文章の書き手の同定における分類法の精度検証の研究

M2011MM047 三品光平
指導教員：松田眞一

1　はじめに

テキストを計量的に分析する研究は 100 年以上前から行われてきており，著者不明の文章の書き手の同定に関する研究の歴史も長い。しかし，日本は欧米に比べて始まるのが遅かったこともあり研究が十分とは言い難い。日本語は英文と違い，分かち書き (単語と単語の間をスペースで区切る) がされておらず，文字の種類も多いため人の手での計量が難しかった。しかし，コンピュータを用いた形態素解析技術の発展により，文章情報を自動で読み取り処理することができるようになった。コンピュータによる解析は完璧ではないが，高速にテキストデータを処理できるようになったことはもちろん，文章に含まれている様々な情報を客観的に集計できるようになった。著者推定の研究は，著作者不明の文献の推定や文献の真贋判定だけでなく，裁判における被告人の上申書と日記の作成者の同一性の検証，さらにはインターネット上のメールや情報分野にまで拡大し，ウェブ文書の推定やスパムメールやウェブスパム (Google のページランクをあげるためのダミーページによる強リンクネットワークの構築) への対策といった実社会の様々な場所で需要がある。

2　研究の目的

金・村上 [13] は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がその際，書き手の特徴が顕著に現れないようにするために変数には単語の相対頻度のみを用いている。本研究では，先行研究で書き手の同定においての有効性が示されている複数の変数を追加し比較検証を行う。また，分類法については金・村上 [13] が有効性を示したランダムフォレスト法と同時期に提案された MART 法を新たに追加する。

3　分析について

3.1　用いる文章データ

金・村上 [13] で分析に用いられたデータは 10 人各 20 編の合計 200 編の小説，11 人 10 タイトルの 110 編の作文，6 人の 10 日間文の日記 60 編である。ただし，小説データにおいては長い作品の場合分割したものを独立した文章として扱っている。
本研究では小説データは青空文庫から金・村上 [13] とできるだけ同様のものを使用した。現在ダウンロードできない作品もあるためそれらの代わりに同じ作者の他の作品を使用し，10 人各 20 編の合計 200 編の小説データを揃えた。また，作文と日記に関しては同様のデータが手に入らなかったためインターネット上のブログから 5 人各 10 編の合計 50 編のブログ記事を使用する。

3.2　文章のクリーニング

文章を分析する上で重要になるのが電子化し，データの形式をそろえたり，不要なものを削除したりする文章のクリーニング作業である。本研究で用いる青空文庫では以下の様な処理を行った。

1. ルビの様な本文以外の内容を削除する。
2. コンピュータ上で正常に表記されない外字を識別可能な字に置き換え，その単語を解析ソフトの辞書に登録する。
3. 地の文以外の会話文などを削除する。

3.3　形態素解析

文章を統計的に分析するためには文章情報を読み取り集計する必要がある。しかし，膨大な量の文章データの処理を人の手で行う事は難しい。そこで，コンピュータを使った自然言語処理技術である形態素解析を用いる。形態素解析とは文を形態素つまり意味の最小の単位に分割することをいう。例えば「本を読んだ」という文は「本」,「を」,「読んだ」と分割できると思うかもしれないが，言語学では「本」,「を」,「読む」,「だ」と分割され,「読んだ」を動詞の「読む」と助動詞の「だ」に分割する。
形態素解析では文を形態素に分割すると同時に，形態素の品詞を特定するところまで行われる。本研究では形態素解析にフリーの形態素解析ソフトである MeCab を用い，データの集計には統計解析ソフト R 上で MeCab を実行し集計することができる RMeCab を用いる。(石田 [8] 参照)

