6R-1

確率分類ベクターマシンを用いた文書分類方式の検討

池上　裕之†,　阿部　洋志†,　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

1. はじめに

文書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など様々な手法により研究されてきた．近年 Chen らにより確率分類ベクターマシンが提案されており，基本的な分類問題に対して従来の種々の手法よりも優れた結果が示されている[3]．そこで本研究では確率分類ベクターマシンを拡張し，文書分類問題に適用することを考える．このときカーネルにベクトル空間モデルの類似度を用いてその有効性の評価を行う．

2. 確率分類ベクターマシン

本稿では 2 値の文書分類問題を対象として，Chen らによる確率分類ベクターマシン（以下 PCVM と略す）に新しいパラメータ γ を導入した形で記述する．まず i 番目の入力ベクトル $\boldsymbol{x_i}$ とそれに対応するラベル $t_i \in \{-1,+1\}$ のN個の組 $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^{N}$を学習データとする．また $\boldsymbol{x_i}$ に依存するカーネル関数を$\boldsymbol{\phi}_i(\boldsymbol{x})$と表し，

$$\boldsymbol{\phi}(\boldsymbol{x}) = (\boldsymbol{\phi}_1(\boldsymbol{x}), \boldsymbol{\phi}_2(\boldsymbol{x}), \cdots, \boldsymbol{\phi}_N(\boldsymbol{x}))^T$$

と定義する．このときパラメータ

$$\boldsymbol{w} = (w_1, w_2, \cdots, w_N)^T$$

と b を用いて線形識別関数 $y(\boldsymbol{x})$ を

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b \tag{1}$$

と定義する．ここで観測できない隠れ関数 $h(\boldsymbol{x})$ を考え，

$$h(\boldsymbol{x}) = y(\boldsymbol{x}) + \varepsilon \tag{2}$$

とする．ただし観測できない雑音 ε は平均 0，精度 γ の正規分布$N\left(0, \frac{1}{\gamma}\right)$に従うものとする．このときラベル t は次式で決定されるモデルを考える．

$$t = \begin{cases} +1, & h(\boldsymbol{x}) \geq 0 \\ -1, & h(\boldsymbol{x}) < 0 \end{cases} \tag{3}$$

ここでプロビット関数 $\psi(x)$ を

$$\psi(x) = \int_{-\infty}^{x} N\left(t \middle| 0, \frac{1}{\gamma}\right) dt \tag{4}$$

と定義すると，

$$P(t = 1|\boldsymbol{x}, \boldsymbol{w}, b) = P(y(\boldsymbol{x}) + \varepsilon \geq 0) = \psi\big(y(\boldsymbol{x})\big) \tag{5}$$

$$P(h(\boldsymbol{x})|\boldsymbol{w}, b) = N\left(h(\boldsymbol{x}) \middle| y(\boldsymbol{x}), \frac{1}{\gamma}\right) \tag{6}$$

が成り立つ．学習データに対して

$$\boldsymbol{\Phi} = \big(\boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_N)\big)^T \tag{7}$$

$$H = \big(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \cdots, h(\boldsymbol{x}_N)\big)^T \tag{8}$$

と定義すると，各学習データは独立なため

$$P(H|\boldsymbol{w}, b) = \prod_{i=1}^{N} N\{h(\boldsymbol{x}_i)|y(\boldsymbol{x}_i), \frac{1}{\gamma}\}$$

$$= \left(\sqrt{\frac{\gamma}{2\pi}}\right)^N \exp\left\{-\frac{\gamma}{2}\|H - \boldsymbol{\Phi}\boldsymbol{w} - bI\|^2\right\} \tag{9}$$

が成り立つ．ただし I は要素が全て 1 のベクトルを表す．ここで $\boldsymbol{w}$ および b に対する事前分布を

$$P(w_i|\alpha_i) = \begin{cases} 2N(w_i|0, {\alpha_i}^{-1}), & y_i x_i \geq 0 \\ 0 \quad , & y_i x_i < 0 \end{cases} \tag{10}$$

$$P(b|\beta) = N(b|0, \beta^{-1}) \tag{11}$$

と仮定する．このとき$A = \mathrm{diag}(\alpha_1, \alpha_2, \cdots, \alpha_N)$とすると，

$$\log P(\boldsymbol{w}, b|H, A, \beta)$$

$$\propto \log P(H|\boldsymbol{w}, b) + \log(\boldsymbol{w}|A) + \log(b|\beta) + \text{定数}$$

