サポートベクタマシンを用いた著者判別における有効素性推定

吉田 篤弘　斎藤 博昭
慶應義塾大学大学院 理工学研究科 開放環境科学専攻
Email: {atsuhiro, hxs}@nak.ics.keio.ac.jp

1 はじめに

近年発展中の自然言語処理分野の一つに著者判別研究がある. この研究は著者各人の癖を示す文章上の特徴 (素性) を決め, 数値化し, 決定木・最大エントロピー法・サポートベクタマシンなど様々な手法を用いて学習し, その結果に基づいて判別というサイクルで行われる. その中で特に重要と考えられる作業の1つが素性選択の問題であるが, 従来の日本国内における研究では主に経験則から素性を決定しており, 相対的な有効性という観点からこの素性選択を論じた研究はほとんどないと言える. その数少ない研究の1つに著者らが行った [9] があるが, 局所解の問題を解決できていないなど実験条件に改善の余地が残されており, 結果をそのまま論じることは難しい.

そこで本稿では, 様々な素性について統計的に比較検討を行い, 従来より高精度な著者判別を可能とする素性の組み合わせについて考察する.

2 素性に関する研究

素性同士を比較・検討した研究は, 海外における研究や, 前述の4サイクルにより行われる点で類似した分野であるカテゴリ分類においてはいくつか為されている. しかし句読点数や品詞など, 文の構造情報に関する素性を使用した方が単語情報を用いるより有効であったとする [1] と, 逆に全ての単語を用いた方が構造情報を用いた場合より高い判別率が得られたとする研究 [2] との結果の相違など, 各々の結論には差が見られる. さらに [3] において素性の追加が判別精度を低下させる例が報告されていること, かつ大量の素性の追加は計算量や効率の面で現実的でないことから, 素性選択にあたっては何らかの指針が求められる. また [4] においては品詞のフィルタリングの有効性, [5] では単語ベースの素性に対する文字列ベースの素性の有効性が述べられているが, 別分野であるカテゴリ分類研究での結果が著者判別分野にも適用できるか否かの判断は実験による検証を必要とする.

以上のことから, 日本語の著者判別研究において有効な素性, 及びそれぞれの素性についての最適な条件の検討が要求される.

3 サポートベクタマシン

サポートベクタマシン (SVM) は近年注目されているパターン識別手法の一種である. d 個の素性で表される訓練事例を d 次元ベクトルとして扱い, この訓練事例の集合を d 次元ユークリッド空間上で2クラスに分類することで学習を行う.

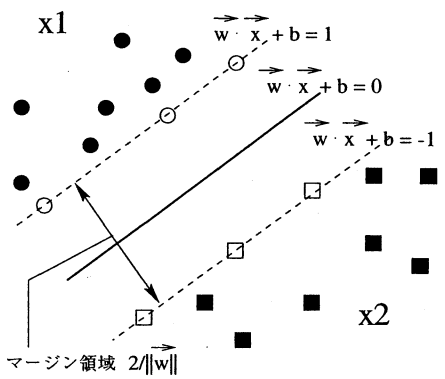

図 1: 分離平面及びマージン

2クラスの分類を行う図1中の分割平面 $\vec{w}\cdot\vec{x_i}+b=0$ は破線 $\vec{w}\cdot\vec{x_i}+b=\pm 1$ 内部の領域 (マージン領域) を最大化するよう一意に求められるため, SVM には局所解の問題がない. さらに SVM はマージン領域への訓練事例への進入の許可やカーネル関数の使用などにより, 図1のような線形分離が不可能な場合においても対応できること, 数万次元もの大量の素性を扱えること, などの長所も有している.

さらに作成した学習器を複数組み合わせることで多値分類にも適用でき, その判別結果も他の学習手法より良好であることが報告されている [2]. このことから, 本研究では TinySVM[7] を使用して学習を

行う.

4 実験概略

本実験においては, 青空文庫 [6] などから取得した近現代日本小説家 22 名の作品群を使用する. これらの著者を一部重複を許して「大正」「昭和初期」「平成」「混合 (明治～昭和初期までの期間)」の 4 グループに分割し, それぞれについて構成されるテキストを変えた学習用データセット 4 つを作成する. 4 つの学習用セットは集積するテキストの構成に違いを持たせ, 取得した作品の内容に左右されない結果を取得できるようにする. また学習後に判別を行わせるテスト用データとして, 各著者の作品全てから構成されるファイルを作成する. 一方, 文章からの取得が容易という観点から表 1 の 10 種類の素性を設定する.

表 1: 本実験で使用する素性

(a)	1 文辺りの長さ
(b)	文字種の割合
(c)	ひらがな各文字の割合
(d)	句点前の各文字の割合
(e)	読点前の各文字の割合
(f)	各品詞の割合
(g)	文頭における各品詞の割合
(h)	文末における各品詞の割合
(i)	文字 3gram の割合
(j)	全形態素の出現割合

そしてこれらの素性とデータ集合を基に素性ベクトルファイルを作成し, ある基準著者の作品を正事例, それ以外の著者の作品を負事例と設定する. その上で SVM による学習及び判別を行い, 基準著者の作品を判別できるか否かの確認を行う. 以上の作業の流れを図 2 に示す.

