統計数理(2007)
第 55 巻 第 2 号 255–268

特集「文化を科学する」
[原著論文]

ランダムフォレスト法による文章の書き手の同定

金 明哲[†]・村上 征勝[†]

(受付 2007 年 1 月 24 日；改訂 2007 年 10 月 30 日)

要　　旨

本稿ではランダムフォレスト法を用いた文章の書き手の同定を提唱し，その有効性を，k 近傍法，サポートベクターマシン法，学習ベクトル量子化法，バギング法，ブースティング法などの分類法との比較分析を通じて示した．比較分析では，10 人が書いた 200 編の小説，11 人が書いた 110 編の作文，6 人が書いた 60 編の日記という 3 種類の異なるタイプの文章を用い，分類法の正解率と学習に用いた標本サイズとの関係に焦点を当て分析した．その結果，ランダムフォレスト法がその他の分類法より正解率が高く，また学習に用いる標本サイズの減少による影響が小さいことが実証された．

キーワード： 著者(書き手)同定，テキスト分類，計量文体学，集団学習法，ランダムフォレスト法．

1. はじめに

我々は読んだ文章が小説であるか，論文であるか，新聞記事であるか，そのジャンルを見分けることができる．これはそれぞれのジャンルの文章の形式(パターン)を学習し，その知識を持っているからである．また特定の作家の作品の愛読者は，文章を読むだけで，その作家の文章であるか否かをある程度見分けることができると言われている．それはその作家の作品を大量に読むうちに，その作家の文章のパターンが愛読者の脳に焼付けられたためであると考えられる．

このような人間の脳におけるパターン認識の仕組みはまだ完全に解明されていないが，人間が行っているパターン処理と認識をコンピュータで実現させる研究が多く行われ，一定の成果を得ている．人間の脳で行われている文体に関するパターン処理も例外ではない．その中の 1 つに，統計的学習法によるパターン認識のアプローチがある．

文体分析の研究で計量的に文章の書き手を推定・同定する本格的な研究が行われるようになったのは 19 世紀の後半からである．例えば，オハイオ州立大学の地球物理学者 Mendenhall (1887)は，単語の長さの分布に書き手の文体の特徴が現れるという研究を『サイエンス』誌に発表した．彼はディケンズ(Dickens, 1812–1870)， サッカレー(Thackeray, 1811–1863)，ミル(Mill, 1806–1873)の文章に使われた単語の長さの分布を調べ，それが作家によって異なることから，作家の文体の特徴になり得ることを示した．1960 年前後には，判別分析の手法が書き手の同定に用いられるようになった．コンピュータが自然言語を自由に扱えない時代には，必要な文体要素を目で確認しながらカウントする原始的な方法を取っていたが，その基本的な考え方は，今日のテキストマイニングの原型であると言えよう． 文章の書き手の同定およびテキス

[†] 同志社大学 文化情報学部：〒610–0394 京都府京田辺市多々羅都谷 1–3

トマイニング分野におけるテキストの分類は文章のパターン分類・認識とみなすことができる．

テキストの分類に関しては，Sebastiani (2002) が近年の主な研究結果について総括を行っている．分類の精度は用いたテキストのコレクション (種類，内容，数など)，用いた特徴ベクトルと関係しているため，報告された分類精度は絶対的なものではないが，k 近傍法 (k nearest neighbor)，SVM 法，ニューラルネット法がよい結果を示している．

テキストマイニングにおけるテキスト分類の 1 つの特徴は，用いるテキストの数が大きいことである．研究事例によっては，学習に用いたテキストの数は約 1 万前後，テストに用いたテキストの数は数千に上る．しかし，書き手が不明である文章の著者を同定する問題では，通常学習およびテストに用いる文章は数十編であり，場合によっては数編しかない．一方，機械的にテキストから書き手の文体の特徴要素 (項目，あるいは変数) を抽出すると，しばしば数百，数千を超えてしまう．このように，文章の書き手の同定とテキストマイニングにおけるテキストの分類におけるデータ構造は同じではない．

書き手の同定に関しては，2000 年までは線形判別モデル，ベイジアン確率モデル，決定木およびルール抽出モデル，ニューラルネットワークモデルなどの方法が多く用いられていたが，近年ではサポートベクターマシン法を用いた試みが行われている (Diederich et al., 2003; Teng et al., 2004)．しかし，どのような方法が，小標本 (少ない文章の数) かつ高次元 (多数の変数) というデータ構造の文体研究に適しているかについての研究は見あたらない．

本研究では，近年提案されたランダムフォレスト法 (Breiman, 2001) を用いた書き手の同定を提唱し，学習に用いる標本サイズと正解率に焦点を絞り，主たる統計的学習法との比較分析を通じてその有効性を示す．

2. 文章と分類手法

2.1 用いた文章

本稿では，表 1 に示した 10 人の 200 編の小説，表 2 に示した 11 人の 10 のタイトルに関する 110 編の作文，表 3 に示した 6 人のワープロと手書きの 60 編の日記を分析に用いた．小説は青空文庫からダウンロードして用いた．小説の選定には，なるべく同年代であることなどに配慮し，また長い作品は青空文庫が分割したサイズをそのまま独立した文章として扱った．用いた文章の長さが均一ではないため，文章から抽出した要素については相対頻度を用いた．作文は 11 人の同年代の大学生が同時期に 10 の異なるタイトルについて書いたものを用いた．日記は 6 人の人が作成した 10 日間の日記である．なお本研究では，分析には文章の中の会話文を除いた地の文のみを用いている．