3.4　変数

金・村上 [13] では分類法の精度と小サンプルにおける書き手の同定に関するアルゴリズムの適応性に焦点をあてており，用いたデータに書き手の特徴が顕著に現れないようにしている。そのため，変数にはノイズが多く含まれていると思われる単語の相対頻度を用いている。また各単語すべてを変数として用いるとデータセット内の値が 0 となってしまい分類手法によっては正常に動かないため頻度がある値以下のものは「その他」の項目にまとめている。
本研究では単語の相対頻度だけではなく，金 [9, 10, 12, 11] で書き手の同定における有効性が示された単語の長さの分布，品詞の n-gram 分布，助詞の分布，読点前の文字の分布などを使用し分類の精度検証を行う。

- 単語の長さ
  金 [10] は単語の長さの分布は品詞別に分けることでより明確に書き手の特徴が現れることを示している。その理由として，書き手の個性が単語の長さに出にくい助詞や文章の内容に依存する名詞などがノイズ

になっていることが挙げられている。最も書き手の特徴が現れる品詞として動詞が挙げられている。

- 品詞の n-gram
  n-gram とは隣接している n 個の文字の共起関係をを現すものであり，品詞の n-gram の場合は形態素解析を行い形態素ごとに分けられ品詞のタグを付けられたものを品詞に関して n-gram をとったものである。例文：今日 < 名詞 > は < 助詞 > 雪 < 名詞 > の < 助詞 > 降る < 動詞 > 寒い < 形容詞 > 日 < 名詞 > な < 助動詞 > ので < 助詞 > 家 < 名詞 > に < 助詞 > い < 動詞 > ます < 助動詞 > 。< 記号 > での $N = 2$ の集計したものを表 1 に示す。

表 1 品詞の n-gram(N=2)

品詞	度数	相対頻度
[形容詞 – 名詞]	1	0.077
[助詞 – 動詞]	2	0.154
[助詞 – 名詞]	2	0.154
[助動詞 – 記号]	1	0.077
[助動詞 – 助詞]	1	0.077
[動詞 – 形容詞]	1	0.077
[動詞 – 助動詞]	1	0.077
[名詞 – 助詞]	3	0.231
[名詞 – 助動詞]	1	0.077

3.5 分析手法

先行研究である金・村上 [13] では分類手法として k 最近傍法，学習ベクトル量子化法，サポートベクターマシン法，Bagging 法，Boosting 法，RandomForest 法が用いられており，RandomForest 法が最もよい結果を示し，その次によい結果を示したのが Bagging 法と Boosting 法 (AdaBoost) であった。本研究ではそれらの三つの手法に加えて，Boosting 法の一種である勾配 Boosting に学習器として CART 型樹木を用い拡張した MART 法を用いる。

4 分類手法

4.1 CART 法

CART 法は Breiman et al.[1] によって提案された樹木に基づく方法（樹木構造接近法）であり，分割規則に従いデータを複数の群に分割するものである。CART 法は次の三つの手順に大きく分かれる。(杉本ら [14] 参照)

1. 前進過程：樹木の成長過程
   規則に従いデータを 2 群（ノード）に分割していき停止規則に達するまで分割を行う。
2. 後退過程：樹木の刈り込み過程成長させた大きな樹木は過剰適合を起こすため，弱い枝を切り落とす。
3. 最適モデル選択過程：最適な樹木の決定最適な樹木を決定には樹木の刈り込み過程において作られた候補となる部分樹木に，樹木の作成に使われていないテストデータを当てはめたとき，誤差率が少ない部分樹木が選ばれる。

4.2 Boosting 法

Boosting 法とは，逐次学習データの調整しながら複数の弱分類器を構築しそれらを組み合わせることによって精度の高い強分類器を構築する方法である。

4.2.1 AdaBoost

AdaBoost は Freund and Schapire[4] によって提案された Boosting 法の一種であり，逐次弱分類器の重み付き誤り率から求めた信頼度を更新していくことで強分類器を構築する手法である。

4.2.2 MART 法

MART 法は Friedman[5, 6] によって提案された Boosting 法の一種である勾配 Boosting に CART 型樹木を分類器に用いた手法であり，勾配 Boosting は逐次損失関数 $L(y, f(\boldsymbol{x}))$ の傾きにより重みを更新していき強分類器を構築する手法である。損失関数とその傾きを表 2 に示す。

