半教師あり学習法に基づく識別・判別問題

九州大学大学院数理学府　川 野 秀 一
九州大学大学院数理学研究院　小 西 貞 則

1. はじめに

識別・判別問題において，群の所属が既にわかっているデータ (ラベルありデータ) と所属が未知であるデータ (ラベルなしデータ) が混在している状況の下，ラベルありデータのみに基づき判別モデルを構築するのではなく，ラベルなしデータを効果的に用いることにより判別モデルを構成し，予測の精度を向上させる手法は半教師あり学習法と呼ばれている (Chapelle *et al.*, 2006). 近年この学習法は大きな注目を集めており，ラベルありデータの取得が高コストである分野，特にバイオインフォマティックスやテキスト分類などの分野で応用されつつある (例えば，Pan *et al.*, 2006). 半教師あり学習法は様々な分野・方面より研究されているが，本発表では特にロジスティックモデルに基づく方法について考えた.

Amini and Gallinari (2002) は，線形ロジスティックモデルを半教師あり学習法に拡張することにより新たな判別方式を得た．しかし，ロジスティックモデルに含まれるパラメータの推定において，従来の最尤法ではしばしば尤度関数が発散，またはその推定量が不安定になる恐れがある．さらに，線形ロジスティックモデルでは，複雑な構造を有する非線形現象を捉えることは困難である.

本発表では，正則化基底展開法に基づく半教師あり非線形ロジスティック判別方式を提案し，提案手法を実データ及び人工データに適用することによりその有効性について検証した.

2. 半教師あり判別方式

いま，p 次元説明変数ベクトル $\boldsymbol{X} = (X_1, \cdots, X_p)^T$ に関するデータ $\boldsymbol{x}$ が，第 1 群に属するならば 1 を，第 2 群に属するならば 0 の値を対応させるラベル変数 Y を導入する．次に，n_1 個のラベルあり観測データ $\{(\boldsymbol{x}_\alpha, y_\alpha); \alpha = 1, \cdots, n_1\}$ と $(n - n_1)$ 個のラベルなし観測データ $\{\boldsymbol{x}_\alpha; \alpha = n_1 + 1, \cdots, n\}$ が得られたとする．これら 2 種類のデータに基づいた非線形ロジスティック判別方式について考察する.

まず，ラベルありデータのみを用いて，非線形ロジスティックモデルを次のように構成する.

$$\log\left\{\frac{\Pr(y_\alpha = 1|\boldsymbol{x}_\alpha)}{\Pr(y_\alpha = 0|\boldsymbol{x}_\alpha)}\right\} = w_0 + \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}_\alpha) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha). \tag{1}$$

ここで，$\phi_j(\boldsymbol{x})$ $(j = 1, \cdots, m)$ は既知の基底関数，$\boldsymbol{\phi}(\boldsymbol{x}) = (1, \phi_1(\boldsymbol{x}), \cdots, \phi_m(\boldsymbol{x}))^T$ は基底関数ベクトル，$\Pr(y_\alpha = k|\boldsymbol{x}_\alpha)$ はデータ $\boldsymbol{x}_\alpha$ を与えたときの事後確率とし，$\boldsymbol{w} = (w_0, w_1, \cdots, w_m)^T$ は未知パラメータベクトルとする．また，基底関数には高次元化が容易なガウス型基底関数を用いる (Ando and Konishi, 2008). 判別方式は，事後確率 $\pi(\boldsymbol{x}_\alpha|\hat{\boldsymbol{w}}) := \hat{\Pr}(y_\alpha = 1|\boldsymbol{x}_\alpha), 1 - \pi(\boldsymbol{x}_\alpha|\hat{\boldsymbol{w}}) := \hat{\Pr}(y_\alpha = 0|\boldsymbol{x}_\alpha)$ が最大となる群に判別する．ここで，$\hat{\boldsymbol{w}}$ は $\boldsymbol{w}$ の推定量とする.

通常，非線形ロジスティックモデル (1) 式に含まれるパラメータ $\boldsymbol{w}$ は，対数尤度関数に正則化項を付与した正則化対数尤度関数

$$\ell_\lambda(\boldsymbol{w}) = \sum_{\alpha=1}^{n_1} \left[y_\alpha \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha) - \log\{1 + \exp(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha))\}\right] - \frac{n_1 \lambda}{2} \boldsymbol{w}^T K \boldsymbol{w}, \tag{2}$$

の最大化に基づき推定する．ここで，λ (> 0) は正則化パラメータであり，K は $(m+1) \times (m+1)$ 型の既知の非負値定符号行列とする.