なお実験に際しては one vs. rest 法を使用, グループに含まれている著者の数だけ識別器を作成し, それぞれの結果を求めるものとする. また, 判別に際しては適合率 (出力正解数/出力数) 及び再現数 (出力正解数/全体正解数) から算出される F 値を評価の上の尺度として使用する.

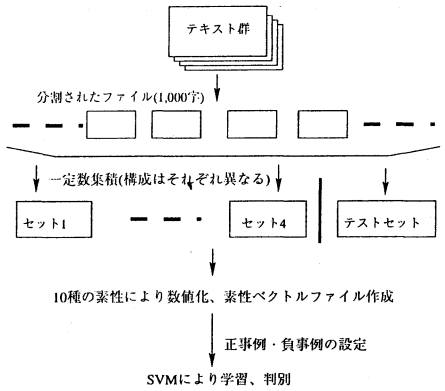

図 2: 学習用及びテスト用ファイル群の生成

5 実験と考察

素性及び品詞を全て使用, また頻度の閾値を 2 としたものを基本条件とした上で以下の実験を行った. なお, それぞれの結果は 4 セットの平均値を表すものとする.

5.1 実験 1 : フィルタリングの効果の有無について

数万個の素性を扱える SVM であっても, あまりに情報量の低い要素まで扱うことは効率及び精度の面で問題であることは自明である. 故に素性 (f)(g)(h)(j) に関して品詞の, (d)(e)(i)(j) に関して頻度のフィルタリングを施し, これらが著者判別において必要であるかどうかの判断及びこれらを削除することの妥当性について検証する. なお, 品詞に関しては表 2 に示す 4 段階を設定し, 各段階においてはそれぞれの品詞のみを使用する. また頻度に関しては閾値 0(全て使用), 2, 5, 10 の 4 段階を設定し比較を行う.

表 2: 品詞によるフィルタリング

(1)	名詞, 動詞
(2)	(1) +形容詞, 形容動詞, 未知語
(3)	(2) +副詞, 連体詞, 助動詞
(4)	全ての品詞

この実験結果を図 3 及び図 4 に示す.

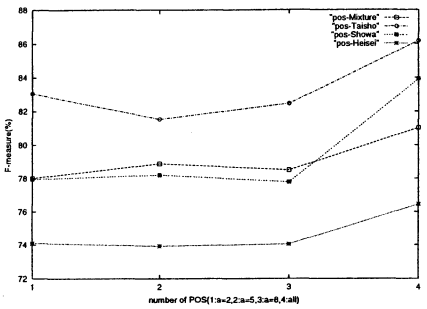

図 3: 品詞フィルタリングの効果 (時代ごと)

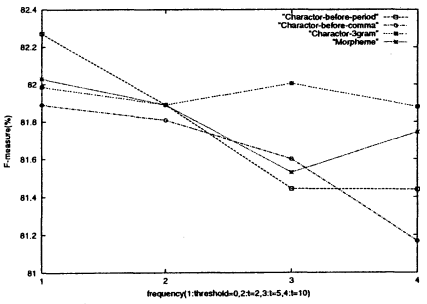

図 4: 頻度フィルタリングの効果の検証

品詞に関しては素性の如何によらず, 品詞数に制限を加えたものに比べて全ての品詞を用いた場合の結果の方が良好であった. 品詞数 8 からさらに増加したものとしては助詞や接頭詞などが挙げられることから, 著者判別に関しては言葉自体には意味のないそれらの品詞も重要な役目を持っており, 品詞ごとの選別は有効な操作でないと言える.

一方, 頻度に関してはフィルタリングの効果が素性 (i) 文字 3gram に関して現われていることが分かる.(i) は 10 種類の中でも生成素性数が特に多く, 不要な要素も比例して多くなることがこの操作が有効であった理由と考えられる. 逆にそれ以外の素性は頻度 0 が最も良好な結果を与えた. これらは素性 (i) に比べて生成数が低く, その分低頻度の素性にも無視できない情報量が含まれていることを示すと言える. このことはフィルタリングを行った素性 4 種の中で生成数の最も少ない (d) 句点前の文字の割合において, 頻度 0 に対する他の閾値設定時の判別率低減の度合が特に大きかったことからも明らかである.

5.2 実験 2 : 各素性の有効性について

前掲の 10 種の素性に対し, それぞれの相対的な有効性がどれほどのものであるかを測定する. これは, ある素性を削除した際の判別結果が全ての素性を使用した際の結果と比べてどの程度変化したかによって把握することが可能である. すなわち, (全素性使用) − (各素性削除) の値が正になれば削除された素性は有効であり, 逆に負値を与えた場合は判別に寄与しない不要な素性と判断できる. この実験結果を表 3 に示す.

