高次元データにおけるブースティング手法の改良: Sparse Learner Boosting

総合研究大学院大学 複合科学研究科　統計科学専攻
博士課程　　プリチャード　真理

1　はじめに

ゲノム解析の技術は飛躍的に向上し，数万のオーダーの遺伝子の情報を 1 度の実験で得ることはもはや日常的に行われることとなった．マイクロアレイを用いた遺伝子発現解析も現在では広く普及し，多くの研究室で利用されている．得られる遺伝子情報が飛躍的に増加した一方，実験に利用可能な患者数や実験動物の数ははるかに少なく，多くて 100 を超えるぐらいである．これは Hastie と Tibshirani (2004) により $p \gg n$ problem と呼ばれ，バイオインフォマティクスにおける重要な課題の一つとして知られている．

遺伝子発現解析のような高次元データを用いた場合，ブースティングでは学習に用いられる弱学習機の数も膨大となる．そのため作成されるモデルも複雑なものとなり，結果として過学習により未知のデータへの判別能力が低下する．本ポスター発表では，過剰な弱学習機の候補を削減する方法を提案し，それにより過学習が抑えられ判別能力が向上することをシミュレーションと実データによる検証結果を用いて報告する。

2　Sparse Learner Boosting

$\boldsymbol{x}$ を j 次元の変数を持つ特徴ベクトルとし，y を ± 1 の 2 値をとるクラスラベルとする．与えられる学習データを $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, N\}$ とし，$\boldsymbol{x}$ に関する関数 $f(\boldsymbol{x})$ を弱学習機と呼ぶ.

2.1　AdaBoost

AdaBoost は Freund と Schapire（1997）により提案されたブースティングであり，判別能力の弱い弱学習機を線形結合し，強力なスコア関数を作成する．弱学習機の集合を

$$\mathcal{F} = \{f_j(\boldsymbol{x}) : j \in (1, \ldots, J)\}. \tag{2.1}$$

とし，最終的なスコア関数を

$$F(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}), \tag{2.2}$$

として表す．ここで α_t と f_t は学習の過程で決定される．AdaBoost は次式で表される指数ロスを F から $F + \alpha f$ と更新させながら，その指数ロスが最小になるような f_t を選び，α_t を計算し，判別モデルを作成する

$$L_{\exp}(F) = \sum_{i=1}^{N} \exp(-y_i F(\boldsymbol{x}_i)). \tag{2.3}$$

2.2 Sparse Lerner Boosting

AdaBoost，η-Boost いずれの手法においても，最終的なスコア関数は（2.2）式に表される弱学習機の線形結合となる．もっとも一般的に使われる弱学習機は決定スタンプであり次式で表される

$$\mathcal{F} = \{f_j(\boldsymbol{x}, a, b) = a \cdot \mathrm{sgn}(x_j - b) : b \in \boldsymbol{R}\}. \tag{2.4}$$

Sparse Learner Boosting は弱学習機の false positive rate $\overline{\mathrm{FP}}(f)$ と false negative rate $\overline{\mathrm{FN}}(f)$ に着目し弱学習機の候補を削減する．$\overline{\mathrm{FP}}(f)$ と $\overline{\mathrm{FN}}(f)$ を

$$\overline{\mathrm{FP}}(f) = \frac{1}{n}\sum_{i=1}^{n} I(f(\boldsymbol{x}_i) \neq y_i | y_i = +1) \tag{2.5}$$

$$\overline{\mathrm{FN}}(f) = \frac{1}{n}\sum_{i=1}^{n} I(f(\boldsymbol{x}_i) \neq y_i | y_i = -1) \tag{2.6}$$

とし，削減された弱学習機のセット $\mathcal{F}_\xi$ を

$$\mathcal{F}_\xi = \{f : \min(\overline{\mathrm{FP}}(f), \overline{\mathrm{FN}}(f)) \leq \xi, f \in \mathcal{F}, \xi \in \mathbb{R}\} \tag{2.7}$$

とする．パラメータ ξ は弱学習機をどの程度スパースにするか決定するパラメータである．パラメータ ξ はクロスバリデーションにより決定することができる．削減した弱学習機のセット $\mathcal{F}_\xi$ を AdaBoost に組み込むことで作成される判別境界面の複雑さが抑えられることを図 1 に示した．詳細はシミュレーション，実データの結果とともにポスター発表にて報告する．

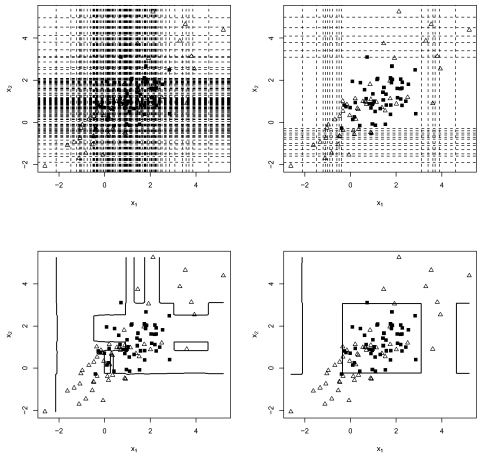

図 1: 2 次元のデータにより作成された弱学習機とその判別境界面．左上図が AdaBoost により作成された弱学習機，左下図が作成された判別境界面．右上図が提案法により作成された弱学習機，右下図が作成された判別境界面．

参考文献

Freund, Y. and Schapire, R. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *Journal of Computer and System Sciences*, **55**, 119-139

Hastie, T. and Tibshirani, R. (2004) Efficient quadratic regularization for expression arrays. *Biostatistics*, **5**, 329-340