6R-8

文脈木重み付け法を用いた文書分類の検討

小畑智広† 池上裕之† 小林 学‡ 坂下善彦‡

†湘南工科大学大学院 工学研究科 電気情報工学専攻
‡湘南工科大学 工学部 情報工学科

1. はじめに

与えられた学習データを用いて，新規データがどのカテゴリに所属するかを自動的に判別する自動分類問題は，ベクトル空間モデルやサポートベクターマシン[1]（以下 SVM と略す）などの手法により大きく発展した．

一方，J.Ziv らは無ひずみデータ圧縮法である LZ アルゴリズムを用いて文書分類を行う手法を提案した[2]．また LZ アルゴリズムよりも圧縮性能の優れた文脈木重み付け法[3]（以下 CTW 法と略す）を文書分類に適用する手法も提案されている[4]．ただしこれらの手法は，各文書の文書長がある程度長いことを想定しており，それぞれの文書が異なる確率モデルから生起することを仮定している．そこで本研究では，各カテゴリに対する文書が同一の確率モデルから生起するという仮定を置く．この仮定の下で，学習文書数と新規テスト文書数が同一の場合と異なる場合について検討を行う．結果的に新聞データを用いた計算機実験により，有効性を考察する．

2. 文脈木重み付け法を用いた文書分類

本節では CTW 法を用いて文書分類を行う手法について述べる．今ある入力系列$\boldsymbol{x}$に対する CTW 法の出力（重み付け確率）を$P_w^{\lambda}(x)$と表記する．また理想符号長$L(x)$を式(1)で定義する．

$$L(x) = -\log_2 P_w^{\lambda}(x). \quad (1)$$

次に文書分類において，カテゴリは 1～C まで存在するものとする．ここでカテゴリ$i \in \{1,2,\cdots,C\}$に所属するN_i個の学習文書をそれぞれ$\tilde{x}_1^{(i)}, \tilde{x}_2^{(i)}, \cdots, \tilde{x}_{N_i}^{(i)}$と書く．さらにカテゴリ i に所属する全ての学習文書を一つの学習データとして連結したものを$\tilde{x}^{(i)} = \tilde{x}_1^{(i)}\tilde{x}_2^{(i)}\cdots\tilde{x}_{N_i}^{(i)}$と表記する．本研究ではこの連結した学習データ$\tilde{x}^{(i)}$を用いて，CTW 法によりカテゴリ i の文脈木を構築する．これは各カテゴリに対して 1 回だけ行えばよく，結果の文脈木を保存しておくものとする．

また CTW 法を用いることにより全ての$i \in \{1,2,\cdots,C\}$に対して$P_w^{\lambda}(\tilde{x}^{(i)})$及び$L(\tilde{x}^{(i)}) = -\log_2 P_w^{\lambda}(\tilde{x}^{(i)})$が得られる．

さて，カテゴリが未知の新規文書を$\boldsymbol{x}$とし，$P_w^{\lambda}(\tilde{x}^{(i)})$の続きから新規文書$\boldsymbol{x}$を結合して，$P_w^{\lambda}(\tilde{x}^{(i)}\boldsymbol{x})$を新たに計算する．このとき$\tilde{x}^{(i)}\boldsymbol{x}$に対する理想符号長は$L(\tilde{x}^{(i)}\boldsymbol{x}) = -\log_2 P_w^{\lambda}(\tilde{x}^{(i)}\boldsymbol{x})$となる．$L(\tilde{x}^{(i)}\boldsymbol{x})$からカテゴリ i における新規文書$\boldsymbol{x}$に対する理想符号長$L_i(\boldsymbol{x})$を次式で定義する．

$$L_i(\boldsymbol{x}) = L(\tilde{x}^{(i)}\boldsymbol{x}) - L(\tilde{x}^{(i)}). \quad (2)$$

最終的に新規文書$\boldsymbol{x}$が所属するカテゴリの推定値は，次式により与えられる．

$$\hat{c} = \arg\min_i L_i(\boldsymbol{x}). \quad (3)$$

すなわち，理想符号長を最も小さくするカテゴリに分類を行うことになる．さて式(1), (2)より

$$\begin{aligned} \mathrm{L}_i(\boldsymbol{x}) &= \mathrm{L}(\tilde{x}^{(i)}\boldsymbol{x}) - \mathrm{L}(\tilde{x}^{(i)}), \\ &= -\log_2 \frac{P_w^{\lambda}(\tilde{x}^{(i)}x)}{P_w^{\lambda}(\tilde{x}^{(i)})} = -\log_2 P_w^{\lambda}(\boldsymbol{x}|\tilde{x}^{(i)}), \end{aligned} \quad (4)$$