$$\propto -\frac{\gamma}{2}\|H - (\boldsymbol{\Phi}\boldsymbol{w} + bI)\|^2 - \frac{1}{2}\boldsymbol{w}^T A\boldsymbol{w}$$

$$-\frac{1}{2}\beta b^2 + \text{定数}$$

$$= \gamma\boldsymbol{w}^T\boldsymbol{\Phi}^T(2H - \boldsymbol{\Phi}\boldsymbol{w}) - 2\gamma bI^T\boldsymbol{\Phi}\boldsymbol{w} + 2\gamma bI^T H$$

$$-\boldsymbol{w}A\boldsymbol{w} - \beta b^2 - \gamma b^2 N + \text{定数} \tag{12}$$

が成り立つ．上式を最大とする $\boldsymbol{w}$ および b を求めるために，EM アルゴリズムを用いる．

①E ステップ

$$Q(\boldsymbol{w}, b|\boldsymbol{w}^{old}, b^{old})$$

$$= E_{H,\alpha,\beta}[\log P(\boldsymbol{w}, b|t, H, \alpha, \beta)|t, \boldsymbol{w}^{old}, b^{old}]$$

$$= 2\gamma(\boldsymbol{w}^T\boldsymbol{\Phi}^T + bI^T)E[H|t, \boldsymbol{w}^{old}, b^{old}]$$

$$-\boldsymbol{w}^T E[A|t, \boldsymbol{w}^{old}, b^{old}]\boldsymbol{w} - b^2 E[\beta|t, \boldsymbol{w}^{old}, b^{old}]$$

$$-\gamma\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} - 2\gamma bI^T\boldsymbol{\Phi}\boldsymbol{w} - \gamma b^2 N$$

$$= 2\gamma(\boldsymbol{w}^T\boldsymbol{\Phi}^T + bI^T)\bar{H} - \boldsymbol{w}^T\bar{A}\boldsymbol{w} - b^2\bar{\beta}$$

$$-\gamma\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} - 2\gamma bI^T\boldsymbol{\Phi}\boldsymbol{w} - \gamma b^2 N \tag{13}$$

ただし$\bar{H} = (\bar{h}_1, \bar{h}_2, \cdots, \bar{h}_N)^T$は $t_i = 1$ のとき

$\bar{h}_i = E[h_i|t_i, \boldsymbol{w}^{old}, b^{old}]$

Probabilistic Classification Vector Machines for Document Classification
Hiroyuki IKEGAMI†, Hiroshi ABE†, Manabu KOBAYASHI† and Yoshihiko SAKASHITA†
†Graduate School of Engineering, Shonan Institute of Technology

$$= \int h_i \cdot P(h_i|t_i, \boldsymbol{w}^{old}, b^{old}) dh_i$$
$$= \int_{y(\boldsymbol{x}_i)+\varepsilon \geq 0} (y(\boldsymbol{x}_i)+\varepsilon) N\left(\varepsilon|0, \frac{1}{\gamma}\right) d\varepsilon$$
$$= y(\boldsymbol{x}_i)\psi\big(y(\boldsymbol{x}_i)\big) + \frac{1}{\gamma} N\left(y|0, \frac{1}{\gamma}\right) \quad (14)$$

となる．また $t_i = -1$ のときは

$$\bar{h}_i = \int_{y(\boldsymbol{x}_i)+\varepsilon \leq 0} (y(\boldsymbol{x}_i)+\varepsilon) N\left(\varepsilon|0, \frac{1}{\gamma}\right) d\varepsilon$$
$$= y(\boldsymbol{x}_i)\psi\big(-y(\boldsymbol{x}_i)\big) - \frac{1}{\gamma} N\left(y|0, \frac{1}{\gamma}\right) \quad (15)$$

である．

②M ステップ

$$\frac{dQ}{d\boldsymbol{w}} = 2\gamma \boldsymbol{\Phi}^T \bar{H} - 2\bar{A}\boldsymbol{w} - 2\gamma \boldsymbol{\Phi}\boldsymbol{\Phi}^T \boldsymbol{w} - 2\gamma b \boldsymbol{\Phi}^T I = 0 \quad (16)$$

として

$$\boldsymbol{w}^{\text{new}} = (2\bar{A} + 2\gamma \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}})^{-1} \cdot \gamma \boldsymbol{\Phi}^T (\bar{H} - bI) \quad (17)$$

