日本オペレーションズ・リサーチ学会
2004年春季研究発表会

0-1損失関数を用いたロバストなブースティングの提案

02203190 筑波大学 *佐野夏樹 SANO Natsuki
01207840 筑波大学 鈴木秀男 SUZUKI Hideo
01105930 筑波大学 香田正人 KODA Masato

1 はじめに

ブースティング (Boosting) は逐次的に学習データの分布を変化させ，その度に弱学習機 (学習アルゴリズム) をあてはめ，これらの弱学習機を組み合わせて精度の高い学習機械を構成する手法である．代表的なものとして AdaBoost [1], Bagging 等が挙げられデータマイニングにおいてニューラルネットワークや決定木等の性能を強化するために用いられる．さらに AdaBoost を一般化させ任意の損失関数からブースティングを導出できるようにしたものに次の MarginBoost [2] がある．

2 MarginBoost

MarginBoost (Mason et al., 2000)

1. Specify a suitable loss function C.
2. Initialize $w_0(i) = 1/n$ for $i = 1, 2, \ldots, n$, and $F_0(x) = 0$.
3. For t = 1,2,...,T, do
   (a) Fit a classifier $f_t(x)$ to the training data weighted by $w_{t-1}(i)$ for $i = 1, 2, \ldots, n$.
   (b) If $\sum_{i=1}^n w_{t-1}(i) y_i f_t(x_i) \geq 0$, then return F_t.
   end If.
   (c) Choose β_t appropriately.
   (d) Let $F_t(x) = F_{t-1}(x) + \beta_t f_t(x)$.
   (e) Set $w_t(i) = C'(y_i F_t(x_i)) / \sum_{i=1}^n C'(y_i F_t(x_i))$ for $i = 1, 2, \ldots, n$.

   end For
4. Output $\mathrm{sign}(F_T(x))$.

MarginBoost は任意の関数 (通常は単調減少関数) を損失関数とし (ボックス中1)，さらにその y $F_t(x)$ に関する微分 $C'(y_i F_t(x_i)) / \sum_{i=1}^n C'(y_i F_t(x_i))$ を重み $w_t(i)$ としている (ボックス中3 (e)) 点で一般化といえる．この $yF(x)$ で表される量が Margin であり，最終的な判別は $\mathrm{sign}(yF_T(x))$ により行われるため次のような関係が成り立つ．

$$\text{if } y = \mathrm{sign}(F_T(x)) \text{ then } yF_T(x) \geq 0 \qquad (1)$$
$$\text{if } y \neq \mathrm{sign}(F_T(x)) \text{ then } yF_T(x) < 0$$

Margin は学習データの判別の難易度を表す量であり，Margin の値が大きいほど判別が易しく Margin の値が小さいほど判別が難しいデータであることを示す．

3 損失関数

AdaBoost は $C = \exp$ とした場合の MarginBoost であり，図1に示される．指数関数は負の Margin を持つデータに対して極端に大きなペナルティーをかけていることがわかる．つまり判別の難しいデータを集中的に学習している．図1には他に2項対数尤度等の損失関数が示されているが，これらは全て 0-1 損失関数を近似したものとみなすことができる．損失関数を 0-1 損失関数とした場合，期待損失を最小化するような識別関数はベイズ識別となることから 0-1 損失関数は最も理想的な損失関数であるといえる．しかしながら 0-1 損失関数は原点において微分不可能であることから次の近似法によって 0-1 損失関数の微分の近似を行う．

4 Stochastic Noise Reaction

Stochastic Noise Reaction(SNR) はノイズを用いて微分の近似を行う方法である．ある関数 C の微分 $\partial C(x)/\partial x_i$ を求めたいとする．SNR はまず変数 x_i にノイズ $\xi_i \in N(0,1)$ を入れる．

$$x_i(j) = x_i + \xi_i(j) \text{ for } j = 1, \ldots, M. \qquad (2)$$

$\xi_i(j)$ は x_i に入れられたノイズの j 番目の実現値である．さらに関数 C と ξ の積の平均を取ることで微分を

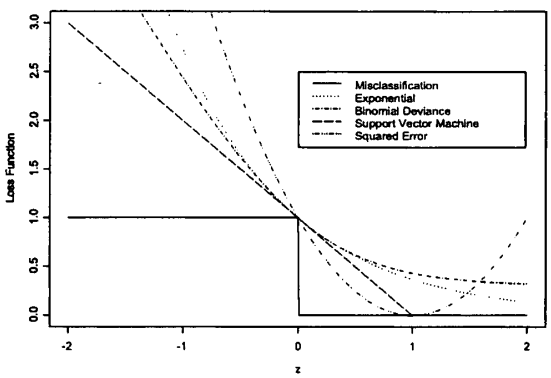

図 1: 種々の損失関数．横軸 z は Margin

近似できる．

$$\frac{\partial C(x)}{\partial x_i} \approx \frac{1}{M}\sum_{j=1}^{M} C(x(j))\xi(j). \tag{3}$$

