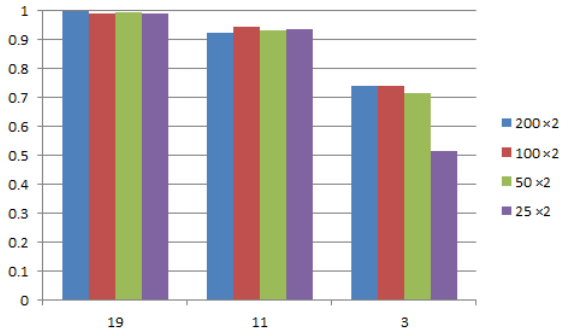


図 6: ユニットの数の違いによる F 値の変化 (小説 n-gram 分布)

数の項目数は n-gram 分布で 158，読点前の文字の分布で 33 と大きく異なっている．n-gram 分布のように 158 項目ある変数をユニット数 25 という少ない次元で表現しようとすると，著者の特徴をうまく表現できず精度が下がってしまった可能性がある．n-gram 分布のユニット数は 100 程度は必要であると考えられる．

7.3　ドロップアウト

岡谷 [12] では過学習を防ぐ方法としてドロップアウトについて述べている．ドロップアウトとは学習時において各層から層へ伝搬する際に一定の確率でユニットを無効化して学習を進める方法である．学習を繰り返すごとに無効化するユニットは異なり，複数の異なるニューラルネットワークを用いて学習を繰り返す動きをするので汎化性能を上げることができ，過適合を防ぐことができるとされている．今回はドロップアウトを有効にしてユニットを無効化する確率をデフォルト値である 0.5 に設定して検証を行う．

前節までに小説データにおける適切な学習回数と層の数が導き出すことができた．本節ではデフォルト値であるユニット数 200 に固定をし，層の数の違いによる F 値の変化も見ながらドロップアウトを適応させていき精度比較を行う．また，標本サイズが 3 の場合にチューニングによる差が出やすかったので標本サイズ 3 の場合のみを対象とした F 値の変化を図 7 に示す．これ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精度が出せた．n-gram 分布で最も高い判別精度は 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%も精度が上がっており，三品・松田 [10] の RandomForest 法の結果 0.845 を上回る結果となった．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．

一方，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．しかし，ドロップアウトの効果は層の数を増やすのと同程度は見られた．

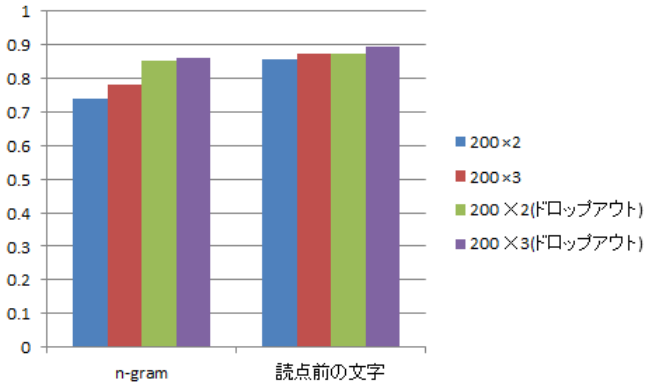


図 7: ドロップアウトと中間層の数の違いによる F 値の変化 (小説)

8 まとめ

三品･松田 [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・松田 [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法であるといえるだろう．

9 おわりに

本論文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パラメータの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，学習方法を変えたりす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] AOKIDS Home Page: 青空文庫のテキストからルビを削除するには, `http://www.aokids.jp/others/delruby.html` (2017/9/11 閲覧).

[2] 青空文庫: `http://www.aozora.gr.jp/` (2017/9/7 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 (2008): 『R によるテキストマイニング入門』, 森北出版.

[5] 金明哲 (1993): 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.

[6] 金明哲 (1996): 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.

[7] 金明哲 (2002): 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.

[8] 金明哲 (2013): 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.

[9] 金明哲・村上征勝 (2007): 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.

[10] 三品光平・松田眞一 (2013): 文章の書き手の同定における分類法の精度比較, 南山大学紀要『アカデミア』情報理工学編, **13**, 35-46.

[11] 村上征勝 (1994): 行動計量学シリーズ『真贋の科学－計量文献学入門』, 朝倉書店.

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

[13] Package'h2o': `https://cran.r-project.org/web/packages/h2o/h2o.pdf` (2017/7/1 閲覧).