分布の異質性を考慮した t 統計量と AUC の一般化

統計数理研究所　　小森 理，江口 真透

二値判別の問題において，古典的でかつ有用な手法に Fisher の線形判別関数がある．群間分散と群内分散の比を最大化するように求めるこの単純な手法が，さまざまな実データ解析において，より複雑な計算を含む機械学習の手法より有用であることが報告されている [1, 2]．そこで t 統計量をもとにして Fisher の線形判別関数の拡張を試みた [3]．そこでは片方の集団には正規性を仮定した上で，もう一方の集団には確率分布の異質性を考慮した判別手法を提案した．さまざまな生成関数 U を考えることで Fisher の線形判別関数，AUC(Area under the ROC curve), Kullback-Leibler ダイバージェンスとの興味深い関係性が明らかにされ，線形判別関数の係数の漸近分散を陽に求めることで，推定精度と判別精度の両方の意味で最適な線形判別関数の構築を行った．

二つの集団両方に分布の異質性を考慮し，上記の手法の拡張を試みる．クラスラベル $y \in \{0, 1\}$ のデータをそれぞれ $\{x_{0i} \in \mathbb{R}^p | i = 1, \ldots, n_0\}$，$\{x_{1j} \in \mathbb{R}^p | j = 1, \ldots, n_1\}$ とし，生成関数を $U : \mathbb{R} \to \mathbb{R}$ とすると，一般化 AUC は以下のように定義される．

$$L_U(\beta) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} U\left\{ \frac{\beta^{\mathrm{T}}(x_{1j} - x_{0i})}{(\beta^{\mathrm{T}} S \beta)^{1/2}} \right\}.$$

ただし $S = 1/n_0 \sum_{i=1}^{n_0} (x_{0i} - \bar{x}_0)(x_{0i} - \bar{x}_0)^\top + 1/n_1 \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)^\top$ とする．一般化 AUC から求まる推定量を

$$\widehat{\beta}_U = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}}\ L_U(\beta)$$

とする．この推定量のクラスの統計的性質として一致性と漸近有効性を考察するために推定したいパラメータを

$$\beta_0 = \frac{\Sigma^{-1}(\mu_1 - \mu_0)}{\{(\mu_1 - \mu_0)^{\mathrm{T}} \Sigma^{-1} (\mu_1 - \mu_0)\}^{1/2}}$$

と定める．このパラメータ β_0 に対して次の二つの仮定を考える．

(A)　　$E_y(g_y \mid w_y = a) = 0 \quad \text{for all } a \in \mathbb{R}, \text{ for } y = 0, 1$

(B)　　$\mathrm{var}_y(g_y \mid w_y = a) = Q \Sigma_y Q^{\mathrm{T}} \quad \text{for all } a \in \mathbb{R}, \text{ for } y = 0, 1.$

ただし，$w_y = \beta_0^{\mathrm{T}} x_y$, $g_y = Q x_y$, $Q = I - \Sigma \beta_0 \beta_0^{\mathrm{T}}$, $\Sigma_y^* = Q \Sigma_y Q^{\mathrm{T}}$ とする．ここで μ_0, μ_1, Σ はそれぞれ $\bar{x}_0$, $\bar{x}_1$, S に対する平均と分散の母集団パラメータとする．これらの設定で $\widehat{\beta}_U$ の漸近分散を最小にする最適な生成関数 U の導出を考察し，[3] との関連性を議論する．またシミュレーションによる最適な U の評価も行う．

[1] Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). *JASA* **97**, 77–87.

[2] Hess, K. R., Anderson, K. *et al.* (2006). *Journal of Clinical Ontology* **24**, 4236–4244.

[3] Komori, O., Eguchi, S. and Copas, J. B. (2015). *Biometrics* **71**, 404–416.