集団学習法による文章の書き手の同定

金 明哲, 村上 征勝
同志社大学文化情報学部

概要：本研究では，文章の書き手の同定における集団学習手法の分類器の精度と学習データの標本サイズとの関連性などについて実証研究を行った．比較分析のため分類器としては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，ランダム森法を用いた．また実証材料としては10人が書いた200の小説と11人が書いた110の作文を用いた．その結果，集団学習法の精度が最も高く，特にランダム森法のパフォマンスの良さが実証された．

Authorship identification with ensemble learning

Mingzhe Jin, Masakatsu Murakami
Faculty of Culture and Information
Doshisha University

ABSTRACT: In this paper we analyzed the relationship between the performance of classifiers for identifying authorship and the size of training data. We employed k-NN(k Nearest Neighbor), SVM(Support Vector Machines), LVQ(Learning Vector Quantization), BAG(Bagging), ADA(Boosting) and RF(Random Forests) classifiers with 200 novels written by 10 great writers and 110 compositions written by 11 undergraduates. The experimental results showed that the RF is more powerful and robust than other classifiers.

Key words: authorship identification, text classification, stylometrics, ensemble learning, Random Forests

1. はじめに

文体分析の研究では，古くから書き手の文体特徴をもとに，計量的に文章の書き手を推定・同定する研究が行われている．例えば，オハイオ州立大学の地球物理学者Mendenhallは1887年に，単語の長さの分布における書き手ごとの文章の特徴について分析した結果を『サイエンス』誌に発表した[11]．彼はディケンズ(Dickens, C., 1812～1870)，サッカレー(Thackeray, W.M., 1811～1863)，ミル(Mill, J.S., 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り，作家の特徴になることを示した．1960年前後には，判別分析の手法が書き手の判別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では，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，その基本的なアイディアは，今日のテキストマイニングの原型であると言えよう．

ところで，われわれ人間は，文章に関する素養をある程度持っていれば，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可能である．これは，われわれがそれぞれのジャンルの文章の形式（パターン）に関する知識を

持っているからである．もしそのような事前の知識がなければ，文章のジャンルを見分けることは不可能であろう．一方，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見分けす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，その作家の何らかの独特な文章のパターンが，その愛読者の脳に焼付けられたためであると考えられる．しかし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解明されていないのが現状である．

近年，人間が行っている情報処理の多くをコンピュータに任せることが可能になりつつあるが，このような人間の脳で行われている文体に関する処理をコンピュータで実現するにはどのようにすればよいかに関してさまざまの試みが行われている．その中の一つが，パターン認識のアプローチによる文体識別である．

機械によるパターン認識は，パターンの特徴の抽出とアルゴリズムによる識別という 2 つのステップに分けられる．著者不明の文章の書き手を同定する場合でもこの一般性は失われない．本研究では，書き手の同定におけるアルゴリズムの適応性及び学習データの標本サイズとの関連性について実証研究を行う．

文章の書き手の同定に関する研究はテキストの分類との類似点が多い．テキストの分類に関して，Sebastiani は近年の 40 を超える研究結果について総括を行った[12]．分類の精度は用いたテキストのコレクション，用いた特徴ベクトルと関係しているので，報告された分類精度に関する評価は絶対的のものではないが，K 近隣法(k nearest neighbor)，SVM 法，ニューラルネット法がよいパホマンスを示している．

テキストマイニングにおけるテキスト分類の 1 つの特徴は大標本である．研究事例を見ると，学習に用いたテキストの数は約 1 万前後で，テストに用いた文書の数は数千に上る．しかし，書き手不明である文章の著者を同定する問題では，通常学習及びテストに用いる文章は数十であり，場合によっては数編しかない．一方，機械的に文章から書き手の文体特徴要素(変数と呼ぶことにする)を集計するとしばしば数百，数千を超えてしまう．したがって，通常のテキストマイニングにおけるデータ構造と文体分析におけるデータ構造は異なる．