により $\boldsymbol{w}$ を更新する．同様に

$$\frac{dQ}{db} = 2\gamma I^T \bar{H} - 2b\bar{\beta} - 2\gamma I^T \boldsymbol{\Phi}\boldsymbol{w} - 2b\gamma N = 0 \quad (18)$$

として

$$b^{\text{new}} = \left(\bar{\beta} + \gamma N\right)^{-1} \gamma I^T (\bar{H} - \boldsymbol{\Phi}\boldsymbol{w}) \quad (19)$$

により b を更新する．

なお本節で $\gamma = 1$ と置くと，Chen らの PCVM と等価となる．

3. 計算機による実験

本節では文書分類問題に PCVM を実際に用いて評価を行う．ここで i 番目の学習文書 $\boldsymbol{x}_i$ を $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,M}) \in \{0,1\}^M$ と表現する．ただし $x_{i,j}$ は i番目の学習文書に j番目の単語が存在すれば 1，そうでなければ 0 とする．新規文書$\boldsymbol{x}$についても同様に$\boldsymbol{x} = (x_1, x_2, \cdots, x_M) \in \{0,1\}^M$と表記する．なお M は総単語数である．このとき本節で実験に用いるカーネル関数を

$$\phi_i(\boldsymbol{x}) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}\|} \quad (20)$$

と定義する．すなわち$\boldsymbol{\phi}_i(x)$は$\boldsymbol{x}_i$と$\boldsymbol{x}$の余弦とする．実験用のデータには CD-毎日新聞 94’データ集[4]を利用した．これには毎日新聞の 1 年分の記事データにラベルを付与したものが収録されている．ここで 2 カテゴリのデータを選択して 2 値分類を行った結果を表 1,2 に示す．表 1 は γ の値を変化させたときの 2 値分類の正分類率(%)を示している．ただし学習データ数を N = 400 とし，評価のためのテストデータ数は 1600 とした．なお PCVM における EM アルゴリズムの繰り返し最大回数は 50 とした．表中カテゴリ Int., So., Sp., Cul., Ec., Ent はそれぞれ国際，社会，スポーツ，文化，経済，芸能を意味している．また非零の w_i の個数を調べるために，表 2 に最大繰り返し回数後に $|w_i| > 10^{-8}$である w_iの個数を示した．

表 1：各 γ に対する 2 値分類の正分類率(%).

γ	0.05	0.1	0.5	1.0	2.0	4.0
Int./So.	84.3	86.5	86.2	86.1	85.8	85.6
So./Sp.	86.1	88.4	89.5	88.4	86.9	85.6
Cul./Ec.	94.9	95.8	96.9	96.9	97.1	97.4
Ec./Ent	95.3	97.9	97.3	96.1	95.0	94.1

表 2：各 γ に対する $|w_i| > 10^{-8}$である w_iの個数.

γ	0.05	0.1	0.5	1.0	2.0	4.0
Int./So.	6	10	33	42	62	88
So./Sp.	6	9	36	61	82	109
Cul./Ec.	7	10	24	36	52	82
Ec./Ent	8	10	30	40	59	81

表 1 を見ると，$\gamma = 0.1$ あるいは 0.5 の正分類率が高いことが分かる．ただしカテゴリによって若干の差が見られ，文化と経済の分類(Cul. / Ec.）では γ が大きいほど分類精度が良い結果となっている．このように与えられた問題によって最適な精度を与える γ に差があることが分かる．一方非零に収束する w_i の個数は γ が小さくなるにつれて単調に小さくなる．新規文書の分類は式(1)の$y(\boldsymbol{x})$の正負により判別するため，分類の計算量は非零の w_i の個数に比例する．従って計算量の点から見ると γ は小さい方が良いことが分かる．本実験による文書分類では γ の値は 0.1～0.5 程度がおおむね良い結果であることが分かる．

4. まとめと今後の課題

本研究では新しいパラメータ γを導入した PCVM を用いて 2 値の文書分類の評価を行った．結果的に文書分類では γ の値は 0.1～0.5 程度がおおむね良い結果であった．多カテゴリに対する PCVM の適用と γ の自動決定法などが今後の課題である．

参考文献

[1] V. N. Vapnik, Statistical Learning Theory. New York: Wiley-Interscience,1998.

[2] M. E. Tipping, “Sparse Bayesian learning and the relevance vector machine,” J. Mach. Learn. Res., vol. 1, pp. 211–244, 2001.

[3] 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

[4] CD-毎日新聞 94’データ集，日外アソシエーツ，1995.