段階的な品詞選択手法を用いたSVMによるテキスト分類

増山 毅司
東京大学大学院総合文化研究科
tak@r.dl.itc.u-tokyo.ac.jp

中川 裕志
東京大学情報基盤センター
nakagawa@dl.itc.u-tokyo.ac.jp

1 はじめに

近年, 新聞記事や特許文書をはじめ多くのテキストが電子化されている. これらの電子テキストを人手で分類することは, 手間とコストがかかり, また, 分類結果の一貫性を保つことが困難なため, 自動テキスト分類技術に対する期待が高まっている.

テキスト分類問題においては, テキストを互いに独立な集合とみなして, 各単語を素性に対応させるアプローチが一般的である. このようにしてテキストから生成された素性空間は, 素性の次元数が非常に大きくスパースとなるため, これまでに, 相互情報量, 情報利得, χ^2 検定等を用いた素性選択 (次元削減) 手法が提案されている. しかし, これらの素性選択手法は, 一般に, 全品詞や名詞のみを対象に行われ, 品詞情報が及ぼす影響についてはあまり報告されていない.

本稿では, テキスト分類で現在最も優れた性能を発揮している Support Vector Machine (SVM) を例にとって, 品詞情報が素性選択手法に及ぼす影響について報告する. また, 単語とカテゴリとの間の相互情報量を基準とした素性選択手法を例にとって, 段階的に品詞を適用した場合の有効性について検証する.

本稿の構成は以下の通りである. 次節で SVM の概要及び SVM を用いた先行研究について述べる. 第三節で, 本稿が提案する可変的な段階的素性選択について説明する. 第四節で, 実験結果を示し, 本手法の有効性について検証する. 最後に第五節で本稿のまとめをする.

2 SVM を用いた先行研究

SVM は, 学習サンプルと分類境界の間隔を最大化するような戦略に基づいて 2 値分類を行う学習アルゴリズムである. SVM は, Kernel 関数を導入することにより非線形のモデル空間を仮定したり, 複数の素性の組合せを考慮して学習を行うことができる.

これまでに多くの研究者が SVM を用いて分類を行っている. しかし, ほとんどの研究者は, 全品詞や名詞のみを対象として実験を行っており, 品詞情報が及ぼす影響については報告していない.

平らは, RWCP コーパス (毎日新聞 (1994 年発行分)) を用いて, 単語とカテゴリとの間の相互情報量を基準とした素性選択手法及び品詞を基準とした素性選択手法の有効性について検証している. そして, 相互情報量を基準とした素性選択手法では, 最適な素性数はカテゴリ毎に大きく異なり先見的に決定するのは難しいと結論付けている. また, 平均での最高精度はすべての単語を使った時に得られたと報告している. 品詞を基準とした素性選択手法については, 最適な品詞はカテゴリによって異なり, 平均での最高精度はすべての品詞を使った時に得られたと報告している (Taira and Haruno, 1999).

本稿では, 平らと同様に相互情報量と品詞を基準とした素性選択手法を用いて実験を行った. しかし, 予め品詞毎に分けられたデータに対して, カテゴリ毎に高い相互情報量を持つ単語を選択し, 品詞による影響を調べている点が異なる.

3 可変的な段階的素性選択

これまでの素性選択手法とは異なり, 本稿では, 可変的な段階的素性選択 (Variable Cascaded Feature Selection, VCFS) を提案する. 本手法 (VCFS) の流れを図 1 に示す. VCFS とは, 1 度分類をして正例集合 (positive set) に属すると判定されたテストデータに対し, 素性 (品詞) を換えてさらに分類を行うという手法である. 第 1 段階 (step1) と第 2 段階 (step2) 共に, 学習時にカテゴリ毎に最適だと判定された品詞を用いて分類を行う. 例えば, テストデータを 'cocoa' に分類する場合, 第 1 段階では { 名詞, 動詞, 形容詞 } が品詞として選ばれ, 第 2 段階では { 全品詞 } が品詞として選ばれる. この場合, { 名詞, 動詞, 形容詞 } では, "seedpod", "seedcase", "pod", "inflorescence", "florescence", "blossoming", "ghana's", "ghanaian" 等が素性として選ばれ, { 全品詞 } では, "chocolate", "hot chocolate", "cocoa", "ghana", "umber", "deep brown", "burnt umber", "cocoas" 等が素性として選ばれる. また, 第 2 段階において最適な品詞

が見つからない場合は, 第1段階のみで分類を行う. 尚, 本稿で対象とした品詞は, { 名詞 }, { 名詞, 動詞 }, { 名詞, 形容詞 }, { 名詞, 副詞 }, { 名詞, 動詞, 形容詞 }, { 名詞, 動詞, 副詞 }, { 名詞, 形容詞, 副詞 }, { 名詞, 動詞, 形容詞, 副詞 }, 及び, { 全品詞 } である.

