

最小分類誤り学習における初期値依存性

1 J-5

佐藤 敦

NEC C&C メディア研究所

e-mail : asato@ccm.CL.nec.co.jp

1 はじめに

最小分類誤り (MCE) とは、ベイズ決定理論における全危険を直接最小化する学習の枠組みである。一般化確率的降下法 (GPD)[1] は MCE に基づく学習の定式化で、学習ベクトル量子化 (LVQ)[2] と同等の学習式が得られるが、参照ベクトルが発散する問題があった。筆者は MCE に基づく LVQ の新しい一般化 (GLVQ) を提案し [3]、参照ベクトルの収束点が存在することを示した [4]。本稿では、参照ベクトルの初期値依存性について理論的考察を行い、シミュレーション実験で GPD と GLVQ の収束性を比較する。

2 最小分類誤り学習

N 個の入力ベクトル \mathbf{x}_i に対する経験損失 R_e は、

$$R_e = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \ell(\omega(\mathbf{x}_i)|\omega_k) 1(\mathbf{x}_i \in \omega_k) \quad (1)$$

と書ける。 $1(\cdot)$ は $1(true) = 1, 1(false) = 0$ なる関数、 $\ell(\omega(\mathbf{x})|\omega_k)$ は識別器の判定 $\omega(\mathbf{x})$ に対する損失で、 K はクラス数を表す。誤分類尺度 $\rho_k(\mathbf{x}; \theta)$ を用いて $\ell(\omega(\mathbf{x})|\omega_k) = \ell(\rho_k(\mathbf{x}; \theta))$ とすれば、識別器のパラメータを $\theta \leftarrow \theta - \epsilon \nabla R_e$ と修正して経験損失を減らすことができる ($\epsilon > 0$)。GPD では、誤分類尺度として

$$\rho_k(\mathbf{x}; \theta) = -g_k(\mathbf{x}; \theta) + g_\ell(\mathbf{x}; \theta) \quad (2)$$

が用いられる。 $g_k(\mathbf{x}; \theta)$ はクラス ω_k の識別関数で、 $g_\ell(\mathbf{x}; \theta) = \max_{i \neq k} g_i(\mathbf{x}; \theta)$ である。GPD を最近傍法に適用すると、クラス ω_k に属する \mathbf{x} に対して、次に示す LVQ2.1 と同等の学習式が得られる。

$$\mathbf{m}_r \leftarrow \mathbf{m}_r + 2\epsilon \ell'(\rho_k)(\mathbf{x} - \mathbf{m}_r) \quad (3)$$

$$\mathbf{m}_s \leftarrow \mathbf{m}_s - 2\epsilon \ell'(\rho_k)(\mathbf{x} - \mathbf{m}_s) \quad (4)$$

ここで、 \mathbf{m}_r はクラス ω_k に属し、かつ \mathbf{x} に最も近い参照ベクトルであり、 \mathbf{m}_s はクラス ω_k 以外に属し、かつ \mathbf{x} に最も近い参照ベクトルである。 ℓ' は損失関数の微

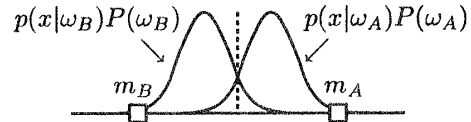


図 1: 2 クラス分類問題の例

分を表し、損失関数としては $\ell(\rho) = 1/\{1 + \exp(-\xi\rho)\}$ なるシグモイド関数がよく用いられる ($\xi > 0$)。一方 GLVQ では、式 (2) の代わりに

$$\rho_k(\mathbf{x}; \theta) = \frac{-g_k(\mathbf{x}; \theta) + g_\ell(\mathbf{x}; \theta)}{|g_k(\mathbf{x}; \theta)| + |g_\ell(\mathbf{x}; \theta)|} \quad (5)$$

