One-Pass AUC Optimization 手法の実務的拡張

（株）金融工学研究所　木村和央 *

1. はじめに

ロジットモデルに代表されるクラス分類を目的とした線型モデルの精度評価指標として，**AUC**，**AR** 値（**Somers' D**）といった順序統計量（序列精度）が重要視されている．一方で，当該モデル構築時におけるパラメータの推計は，一般には最尤法によりなされ，直接的に評価指標の最大化を目的とはしていない．

これに対し，山下・三浦 [1] はシグモイド関数を用いた近似 **AR** 値を直接最大化するという強力な推計手法を提案したが，計算には $\mathcal{O}(N^2)$（N はサンプル数）の記憶容量を必要とした．一方で，Gao et al. [2] はサンプルを 1 度だけ読込み逐次処理するオンライン学習で，最小二乗型損失を用いたことで記憶容量は $\mathcal{O}(d^2)$（d は説明変数の数）で済む **One-Pass AUC Optimization**（**OPAUC**）手法を提案した．

本稿では，[2] の **OPAUC** 手法をベースに，サンプル重みとパラメータ制約を考慮しつつ，順序のある複数クラス問題へも適用可能となるよう，実務を意識しつつ自然な拡張を試みた．

2. OPAUC 手法の拡張

基本的な枠組みは [2] と同様であるので原論文を参照いただきたい．t 番目のサンプルを $(\mathbf{x}_t, y_t)$，重みを $\omega_t > 0$ とする．ここで，$\mathbf{x}_t \in \mathbb{R}^d$ は説明変数，$y_t \in \{0, 1, \cdots, m\}$ はクラスである．また，t 番目までのサンプルで推計された線型モデルを $f_t(\mathbf{x}) = \mathbf{w}_t \cdot \mathbf{x}$ とする．サンプルを得るたび，$\mathbf{w}_t$ を次式にて更新する．$\mathbf{w}_t = \max(\min(\mathbf{w}_{t-1} - \eta \nabla \mathcal{L}_t(\mathbf{w}_{t-1}), \mathbf{b}_{\max}), \mathbf{b}_{\min})$，ここで，$\mathbf{b}_{\min} \leq \mathbf{w} \leq \mathbf{b}_{\max}$ は制約条件，η はステップ幅，$\nabla \mathcal{L}_t$ は t 番目のサンプルとそれとは異なるクラスに属す $t-1$ 番目までのサンプルから計算される損失関数の勾配（$\mathbf{w}$-微分）である．サンプルの重み総和を T_t，クラス k に属するサンプルの重み和を T_t^k，説明変数の平均と共分散行列を $\mathbf{c}_t^k$，$S_t^k \sim \mathcal{O}(d^2)$ とすると，$\nabla \mathcal{L}_t$ は，

$$\nabla \mathcal{L}_t(\mathbf{w}) = \omega_t \left[\lambda \mathbf{w} + \sum_{\substack{k=0, \\ k \neq y_t}}^{m} \frac{T_t^k}{T_t - T_t^{y_t}} \left(- \left(\mathbb{I}_{(y_t > k)} - \mathbb{I}_{(y_t < k)} \right) (\mathbf{x}_t - \mathbf{c}_t^k) + (\mathbf{x}_t - \mathbf{c}_t^k)(\mathbf{x}_t - \mathbf{c}_t^k)^{\mathrm{T}} \mathbf{w} + S_t^k \mathbf{w} \right) \right]$$

と簡略化される．λ は正則化パラメータ，$\mathbb{I}_\pi$ は指示関数（π が真は 1，偽は 0），T は転置記号である．

3. 数値実験

格付モデル構築に，近似 **AR** 値最大化 [1]，最尤法による順序ロジット，拡張 **OPAUC** の各手法を適用し，**Somers' D** により精度を比較した．本決算データから作成した財務指標を説明変数（$d = 76$）に，6 か月後の R&I 格付（AA+以上/BB+以下を集約，$m = 10$）をクラスに用いた．格付別サンプル数の偏りを補正するため逆数にてサンプル重みを，また説明変数には符号条件を設定した．$Y-2$，$Y-1$ 年のデータ（$N \sim 800$）にて学習し，Y 年のデータ（$N \sim 400$）にてテストした．拡張 **OPAUC** の精度は劣後するものの，計算コストを考慮すれば良好な結果が得られた．また，オンライン学習ではなくなるものの，複数回処理を繰り返すことで，精度が向上することも確認できた（拡張 **TPAUC** は **Two-Pass AUC Optimization**）．

モデル構築手法	**Somers' D** 2007-15 年平均	
	学習データ	テストデータ
近似 **AR** 値最大化	0.834	0.816
最尤法順序ロジット	0.832	0.818
拡張 **OPAUC** 手法	0.818	0.807
拡張 **TPAUC** 手法	0.827	0.814

参考文献

[1] 山下智志, 三浦翔 (2011). 信用リスクモデルの予測精度- AR 値と評価指標- . 朝倉書店.

[2] Gao, W., Wang, L., Jin, R., Zhu, S., Zhou, Z.-H. (2016). One-Pass AUC Optimization. Artificial Intelligence, 236, 1-29.

*本稿の内容は筆者に属し，所属組織の見解ではない．