が成り立つ．従って式(3)による分類法は，カテゴリ i に所属する文書は同一の確率モデルから生起することを仮定した下で，新規文書の条件付き事後確率を最大とするカテゴリに分類していることに相当する．

3. 新聞データを用いた計算機実験による評価

本節では，実際の新聞データを用いて 2 節の分類手法に対する文書分類を行った結果を示す．新聞データには，毎日新聞の 94 年の一年分の記事データを用いた[6]．これには，1 年分の記事データがカテゴリ別に収録されている．

実験では新聞のカテゴリとして「文化」，「経済」，「国際」，「芸能」，「スポーツ」，「社会」，「家庭」の 7 カテゴリを対象とし，各カテゴリに対して学習データと新規テストデータを用意した．まず，各カテゴリに対して，学習文書数を全て同一の$N_i = 200,500,1000$とした正分類率を表 1 に示す．ただし新規テスト文書数は各カテゴリに対して 200 とした．なお文書はランダムに選択し，10 回行った平均を示している．

次に，学習文書数及び新規テスト文書数がカテゴリごとに同一の場合と異なる場合で性能にどのような影響を与えるかを検討する．そこで学習文書数ベクトルを$\boldsymbol{N} = (N_1, N_2, \ldots, N_C)$とし，同様に新規テスト文書数ベクトルを$\boldsymbol{M} = (M_1, M_2, \ldots, M_C)$と定義する．まず$\boldsymbol{X}^{(1)} = (200,200,200,200,200,200,200)$とし，$\boldsymbol{X}^{(2)} = (40,200,60,265,310,445,80)$と定義する．なお$\boldsymbol{X}^{(2)}$

Examination of Text Classification Using Context-Tree Weighting Algorithm
Tomohiro OBATA†, Hiroyuki IKEGAMI†, Manabu KOBAYASHI† and Yoshihiko SAKASHITA†
†Graduate School of Engineering, Shonan Institute of Technology, ‡Shonan Institute of Technology

は各カテゴリの全記事数におおよそ比例するようにとっている．このとき様々な文書数の組み合わせ結果を表 2 に示す．また表 3 にはカテゴリごとの正分類率を示した．また表 4 に各カテゴリの新規テスト文書が社会へ分類される割合を示した．表 5 には，$\boldsymbol{N}=\boldsymbol{X}^{(1)}$におけるカテゴリごとの学習データの平均ファイル容量を示す．

比較のため，一般的に広く用いられている SVM の結果を表 6 に示す．SVM による文書分類は，まず各学習文書に対して形態素解析を行い，助詞，助動詞を除いた単語に区切る．次に各単語が出現していれば 1，していなければ 0 とした文書ベクトルを構成した．このときカーネルとしてベクトル同士の余弦を用いた．本実験では 7 カテゴリの分類を行うため，SVM による 2 値分類を多値分類に拡張する Pairwise 法を用いた．

表 1 の結果を見ると，学習文書数が 500 程度で約 84%に収束していることが分かる．次に表 2 の結果をみると$(\boldsymbol{N},\boldsymbol{M})=(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(1)})$ 以外はある程度高い正分類率を示した．一方表 3 をみると，$(\boldsymbol{N},\boldsymbol{M})=(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(2)})$の結果はカテゴリごとにばらつきが大変大きいことが分かる．具体的には学習文書数の少ないカテゴリである「文化」，「国際」，「家庭」の正分類率が極端に低く，学習文書数の最も多い「社会」の正分類率が高い．そこで表 4 の「社会」へ分類率をみると，$(\boldsymbol{N},\boldsymbol{M})=(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(1)})$及び$(\boldsymbol{N},\boldsymbol{M})=(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(2)})$では，学習文書数の少ないカテゴリの新規テスト文書の 5 割以上が「社会」へ誤分類されている．従って表 1~4 の結果からは学習文書数は全体の記事数に応じて変化させるよりも，すべてのカテゴリで同程度の文書数にすべきであることが分かる．次に表 6 の結果をみると CTW 法を用いた分類法よりも SVM の正分類率が約 4%程度高かった． これは$\boldsymbol{N}=\boldsymbol{X}^{(1)}$のときに SVM ではカテゴリ「社会」の正分類率が CTW 法よりも十分に大きいためである．この原因を考えると，CTW 法を用いた分類において表 5 から「社会」の学習文書のファイル容量が他に比べ少ないことが影響を与えているものと思われる．以上から，CTW 法を用いた文書分類では，学習文書のファイル容量を各カテゴリで同程度にするのが良いものと思われる．

