Bregman/Power divergenceによるロバスト能動学習

日野 英逸　　モデリング研究系 教授

【概要】 能動学習の方法論として，多数の予測器の合意に基づくQuery By Committee(QBC)が古くから用いられている．予測器が「外れ値」的な振る舞いをする場合，あるいは説明変数に外れ値が含まれている場合でも大きく精度を落とすことのないQBC手法を，Bregmanダイバージェンス及びPowerダイバージェンスを用いて実現し，そのロバスト性を解析した．

【動機】

・能動学習（Active Learning）は予測モデルの予測精度を向上するために適した学習サンプルを逐次的に選択する手法である．特に，少数のラベル付き学習データ集合と，多数のラベルなしデータ集合（プール）が与えられ，プールの中からラベルを付けるべきサンプルを選択する問題設定を考える．

・現状の予測モデルとラベル付き学習データ集合の情報から，次にどのサンプルのラベル付けするかは，獲得関数によって決定する．

・獲得関数として，複数の予測器の合意に基づくQBCが古くから知られている

・能動学習はサンプル数が少ない状況で実行するため，予測器たちには非常に質の悪いものが混入しうる．また，外れ値的なデータを積極的に選択する傾向がある．

・QBCをダイバージェンスを用いて理解し，ロバストなダイバージェンスを用いることで安定したQBCアルゴリズムを実現する．

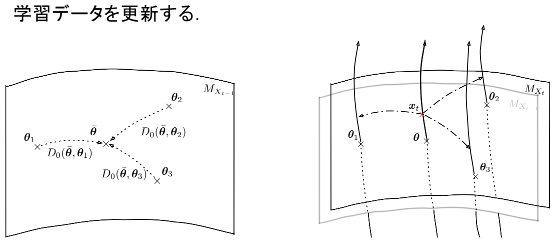
【QBC】

・本研究では一般化線形モデルによる予測を考える．

$$p(y|\xi(\boldsymbol{x})) = \exp\left(\frac{y\xi(\boldsymbol{x}) - \psi(\xi(\boldsymbol{x}))}{\phi} + c(y,\phi)\right)$$

QBCアルゴリズム：

1. C個の予測モデルを（例えばBaggingなどを用いて）学習する．これらのモデルパラメタを $\boldsymbol{\theta}_{c,t-1}, c = 1,\dots,C$ とする．
2. データプールから獲得関数 $a_0(\boldsymbol{x}) = \sum_{c=1}^{C} w_c D_0(p(y|\xi_{c,t-1}(\boldsymbol{x})), p(y|\bar{\xi}(\boldsymbol{x})))$ を用いてラベル付け対象のデータを選択し，対応する応答変数値を得る．ここで D_0 はKL-divergence，$\bar{\xi}(\boldsymbol{x})$ は
$$\bar{\xi}(\boldsymbol{x}) = \arg\min_{\xi(\boldsymbol{x})} \sum_{c=1}^{C} w_c D_0(p(y|\xi(\boldsymbol{x})), p(y|\xi_c(\boldsymbol{x})))$$
で定義されるconsensus model．$\xi_{c,t-1}(\boldsymbol{x}) = \langle \boldsymbol{\theta}_{c,t-1}, \boldsymbol{x} \rangle$ である
3. 学習データを更新する．



Remark: Consensus modelも獲得関数もKL-divergenceを用いて定義される

【Bregman divergence / u-mixture】

KL-divergenceやL2距離を特殊な場合として含む．

$$D_U(p,q) = \int d_U(p(y),q(y))\mathrm{d}\Lambda(y) = \int d_U(p,q)\mathrm{d}\Lambda, \qquad d_U(f,g) = U^*(f) + U(\breve{g}) - f\breve{g},$$

$$u^* = u^{-1} \quad \breve{f} = u^*(f)$$

ここでは β -divergenceを考える．
$$U(z) = \frac{1}{\beta+1}(\beta z+1)^{\frac{\beta+1}{\beta}}, \quad U^*(\zeta) = \frac{\zeta^{\beta+1}}{\beta(\beta+1)} - \frac{\zeta}{\beta},$$
$$u(z) = (\beta z+1)^{1/\beta}, \qquad u^*(\zeta) = \frac{\zeta^\beta - 1}{\beta}.$$

$$\begin{aligned} D_\beta(p,q) =& \frac{1}{\beta+1}\int q^{\beta+1}\mathrm{d}\Lambda - \frac{1}{\beta+1}\int p^{\beta+1}\mathrm{d}\Lambda - \frac{1}{\beta}\int p(q^\beta - p^\beta)\mathrm{d}\Lambda \\ =& \frac{1}{\beta(1+\beta)}\int p^{\beta+1}\mathrm{d}\Lambda - \frac{1}{\beta}\int pq^\beta\mathrm{d}\Lambda + \frac{1}{\beta+1}\int q^{\beta+1}\mathrm{d}\Lambda. \end{aligned}$$