表 2 損失関数とその傾き

種類	損失関数	傾き
回帰	$\frac{1}{2}[y - f(\boldsymbol{x})]^2$	$y - f(\boldsymbol{x})$
回帰	$\lvert y - f(\boldsymbol{x})\rvert$	$\mathrm{sign}[y - f(\boldsymbol{x})]$
2 分類 (2 項ロジット)	$\log(1 + \exp(-2yf(\boldsymbol{x})))$	$\frac{2y}{(1+\exp(2yf(\boldsymbol{x})))}$
多分類 (S クラス)	$-\sum_{s=1}^{S} y_s \log p_s(\boldsymbol{x})$	$y_s - p_s(\boldsymbol{x})$

本研究では多分類に対応した損失関数を用いた。

4.3 Bagging 法

Bagging 法とはアンサンブル学習法の一種であり，Breiman[2] によって提案された。Bagging 法はブートストラップと呼ばれる復元抽出法で複数の学習データセットを作成し，各学習データで分類器を作成し，多数決を取ることで精度の向上を図っている手法である。

4.4 RandomForest 法

RandomForest 法とはアンサンブル学習法の一種であり，Breiman[3] によって提案された。RandomForest 法も Bagging 法と同様に復元抽出によりサンプリングされた複数の学習データセットを作成しそれらから分類器を作成し，多数決を取る。Bagging 法との違いは変数もサンプリング (非復元抽出) されたものを用いる点である。また，Bagging では分類器を作成する際，樹木の刈り込み過程があったが，RandomForest 法では刈り込みを行わず最大の樹木を用いる。

5 分析結果

5.1 検証方法について

小説データ作者 10 人各 20 編とブログデータ 5 人各 10 編のデータに関して，それぞれ分類器の学習に用いる学習データと分類器の評価に用いるテストデータにサンプリングを行う。標本サイズの違いによる判別精度をみるために，著者一人あたりの標本サイズを S としたとき学

習用データを $(S-1, S-2, ..., 3)$ 個ずつ各著者からランダムサンプリングを行いそれ以外をテストデータとした。分類器内で用いられている乱数やサンプリングされる標本データの違いにより判別精度が違ってしまうので，評価には実験を 100 回繰り返した評価指標の平均を用いることとした。

5.2 再現率・精度

評価指標として再現率 (recall) と精度 (precision)，そして，それらから求められる調和平均 F を用いる。著者 $i(i = 1, 2, ..., n)$ とその他の著者を A,B とラベルをつけたグループを G_i とし A を正しく分類したい場合を例とすると，再現率 R_i は A と判断されるべきものの内どれだけ正しく A と判断されたかを表し，精度 P_i は A と判断されたものの内どれだけ A と正しく判断されたかを表す。同定結果を表 3 とおいたとき再現率・精度・F 値は下記の式で表される。一般的に精度は (1) 式の計算で求めるが，本研究の場合，小さい標本サイズで精度を出すことがあり，その際 a_i,b_i ともに 0 となり計算ができないことがでてくるため，精度を (2) 式で求めることとする。

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

$$\text{精度 } 1 : P_i = \frac{a_i}{a_i + b_i} \quad (1)$$

$$\text{精度 } 2 : P_i = \frac{a_i + d_i}{a_i + b_i + c_i + d_i} \quad (2)$$

多数の分類の場合は再現率と精度の平均を用いる。式は次の式で定義される。

$$\text{再現率} : \hat{R} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{a_i + c_i}, \text{ 精度} : \hat{P} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i + d_i}{a_i + b_i + c_i + d_i}$$