2.2 ランダムフォレスト法

ランダムフォレスト (RF; random forest) 法は，集団学習法の一種である．集団学習 (ensemble learning，アンサンブル学習とも呼ぶ) 法は決して精度が高いとは言えない複数の結果を組み合わせ，精度を向上させる方法である．いわば，「三人寄れば文殊の知恵」である．集団学習法の中の代表的な方法としてはバギング法，ブースティング法，ランダムフォレスト法がある．ランダムフォレスト法は Breiman (2001) が拡張・発展させた多数の決定木を用いた集団学習法である．ランダムフォレスト法のアルゴリズムを次に示す．なお本稿では，分類アルゴリズムを分類法 (あるいは分類器，classifier) と呼ぶ．

ステップ 1: 与えられたデータセットから N 組のブートストラップサンプル $B_1, B_2, \ldots, B_i, \ldots, B_N$ を作成する．ブートストラップサンプルは変数をサンプリングして作成する．その際，用いるデータセットの約 3 分の 1 はテスト用として取り除き，残りを学習用とする．テスト用として取り除いたデータを OOB (Out-Of-Bag) データと呼ぶ．

表 1.　分析に用いた文学作品（小説）のリスト（青空文庫からダウンロード）.

著者名	作品名
芥川　竜之介 (1892-1927)	或阿呆の一生，玄鶴山房，歯車，芋粥，煙管，或日の大石内蔵助，偸盗，地獄変，毛利先生，路上，お律と子等と，奇怪な再会，杜子春，将軍，母，おぎん，保吉の手帳から，少年，春，彼
菊池　寛 (1888-1948)	芥川の事ども，仇討禁止令，仇討三態，青木の出京，勲章を貰う話，身投げ救助業，三浦右衛門の最後，M侯爵と写真師，無名作家の日記，大島が出来る話，恩を返す話，恩讐の彼方に，乱世，船医の立場，俊寛，勝負事，出世，忠直卿行状記，若杉裁判長，ゼラール中尉
夏目　漱石 (1867-1916)	それから，一夜，三四郎，倫敦塔，吾輩は猫である 1，吾輩は猫である 2，吾輩は猫である 3，坊っちゃん，幻影の盾，彼岸過迄 1，彼岸過迄 2，琴のそら音，草枕，薤露行，虞美人草 1，虞美人草 2，行人 1，行人 2，趣味の遺伝，門
森　鴎外 (1862-1922)	かのように，じいさんばあさん，カズイスチカ，キタ・セクスアリス，二人の友，余興，堺事件，妄想，寒山拾得，山椒大夫，普請中，最後の一句，杯，百物語，護持院原の敵討，阿部一族，雁，青年，高瀬舟，鶏
島崎　藤村 (1872-1943)	ある女の生涯，三人，並木，伸び支度，分配，刺繍，千曲川のスケッチ，家-上巻，岩石の間，嵐，旧主人，春，桃の雫，桜の実の熟する時，海へ，熱海土産，船，芽生，藁草履，食堂
泉　鏡花 (1873-1939)	七宝の柱，伯爵の釵，化鳥，半島一奇抄，国貞えがく，売色鴨南蛮，女客，婦系図， 小春の狐，怨霊借用，木の子説法，歌行燈，眉かくしの霊，絵本の春，縁結び，草迷宮，薬草取，遺稿，高野聖，鷭狩
岡本 綺堂 (1872-1939)	ゆず湯，宣室志（唐），捜神後記（六朝），捜神記（六朝），白猿伝・其他，酉陽雑俎（唐），お化け師匠，お文の魂，勘平の死，半鐘の怪，湯屋の二階，石燈籠，寄席と芝居と，影を踏まれた女，心中浪華の春雨，異妖編，穴，箕輪心中，青蛙堂鬼談，鳥辺山心中
海野　十三 (1897-1949)	あの世から便りをする話，ある宇宙塵の秘密，奇賊は支払う，奇賊悲願，宇宙の迷子，宇宙尖兵，宇宙戦隊，怪星ガン，恐しき通夜，暗号の役割，暗号音盤事件，海底都市，火薬船，生きている腸，科学者と夜店商人，英本土上陸作戦の前夜，鍵から抜け出した女，鞄らしくない鞄，骸骨館，鬼仏洞事件
佐々木 味津三 (1896-1934)	なぞの八卦見，へび使い小町，七化け役者，京人形大尽，千柿の鐸，卍のいれずみ，南蛮幽霊，明月一夜騒動，曲芸三人娘，村正騒動，毒色のくちびる，生首の進物，笛の秘密，耳のない浪人，血染めの手形，袈裟切り太夫，足のある幽霊，身代わり花嫁，達磨を好く遊女，青眉の女
太宰 治 (1909-1948)	二十世紀旗手，八十八夜，愛と美について，小さいアルバム，老ハイデルベルヒ，兄たち，美少女，地球図，千代女，断崖の錯覚，男女同権，誰，誰も知らぬ，服装に就いて，玩具，逆行，恥，花吹雪，春の盗賊，皮膚と心，富嶽百景

表 2.　分析に用いた 11 人の作文のサイズ（単位は文字数）.