第2段階で正例集合に属すると判定されたテストデータを絞り込む理由としては, 近年の適合率へのニーズが挙げられる (Yang, 2001). 例えば, 検索エンジンを利用する場合に, 検索結果の最初の数ページにしか目を通す余裕がないことやテキストに付与されているカテゴリ名から関連するテキストを探す場合に, 特定のカテゴリ名しか参照しないことからもこのニーズの重要性がわかる. 本稿では, VCFS により, F_1 値を一定または向上させつつ, 適合率を向上させることでこのニーズに応えることを目的としている.

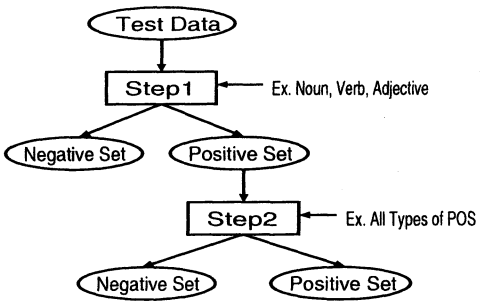

図 1: 本手法の流れ

最適な品詞選択

本稿では, 単語 F とカテゴリ C との間の相互情報量 (Mutual Information) を基準とした素性選択手法において, 品詞により上位にくる単語が異なることを利用し, カテゴリ毎に最適な品詞を選択する. 具体的には, 5回の交差検定 (Yang, 2001) により, ({ 第1段階 },{ 第2段階 }) > { 第1段階 } (F_1 値) となるような最適な品詞を学習により決定する. また, この条件を満たさない場合は, 第1段階において最適な品詞のみを F_1 値により決定する. 尚, 本稿では, カテゴリ毎に相互情報量の高い上位 {100, 200, 300, 400, 500, 600} 単語を選択して実験を行った.

本稿で用いた相互情報量 (MI) を式 (1) 示す (Taira and Haruno, 1999),(Dumais and Chen, 2000).

$$MI(F,C) = \sum_{F\in\{f,\bar{f}\}} \sum_{C\in\{c,\bar{c}\}} P(F,C) log \frac{P(F,C)}{P(F)P(C)} \quad (1)$$

ここで, 式 (1) の $P(F)$, $P(C)$, $P(F,C)$ はそれぞれ, 訓練データ中の単語 F を含む記事の割合, カテゴリ C に属する記事の割合, 単語 F を含みかつカテゴリ C に属する記事の割合, を示している.

相互情報量は, 単語 F の出現頻度がカテゴリ C とその他のカテゴリとの間で偏りがあるときに大きな値をとる. したがって, 相互情報量の高い単語は, カテゴリ C に分類するために必要な素性になると考えられる.

4 実験

4.1 実験設定

本稿では, 実験に Reuters-21578(Apte 分割) という英字新聞を用いた. そして, 7,769 記事を訓練データとし, 3,019 記事をテストデータとした. また, カテゴリには訓練データとテストデータの両方に付与されている 90 カテゴリを用いた. 尚, 最大 15 カテゴリが 1 記事に付与されていて, 1 記事あたりの平均カテゴリ数は 1.3 であった.

Reuters-21578 の特徴としては, 記事中の単語とカテゴリ名が密接に結び付いていることが挙げられる. 例えば, 'nickel' というカテゴリに記事を分類する場合, 記事中の "nickel" という単語はとても重要になる. 'grain' や 'crude' 等についても同様である.

もう 1 つの特徴としては, 各カテゴリに割り当てられている記事数が著しく不均一となっていることが挙げられる. 訓練データにおいて 100 記事以上割り当てられていたカテゴリは全体の 18%で, 10 記事以上 100 記事未満のカテゴリが 49%, 10 記事未満のカテゴリが 33%となっており, 全体の 82%のカテゴリが 100 記事未満であった (テストデータもほぼ同様の比率であった). 高村らは, SVM を用いたテキスト分類について, 訓練記事数が少ないと良い分類結果が得られにくいと報告している (Takamura and Matsumoto, 2001). また, Yang らは, 先行研究について, 訓練記事数が少ないカテゴリの分析がほとんどされていないと報告している (Yang and Liu, 1999).