書き手の同定に関しては，線形判別モデル，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用いられている先行研究はあるが，選別された少量の変数を用いるのがほとんどである．変数を選別せずに，書き手の同定を行う方法としては，サポートベクターマシンが注目を集められている．どのような方法が文体研究における小標本，大量の変数のデータ構造に適しているかに関する総合的な比較研究は見あたらない．そこで，本研究では，近年パターン認識で用いられている主たる統計方法及び機械学習法(k 近傍法，サポートベクターマシン，学習ベクトル量子化，バギング，ブースティング，ランダム森)の書き手の同定におけるパフォーマンス及び学習に用いる標本サイズとの関連性などについて比較分析を行う．

2. データと分類手法

2.1 用いたデータ

本稿では，表 1 に示す 10 人が書いた 200 の小説，表 2 に示す 11 人が 10 のタイトルについて書いた 110 の作文を用いた．文学作品においては，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．用いた小説の文章の選定に関しては，旧漢字が少ないこと，なるべく同年代であることなどに配慮した．用いた文章の長さはアンバランスになっているので，文章から抽出した要素については相対頻度を用いた．本研究では，文章の中の地の文のみを用いた．書き手の同定を行う際には，文章の中のどのような要素を用いるかが一つの鍵となる．日本現代文における非特定の書き手の特徴に関しては，いくつかの実証研究が報告されている[18-33]．

本研究は，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おいているので，ノイズが多く含まれていると思われるタグ付き単語の n-gram を用いた．

表１．　分析に用いた文章のリスト(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，ヰタ・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，伯爵の叙，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，鶴狩
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，千柿の鍔，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　分析に用いた作文のサイズ

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;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

単語の品詞タグ付けは，形態素解析ソフト「茶筅」を用いた．茶筅の品詞情報は階層的になっている．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．

本研究では，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$P_i=(p_{i1},p_{i2},\cdots,p_{ij},\cdots,p_{in})$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．

2.2　用いた分類器

本稿では，分類アルゴリズムを分類器(classifier)と呼び，近年注目されているサポートベクターマシン，学習ベクトル量子化，バギング，ブースティング，ランダム森を中心とする．ただし，比較のために古典的な分類法ｋ近傍法も用いる．古典的分類器としては，ベイジアンのアプローチによるナイーブベイズ分類器(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な結果が得られたので本報告では比較対象から取り除いた．

2.2.1　ｋ最近隣法

ｋ最近隣(k-NN: k Nearest Neighbor)法は伝統的なパターン分類器である[3]．ｋ最近隣法は，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．

2.2.2　学習ベクトル量子化

人間の脳で行う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口ニューラルネットワーク(ANN: Artificial Neural Network)と呼ぶ．ANNにはさまざまなタイプがあるが，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いのが一つの短所である．階層的ニューラルネットワークモデルに基づいた文章の著者の同定に関する研究事例としては[6], [8], [10], [14], [16]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が実用的ではないため，これらの研究では変数を選別して用いている．

本研究では，機械的に抽出したデータの変数を選別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．

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教師データを用いるパターン認識の手法である．その特長は，伝統的な統計的手法より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[9]．

LVQ は一定の数のコードブック・ベクトル(codebook vector) m_i ($i=1,2,3,\cdots,k$)の中から個体$\mathbf{x}$に最も近似するコードブック・ベクトル$\mathbf{m}_c$を見つける近似法である．

$$\|\mathbf{x}-\mathbf{m}_c\|=\min_i\|\mathbf{x}-\mathbf{m}_i\|$$

LVQ は，LVQ1，LVQ2，LVQ3 や OLVQ1 のようなアルゴリズムの集合の総称である．これらの各々のアルゴリズムは，すべてが教師データありの学習法である．LVQ1 は，学習の収束を早めるのに工夫されていることに対し，LVQ2，LVQ3 は，よりロバストに設計されている．OLVQ1 は LVQ1 の収束性を早めるように工