を誤分類尺度として採用している。これを最近傍法に適用すると、次の学習式が得られる。

$$\mathbf{m}_r \leftarrow \mathbf{m}_r + 4\epsilon \ell'(\rho_k) \frac{d_s}{(d_r + d_s)^2} (\mathbf{x} - \mathbf{m}_r) \quad (6)$$

$$\mathbf{m}_s \leftarrow \mathbf{m}_s - 4\epsilon \ell'(\rho_k) \frac{d_r}{(d_r + d_s)^2} (\mathbf{x} - \mathbf{m}_s) \quad (7)$$

ただし、 $d_r = \|\mathbf{x} - \mathbf{m}_r\|^2, d_s = \|\mathbf{x} - \mathbf{m}_s\|^2$ である。

3 初期値依存性

図 1 に示す ω_A, ω_B の 2 クラス分類問題を考える。 $P(\omega_A), P(\omega_B)$ は事前確率、 $p(x|\omega_A), p(x|\omega_B)$ は確率密度関数で、 $Q(x) = p(x|\omega_A)P(\omega_A) - p(x|\omega_B)P(\omega_B)$ とすると $x > 0$ で $Q(x) > 0, x < 0$ で $Q(x) < 0$ である。以後、簡単のために $\ell'(\rho) = 1$ として議論する。

3.1 GPD

GPD による参照ベクトル \mathbf{m}_A の変位の期待値は、

$$E\{\Delta \mathbf{m}_A\} = 2\epsilon \int (\mathbf{x} - \mathbf{m}_A) Q(x) dx \quad (8)$$

となる。 $E\{\Delta \mathbf{m}_A\} = 0$ で定義される収束点は $\mathbf{m}_A = \int \mathbf{x} Q(x) dx / \{P(\omega_A) - P(\omega_B)\}$ として求まり、2つの事前確率が等しい場合は無限遠が収束点となる。つまり、初期値に依らず参照ベクトルは常に発散する。

3.2 GLVQ

GLVQ による参照ベクトル \mathbf{m}_A の変位の期待値は、

$$E\{\Delta \mathbf{m}_A\} = 4\epsilon \int \frac{(\mathbf{x} - \mathbf{m}_B)^2 (\mathbf{x} - \mathbf{m}_A) Q(x)}{\{(\mathbf{x} - \mathbf{m}_A)^2 + (\mathbf{x} - \mathbf{m}_B)^2\}^2} dx \quad (9)$$

となる。確率密度関数が原点について対称な場合、つまり $Q(-x) = -Q(x)$ とすると、 $m_B = -m_A$ と仮定できる。これを式(9)に代入すると、 $m_A = 0$ の時は

$$E\{\Delta m_A\} = 2\epsilon \int_0^{\infty} \frac{Q(x)}{x} dx > 0 \quad (10)$$

となり、 m_A が十分大きい時は

$$E\{\Delta m_A\} \simeq -\frac{2\epsilon}{m_A^2} \int_0^{\infty} xQ(x)dx < 0 \quad (11)$$

となる。これは、 $m_A > 0$ で少なくとも1つの収束点が存在することを意味する。しかし、 $m_A \rightarrow \infty$ で $E\{\Delta m_A\} \rightarrow 0$ となり、収束点から離れた位置で修正量が微小になる不都合がある。つまり、参照ベクトルの初期値に応じて ϵ を試行錯誤で調整する必要がある。

3.3 修正 GLVQ

式(6) (7) の分母の2乗を1乗にした場合、式(9)に対応する式は

$$E\{\Delta m_A\} = 4\epsilon \int \frac{(x - m_B)^2(x - m_A)Q(x)}{(x - m_A)^2 + (x - m_B)^2} dx \quad (12)$$

として得られる。前述したように $m_B = -m_A$ とすると、 $m_A = 0$ の時は

$$E\{\Delta m_A\} = 4\epsilon \int_0^{\infty} xQ(x)dx > 0 \quad (13)$$

となり、 $m_A \rightarrow \infty$ の時は

$$E\{\Delta m_A\} = -4\epsilon \int_0^{\infty} xQ(x)dx < 0 \quad (14)$$

となる。このことから、 $m_A > 0$ で少なくとも1つの収束点が存在するとともに、 $m_A \rightarrow \infty$ において $E\{\Delta m_A\}$ は0でない一定値になることがわかる。すなわち、収束点から離れた位置で修正量が一定値になるため、参照ベクトルの初期値に応じて ϵ を調整する必要がなく、良好に収束すると考えられる。以上、 m_A についてのみ議論したが、 m_B についても同様である。