本稿では, 訓練データとテストデータの両方について WordNet 1.7 により名詞の類義語を加えた. 尚, 形態素解析に Brill Tagger を使用したが, ストップワード処理や派生処理は行っていない. また, 分類結果の評価とし

ては, 標準的な方法になっている, マイクロ平均による適合率 (Pr), 再現率 (Re), F_1 値 ($2PrRe/(Pr+Re)$) を用いた.

4.2 本手法を適用した場合の実験結果

本手法を適用した場合の第 1 段階と第 2 段階での精度比較を表 1 に示す. 尚, 数値は, カテゴリ毎に MI 値の高い上位 400 語を選択した場合の精度を示している. 表 1 の第 1 列は何段階目で適用したかを示し, 第 2, 第 3, 第 4 列はそれぞれ適合率 (Pr), 再現率 (Re), F_1 値を示している. 適合率と F_1 値の向上が僅かであるが, 精度が向上した上位 7 カテゴリを示した表 2 を見るとその僅かの差が持つ意義がわかる. 尚, 表 2 の第 1 列はカテゴリ名とそのカテゴリに割り当てられている訓練記事数 (Tr) を示している.

表 2 より, 訓練記事数が少ないカテゴリにおいて適合率と F_1 値の向上が大きいことがわかる. ここで, 本手法の特徴をよく示しているのが 'sorghum' と 'cocoa' である. 'sorghum' では, 第 1 段階の適合率が再現率に比べて低いのに対し, 第 2 段階では, 適合率が再現率を上回るだけでなく, F_1 値も向上していることがわかる. また, 'cocoa' では, 第 2 段階において適合率が 50% も向上していることがわかる. これは, 第 1 段階では現れない "chocolate", "hot chocolate", "cocoa" 等といった単語が, 正例記事を絞り込むのに有効に働いたためと考える.

次に, 閾値毎の第 1 段階から第 2 段階において F_1 値が向上したカテゴリ数と閾値毎の各段階における F_1 値 (%) の違いをそれぞれ表 3 と表 4 に示す. 表 3 より, どの閾値においてもおよそ 17 カテゴリで F_1 値が向上していることがわかる. 尚, 他のカテゴリは, F_1 値が変わらなかった. また, 表 4 より, どの閾値においても第 2 段階の方が第 1 段階よりも上回っていることがわかる.

本稿では, 第 2 段階で全カテゴリに対して分類を行った場合との精度比較も行った. この手法を CFS と呼ぶ. CFS の第 1 段階と第 2 段階には, それぞれ { 名詞, 動詞, 形容詞 } と { 全品詞 } を選択した. 本手法と CFS の精度比較を示したのが表 5 である. 尚, 数値は, F_1 値 (%) を示している. 表 5 より, どの閾値についても本手法の方が上回っていることがわかる.

4.3 本手法を適用しない場合との精度比較

テキスト分類問題では, 全品詞や名詞のみを対象として素性選択を行うことが一般的であるため, 全品詞や名詞のみを用いた場合との精度比較を行った. 尚, 全品詞や名詞のみを用いた場合についても MI 値により素性選択を行っている.

表 1: 段階的素性選択を適用した場合の実験結果

段階	Pr(%)	Re(%)	F_1 値 (%)
第 1 段階	91.9	75.8	83.1
第 2 段階	92.5	75.8	83.3

表 2: 本手法により F_1 値が向上した上位 7 カテゴリ (第 2 段階/第 1 段階)

カテゴリ名 (Tr)	Pr(%)	Re(%)	F_1 値 (%)
sorghum(24)	71.4/55.6	62.5/62.5	66.7/58.8
orange(16)	100/85.7	54.5/54.5	70.6/66.7
gas(37)	100/90.9	62.5/62.5	76.9/74.1
bop(75)	85.7/78.3	60.0/60.0	70.6/67.9
copper(47)	100/93.8	83.3/83.3	90.9/88.2
ipi(41)	71.4/62.5	41.7/41.7	52.6/50.0
cocoa(55)	100/50.0	12.5/12.5	22.2/20.0

表 6, 表 7, 表 8 は, 全品詞及び名詞のみを用いた場合との精度比較を示している. 尚, 表 6 の数値は, F_1 値 (%) を示し, 表 7 と表 8 の数値は, カテゴリ毎に MI 値の高い上位 400 語を選択した場合の精度 (%) を示している.

表 6 より, どの閾値においても, 本手法の方が全品詞及び名詞のみを用いた場合より上回っていることがわかる. また, 表 7 と表 8 より, 特に訓練記事数が少ないカテゴリにおいて本手法の有効性が顕著であることがわかる.

