A-054

多値分類問題に対するブースティングの困難さについて
On the Difficulty of Boosting for Multi-class Classification Problems

田中 恭 † 瀧本 英二 ‡ 丸岡 章 ‡
Kyo Tanaka Eiji Takimoto Akira Maruoka

1. はじめに

ブースティングとは，精度の低い複数の仮説（弱仮説）を統合することによって精度の高い仮説（統合仮説）を作り出す手法である．これまで多くのアルゴリズムが提案されているが，それらのほとんどは本質的に 2 値分類問題に対して設計されており，多値分類問題に適用するためには，問題を 2 値分類問題に還元する必要がある．その背景には，そもそも弱仮説の概念が確立されていないという事情がある．

一方，仮説の精度を情報量の概念を用いて評価し，弱仮説を学習対象に関して保持する情報の量が 0 ではないこととする定義が提案され，この定義の下で決定木生成アルゴリズムが多値分類問題に対するブースティングを実現していることが示されている [2]．ただし，この方法は効率が悪く，統合仮説の精度を十分高くするには，非常に多くの弱仮説を必要とするという欠点がある．

本稿では，この情報量に基づく定式化の下で，多値分類問題に対するブースティングの困難さを示唆する 3 つの結果を与える．

2. 準備

X を事例空間，Y をラベル空間とする．特に N 値分類問題に対しては $Y = \{0, \ldots, N-1\}$ とする．学習アルゴリズムとはサンプル $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq X \times Y$ が与えられたとき，そのサンプルの背後にある規則を推測し，仮説 $h : X \to Y$ を出力する過程である．ブースティングとは，性能のあまり良くない学習アルゴリズム（弱学習者）WL と，これを用いて精度の高い仮説を作ることを目指すブースターからなるスキームで，次のように動作する．各ラウンド $t\,(1 \le t \le T)$ において，(1) ブースターはサンプル S 上の確率分布 D_t を作り，これを S と共に弱学習者 WL に与える．(2)WL は弱仮説 h_t をブースターに返す（$h_t = \mathrm{WL}(S, D_t)$ と記す）．(3) ブースターは，分布を D_{t+1} に更新する．T ラウンド終了後，ブースターはこれまで得られた弱仮説 $h_1, \ldots, h_T$ を統合して仮説 F_T として出力する．

本稿では，仮説の性能を測る自然な尺度として，情報量の概念を用いる．そのために，エントロピー関数として次の条件を満たす任意の関数 $G : \Delta_N \to [0,1]$ を考える．ただし，Δ_N は N 次元確率ベクトル，すなわち，$\Delta_N = \left\{(q_0, \ldots, q_{N-1}) \in [0,1]^N \;\middle|\; \sum_j q_j = 1\right\}$ である．

1. $G(q_0, \ldots, q_{N-1}) \in [0,1]$
2. $G(q_0, \ldots, q_{N-1})$ は上に凸
3. $\exists i, q_i = 1 \Leftrightarrow G(q_0, \ldots, q_{N-1}) = 0$

† （株）NTT データ
‡ 東北大学大学院情報科学研究科

シャノンのエントロピー関数や，InfoBoost[1] で用いられているエントロピー関数 $G(q_0, q_1) = 2\sqrt{q_0 q_1}$ はこれらの条件を満たす．

各 t ラウンドにおいて，(X, Y) を確率 $D_t(i)$ で値 (x_i, y_i) を取る確率変数とみなす．

$$q_j = \Pr_{D_t}[Y = j], \quad q_{j|z} = \Pr_{D_t}[Y = j \mid h_t(X) = z]$$

とすると，分布 D_t の下でのラベル Y に関する不確実さ（エントロピー）は

$$H_{D_t}(Y) = G(q_0, \ldots, q_{N-1}),$$

仮説 h_t を受け取った後に残る Y の不確実さ（条件つきエントロピー）は

$$H_{D_t}(Y|h_t) = \sum_{z=0}^{N-1} \Pr_{D_t}[h_t(X) = z] G(q_{0|z}, \ldots, q_{N-1|z})$$