Bregman-divergenceに関するconsensus modelは，u-mixtureと呼ばれ，次の定理で特徴づけられる：

定理（u-mixtureの特徴づけ）

C個の確率モデルの重み付きu-mixture（重み付きu-射影の総和最小のモデル）

$$\arg\min_q \sum_{c=1}^{C} w_c D_U(q, p_c). = p_u(y;w)$$

は

$$p_u(y;w) = u\left(\sum_{c=1}^{C} w_c \breve{p}_c(y) - b\right)$$

で与えられる．bは正規化定数．

命題

指数型分布族のu-mixtureモデルの影響関数は，u-divergenceがKL-divergenceの場合には非有界である．一方，u-divergenceがβ-divergence（β>0）の場合，影響関数は有界である．

Remark: u-mixtureは，正規化因子bが一般には求まらない（離散分布なら常に求まる）．

【Dual γ-power divergence/ dual γ-mixture】

γ-power divergenceは第2引数に関するscale不変性を持つdivergenceであり，非正規化統計モデルに関する推論によく用いられる．

$$D_\gamma(p,q) = -\frac{\int pq^\gamma \mathrm{d}\Lambda}{(\int q^{\gamma+1}\mathrm{d}\Lambda)^{\frac{\gamma}{\gamma+1}}} + \left(\int p^{\gamma+1}\mathrm{d}\Lambda\right)^{\frac{1}{\gamma+1}}$$

ここで，dual γ-power divergenceを考える（第一引数についてscale不変）：

$$D^*_\gamma(p,q) = -\frac{\int pq^\gamma \mathrm{d}\Lambda}{(\int p^{\gamma+1}\mathrm{d}\Lambda)^{\frac{1}{\gamma+1}}} + \left(\int q^{\gamma+1}\mathrm{d}\Lambda\right)^{\frac{\gamma}{\gamma+1}}$$

命題（dual γ-power mixtureの特徴づけ）

C個の確率モデルの重み付きdual γ-power mixture（重み付きdual γ-power射影の総和最小モデル）

$$\arg\min_q \sum_{c=1}^{C} w_c D^*_\gamma(q, p_c) = p_\gamma(y;w)$$

は

$$p_\gamma(y;w) = \frac{1}{z(w)}\left(\sum_{c=1}^{C} w_c p_c(y)^\gamma\right)^{\frac{1}{\gamma}}$$

で与えられる．ここで

$$z(w) = \int \left(\sum_{c=1}^{C} w_c p_c(y)^\gamma\right)^{\frac{1}{\gamma}} \mathrm{d}\Lambda(y)$$

Remark: dual γ-power mixtureは常に定義可能

命題

指数型分布族のdual γ-mixtureモデルの影響関数は，γ>0の場合，有界である

【QBCの獲得関数と影響関数】

Consensus modelと同様に，β-divergence及びdual γ-power divergenceに基づく獲得関数を考える：

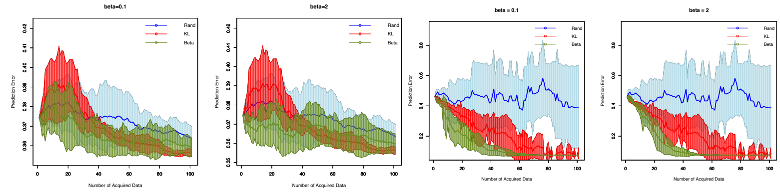
$$a_\beta(\boldsymbol{x};w) = \sum_{c=1}^{C} w_c D_\beta(p(y;\xi_c(\boldsymbol{x}), p_\beta(y|\boldsymbol{x}))) \qquad a_\gamma(\boldsymbol{x};w) = \sum_{c=1}^{C} w_c D^*_\gamma(p(y;\xi_c(\boldsymbol{x}), p_\gamma(y|\boldsymbol{x})))$$

命題

$a_0(\boldsymbol{x})$ の影響関数は非有界で，$a_\beta(\boldsymbol{x})$ と $a_\gamma(\boldsymbol{x})$ の影響関数は有界である．

【数値実験】

ロジスティック回帰モデルで，β-divergence，KL-divergenceに基づくQBCとランダムサンプリングを比較．



2次元人工データ　　　　IJCNNデータ

本研究成果は統計数理研究所　江口真透氏との共同研究に基づくものです．