書き手	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	平均
WA	1065	1168	1582	1053	1208	1049	1065	1299	1089	1006	1159
WB	1097	1157	978	1270	1374	1295	1167	1126	1235	1054	1175
WC	1068	1761	1155	1414	1114	1017	1242	1292	1229	1102	1240
WD	1102	1035	1129	1032	1007	1089	1046	1054	1051	993	1054
WE	1032	1063	1266	1173	1018	1178	1081	1061	1101	1126	1110
WF	1066	1105	1069	1075	1039	1077	1100	1125	1045	1164	1087
WG	1060	1261	1438	1300	1170	1068	1184	1471	1032	1170	1216
WH	998	1045	1187	1133	1168	1030	1230	993	1238	1194	1122
WI	1046	1060	1047	1113	1109	1111	1044	1042	1101	1090	1076
WJ	1077	1026	1045	1044	1063	1025	1060	1081	1033	1089	1055
WK	1392	1042	1013	1006	1009	1015	1052	1029	1135	1012	1071
平均	1091	1157	1174	1147	1116	1087	1115	1143	1117	1091	1124

注：T1〜T10 は作文のタイトル(T1;住まい，T2;家族，T3;友達，T4;学校，T5;スポーツ，T6;旅行，T7;車，T8;アルバイト，T9;映画は映画館で見るかビデオで見るか，T10;日本食)である

表 3. 6 人の書き手の日記のサイズ(単位は文字数).

書き手	0	1	2	3	4	5	6	7	8	9	平均
A	290	268	405	355	325	397	499	577	428	452	400
B	399	412	418	411	402	403	400	608	439	445	434
C	502	549	610	636	581	696	673	697	754	516	622
D	585	567	475	686	546	491	561	567	517	458	546
E	521	555	468	486	623	668	726	693	1244	778	676
F	437	495	503	484	502	470	510	562	492	524	498

ステップ 2: 各々のブートストラップサンプル B_i を用いて未剪定の最大の決定木 T_i を生成し，木の生成に用いていない OOB データを用いてテストを行う．その誤り率を OOB 推定値と呼ぶ．T_i の構築を行う際の各分岐ノードは，異なる木を多数生成するため，ランダムに m_{try} 個変数をサンプリングし，その中からもっとも分岐がよい変数を用いる．

ステップ 3: 分類器は，すべてのブートストラップサンプル B_i の OOB 推定値に基づいて多数決をとる．

ランダムフォレスト法のパフォマンスは，バギング法，ブースティング法より優れているとの報告(Breiman, 2001; 金, 2005)はあるが，テキストマイニングや書き手の同定などに用いた研究は見あたらない．

2.3 比較に用いた分類法

本研究では，ランダムフォレスト法の有効性を実証するため，近年注目されているサポートベクターマシン法，学習ベクトル量子化法，バギング法，ブースティング法などの分類方法を用いて比較分析を行った．また，代表的な古典的分類法の k 近傍法も用いる．k 近傍法の他に古典的分類法としては，ベイジアンのアプローチによるナイーブベイズ分類法(naive Bayes classifier)が広く知られているが，テキスト分類における先行研究では，k 近傍法より高い精度が得られていないケースが殆どである．本研究の予備実験でも先行研究と類似の結果が得られたので本研究ではナイーブベイズ法は用いない．比較に用いた分類法を以下に簡単に説明する．

2.3.1 k 最近傍法

k 近傍法(k-NN: k Nearest Neighbor)は伝統的なパターン分類法である(Cover and Hart, 1967)．k 近傍法は，判別すべき個体の周辺の最も近いものを k 個見つけ，その k 個の多数決により，どのグループに属するかを判断する．距離の測度としては一般的にはユークリッド距離が使用されている．本研究では近傍の個体の数 k は，予備実験の結果を踏まえて $k=5$ とした．

2.3.2 学習ベクトル量子化法

人間の脳で行っている情報処理をコンピュータで行おうという試みとして，人間の脳のニューロンをモデル化したものを人工ニューラルネットワーク(ANN: Artificial Neural Network)と呼ぶ．ANN には幾つかのタイプがあるが，パターン認識では隠れ層を持つ(HLNN: Hidden-layer Neural Network)階層型ニューラルネットワークモデルが多用されている．階層型ニューラルネットワークは，入力層，隠れ層，出力層により構成され，層の数とニューロンの数が多くなると，計算量の負荷が大きくなるのが 1 つの短所である．階層的ニューラルネットワークモデルを用いた，文章の著者の同定に関する研究事例としては Matthews and Merriam (1993)，Kjell (1994)，Tweedie et al. (1996)，Hoorn et al. (1999)，Waugh et al. (2000) などがある．階層的ニューラルネットワークモデルでは，大量の変数を用いるのは実用的ではないため，これらの研究では変数を選択して用いている．

本研究では，機械的に抽出した変数を選択せずに用いることを前提としているので，より柔軟性を持つニューラルネットワークモデル LVQ (Learning Vector Quantization) を用いることにする (Kohonen, 1985)．

