AUCのブースティングとその遺伝子データへの応用

総合研究大学院大学　　小森　理
統計数理研究所　　江口　真透

Receiver Operating Characteristic (ROC) カーブは信号検出理論から考え出され，今日では心理学，医学その他さまざまな分野で使われている．その一つの理由としてROCカーブの下側面積（AUC）が判別の良さを測る指標として有用であることが挙げられる．誤判別には第一種の誤りと第二種の誤りがあるが，AUCはその両者を考慮した指標であることが通常のエラーレイトとは異なる点である．医療の現場では健常者を誤って病気であると判断する場合と，病人を誤って健常と判断する場合とではそれによって被るコストがそれぞれ異なる．その診断によって手術などが行われる場合には誤って病気と誤判別する確率をできるだけ抑え，正しく病状を判断する必要にせまられる．

ROCカーブには主に三つの特徴がある．一つ目はロジスティック回帰による解析や正規分布を仮定した線形判別による判別などとは違い，特定の確率分布をなんら仮定せずに解析をすることができること．二つ目は想定している二つの母集団の事前確率に無関係であること．よってケース・コントロール研究にも適用可能となる．三つ目はAUCが，判別をする者が用いる閾値によらない指標であることである．エラーレイトやオッヅ比とは異なる特徴である (Pepe et al, 2004)．

さまざまな変量（検査値やマーカーの値）を組み合わせ判別の精度を上げる試みは今までいくつかなされてきた．Pepe and Thompson (2000) はAUCを最大にする手法を変量数が小さい場合に提案し，Cai and Longton (2006) はAUC最大化によって求めた変量の組み合わせ方がロジスティック回帰によって求めたものとかなり食い違うことがあり得ることを示した．またMa and Huang (2005) は高次元データへに対応できるよう，これらの手法を拡張した．しかしながら，これらの手法は全て変量の線形結合に限っていることが問題として挙げられる．現実世界のデータに内在する非線形性を考慮したより実際的な手法が望まれる．

さらに，Eguchi and Copas (2002) とMcIntosh and Pepe (2002) によりAUCを最大にするスコア関数（変量の組み合わせ方）は二つの母集団に対する尤度比の単調増加関数の形で表現できることが示されている．つまり，両母集団の確率分布に正規分布を仮定した単純なモデルにおいても，分散が異なる場合はAUCを最大にする最適なスコア関数は線形ではなく非線形となる．

我々は今回，機会学習の分野の一手法であるブースティングを用いてAUCを最大化する統計的手法を提案する．ブースティングは弱判別機を呼ばれるごく単純な関数を重みつき線形結合で組み合わせ，最終的に非常に柔軟なスコア関数を構築する手法である．これはFreund and Schapire (1997) のAdaBoostからはじまり，その他ロジスティック回帰の考えをブースティングに取り入れたLogitBoost (Hastie and Tibshirani, 2000), AdaBoostの過剰学習をnaive lossを導入することで改良したη-Boost (Takenouch and Eguchi,2004) などが提案されている．またAdaBoostの拡張として$\mathcal{U}$-BoostがMrata et al (2004) により提案され，その統計的性質が議論されている．

まずはじめに特徴量として p 個のマーカー $\boldsymbol{x} \in \mathbf{R}^p$ を考え，それぞれに対し弱判別機 $f(\boldsymbol{x})$ の集合として，

$$\mathcal{F}_k = \left\{ f(\boldsymbol{x}) = a\mathrm{H}(x_k - b) + (1-a)/2 \mid a \in \{-1, 1\},\ b \in \mathcal{B}_k \right\},\ k = 1, \ldots, p,$$

を考える．但し, H は Heaviside 関数，$\mathcal{B}_k$ は実数空間上のある集合とする．そして全体の弱判別機の集合を $\mathcal{F} = \bigcup_{k=1}^{p} \mathcal{F}_k$ とする．つまり，ブースティングにより構築されるスコア関数 $F(\boldsymbol{x})$ は上記のような 0 または 1 しか値を取らない $f(\boldsymbol{x})$ の組み合わせとして得られる．このような設置のもとで，スコア関数は以下のように書き表わせる．

$$\begin{aligned} F(\boldsymbol{x}) &= \sum_{f \in \mathcal{F}_1} \alpha_f f(\boldsymbol{x}) + \cdots + \sum_{f \in \mathcal{F}_p} \alpha_f f(\boldsymbol{x}) \\ &= F_1(x_1) + \cdots + F_p(x_p). \end{aligned}$$

但し，α_f は弱判判別機 f に対する重みとする．このようにスコア関数のモデルは一般化加法モデル (GAM) と同様なものを想定する．このモデルは変量間の相互作用を考慮しないかわりに各変量がどのように判別に寄与するかの解釈を容易にする．x_k に対する $F_k(x_k)$ のプロットをスコアプロットと呼び，変量が持つ非線形性の視覚的な理解に非常に有用なものである．

また，このスコア関数 $F(\boldsymbol{x})$ を用いると，我々が提案する目的関数は

$$\overline{\mathrm{AUC}}_\lambda(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \Phi(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) - \lambda \sum_{k=1}^{p} \sum_{x_k \in \mathcal{B}_k} \left\{ F_k^{(2)}(x_k) \right\}^2,$$

と書き表わすことができる．但し，n_0, n_1 はそれぞれの群からの標本の数であり，第一項は標準正規分布の分布関数 Φ で近似した AUC に相当する．第二項はスコア関数の滑らかさを平滑化パラメータ λ で調整する項である．この平滑化はスコア関数のデータへの過剰な当てはまりを防ぐだけでなく，変量効果の単調性を重視する医療や疫学の分野において解析後の変量解釈を有意義なものとする．

最後に今回提案する AUCBoost を遺伝子発現量のような高次元データに応用することを考える．一般的に高次元データでは “$p >> n$” が問題となる．つまり標本数 n に対し変量の数 p が非常に小さいため，過剰学習や擬陽性が引き起こされる．Takenouch et al (2007) はブースティングの各ステップで弱判別機を集団として選ぶことにより，過剰学習を抑えたよりロバストな手法を提案した．我々は今回弱判別機を上記の stump から polynomial（多項式）に変えることでこの問題に取り組んだ．

$$\mathcal{P}_k = \left\{ f(\boldsymbol{x}) = (x_k - \mu_k)^a \mid a = 1, 2, 3 \ldots \right\},\ k = 1, \ldots, p,$$

但し μ_k は各変量の平均とする．あるシミュレーションにおいては擬陽性の問題が stump を用いる場合と比べ改善されることが分かった．この考えは実データの解析でも有用と思われる．