夫したものである．学習時間に制限がある場合は，OLVQ1 の学習のみで十分であるが，データによっては OLVQ1 が返したコードブック・ベクトルを用いて LVQ1，LVQ2，LVQ3 で学習を続けることで精度を上げることが可能である場合もある．LVQ の開発者は，学習の最適化は常に収束性が早い LVQ1(OLVQ1)のアルゴリズムで始めるべきであることを勧めている．LVQ をテキスト分類に適応した研究事例は見あたらない．

2.2.3 サポートベクターマシン

サポートベクターマシン(SVM：Support Vector Machin)は，Vapnik が 1990 年代の中頃に提案したパターン分類の方法である[15]．SVM はグループを分ける無限に存在する超平面の中から，最もグループ分けが良い平面をマージン最大の基準で求める方法である．

学習データ集合($\mathbf{x}_1, y_1$)，($\mathbf{x}_2, y_2$)，・・・($\mathbf{x}_m, y_m$)があるとする．ここの $\mathbf{x}=(x_1, x_2, \cdots, x_n)$ は個体の特徴ベクトルで，y は目的変数で，回帰問題では数値，分類問題ではクラスのラベルである．線形回帰と線形判別の問題では次に示す線形モデルを用いる．

$$y = f(\mathbf{x}) = \sum_{i-1}^{n} w_i x_i + b$$

初期の SVM は，2 群線形分類器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器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 SVM である．

$$y = f(\mathbf{x}) = \sum_{i=1}^{N} w_i K(\mathbf{x}_i, \mathbf{x}) + b$$

SVM をテキスト分類に適応した研究事例としては[4]，[7]，[13]，[17]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．

2.2.4 集団学習法

集団学習とは決して精度が高くとは言えない複数の結果を組み合わせる方法で，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．機械学習方法としてバギング，ブースティング，ランダム森法がある．

(1) バギング

バギング(bagging)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．バギングは 1996 年 Breiman によって提案された[1]．バギング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作成した回帰・分類結果を統合・組み合わせる方法で精度を向上させる．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．

(2) ブースティング

ブースティング(boosting)は与えた教師付き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[5]．そのパフォマンスは高く評価されている．しかし，計算機の資源が十分ではない個人ユーザにとっては，大きいサイズのデータを解析することには困難が伴うであろう．ちなみに，本研究で用いたデータ 200×2103 のマトリックスを用いるのに 17284 Kb のベクトルを割り当てるメモリが求められる．

(3) ランダム森

ランダム森(RF; random forest)はバギング提案者 Breiman が，今世紀の初めに提案した大量の決定木による集団学習アルゴリズムである [2]．ランダム森とバギングとの大きい違いは，バギングは全ての変数を用いるが，

ランダム森は変数もランダムサンプリングしたサブセットを用いるので，高次元データ解析に向いている．ランダムサンプリングする変数の数Mはユーザが自由に設定することができる．Breimanは，Mはデータセットの変数の数の正の平方根を取ることを勧めている．

ランダム森のアルゴリズムを次に示す．

① 与えられたデータセットからB組のブートストラップサンプルを作成する．

② 各々のブートストラップサンプルデータを用いて未剪定の最大の決定・回帰木を生成する．ただし，分岐のノードはランダムサンプリングされた変数の中の最善のものを用いる．

③ すべての結果を統合・組み合わせ(回帰の問題では平均，分類の問題では多数決)，新しい予測・分類器を構築する．

ランダム森のパフォマンスは，バギング，ブースティングより優れているとの報告はあるが，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．

3. 分類の比較

3.1 学習とテスト

教師データありの分類の問題では，学習の結果を学習に用いていないデータでテストを行うことが必要である．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均等にN等分にし，その中のN-1等分を学習用とし，残りの1等分をテスト用とする．このような学習テストを交差しながらN等分のすべてについてテストを行い，その平均を結果とする．このような方法は，サンプルサイズが小さいときにはランダムに分割を行う際の偏りの影響が大きい．