調和平均 F は次の式で定義される。

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

表 3 同定結果のクロス表

G_i		分類法の結果	
		A	B
データ	A	a	c
	B	b	d

5.3 単語の相対頻度

小説データでは出現頻度が 50 以上を基準とし，それより下の単語はその他の項目にまとめたところ 587 項目となり，ブログデータでは出現頻度が 10 以上を基準とし，それより下の単語はその他の項目にまとめたところ 56 項目となった。これらの変数を用いて各分類法で分析を行う。また，F 値は各著者とその他の 2 値に分類し判別・同定し求めたものと各著者を同時に (多分類) 判別・同定し求めた結果から求めたもの 2 つの結果を図 1，2，3，4 に示す。まず，図 1，2，4 の adaboost の分析結果が途中から途切れていることについて述べる。金・村上 [13] でデータセット内の値が 0 となり正常に動かない分類法があることが示されていたので，データセット内の値に 0 を含まないデータセットでも試したが上手くいかず，他の変数での分析でも同様の標本サイズで正常に動かなくなることから標本サイズが関係していると思われる。ある程度，標本サイズのある状態では判別が正常に行われているのでその点で比較を行う。

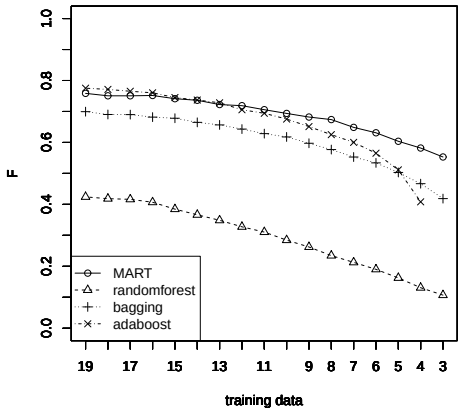

図 1 小説データの F 値の平均 (2 値判別)

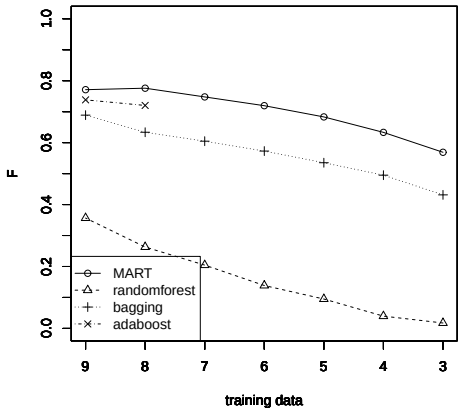

図 2 ブログデータの F 値の平均 (2 値判別)

2 値判別での F 値の図 1，2 から小説・ブログデータともに，MART，adaboost がほぼ同じで bagging がそれに続き，そして RandomForest という順番になっている。特に RandomForest 法が他の分類手法よりも低くなっていることがわかる。これは RandomForest 法の欠点である正例と負例の数に差がある場合識別精度が下がるという性質が原因と考えられる。正例と負例の比は小説データ 1 : 9，ブログデータ 1 : 4 となる。(sfchaos[7] 参照)

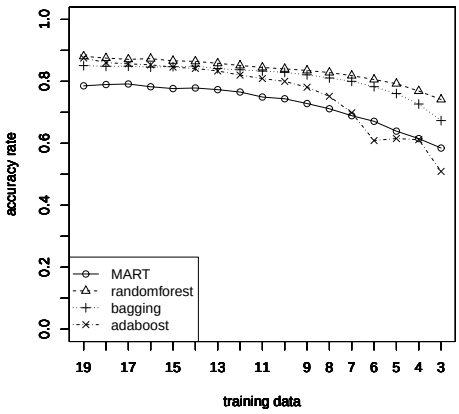

図 3 小説データの F 値の平均 (多値判別)

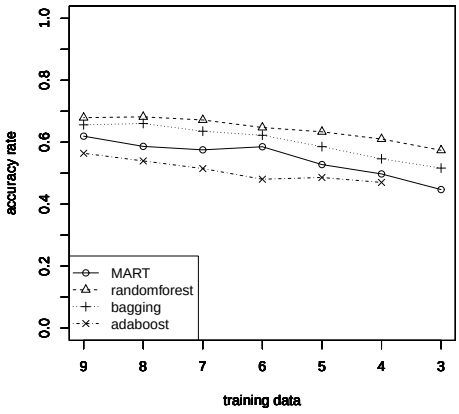

図 4 ブログデータの F 値の平均 (多値判別)