LVQ はノイズが多い高次元の確率データを取り扱うことを意識して開発された，競合学習型ニューラルネットワークモデルである．その特長は，計算量を減少させると同時に最適な認識精度を得ることが可能であるとコホネンは主張している．LVQ は入力層と競合層により構成されている．入力層のニューロンの数は入力の変数の数に等しく，競合層で入力されたデータの分類を行う．

LVQ は，LVQ1 や OLVQ1 などのアルゴリズムの集合の総称である． LVQ の代表として，LVQ1 のアルゴリズムの主な部分を示す．

ステップ 1: 入力層と競合層とを結合する重みベクトルを与える．

ステップ 2: 競合層の区域と学習データのクラスと対応するクラステーブルを作成する．

ステップ 3: 入力データ $x \in R^n$ と最も距離が近い競合層のニューロン $m_i \in R^n$ を見つけ出す．$c = \arg\min_i \|x - m_i\|$

ステップ 4: 最も距離が近いニューロン m_c との結合の重みを更新する．

$$m_c(t+1) = m_c(t) + \alpha(t)[x(t) - m_c(t)]$$

t は時間に関する離散変数であり，$\alpha(t)$ は単調減少関数である．式の中の $\alpha(t)$ は，もし $x(t)$ と $m_c(t)$ が同じクラスに属するなら符号は正，異なるクラスに属するなら負をとる． ここでのクラスは同じ質的な外的規準 (グループの属性) を持つ集合である．

教師データなしの競合学習法 (自己組織化マップ) を用いた書き手ごとのテキスト分類などに関する研究 (金, 2003) はあるが，教師データありの競合学習法 LVQ をテキスト分類に適応した研究事例は見あたらない．本研究では OLVQ1 が返したコードブック・ベクトルを用いて，LVQ1 でさらに学習を行った．

2.3.3 サポートベクターマシン法

サポートベクターマシン (SVM：Support Vector Machine) 法は，Vapnik (1995) が考案したパターン分類の方法である．サポートベクターマシン法はグループを分ける無限に存在する超平面の中から，最もグループ分けの良い平面をマージン最大の基準で求める方法である．

いま学習データ集合 $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)$ があるとする．ここの $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ は個体の特徴ベクトルであり，y は目的変数である．線形のサポートベクターマシンは次の式で表される． 参考のため図 1 に 2 群判別のサポートベクターマシンの分類イメージを示す．

$$f(\boldsymbol{x}) = \sum_{i=1}^{p} w_i x_i + b$$

初期のサポートベクターマシンは，2 群線形分類法として提案されたが，幾つかのアプローチで非線形の多群判別法として改良が進められている．その中の 1 つが，次のモデルを最適化するカーネル法によるサポートベクターマシンである．カーネルサポートベクターマシンは次の識別関数で表される．

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

式の中の $K(\boldsymbol{x}_i, \boldsymbol{x})$ はカーネル関数である (大北 2005)．サポートベクターマシンをテキスト分類および書き手の同定に適応した研究事例としては Diederich et al. (2003)，Joachims (1998)，

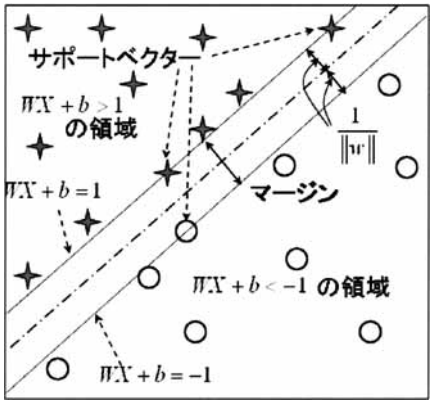

図 1. サポートベクターマシンのイメージ.

Teng et al. (2004), Zheng et al. (2005) などがある. 多くの先行研究では SVM は k-NN 法, ナイーブベイズ法より分類性能が優れていると報告されている.

2.3.4 バギング法

バギング (Bagging) 法の Bagging は, bootstrap aggregating の頭部分の文字列を組み合わせた造語である. バギング法は集団学習法の一種であり, Breiman (1996) によって提案された. バギング法は, 与えられたデータセットから, ブートストラップ (bootstrap) と呼ばれているリサンプリング法で複数の学習データセットを作成し, そのデータをもとに得られた結果を統合・組み合わせることで分類器の精度を向上させる. ブートストラップサンプルはそれぞれ独立であり, 学習は並列に行うことができる. バギングのアルゴリズムの流れを次に示す.

ステップ 1: 教師付きデータから復元抽出方法で m 回抽出を行い, 新たな訓練サブデータセットを作成し, モデルを構築する.
ステップ 2: ステップ (1) を B 回繰り返し, B 個のモデルを構築する.
ステップ 3: B 個のモデルから多数決による分類器を構築する.

2.3.5 ブースティング法

ブースティング (boosting) 法は与えた教師付き学習データを用いて学習を行い, その学習結果を踏まえて逐次に重みの調整を繰り返すことで複数の学習結果を求め, その結果を統合・組み合わせ, 精度を向上させる方法である (Freund and Schapire, 1996). その代表的なアルゴリズムとしては AdaBoost がある. そのアルゴリズムの流れを次に示す.