で表すことができる．WL が弱学習者であるとは，ある定数 $\gamma > 0$ が存在して，任意の分布 D_t が与えられたとき

$$H_{D_t}(Y|h_t) \le (1-\gamma) H_{D_t}(Y) \tag{1}$$

を満たす仮説 $h_t = \mathrm{WL}(S, D_t)$ を返すこととする．これは，仮説 h_t が与えられた分布 D_t の下で何らかの情報を保持していることを意味する．

3. 2 値分類問題への還元の困難さ

多値分類問題に対して，従来 2 値分類問題に還元してブースティングを行う手法が知られている．まず，その手順を紹介する．

1. 与えられたサンプル $S \subseteq X \times Y$ を
$$\tilde{S} = \bigcup_{\ell=0}^{N-1} \left\{ \left\langle (x_i, \ell), [[y_i = \ell]] \right\rangle \mid (x_i, y_i) \in S \right\}$$
により 2 値のサンプル $\tilde{S} \subseteq (X \times Y) \times \tilde{Y}$ に変換する．ここで，$\tilde{Y} = \{0, 1\}$ で，$[[\pi]]$ は命題 π が真のとき 1，偽のとき 0 を値として取る関数である．
2. 2 値分類問題に対する弱学習者 $\widetilde{\mathrm{WL}}$ の存在を仮定し，2 値のブースティングアルゴリズム（例えば InfoBoost）を適用する．即ち，ブースタは各ラウンド t において $\tilde{S}$ 上の分布 $\tilde{D}_t$ を $\widetilde{\mathrm{WL}}$ に与え，2 値の弱仮説 $\tilde{h}_t : X \times Y \to \tilde{Y}$ を得る．
3. T ラウンド終了後，ブースタは 2 値の統合仮説 $\tilde{F}_T$ を出力する．
4. $\tilde{F}_T$ を変換し，多値分類問題に対する最終仮説 F_T を出力する．

しかし，この手法は多値分類問題に対する弱学習者を用いていないので，厳密な意味で還元を実現しているとは言えない．そこで，S に対する弱学習者 WL の存在を仮定し，それを用いて $\tilde{S}$ に対する弱学習者 $\widetilde{\mathrm{WL}}$ を模倣することによる還元について考える．即ち，手順 2 を以下のように変更する．

2' 各ラウンド t において，ブースタが作った $\tilde{S}$ 上の分布 $\tilde{D}_t$ を (a)S 上の分布 D_t に変換 し，$h_t = WL(S, D_t)$ により多値の弱仮説 h_t を得る．そして，h_t を (b)2 値の弱仮説 $\tilde{h}_t$ に変換 してブースタに返す．

明らかに，還元の成否は (a) と (b) で行なう変換に強く依存する．まず，$\tilde{D}_t$ から D_t への変換として，最も自然と思われる $D_t(i) = \sum_{\ell=0}^{N-1} \tilde{D}_t(i, \ell)$ を考える．次に，h_tから$\tilde{h}_t$ への変換として，$g_t : \{0, \ldots, N-1\} \times \{0, \ldots, N-1\} \to \{0, 1\}$ なる任意の関数 g_t を用いて

$$\tilde{h}_t(x_i, \ell) = g_t\left(h_t(x_i), \ell\right)$$

とするものを考える．これは，h_t を最大限利用することができる最も強力なものである．このとき，次の定理を得る．

定理 1 $\exists \tilde{D}_t, \exists h_t,\ H_{D_t}(Y|h_t) \ll 1, \quad \tilde{H}_{\tilde{D}_t}(\tilde{Y}|\tilde{h}_t) = 1.$

これは，WL が弱学習者の条件を満たしたとしても，この還元によって模倣される $\widetilde{WL}$ が弱学習者の条件を満たさず，ブースティングに失敗することがあり得ることを示す．

4. InfoBoost の多値への拡張

ここでは，情報量の概念を用いて提案された InfoBoost の多値分類問題への自然な拡張を試みる．まず，アルゴリズムを行列表現を用いて多値分類問題を扱える形式に書き直し，次にその性能を評価する．以下にアルゴリズムの概要を与える．整数 $\ell \in \{0, \ldots, N-1\}$ に対し，$vec(\ell)$ を第 $\ell+1$ 成分が 1 で残りの成分が 0 の N 次元列ベクトルとする．以下のアルゴリズムにおいて，$N = 2$ とすると従来の InfoBoost が得られる．