4. まとめ

本研究では，CTW 法を用いた文書分類手法について検討を行った．結果的に学習文書数及び学習文書のファイル容量を同程度にすると，正分類率がカテゴリ間で偏ることなくうまく分類できることが分かった．

表 1：様々な学習文書数に対する正分類率 (%)

学習文書数N_i	200	500	1000
正分類率	79.5	84.2	85.7

表 2：$\boldsymbol{N}$及び$\boldsymbol{M}$の組合せに対する正分類率(%)

$(\boldsymbol{N},\boldsymbol{M})$	$(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(1)})$	$(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)})$	$(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(2)})$
正分類率	57.2	77.7	76.4

表 3：カテゴリごとの正分類率(%)

$(\boldsymbol{N},\boldsymbol{M})$	$(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(1)})$	$(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(1)})$	$(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)})$	$(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(2)})$
文化	74.3	11.6	81.0	11.8
経済	85.6	71.6	82.4	71.4
国際	89.6	39.1	88.2	33.2
芸能	70.2	76.6	77.7	76.2
スポーツ	87.2	88.0	86.7	87.5
社会	62.7	93.7	65.9	92.4
家庭	87.3	19.9	87.3	22.0

表 4：カテゴリ「社会」への分類率(%)

$(\boldsymbol{N},\boldsymbol{M})$	$(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(1)})$	$(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(1)})$	$(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)})$	$(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(2)})$
文化	0.4	77.3	0.5	75.0
経済	1.4	21.3	2.7	19.9
国際	0.3	50.4	0.2	51.3
芸能	2.6	12.8	3.7	14.0
スポーツ	3.4	10.3	3.7	11.1
社会	62.7	93.7	65.9	92.4
家庭	0.5	77.2	0.3	74.8

表 5：各カテゴリの学習データの平均容量(KB)

$(\boldsymbol{N},\boldsymbol{M})$	文化	経済	国際	芸能	スポーツ	社会	家庭
$(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(1)})$	319	188	264	199	192	152	276

表 6: SVM の正分類率(%)

$(\boldsymbol{N},\boldsymbol{M})$	$(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(1)})$	$(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(1)})$	$(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)})$	$(\boldsymbol{X}^{(2)},\boldsymbol{X}^{(2)})$
正分類率	83.1	68.4	82.2	80.7

参考文献

[1] C.M.ビショップ，パターン認識と機械学習，Springer，2008.

[2] J.Ziv and N.Merhav, ”A Measure of Relative Entropy Between Individual Sequences with Application to Universal Classification,” IEEE Trans. On Information Theory, vor.39, no.4, pp.1270-1279, July 1993.

[3] F.M.J.Willems,Y.M.Shtarkov and T.J.Tjalkens, ”The Conectext Tree Weighting Method: Basic Properties,” IEEE Trans. On Information Theory, vol.41, no.3, pp.653-664, May 1995.

[4] Z.Dawy, J.Hagenauer and A.Hoffmann, ”Implementing the context tree weighting method for context recognition,” Proc. Of the IEEE Data Compression Conference, p.536, March 2004.

[5] H.Cai, S.R.Kulkarni and S.Verdu, “Universal divergence estimation for finite-alphabet sources,” IEEE Trans. On Information Theory, vol.52 no.8, pp.3456-3475, Aug. 2006.

[6] CD-毎日新聞 94’データ集，日外アソシエーツ，1995.

[7] H. Chen, P. Ti and X. Yao, “Probabilistic Classification Vector Machines” IEEE Trans. On Neural Networks, vol. 20, no. 6,pp.902-914, JUNE 2009.