表 3: 各素性削除時の判別結果 (単位:F 値 (%))

削除素性	混合	大正	昭和	平成
なし	81.0	86.2	83.9	76.4
(a)	81.0	85.0	81.2	71.4
(b)	80.1	85.7	82.5	75.7
(c)	79.8	86.2	83.6	76.4
(d)	79.4	86.1	82.7	76.4
(e)	80.7	83.6	82.3	74.1
(f)	80.0	82.0	80.7	76.6
(g)	80.3	82.6	82.8	75.9
(h)	84.0	84.6	84.7	73.3
(i)	80.7	84.1	83.3	76.6
(j)	79.0	81.7	82.3	75.9

表 3 において高い値を示したのは素性 (a)(e)(f)(j) などであった. 逆に (c)(h) などは全般的に値が低く, あまり判別に影響を与えていない素性と言える. しかし, グループごとの平均のみでは特定のセットの値に全体の結果が左右される可能性を考慮することができない. 故に実験 2 の別の評価方法として, 各グループにおいて負値を与えたセット数がどれほど存在したかを測定する. これを表 4 に示す.

表 4 から, 表 3 からは有効と判断できた素性 (a) はおよそ半分のセットにおいて負値を与えており, 常に有効であるとは言いがたいことが分かる. 一方素性 (e)(f)(j) は負値を与えることが少なく, 設定によらず安定して判別に影響を与えており, 表 3 と合わせて時代によらず安定して判別に高い影響を与える有効な素性と判断できる. また素性 (c)(h) は負値になる割合も高く, 全素性を使用した場合との差も比較的低くなった. この結果を鑑みると, これらの素性は削除した方が精度の面でも効率の面でも良いと判断できる.

表 4: 負値を与えたセット数

削除素性	混合	大正	昭和	平成	計
(a)	2	2	2	1	7
(b)	0	2	0	1	3
(c)	0	3	2	2	7
(d)	0	2	1	2	5
(e)	1	0	0	0	1
(f)	0	0	0	2	2
(g)	2	2	1	0	5
(h)	4	2	2	0	8
(i)	1	2	0	2	5
(j)	0	1	0	1	2

5.3 実験 3 : 検証実験

実験 1・2 の結果から最適と考えられる条件を「品詞全使用」「閾値 0(ただし (i) のみ 5)」「素性 (c)(h) 削除」と設定, 新規に作成した著者 13 名のグループに対して比較実験を行った. その結果を図 5 に示す.

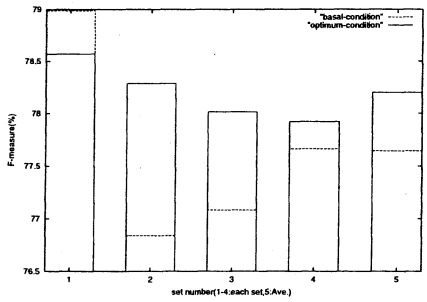

図 5: 検証実験の結果

図 5 から, 設定した条件が有効であることが確認された. セット 1 に関してのみ基本条件の結果が上回っているが, これは一部の著者において極度に値が低下したためであり, 設定条件により値が向上した著者数は低下した数を上回っていることから全体としてはこのセットに関しても有効であったと言える.

6 終わりに

様々な設定条件のもとでの比較及び検証実験により, 各種素性の品詞及び頻度へのフィルタリングの有効性を確認することができた. また日本語の著者判別において有効と考えられる素性を見出すことができた. これらの条件及び素性の適用により常に一定の成果を期待でき, さらに新規に有効な素性の組み合わせを使用してのより高精度な判別も可能になったと言える.

参考文献

[1] E.Stamatatos, N.Fakotakis, G.Kokkinakis : " Computer-Based Authorship Attribution Without Lexical Measures", COMPUTERS AND THE HUMANITIES, Vol.35, No.2, pp.193–214, 2001

[2] Joachim Diedrich, Jorg Kindermann, Edda Leopold, Gerhard Paass :" Authorship Attribution with Support Vector Machines ", Poster presented at the Learning Workshop, pp.1–17, 2000

[3] O.de Vel, A.Anderson, M.Corney, G.Mohay :" Mining E-mail Content for Author Identification Forensics ", SIGMOD Record, Vol.30, No.4, pp.55–64, 2001

[4] 平 博順, 春野 雅彦 :" Support Vector Machine によるテキスト分類における属性選択 ", 情報処理学会論文誌, Vol.41, No.4, pp.1113–1123, 2000

[5] 石田 栄美, 辻 慶太 :" 日本語テキストの自動分類のための特徴素抽出手法の比較 ", 情報処理学会自然言語処理研究会報告, NL Vol.151, No.7, pp.81–86, 2002

[6] 青空文庫, http://www.aozora.gr.jp

[7] Taku Kudo : TinySVM, http://cl.aist-nara.ac.jp/\~taku-ku/software /TinySVM, 2001

[8] 松浦 司, 金田 康正 :" 近代日本小説家 8 人による文章の n-gram 分布を用いた著者判別", 情報処理学会自然言語処理研究会報告, NL Vol.137, No.1, pp.1–8, 2000

[9] 吉田 篤弘, 延澤 志保, 平石 智宣, 斎藤 博昭: 著者判別に有効な特徴量の推定, 自然言語処理学会研究報告, NL Vol.145, No.13, pp80 - 90, 2001