6R-7

雑音の分散を考慮した確率分類ベクターマシンによる自動分類法

池上裕之†，小畑智広†，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

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など様々な手法により研究されてきた[1][2]．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][4]．本研究では確率分類ベクターマシンにおける雑音の分散をパラメータ化して拡張する．また分散のパラメータに事前分布を導入し， EM アルゴリズムを用いて最適化する手法を示す．さらに提案手法を文書分類問題に適用し，分類精度及び分類速度について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 及び λ を導入した形で記述する．まず i 番目の入力ベクトル $\boldsymbol{x_i}$ とそれに対応するカテゴリのラベル $t_i \in \{-1,+1\}$ 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，

$$\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}),\phi_2(\boldsymbol{x}),\cdots,\phi_N(\boldsymbol{x}))^T$$

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 $h(\boldsymbol{x}_i)$ を考え，

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

とする．ただし観測できない雑音 ε_i 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_i は次式で決定されていると仮定するモデルを考える．

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

このとき，

$$N(z|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \tag{4}$$

とし，プロビット関数 $\psi(x)$ を

$$\psi(x) = \int_{-\infty}^{x} N\left(t\middle|0,\frac{1}{\gamma}\right)dt \tag{5}$$

と定義すると，

$$P(t_i = 1|\boldsymbol{x}_i,\boldsymbol{w},b,\gamma) = P(y(\boldsymbol{x}_i)+\varepsilon \geq 0) = \psi\big(y(\boldsymbol{x}_i)\big) \tag{6}$$

$$P(h(\boldsymbol{x}_i)|\boldsymbol{w},b,\gamma) = N\left(h(\boldsymbol{x}_i)\middle|y(\boldsymbol{x}_i),\frac{1}{\gamma}\right) \tag{7}$$

が成り立つ．学習データに対して

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

$$\boldsymbol{H} = \big(h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),\cdots,h(\boldsymbol{x}_N)\big)^T \tag{9}$$

と定義すると，各学習データは独立なため

$$P(\boldsymbol{H}|\boldsymbol{w},b,\gamma) = \prod_{i=1}^{N} N\left(h(\boldsymbol{x}_i)|y(\boldsymbol{x}_i),\frac{1}{\gamma}\right) = \left(\sqrt{\frac{\gamma}{2\pi}}\right)^N \exp\left\{-\frac{\gamma}{2}\|\boldsymbol{H}-(\boldsymbol{\Phi}\boldsymbol{w}+bI)\|^2\right\} \tag{10}$$

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 $\boldsymbol{w}$，b 及び γ に対する事前分布を

$$P(w_i|\alpha_i) = \begin{cases} 2N(w_i|0,{\alpha_i}^{-1}), & t_i w_i \geq 0 \\ 0, & t_i w_i < 0 \end{cases} \tag{11}$$

$$P(b|\beta) = N(b|0,\beta^{-1}) \tag{12}$$

$$P(\gamma|\lambda) = \begin{cases} 2N(\gamma|0,\lambda^{-1}), & \gamma > 0 \\ 0, & \gamma \leq 0 \end{cases} \tag{13}$$

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

$$\begin{aligned}&\log P(\boldsymbol{w},b,\gamma|\boldsymbol{H},A,\beta,\lambda)\\ &= \log P(\boldsymbol{H}|\boldsymbol{w},b,\gamma) + \log P(\boldsymbol{w}|A) + \log P(b|\beta) + \log P(\gamma|\lambda) + \text{定数}\\ &\propto \gamma\boldsymbol{w}^T\boldsymbol{\Phi}^T(2\boldsymbol{H}-\boldsymbol{\Phi}\boldsymbol{w}) - 2\gamma bI^T\boldsymbol{\Phi}\boldsymbol{w} + 2\gamma bI^T\boldsymbol{H} - \boldsymbol{w}A\boldsymbol{w} - \beta b^2 - \gamma b^2 N - \lambda\gamma^2 + N\log\gamma + \text{定数}\end{aligned} \tag{14}$$

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

①E ステップ

$$\begin{aligned}&Q(\boldsymbol{w},b,\gamma|\boldsymbol{w}^{old},b^{old},\gamma^{old})\\ &= E_{\boldsymbol{H},A,\beta}[2\log P(\boldsymbol{w},b,\gamma|\boldsymbol{H},A,\beta,\lambda)|t,\boldsymbol{w}^{old},b^{old},\gamma^{old}]\\ &= 2\gamma(\boldsymbol{w}^T\boldsymbol{\Phi}^T + bI^T)\bar{\boldsymbol{H}} - \boldsymbol{w}^T\bar{A}\boldsymbol{w} - b^2\bar{\beta} - \gamma\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} - 2\gamma bI^T\boldsymbol{\Phi}\boldsymbol{w} - \lambda\gamma^2 - \gamma b^2 N + N\log\gamma + \text{定数}\end{aligned} \tag{15}$$

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

$$\bar{h}_i = E[h_i|t_i,\boldsymbol{w}^{old},b^{old},\gamma^{old}] = y(\boldsymbol{x}_i)\psi\big(y(\boldsymbol{x}_i)\big) + \frac{1}{\gamma^{old}}N\left(y(\boldsymbol{x}_i)|0,\frac{1}{\gamma^{old}}\right) \tag{16}$$