ステップ 1: 重みの初期値 w_{1i} を生成する.
ステップ 2: 重みを更新しながらモデルを繰り返し構築する ($t=1,2,\ldots,T$).
(a) 重み w_{ti} を用いてモデルを構築する.
(b) 誤り率を計算する.
(c) 誤り率を用いて, 結果の信頼度を計算する.
(d) 重みを更新する. $w_{(t+1)i}=g(w_{ti})$
ステップ 3: 分類器の場合は, T 個の結果を信頼度で重み付けて多数決をとる.

3. 分類法の比較

3.1 学習とテスト

教師付きデータの分類の問題では学習の結果を，学習に用いていないデータを用いてテストを行う．1つのデータセットにおける学習結果の評価にはN分割交差確認(N-fold cross-validation)法がある．N分割交差確認法は，データセットをランダムに N 等分し，その中の $N-1$ 個を学習用とし，残りの1個をテスト用とする．本研究では，学習のサンプルサイズと分類法の精度について各手法を比較するのが1つの目的であるので，データセットからランダムに「著者数$\times s$」($s=1,2,3,\ldots,S-1$) の個体を抽出して学習用とし，残りをテスト用とする．S は各データセットにおける著者別の文章の数である．本研究で用いた3種類のデータセットでは，各データセットにおける著者別の文章の数は一致している．

著者 i ($i=1,2,\ldots,g$) とそれ以外の著者にラベルをつけたグループをクラス C_i とすると，C_i における判別・同定の結果は表4に示すクロス表で表すことができる．

分類結果の評価には，再現率(recall)や精度(precision)などが多く用いられている．再現率 R_i は，分類法がどれぐらい「漏れ」なく正しく判別しているかに関する度合であり，精度 P_i は分類法の分離結果に混入された「ゴミ」に対する的中率である．それぞれの定義を次に示す．

$$\text{再現率}: R_i = \frac{a_i}{a_i + c_i}, \quad \text{精度}: P_i = \frac{a_i}{a_i + b_i}$$

多群の分類問題では，評価指標として再現率と精度のマクロ平均(macro average)，マイクロ平均(micro average)を用いることがある．本研究では，次に示すマクロ平均を用いている．マクロ平均はクラス $C_i, i=1,2,\ldots,g$ における次の式で定義されている．

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

本研究では，分類法の評価は再現率と精度の調和平均 F を用い，F 値が大きければ，分類性能がよい(正解率が高い)と評価する．

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

多くの機械学習法では乱数を使用しているため，同一のデータセットに対して分類を繰り返しても分類の精度が同じになるとは限らない．そこで，評価には実験を100回繰り返した F 値の平均と信頼区間を用いることにした．

3.2 書き手の特徴の抽出

書き手の同定を行う際には，文章に関するどのような変数を用いるかが1つの鍵となる．日本の現代文において，どのようなところに書き手の特徴が現れるかに関しては，幾つかの研究が報告されている(安本・本多, 1988; Jin and Murakami, 1993; 金 他, 1993; 松浦・金田, 2000; 金, 1994, 1995, 1997, 2002, 2003, 2004)．

表 4. クラス C_i の同定の結果のクロス表.

クラスC_i		分類法の結果	
		Yes	No
データ	Yes	a_i	c_i
	No	b_i	d_i

本研究は，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いる．用いたデータに書き手の特徴が顕著に現れている場合は，分類法の違いによる精度の差が見られない恐れもあるので，本研究では，変数としてノイズが多く含まれていると思われるタグ付き単語の相対頻度を用いた．タグ付き単語とは，文を単語単位に分割し，各単語に品詞の情報を付けたものである．単語の品詞タグ付けは，形態素解析ソフト『茶筅』(松本 他, 2003)を用いた．機械による形態素解析には誤りがともなうが，修正は行わなかった．品詞のタグ付けはすべて同一の基準で行われるので，形態素解析に含まれる誤りが分類の精度の比較に影響を与えることはないと判断した．『茶筅』の品詞情報は階層的になっているが，本研究では，助詞と記号に関しては第 2 層まで，それ以外は第 1 層のみの情報を用いた．

文章に出現する全ての異なるタグ付きの単語をそれぞれ 1 つの変数として用いるとデータセットの中の値がゼロとなるものが多く，分類器によっては正常に作動しないケースが頻繁に起きる．そこで，本研究では文章に出現する単語の頻度がある値以下のものは「その他」の項目にまとめた．

分析に用いたテキストの長さは同じではないので，文章 i におけるタグ付き単語 j の相対頻度を p_{ij} としたベクトル $P_i=(p_{i1},p_{i2},\ldots,p_{ij},\ldots,p_{in})$ をその文章の特徴ベクトルとして用いる．ただし $\sum_{j=1}^{n} p_{ij}=1$ である．

3.3 書き手の同定結果

3.3.1 文学作品

表 1 に示した文学作品においては，1 編の作品に出現する単語の頻度を，平均 1 回(合計 200 回)を規準とし，出現頻度がそれ以下の単語は「その他」の項目にまとめると合計 496 項目(変数)になる．この 496 変数を用いて 10 人の 200 編の作品の書き手の同定を行った場合の F 値の平均プロットを図 2 に示す．図 2 の横軸は書き手ごとの 20 編の作品から抽出して用いた学習データの標本サイズである．ここでは，サンプルサイズが s の標本を用いた学習の結果を用い，残りの $20-s$ のテストデータの分類を行った．縦軸は F 値の平均値であり，エラーバーは 95%の信頼区間である．