そこで本研究では，データセットからランダムに「著者数×k」($k=1,2,3,\cdots,s$)の個体を抽出してテスト用とし，残りを学習用のデータとする．学習およびテストを繰り返した算術平均を結果とする．

クラスc_i(本研究では書き手)における判別・同定の結果は交差行列で表すことができる．

クラスc_i		分類器の結果	
		Yes	No
データ	Yes	a_i	c_i
	No	b_i	d_i

結果評価には，再現率(recall)や精度(precision)などが多く用いられている．再現率は，分類器がどれぐらい「漏れ」なく正しく判別しているかに関する度合であり，精度は分類器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．

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

マルチクラスの分類の問題では，評価指標として再現率，精度のマクロ平均(macro average)とマイクロ平均(micro average)がある．本研究では，次に示すマクロ平均を用いる．マクロ平均はクラス$c_1,c_2,\cdots,c_i,\cdots,c_m$における次の式で定義されている．

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

また，再現率と精度の要約の測度としてF_β値がある．F_β値は次のように定義されているが，通常F_1が多用されている．

$$F_\beta=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}$$

3.2 実験結果

3.2.1 文学作品

表1に示す文学作品について，1つの文章に現れる単語の頻度が平均1回以上である単

語は495種類である．本研究では，それ以外のものは「その他」の項目にまとめた．合計496項目を用いた10人の200文章のタグ付き単語の相対頻度分布を用いて書き手の同定を繰り返したマクロ平均のF_1値を図1に示す．図1の横軸は20文章から抽出して用いた学習のデータの標本サイズであり，縦軸はマクロ平均のF_1である．サンプルサイズがkのとき，テストの標本サイズは20-kである．エラーバーは95%信頼区間である．

図1からわかるように，集団学習法の精度が全体的に高く，最も高いのはランダム森法であり，標本サイズの影響も最も小さい．学習データの標本サイズの減少に伴い，SVMとランダム森法の差が大きくなる点が興味深い．

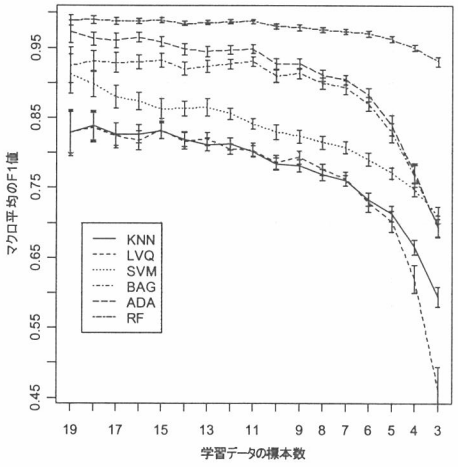

図1　表1の文学作品の同定結果

3.2.2 作文データ

表2の作文におけるタグ付き単語についても，1つの文章に平均1回現れることを基準とし，それ以外の単語は「その他」の項目にまとめると，項目の数は合計90になる．学習・テストを繰り返したマクロ平均のF_1値の平均プロットを図2に示す．エラーバーは，95%信頼区間である．横軸は書き手別の10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kのとき，テストの標本サイズは10-kである．

図2からわかるように，文学作品の場合と同じくランダム森(RF)，ブースティング(ADA)，バギング(BAG)の精度が高い．最も高いのがランダム森法で，学習データの標本サイズの影響も最も小さい．

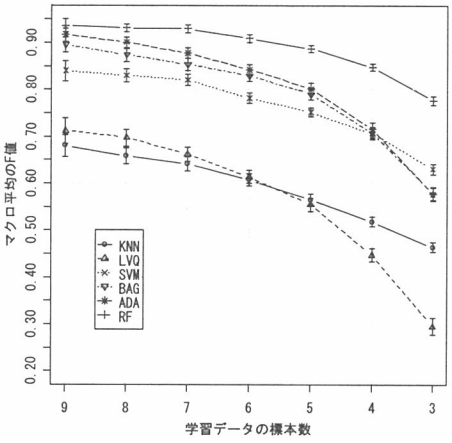

図2　表2の作文の書き手の同定結果

4 おわりに

本研究では，集団学習法に基づいた分類器と近年注目されている幾つかの分類器について，書き手の同定における精度と学習データの標本サイズとの影響について，実証研究を行った．その結果，集団学習法による分類器の精度が用いたその他の方法よりよい．その中でもBreimanのランダム森法が最もよい結果を示した．また，ランダム森法の分類器はバギング，ブースティングよりロバストであり，ブースティングと比べると計算量も少ない．

参考文献

[1] Breiman, L.(1996); Bagging Predictors, Machine Learning, 24, 123-140.