多値判別での F 値の図 3，4 から小説データでは RandomForest，baagging，adaboost，MART，ブログデータでは RandomForest，baagging，MART，adaboost，の順に良く，ともに RandomForest 法が一番良い結果となった。また，小説データでの標本サイズの減少に伴う adaboost の F 値の変化をみてみると 2 値判別，多値判別ともに他の分類法よりも標本サイズの減少による判別精度の低下が大きいことがわかる。以上のことから正例と負例に差のある 2 値判別においては MART 法と adaboost が有効であるが，標本サイズの減少に対して影響を受けにくい MART 法の方が有効だといえる。また均衡データでの多値判別では RandomForest 法が最も有効だといえる。

6 まとめ

他の変数での分析結果も単語の相対頻度と概ね同じになり，やはり MART 法が正例と負例に差がある 2 値判別に強いという結果となり，RandomForest 法が均衡データでの多値判別に強いという結果となった。また，各変数での F 値は助詞の分布と単語の長さの分布 (動詞) はあまり良くなかったが，品詞の n-gram と読点前の文字の分布は単語の相対頻度を上回る良い結果となった。特に RandomForest 法での多値判別の F 値が良く，品詞の n-gram の小説データで最大で 0.94，学習用の標本を減らしていっても，各著者の標本サイズが 6 になるまで 0.9 以上，標本サイズ 3 でも 0.85 という高い判別精度を示した。また，小説データよりも短い文章であるブログデータでも，最大 0.92，標本サイズ 4 まで 0.8 以上，標本サイズ 3 でも 0.78 と短い文章でも比較的高い判別精度を示した。読点前の文字の分布に関してもこれに近い良い結果となった。品詞の n-gram の結果に関しては金 [12] で示されている品詞の n-gram の短い文章での判別の有効性にも合致する結果となった。

7 おわりに

本研究では，正例と負例に差がある場合でも MART 法は高い判別精度を持つことと，均衡データでの多値判別における RandomForest 法の判別精度の高さを示すことができた。しかし，adaboost が正常に動かない問題を解決できずに終えてしまった。また，テキストのクリーニング作業は文献やインターネット上の情報を基に独学で行っているため完璧とはいえない。より，正確にテキストクリーニングを行うことで今回あまり良い結果とならなかった変数の判別結果も変わってくる可能性がある。

参考文献

[1] Breiman, L.,Friedman,J.H,Olshen,R.A. & Stone,C.j.: Classification And Regression Trees, Wadsworth, 1984.

[2] Breiman,L.: Bagging predictors, Machine Learning, 26(2), 123-140, 1996.

[3] Breiman,L.: Random Forests, Machine Learning, 45(1), 5-32, 2001.

[4] Freund,Y.and Schapire,R.E.: Experiments with a new boosting algorithm, Machine Learning, Proceedings of the Thirteen International Conference, 148-156, 1996.

[5] Friedman,J.H.: Greedy function approximation: a gradient boosting machine, The Annals of Statistics, 29(5), 1189-1232, 2001.

[6] Friedman,J.H.: Stochastic gradient boosting: Nonlinear methods and data mining, Computational Statistics and Data Analysis, 38, 367-378, 2002.

[7] sfchaos: 不均衡データのクラス分類, `www.slideshare.net/sfchaos/ss-11307051` , 2012.

[8] 石田基広：RMeCab の使い方, `rmecab.jp/wiki/index.php?plugin=attach&refer=RMeCab&openfile=manual.pdf`, 2008.

[9] 金明哲：読点の情報に基づく文献の分類，全国大会講演論文集 第 46 回平成 5 年前期 (3)，131-132，1993.

[10] 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.

[11] 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), 15-2，2002.

[12] 金明哲：品詞のマルコフ遷移の情報を用いた書き手の同定, 日本行動計量学会大会発表論文抄録集 32, 384-385，2004.

[13] 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.

[14] 杉本知之, 下川敏雄, 後藤昌司：樹木構造接近法と最近の発展, 計算機統計学 18(2), 123-164，2005.