図 2 からわかるように，集団学習法の正解率が全体的に高く，最も高いのはランダムフォレスト法(RF)であり，学習データの標本サイズが 4 までは 95%の正解率が得られている．学習データの標本サイズの減少に伴う正解率の影響も小さいことが読み取られる．

3.3.2 作文データ

表 2 に示す作文のタグ付き単語について，1 編の文章に平均 1 回現れることを基準とし，出現頻度がそれ以下の単語は「その他」の項目にまとめると，項目の数は合計 90 になる．その同定結果の F 値の平均プロットを図 3 に示す．横軸は書き手別の 10 編の作文から抽出して学習に用いたサンプルのサイズである．サンプルサイズが s のとき，書き手ごとのテストの標本サイズは $10-s$ である．

図 3 からわかるように，文学作品の場合と同じくランダムフォレスト法，ブースティング法，バギング法の正解率が高い．その中で最も高いのがランダムフォレスト法であり，学習の標本サイズが 6 までの正解率は 90%を超えている．しかし，図 2 と比較すると学習データの標本サイズの減少に伴う正解率の低下は作文データの方が著しい．

3.3.3 日記データ

表 3 に示す日記文においては，文章が短いためより多くの項目を抽出するため，単語の合計頻度が 10 以上の項目をそれぞれ 1 つの項目とし，10 以下の項目を「その他」にまとめて用いることにした．このようにした場合，項目数の合計は 241 である．図 4 にその F 値の平均プ

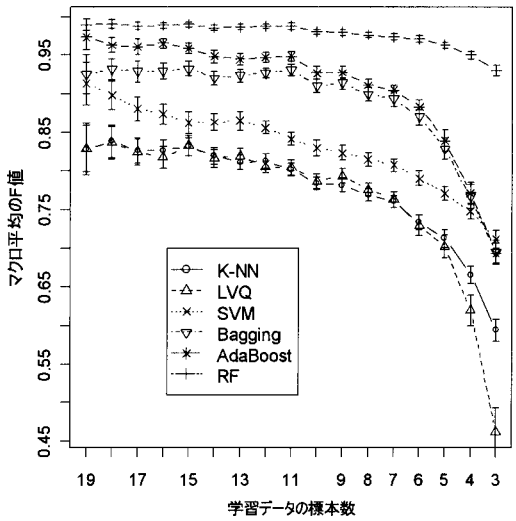

図 2. 文学作品の書き手の同定結果.

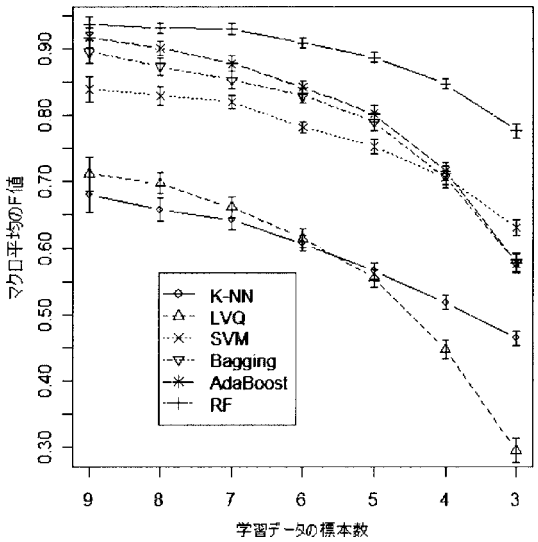

図 3. 作文の書き手の同定結果.

ロットを示す．横軸は書き手別の 10 編の日記から抽出して学習に用いたサンプルのサイズである．サンプルサイズが s のとき，書き手ごとのテストの標本サイズは $10 - s$ である．

図 4 からわかるように，日記文における正解率が最も高いのはランダムフォレスト法であり，正解率 90%以上を得るために必要な学習データの標本サイズは 7 以上である．学習データの標本サイズの影響はサポートベクターマシンとほぼ同じ程度で，最もよい．

ランダムフォレスト法の F 値の平均と標準誤差を表 5 に示す．ただし，文学作品について

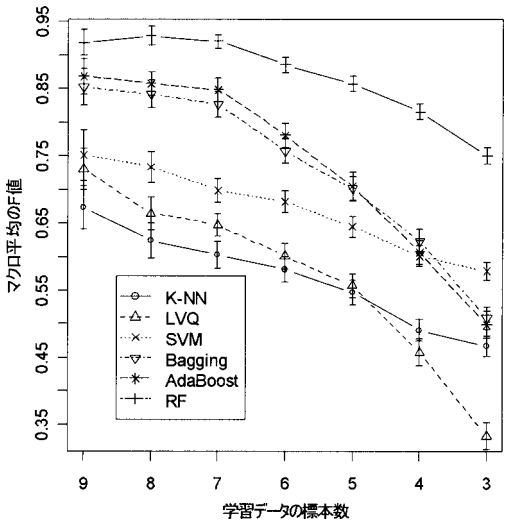

図 4. 日記文の書き手の同定結果.