4 シミュレーション実験

図2(a)に示す2次元の人工データを対象にGPDとGLVQの収束性を比較した。クラスA, Bそれぞれ1個の参照ベクトルを用意し、初期値は $[-3, 3]$ の一様乱数で設定した。10セットの初期値に対する学習中の参照ベクトルの軌跡を重ね書きしたものを図2(b)~(d)に示す。ただし、 $\epsilon, \ell(\rho_k)$ は一定値とした。GPDでは、参照ベクトルは右方向、左方向へそれぞれ発散している(図2(b))。オリジナルのGLVQでは、初期位置のまま変化がなく、実験に用いた ϵ の値が小さ過ぎたと考えられる(図2(c))。修正GLVQでは、参照ベクトルは初

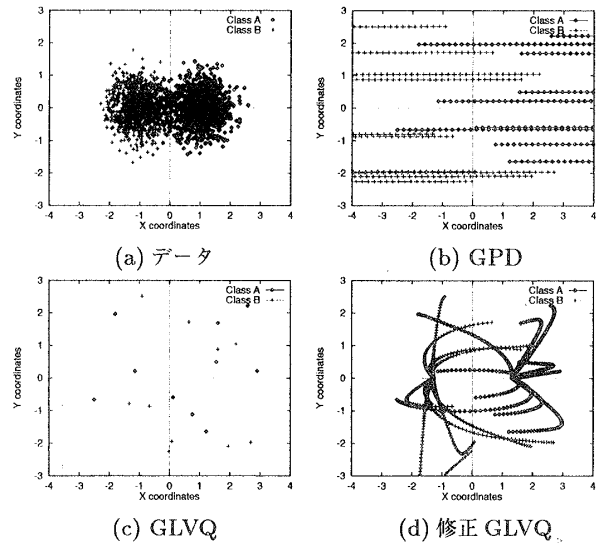


図2: 実験に用いたデータと参照ベクトルの軌跡

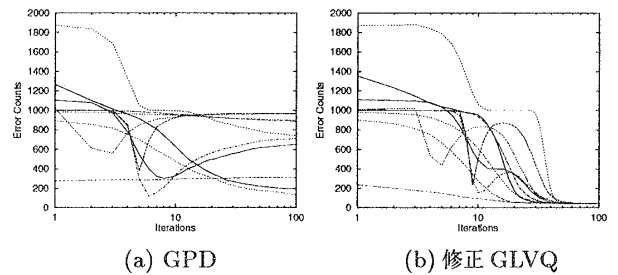


図3: 学習中のエラー数の推移

期値に依らず、それぞれが属するクラスのある1点に収束している(図2(d))。GPDと修正GLVQによる学習中のエラー数の推移を、それぞれ図3(a), (b)に示す(横軸は学習回数、縦軸はエラー数)。GPDでは初期値に強く依存しているが、修正GLVQでは初期値に依らず、いずれの場合もエラー数が最小値に到達している。

5 まとめ

最小分類誤りに基づくGPDとGLVQの収束性を議論し、修正GLVQは参照ベクトルの初期値に頑健であり、従来のGPDよりも収束性及び識別性能において優れていることを明らかにした。

参考文献

- [1] B.H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. on Signal Proc.*, Vol.40, No.12, pp.3043-3054, 1992.
- [2] T. Kohonen, "Self-Organizing Maps," Springer-Verlag, 1995.
- [3] A. Sato and K. Yamada, "Generalized Learning Vector Quantization," *NIPS 8*, pp.423-429, 1996.
- [4] A. Sato and K. Yamada, "An Analysis of Convergence in Generalized LVQ," *ICANN'98*, Vol.1, pp.171-176, 1998.