SNR は微分不可能な関数の微分を求めたい場合に有効であり，ニューラルネットワーク [3] や巡回セールスマン問題に適用した研究がある．

4.1 SNR の導出

σ_i はノイズ $\xi_i \in N(0,\sigma_i^2)$ の標準偏差であり，$\langle\rangle$ が期待値演算を表すとすると次のような関係が成り立つ．

$$\langle \xi_i \xi_j \rangle = \sigma_i^2 \delta_{ij}, \quad \langle \xi_i^{2r+1} \rangle = 0, \quad \langle \xi_i^{2r} \rangle = \sigma_i^{2r}\frac{(2r)!}{2^r r!}, \tag{4}$$

δ_{ij} はクロネッカーのデルタを表し，r は非負の整数を表す．(4) 式を用いて $f(x+\xi)$ を x の周りでテイラー展開すると次のような関係が成り立つ．

$$\begin{aligned}
&\frac{1}{\sigma_i^2}\langle f(x+\xi)\xi_i\rangle \\
&= \frac{1}{\sigma_i^2}\langle \{ f(x) + \sum_{k=1}^{\infty}\frac{1}{k!}\sum_{j=1}^{N}(\xi_j \frac{\partial}{\partial x_j})^k f(x)\}\xi_i \rangle \\
&= \frac{1}{\sigma_i^2}\{ f(x)\langle\xi_i\rangle + \sum_{k=1}^{\infty}\frac{1}{k!}\sum_{j=1}^{N}\langle \xi_i(\xi_j \frac{\partial}{\partial x_j})^k\rangle f(x)\} \\
&= \frac{1}{\sigma_i^2}\sum_{k=1}^{\infty}\frac{1}{k!}\langle\xi_i^{k+1}\rangle\frac{\partial^k}{\partial x_i^k}f(x) \\
&= \frac{1}{\sigma_i^2}\sum_{k=2,4,6,\dots}\frac{1}{(k-1)!}\langle\xi_i^k\rangle\frac{\partial^{k-1}}{\partial x_i^{k-1}}f(x) \\
&= \frac{\langle\xi_i^2\rangle}{\sigma_i^2}\frac{\partial f(x)}{\partial x_i} + \sum_{k=4,6,8,\dots}\frac{1}{(k-1)!}\frac{\langle\xi_i^k\rangle}{\sigma_i^2}\frac{\partial^{k-1}}{\partial x_i^{k-1}}f(x) \\
&= \frac{\partial f(x)}{\partial x_i} + \sum_{r=2}^{\infty}\frac{1}{(r-1)!}\left(\frac{\sigma_i}{\sqrt{2}}\right)^{2(r-1)}\frac{\partial^{2r-1}}{\partial x_i^{2r-1}}f(x) \\
&\approx \frac{\partial f(x)}{\partial x_i},
\end{aligned} \tag{5}$$

最後の式の近似は $\sigma_i < \sqrt{2}$ の場合に成り立ち，$\sigma_i \to 0$ となるほど良い近似となる． 簡便のために $\sigma_i = 1$ とし， 期待値はサンプル平均に置き換えられている．M を大きくするほど，やはり近似の精度は上がる．

5 SNRBoost

SNR により 0-1 損失関数の微分を近似し，MarginBoost からブースティングを導出したのが SNRBoost である．次のボックス中にアルゴリズムを示す．発表では数値実験による SNRBoost のロバスト性の検証を行う．

SNRBoost

1. Initialize $w_0(i) = 1/n$ for $i = 1, 2, \dots, n$, and $F_0(\mathbf{x}) = 0$.
2. For $t = 1, 2, \dots, T$, do
   - (a) Fit a classifier $f_t(x)$ to the training data weighted by $w_{t-1}(i)$ for $i = 1, 2, \dots, n$.
   - (b) Set $\beta_t = 1/t$.
   - (c) $F_t(x) = F_{t-1}(x) + \beta_t f_t(x)$.
   - (d) Compute $z_t(i) = y_i F_t(x_i)$.
   - (e) Compute approximated derivatives for zero-one loss function $d_t(i)$ at $z_t(i)$ using SNR for $i = 1, 2, \dots, n$.
   - (f) Compute weights $w_t(i) = d_t(i)/\sum_{i=1}^{n} d_t(i)$ for $i = 1, 2, \dots, n$.
   - (g) If $w_t(i) \leq 0$ then $w_t(i) = 0$. end If.

   end For
3. Output $\mathrm{sign}(F_T(x))$.

参考文献

[1] Freund, Y. and Schapire, R. E. "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Comp. and System Sci.*, **55**, 119-139. (1997)

[2] Mason, L., Baxter, J., Bartlett, P. L., and Frean, M. (2000): Functional gradient techniques for combining hypotheses. In A.J.Smola, P.L.Bartlet, B.Scholköpt, and D. Schuurmans, editors, *Advances in Large Margin Classifiers,* pages 221-246. MIT Press, Cambridge, MA, 2000.

[3] Koda, M. and Okano, H. (2000): A new stochastic learning algorithm for neural networks. *Journal of the Operations Research Society of Japan*, **43** (4), 469-485.