は作文と日記に合わせた学習用の標本サイズ部分のみを示す．このように用いた文学作品，作文，日記 3 種類のテキストデータにおいて，すべてランダムフォレスト法の正解率が最も高く，標本誤差も最も小さい．

表 5 から分かるように文学作品に比べ，作文，日記の F 値は標本サイズの影響が大きい．その主な原因は，用いたテキストの長さにあり，テキストが短いほど用いたデータは安定性に欠け，そのデータの質が精度に影響を与えていると考えられる．もちろん，テキストのジャンルの影響を排除するものではない．

作文と日記の場合は，学習サンプル数が 9 のときの F 値が学習サンプル数が 8 のときより若干低くなっている．これは学習サンプルが 9 のときのテストデータが 1 つであるのが原因であると考えられる．これらに関しては，さらなる実証研究を重ねることが必要である．

書き手の同定の精度は，用いた変数の数(ベクトルの長さ)にも依存する．ちなみに，作文における項目の頻度が合計 10 以下のものを「その他」としてまとめた 727 変数を用いた場合と表 5 に示した 90 変数を用いた場合の結果を表 6 に示す．表 6 から分かるように 727 変数を用

表 5. ランダムフォレスト法のマクロ平均の F 値.

学習データの標本サイズ		9	8	7	6	5	4	3
文学作品	平均	0.9797	0.9759	0.9744	0.9698	0.9599	0.9504	0.9286
	標準誤差	0.0021	0.0017	0.0021	0.0021	0.0022	0.0024	0.0035
作文	平均	0.9346	0.9430	0.9202	0.9114	0.8886	0.8604	0.7753
	標準誤差	0.0061	0.0039	0.0041	0.0039	0.0039	0.0040	0.0055
日記	平均	0.9188	0.9289	0.9207	0.8854	0.8570	0.8155	0.7498
	標準誤差	0.0100	0.0073	0.0052	0.0058	0.0059	0.0061	0.0065

表 6. 作文データにおける学習に用いた標本サイズごとのマクロ平均の F 値.

学習データの標本サイズ		9	8	7	6	5	4	3
90 変数	平均	0.9346	0.9430	0.9202	0.9114	0.8886	0.8604	0.7753
	標準誤差	0.0061	0.0039	0.0041	0.0039	0.0039	0.0040	0.0055
727 変数	平均	0.9658	0.9663	0.9588	0.9506	0.9350	0.9121	0.8604
	標準誤差	0.0048	0.0033	0.0033	0.0027	0.0030	0.0031	0.0038

表 7. 日記文の書き手の同定の混同マトリックス (100 回の平均, 主対角線が正解率).

	A	B	C	D	E	F	誤り率
A	0.939	0.020	0.020	0.005	0.016	0.000	0.061
B	0.039	0.689	0.000	0.006	0.000	0.266	0.311
C	0.018	0.048	0.825	0.033	0.076	0.000	0.175
D	0.000	0.005	0.007	0.971	0.000	0.017	0.029
E	0.000	0.000	0.001	0.003	0.996	0.000	0.004
F	0.000	0.103	0.000	0.008	0.000	0.889	0.111

いた場合の精度は 90 変数を用いた場合より高い.

また, 書き手の同定の精度は書き手によってバラツキが大きい場合がある. ちなみに, ランダムフォレスト法による日記文の書き手の同定結果を表 7 に示す. この結果は 100 回の OOB データによるテストの平均である. 表 7 から分かるように正解率が 70%未満から 99%以上までばらついている.

4. おわりに

本研究では, ランダムフォレスト法による書き手の同定を提唱し, その有効性を示すため, 広く知られている分類法との比較分析を行った. 実証研究には, 3 種類の文章を用いた. その結果, 3 種類の文章の全てにおいて, ランダムフォレスト法が最もよい結果を示した. その次によい結果を示したのはブースティング法とバギング法である. この 3 つの方法は, 全て複数の決定木を用いた集団学習を行うアルゴリズムである. これらの方法のアルゴリズム内部では, データセットの中のから分類に有効となる変数が選択されて用いられている. ランダムフォレスト法がバギング法, ブースティング法よりよい結果を示しているのは, バギング法とブースティング法が作成する決定木は全ての変数の中から幾つか選択して用いているのに対し, ランダムフォレスト法が作成する決定木は全ての変数の中からサンプリングしたブートストラップサンプルを用いているためであると考えられる.

比較に用いた分類法の中には, 変数を選択せずに全ての変数を用いるのもある. したがって, 同じの変数を用いた場合の分類法の正解率などに関しては, さらなる実証研究が必要である. これらに関しては別紙に譲ることにする.

本研究での 3 種類のデータセットの正解率は「日記<作文<文学作品」であり, この正解率の順位は文章の平均長さ「日記<作文<文学作品」と対応する. しかし, 用いる文章のジャンルが異なるため, 同定精度の差が文章の長さの影響であるとは言い切れない点が問題として残

る．文章のジャンルが書き手の同定の精度に与える影響に関しても今後の課題である．

また，ランダムフォレスト法が得意・不得意なデータ構造などに関しても更なる実証研究が必要である．

最後に，貴重な時間を割いて丁寧に査読をしていただき，建設的なご助言を下さった査読者に感謝致する次第である．

参 考 文 献

Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–23.