[2] Breiman, L.(2001); Random Forests, Machine Learning, 45, 5-23.

[3] Cover, T. M. and Hart, P. E. (1967). Nearest Neighbor Pattern Classification. IEEE Transaction on Information theory, IT-B(1), 21-27.

[4] Diederich J., Kindermann J., Leopold E., Paass G.(2003). Authorship Attribution

with Support Vector Machines, Journal Applied Intelligenc, Vol. 19, No. 1-2, 109-123.
[5] Freund, Y. and Schapire, R.E. (1996): Experiments with a New Boosting Algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148-156, Morgan Kaufmann.
[6] 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.
[7] Joachims, T. (1998). Text categorization with support vector machines. In Proceedings of ICML-99, 16th International conference on Machine Learning (Bled, SL), 200-209.
[8] Kjell, B. (1994). Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers, Literary and Linguistic Computing, 9(2), 119-124.
[9] Kohonen, T. (1985). Self-Organizing Maps and Associative Memory. Springer Series in information Science 30. Springer-Verlag. (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)
[10] 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.
[11] Mendenhall , T.C. (1887). The Characteristics Curves of Composition. Science, IX, 237-249.
[12] Sebastiani, F. (2002). Machine Learning in Automated Text Categorisation. *ACM Computing Surveys*. Vol. 34, No. 1, 1-47.
[13] Teng, G., Lai, M., Ma, J., Li, Y. (2004). E-mail authorship mining based on SVM for computer forensic, Machine Learning and Cybernetics, Proceedings of 2004 International Conference on, Vo. 2, 1204 - 1207
[14] 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.
[15] Vapnic, V. (1995). The Nature of Statistical Learning theory. Springer, New York.
[16] Waugh, S., Adams, A. and Tweedie, F. (2000). Computational stylistics using artificial neural networks, Literary and Linguistic Computing, 15(2), 187-198.
[17] Zheng, R., Li, J., Chen, H., 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, Vol. 57, Issue 3, 378-393
[18] 金 明哲(1994)読点の打ち方と文章の分類, 計量国語学, 19巻7号, 317-330.
[19] 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及び合成語の比率, 自然言語処理, Vol. 2, No. 1, 57-75.
[20] 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, Vol. 5, No. 2, 13-21.
[21] 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巻8号, 382-391.
[22] 金 明哲(1994). 読点の打ち方と著者の文体特徴, 計量国語学, 19(7), 317-330.
[23] 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, Vol. 2, No. 1, 57-75.
[24] 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.
[25] 金 明哲(2002a). 助詞分布における書き手の特徴に関する計量分析, 社会情報, Vol. 11, no. 2, 15-23.
[26] 金 明哲(2002b). 助詞n-gram分布を用いた書き手の識別, 計量国語, 23(5), 225-240.
[27] 金 明哲(2003a). 自己組織化マップと助詞分布を用いた書き手の同定及びその特徴分析, 計量国語学, 23巻8号, 369-386.
[28] 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.
[29] 金 明哲・他(2003). 統計科学のフロンティア「言語と心理の統計」, 岩波書店.
[30] 村上 征勝(2002) 文化を計る－文化計量学序説－, 朝倉書店
[31] 村上 征勝 (2004). シェークスピアは誰ですか？－計量文献学の世界－, 文芸春秋
[32] 松村 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.
[33] 安本 美典・本多 正久(1988). 因子分析法, 培風館.