しかし，本発表ではラベルなしデータの情報を活用した半教師あり学習法を考えているため，通常の正則化対数尤度関数にラベルなしデータの情報を組み込んだ次の正則化対数尤度関数

$$\begin{aligned}\ell_\lambda^*(\boldsymbol{w}) &= \sum_{\alpha=1}^{n_1}\left[y_\alpha \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha) - \log\{1+\exp(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha))\}\right] \\ &\quad + \sum_{\alpha=n_1+1}^{n}\left[t_\alpha \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha) - \log\{1+\exp(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha))\}\right] - \frac{n\lambda}{2}\boldsymbol{w}^T K\boldsymbol{w}, \qquad (3)\end{aligned}$$

の最大化に基づきパラメータ $\boldsymbol{w}$ を推定する．ここで，t_α は α 番目のデータが第 1 群に属するならば 1 を，第 2 群に属するならば 0 の値をとるダミー変数とする．しかし，変数 t_α $(\alpha = n_1+1,\cdots,n)$ の値は未知であるため，(3) 式を最大化することは一般に困難である．そこで，変数 t_α を潜在変数と捉え，EM アルゴリズムを適用することにより (3) 式を最大化する．

3. モデル評価基準

構築した統計モデルは，基底関数の個数，正則化パラメータ，基底関数の広がりを調節するハイパーパラメータの値に依存しているため，これら 3 つのパラメータの値を適切に選択する必要がある．本発表では，これらのパラメータの値を一般化ベイズ型モデル評価基準 (Konishi *et al.*, 2004) に基づき客観的に選択することを考える．Konishi *et al.* (2004) により得られた結果を，今回のガウス型基底関数に基づく半教師あり非線形ロジスティックモデルに適用することにより，次式のモデル評価基準が導かれる．

$$\begin{aligned}\mathrm{GBIC} &= -2\sum_{\alpha=1}^{n}\left[z_\alpha \log\pi(\boldsymbol{x}_\alpha|\hat{\boldsymbol{w}}) + (1-z_\alpha)\log\{1-\pi(\boldsymbol{x}_\alpha|\hat{\boldsymbol{w}})\}\right] + n\lambda\hat{\boldsymbol{w}}^T K\hat{\boldsymbol{w}} \\ &\quad -(m+1-d)\log\left(\frac{2\pi}{n}\right) + \log|R| - \log|K|_+ - d\log\lambda. \qquad (4)\end{aligned}$$

ここで，$\boldsymbol{z} = (y_1,\cdots,y_{n_1},t_{n_1+1},\cdots,t_n)^T$，$d$ は行列 K の階数，$|K|_+$ は行列 K の 0 でない固有値の積であり，R は次で与えられる $(m+1)\times(m+1)$ 型行列である．

$$R = \frac{1}{n}\Phi^T\hat{\Pi}(I_n - \hat{\Pi})\Phi + \lambda K. \qquad (5)$$

ここで，$\hat{\Pi} = \mathrm{diag}[\pi(\boldsymbol{x}_1|\hat{\boldsymbol{w}}),\cdots,\pi(\boldsymbol{x}_n|\hat{\boldsymbol{w}})]$，$\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_n))^T$ である．

モデル評価基準 GBIC の値を最小にする基底関数の個数，正則化パラメータ，ハイパーパラメータの値を選択し，対応するモデルを最適なモデルとする．

参考文献

Amini, M-R. and Gallinari, P. (2002): Semi-Supervised Logistic Regression. *Proceedings of the 15th European Conference on Artificial Intelligence.* 390–394.

Ando, T. and Konishi, S. (2008): Nonlinear logistic discrimination via regularized radial basis functions for classifying high-dimensional data. *Ann. Inst. Statist. Math*, (to appear).

Chapelle, O., Schölkopf, B. and Zien, A. (2006): *Semi-Supervised Learning.* MIT Press.

Konishi, S., Ando, T. and Imoto, S. (2004): Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika.* **91**, 27–43.

Pan, W., Shen, X., Jiang, A. and Hebbel, R.P. (2006): Semi-supervised learning via penalized mixture model with application to microarray sample classification. *Bioinformatics.* **22**, 2388–2395.