Cover, T. M. and Hart, P. E. (1967). Nearest Neighbor Pattern Classification, *IEEE Transaction on Information Theory*, **IT-B**(1), 21–27.

Cristianini, N. and Shawe-Taylor, J. (2005). 『サポートベクターマシン入門』(大北 剛 訳), 共立出版, 東京.

Diederich, J., Kindermann, J., Leopold, E. and Paass, G. (2003). Authorship attribution with support vector machines, *Applied Intelligence*, **19**(1–2), 109–123.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm, *Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156, Morgan Kaufmann, San Fransisco.

Hoorn, J. F., Frank, S. L., Kowalczyk, W. and Ham, F. (1999). Neural network identification of poets using letter sequences, *Literary and Linguistic Computing*, **14**(3), 311–338.

Jin, M. and Murakami, M. (1993). Author's characteristic writing styles as seen through their use of commas, *Behaviormetrika*, **20**(1), 63–76.

Joachims, T. (1998). Text categorization with support vector machines, *Proceedings of ICML-99, 16^{th} International Conference on Machine Learning* (*Bled, SL*), 200–209.

金 明哲 (1994). 読点の打ち方と文章の分類, 計量国語学, **19**(7), 317–330.

金 明哲 (1995). 動詞の長さの分布に基づいた文章の分類と和語および合成語の比率, 自然言語処理, **2**(1), 57–75.

金 明哲 (1997). 助詞の分布に基づいた日記の書き手の認識, 計量国語学, **20**(8), 357–367.

金 明哲 (2002). 助詞 n-gram 分布を用いた書き手の識別, 計量国語学, **23**(5), 225–240.

金 明哲 (2003). 自己組織化マップと助詞分布を用いた書き手の同定およびその特徴分析, 計量国語学, **23**(8), 369–386.

金 明哲 (2004). 品詞のマルコフ遷移の情報を用いた書き手の同定, 日本行動計量学会第 32 回大会.

金 明哲 (2005). 決定木と集団学習, *ESTRELA*, No.133, 62–67.

金 明哲, 樺島忠夫, 村上征勝 (1993). 読点と書き手の個性, 計量国語学, **18**(8), 382–391.

Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifiers, *Literary and Linguistic Computing*, **9**(2), 119–124.

Kohonen, T. (1985). *Self-Organizing Maps and Associative Memory*, Springer Series in Information Science, **30**, Springer-Verlag, Berlin, New York. (徳高平蔵 他 共訳 (1996). 『自己組織化マップ』, シュプリンガー・フェアラーク東京)

松浦 司, 金田康正 (2000). n-gram の分布を利用した近代日本文の著者推定, 計量国語学, **22**(6), 225–238.

松本裕治, 北内 啓, 山下達雄, 平野善隆, 松田 寛, 高岡一馬, 浅原正幸 (2003). 形態素解析システム『茶筌』version 2.3.3 使用説明書, 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座.

Matthews, R. A. J. and Merriam, T. V. N. (1993). Neural computation in stylometry I: An application

to the works of Shakespeare and Fletcher, *Literary and Linguistic Computing*, **8**(4), 203–210.
Mendenhall, T. C. (1887). The characteristics curves of composition, *Science*, **IX**, 237–249.
村上征勝(2002).『文化を計る―文化計量学序説―』, 朝倉書店, 東京.
Sebastiani, F. (2002). Machine learning in automated text categorisation, *ACM Computing Surveys*, **34**(1), 1–47.
Teng, G., Lai, M., Ma, J. and Li, Y. (2004). E-mail authorship mining based on SVM for computer forensic, *Machine Learning and Cybernetics*, Proceedings of 2004 International Conference on, Vol.2, 1204–1207.
Tweedie, F. J., Singh, S. and Holmes, D. I. (1996). Neural network application in stylometry: The federalist papers, *Computer and the Humanities*, **30**, 1–10.
Vapnic, V. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.
Waugh, S., Adams, A. and Tweedie, F. (2000). Computational stylistics using artificial neural networks, *Literary and Linguistic Computing*, **15**(2), 187–198.
安本美典, 本多正久(1988).『因子分析法』, 培風館, 東京.
Zheng, R., Li, J., Chen, H. and Huang, Z. (2005). A framework for authorship identification of online messages: Writing-style features and classification techniques, *Journal of the American Society for Information Science and Technology*, **57**(3), 378–393.

268 Proceedings of the Institute of Statistical Mathematics Vol. 55, No. 2, 255–268 (2007)

Authorship Identification Using Random Forests

Mingzhe Jin and Masakatsu Murakami

Faculty of Culture and Information, Doshisha University

This paper proposes the use of Random Forests (RF) for authorship identification. It also reports a comparative study between RF and the following classifiers: k Nearest Neighbor, Support Vector Machines, Learning Vector Quantization, Bagging, and Boosting (AdaBoosting). We focused on the relationship between the performance of the classifiers in authorship identification and the size of training data. In this study, the following three different styles of text were used: 200 novels written by 10 great writers, 110 compositions written by 11 undergraduates, and 60 diaries written by 6 non-eminent writers. It is shown the that Random Forests algorithm is more effective and stable than the other classifiers.

Key words: Authorship identification, text classification, stylometrics, Random Forests.