となる．また $t_i = -1$ のときは次式となる．

$$\bar{h}_i = y(\boldsymbol{x}_i)\psi\big(-y(\boldsymbol{x}_i)\big) - \frac{1}{\gamma^{old}}N\left(y(\boldsymbol{x}_i)|0,\frac{1}{\gamma^{old}}\right) \tag{17}$$

Probabilistic Classification Vector Machines considered variance of noise
Hiroyuki IKEGAMI†, Tomohiro OBATA†, Manabu KOBAYASHI‡ and Yoshihiko SAKASHITA‡
†Graduate School of Engineering, Shonan Institute of Technology

②M ステップ

$\frac{dQ}{dw}=0$として

$$\boldsymbol{w}^{\text{new}}=(\bar{A}+\gamma\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\gamma\boldsymbol{\Phi}^T(\bar{\boldsymbol{H}}-bI^T) \quad (18)$$

により $\boldsymbol{w}$ を更新する．また$\frac{dQ}{db}=0$として

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

により b を更新する．同様に$\frac{dQ}{d\gamma}=0$として

$$\gamma^{\text{new}}=\frac{C+\sqrt{C^2+8\lambda N}}{4\lambda} \quad (20)$$

により γ を更新する．ただし$C=2(\boldsymbol{w}^T\boldsymbol{\Phi}^T+bI^T)\bar{\boldsymbol{H}}-\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi w}-2bI^T\boldsymbol{\Phi w}-b^2N$である．

なお本節ではγを確率変数として扱ったが，$\gamma=1$ と固定値とした場合は Chen らの PCVM と等価となる．

3. 計算機による実験

本節では前節の拡張 PCVM を実際の文書分類問題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．ここで本節で実験に用いるカーネル関数を

$$\phi_i(\boldsymbol{x})=\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}\|} \quad (21)$$

と定義する．すなわち$\phi_i(\boldsymbol{x})$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[5]を利用した．これには毎日新聞の 1 年分の記事データにカテゴリの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 1,2,3 に示す．表 1 は λ の値を変化させたときの正分類率(%)を示している．比較のため同じデータについて$\gamma=1$とした従来の PCVM と，SVM を用いた場合の正分類率を示した．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 800 とした．表中カテゴリ Cu., Ec., Sp., In., En.はそれぞれ文化，経済，スポーツ，国際，芸能を意味している．表 2 は各 λ に対して γ の収束した値を示している．また非零の w_i の個数を調べるために，表 3 に$|w_i|>10^{-8}$である w_iの個数を示した．

表 1 を見ると，$\lambda=40000$あるいは 20000 の正分類率が比較的高いが，λ の値を変化させても正分類率は 2%程度以内と大きな変動はない．一方表 2 を見ると，λ が大きくなると γ は小さな値に収束する傾向がある．それに伴い，非零に収束する w_i の個数は，表 3 の様に λ が大き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λ は大きい方が良いことが分かる．本実験による文書分類では λ の値は 20000～40000 程度がおおむね良い結果であることが分かる．また SVM と比較すると拡張 PCVM の方が正分類率が 2～3%ほど低いが，サポートベクトルの数は 1/10～1/20 程度と大幅に小さく済む．よって PCVM は SVM よりも新規文書の分類に必要な計算量を大幅に抑えることが可能であることが分かる．

表 1： 2 値分類の正分類率(%).

λ	10000	20000	40000	80000	$\gamma=1$	SVM
Cu./Ec.	95.9	96.4	96.6	96.1	95.0	98.6
Cu./Sp.	93.9	94.1	94.6	94.5	92.9	97.7
Ec./In.	89.0	89.1	87.6	87.4	87.6	91.2
En./In.	96.5	96.6	97.1	95.3	95.3	98.3

表 2：各 λ に対する γ の値.

λ	10000	20000	40000	80000
Cu./Ec.	0.95	0.45	0.23	0.14
Cu./Sp.	0.83	0.38	0.20	0.12
Ec./In.	0.39	0.23	0.15	0.10
En./In.	0.93	0.42	0.21	0.12

表 3： $|w_i|>10^{-8}$である w_iの個数.

λ	10000	20000	40000	80000	$\gamma=1$	SVM
Cu./Ec.	38	25	19	14	37	278
Cu./Sp.	45	27	18	15	47	292
Ec./In.	22	15	12	9	40	294
En./In.	45	29	20	12	46	278

4. まとめと今後の課題

本研究では新しいパラメータ γとλ を導入した PCVM を用いて，2 値の文書分類の評価を行った．今回の実験方法による文書分類では，λ の値は 20000～40000 程度がおおむね良い結果であった．うまく適合する λ についての確率モデルの提案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] 池上裕之，阿部洋志，小林　学，坂下善彦，“確率分類ベクターマシンを用いた文書分類方式の検討”，第 73 回情報処理学会全国大会予稿論文集，pp-2-313，March 2011.

[5] CD-毎日新聞 94’データ集，日外アソシエーツ，1995.