Input: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq X \times Y$
Initialize $D_1(i) = 1/m$;
For $t = 1$ to T do
　$h_t = WL(S, D_t)$;
　Let $\vec{h}_t(x_i) = vec\big(h_t(x_i)\big)$, $\vec{y}_i = vec(y_i)$;
　Choose an $N \times N$ matrix A_t;
　Update: $D_{t+1}(i) = D_t(i) \exp\big(-\vec{y}_i' A_t \vec{h}_t(x_i)\big)/Z_t$;
　　(Z_t is for normalization)
Let $\vec{f}_T(x) = \big(f_T^0(x) \ldots f_T^{N-1}(x)\big)' = \sum_t A_t \vec{h}_t(x)$;
Output: $C_T(x) = \{j \mid f_T^j(x) > 0\}$

ここで，以下のような関数 G を定義する．

$$G(q_0, \ldots, q_{N-1}) = N \left(\prod_{j=0}^{N-1} q_j \right)^{\frac{1}{N}}.$$

関数 G は，$N = 2$ のときはエントロピー関数の定義を満たすが，$N \geq 3$ のときは満たさない．しかし，この G を用いることにより，統合仮説 C_T の性能を以下のように評価することができる．U を S 上の一様分布とする．

定理 2 $|C_T(x)| \leq N - 1$,
$\Pr_U\left[Y \notin C_T(X)\right] \leq \prod_t H_{D_t}(Y|h_t)$,
$H_U(Y|h_1, \ldots, h_T) \leq \prod_t H_{D_t}(Y|h_t)$.

つまり，統合仮説は，与えられた事例 x のラベルの候補（$C_T(x)$）を $N-1$ 種類以下に絞り，h_t が関数 G の下で弱仮説の条件 (1) を満たすならば，真のラベルは高い確率でその候補に存在する．このように，一般化された InfoBoost は，いわば消去法によって正しくないラベル候補を除外するという点で，精度の向上を達成しているとみなすことができる．しかし，関数 G の性質（$\exists q_i = 0 \Leftrightarrow G = 0$）より，一般に $|C_T| \leq N - 2$ とすることができないことを示すことができる．このような精度向上の限界があるにも関わらず，定理はまた $H_U(Y|h_1, \ldots, h_T) \to 0$ となることを示している．このことからも，この G は $N \geq 3$ のときはエントロピー関数として不適格であることが分かる．

5. 乗算型分布更新規則を持つブースターの限界

2 章で示した 3 つの条件を満たすエントロピー関数 G に対し，以下のような乗算型の分布更新規則を持つ任意のブースターを考える．

$$D_{t+1}(i) = \frac{1}{Z_t} D_t(i) C_{t,i},$$

ただし，$C_{t,i}$ は t と i で定まる値である．InfoBoost を含め，これまで提案された多くのアルゴリズムがこのような乗算型の分布更新規則を持つ．さて，統合仮説を F_T の訓練誤差を 0 とする（$\Pr_U(F_T(X) \neq Y) \to 0$）ためには以下の 2 つの必要条件を満たさなくてはいけないことが知られている．

C1. $H_U(Y|h_1, \ldots, h_T) \to 0$

C2. $H_{D_{t+1}}(Y|h_t) = H_{D_{t+1}}(Y)$.（分布 D_{t+1} に対する条件）

しかし，次の定理はこれについて否定的な結果を示す．

定理 3 弱仮説の条件 (1) を満たす仮説 h_t を返すにもかかわらず，必要条件 C1 と C2 を同時に満たさない弱学習者が存在する．

すなわち，乗算型の分布更新規則を持つブースターでは，弱学習者 WL から訓練誤差を 0 に近づける統合仮説を構成することは不可能なのである．

参考文献

[1] Javed A. Aslam. Improving Algorithms for Boosting. *13th COLT*, 200-207, 2000.

[2] E. Takimoto and A. Maruoka. Top-down decision tree learning as information based boosting. *Theoretical Computer Science*, 292(2):447–464, 2003.