全品詞と比較した場合に, F_1 値で上回ったのが 46 カテゴリ, 変わらなかったのが 36 カテゴリ, 下回ったのが 8 カテゴリであった. また, 名詞の場合は, 27 カテゴリで上回り, 50 カテゴリで一定, 13 カテゴリで下回った.

F_1 値で下回った原因としては, 第 1 段階において最適な品詞を選択できなかったことが挙げられる. しかし, その下回ったカテゴリについても, 第 2 段階により,

表 3: 各閾値における F_1 値が向上したカテゴリ数

閾値	100	200	300	400	500	600
カテゴリ数	19	20	20	12	16	16

表 4: 閾値毎の各段階における F_1 値の変化

閾値	100	200	300	400	500	600
第 1 段階	79.2	81.9	82.5	83.1	82.8	82.7
第 2 段階	80.3	82.5	82.8	83.3	83.1	83.1

表 5: 閾値毎の本手法 (VCFS) と CFS との F_1 値比較

閾値	100	200	300	400	500	600
VCFS	80.3	82.5	82.8	83.3	83.1	83.1
CFS	75.3	78.9	79.1	79.5	79.9	79.7

全品詞と名詞の場合でそれぞれ 4 カテゴリと 3 カテゴリを補えている.

表 6: 本手法と全品詞及び名詞のみを用いた場合との F_1 値比較

閾値	100	200	300	400	500	600
本手法	80.3	82.5	82.8	83.3	83.1	83.1
全品詞	75.5	78.3	79.1	79.4	79.5	79.6
名詞	78.7	81.1	81.5	82.0	82.0	81.9

表 7: F_1 値の向上が大きかった上位 7 カテゴリ (本手法/全品詞)

カテゴリ名 (Tr)	Pr(%)	Re(%)	F_1(%)
nickel(8)	100/0.0	100/0.0	100/0.0
lei(12)	100/0.0	66.7/0.0	80.0/0.0
rapeseed(18)	100/0.0	62.5/0.0	76.9/0.0
rice(35)	100/100	52.2/8.7	68.6/16.0
soy-meal(13)	100/0.0	33.3/0.0	50.0/0.0
tea(9)	100/0.0	33.3/0.0	50.0/0.0
meal-feed(30)	100/100	56.3/12.5	72.0/22.2

表 8: F_1 値の向上が大きかった上位 7 カテゴリ (本手法/名詞)

カテゴリ名 (Tr)	Pr(%)	Re(%)	F_1(%)
ipi(41)	71.4/16.7	41.7/10.0	52.6/12.5
rapeseed(18)	100/100	62.5/25.0	76.9/40.0
oilseed(124)	83.3/100	45.5/21.1	58.8/34.8
income(9)	50.0/0.0	14.3/0.0	22.2/0.0
lei(12)	100/66.7	66.7/66.7	80.0/66.7
hog(16)	100/100	50.0/33.3	66.7/50.0
meal-feed(30)	100/87.5	56.3/43.8	72.0/58.3

5 おわりに

本稿では, SVM を用いたテキスト分類において, 品詞情報が素性選択手法に及ぼす影響に注目し, 可変的な段階的素性選択を提案した. そして, 先行研究でよく対象となる, 全品詞や名詞のみを用いた場合よりも本手法が有効であることを示した. また, 訓練記事数が少ないカテゴリおいて精度向上が著しいことも示した.

本稿では, 全品詞や名詞のみを用いた場合との精度比較について報告したが, 他の品詞と比較した場合についても本手法の有効性が示せている. また, 相互情報量以外の素性選択手法 (χ^2 検定等) を用いた場合についても本手法の有効性が示せている.

今後は, Reuters(1996 年度版) のような超大規模コーパスや異なる種類のコーパスに本手法を適用することで, さらに定量的な評価を行う予定である.

参考文献

S. Dumais and H. Chen. 2000. Hierarchical classification of Web Content. *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 256–263.

H. Taira and M. Haruno. 1999. Feature Selection in SVM Text Categorization. *Proc. 16th National Conference on Artificial Intelligence*, 480–486.

H. Takamura and Y. Matsumoto. 2001. Feature Space Restructuring for SVMs with Application to Text Categorization. *Proc. 2001 Conference on Empirical Methods in Natural Language Processing*, 51–57.

Y. Yang. 2001. A Study on Thresholding Strategies for Text Categorization. *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 137–145.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 42–49.