LA-5

決定ダイアグラムに基づくブースティング
A Boosting Algorithm Based on Decision Diagrams

松本 淳* 瀧本 英二* 丸岡 章*
Atsushi Matsumoto Eiji Takimoto Akira Maruoka

1 はじめに

ブースティングは，予測精度が低い複数の仮説（ベース仮説）を統合することによって予測精度の高い仮説（最終仮説）を得る一般的な技術である．ベース仮説がランダムな予測を行なうものでない限り，最終仮説の性能は向上することが保証されている．

CART や C4.5 のようなトップダウン型の決定木生成アルゴリズムは，内部頂点にベース仮説を割り当てることによって決定木を構成し，これを最終仮説とするブースティングアルゴリズムとみなすことができる．しかし，性能の良い決定木を得ようとすると木の大きさが指数的になり，過適合を起こして予測精度が落ちてしまうという問題があった．Mansour らは，決定木の予測性能を落とさずに，いくつかの頂点を融合 (マージ) することによって最終仮説の大きさを多項式程度に抑えるアルゴリズムを提案し，この問題を解決した [1]．ここで最終仮説の形は決定ダイアグラムとなる．

しかし，Mansour らの与えたマージ基準（どの頂点をマージすべきかを決める基準）は，2 値分類問題についてのみ適用可能であった．本稿では，多値分類問題に対する一般的なマージ基準を与え，これを 3 値分類問題に適用したときの最終仮説の大きさを評価する．

2 決定木に基づくブースティング

N 値分類問題について考える．事例空間を X，ラベル空間を $\{1,\dots,N\}$ とする．簡単のため，ベース仮説は 2 値分類関数のあるクラス $H \subseteq \{h : X \to \{0,1\}\}$ から選ばれるものとする．また，決定木 T に対し，T の葉の集合を $L(T)$ で表す．

決定木に基づくブースティングアルゴリズムDT の概要を以下に述べる．アルゴリズムは，入力として，事例とラベルのいくつかのペアからなるサンプル $S = \{(x_1,y_1),\dots,(x_m,y_m)\} \subseteq X \times \{1,\dots,N\}$ を受け取り，事例とラベル間の関係を表す決定木を最終仮説として出力する．

アルゴリズム DT
入力：サンプル S，深さ D
出力：決定木 T_D
T_0 をただ 1 つの葉からなる決定木とする．
for $d=1$ **to** D **do**
　forall $\ell \in L(T_{d-1})$ **do**
　　ベース仮説 $h_\ell \in H$ を選ぶ．
　　葉 ℓ を h_ℓ でラベルづけされた内部頂点で置き換え，
　　その 2 つの子を新しい葉 ℓ_0，ℓ_1 として加える．
　end
　こうして得られた深さ d の決定木を T_d とする
end.

*東北大学大学院情報科学研究科，GSIS Tohoku University

葉 ℓ に到達する事例からなる部分サンプルを S_ℓ，

$$q_{i|\ell} = \left|\{(x_j,y_j) \in S_\ell \mid y = i\}\right| \,/\, |S_\ell|$$

とし，葉 ℓ の様相をベクトル $\vec{q}_\ell = (q_{1|\ell},\dots,q_{N|\ell})$ と定義する．ラベルの値を取る確率変数 Y を考える．S_ℓ に基づく Y のエントロピーは，

$$H_{S_\ell}(Y) = G(\vec{q}_\ell) = -\frac{1}{\log N}\sum_i^N q_{i|\ell}\log q_{i|\ell},$$

また，仮説 h_ℓ が与えられたという条件のもとでの S_ℓ に基づく Y のエントロピーは，

$$H_{S_\ell}(Y|h_\ell) = \frac{|S_{\ell_0}|}{|S_\ell|}G(\vec{q}_{\ell_0}) + \frac{|S_{\ell_1}|}{|S_\ell|}G(\vec{q}_{\ell_1})$$

で与えられる．一方，決定木 T_d は事例 x を x が到達する葉に写像する関数 $T_d : X \to L(T_d)$ とみなせることに注意すると，決定木 T_d が与えられたときの S に基づく Y のエントロピーは，

$$H_S(Y|T_d) = \sum_{\ell\in L(T_d)} \frac{|S_\ell|}{|S|}G(\vec{q}_\ell)$$

と表される．目標は，最終仮説 T_D の与える条件つきエントロピー $H_S(Y|T_D)$ を小さくすることである．

以下では，ベース仮説 h_ℓ の与える情報量は 0 ではない，すなわちある定数 $\gamma > 0$ に対して

$$H_{S_\ell}(Y|h_\ell) \le (1-\gamma)H_{S_\ell}(Y) \qquad (1)$$

が成立すると仮定する．このとき，

$$\begin{aligned} H_S(Y|T_d) &\le (1-\gamma)H_S(Y|T_{d-1}) \\ &\le (1-\gamma)^d \end{aligned}$$

となることが示されている [2]．これより，深さ $D = O\left(\frac{1}{\gamma}\ln\frac{1}{\varepsilon}\right)$ で $H_S(Y|T_D) \le \varepsilon$ を達成する．一方，決定木 T_D の大きさは $|T_D| = O\left(2^D\right) = \left(\frac{1}{\varepsilon}\right)^{O(1/\gamma)}$，すなわち，$1/\gamma$ に関して指数的となってしまう．

3 決定ダイアグラムに基づくブースティング

決定ダイアグラムに基づくブースティングアルゴリズムDD は，各ラウンドにおいて，まずDT と同様に各葉にベース仮説を割り当てることによって葉の数を 2 倍にし，その後，いくつかの葉をマージすることによって葉の数をある一定数に抑えるものである．DD の概要を以下に示す．

アルゴリズム DD
入力 : サンプル S, 深さ D
出力 : 決定ダイアグラム T_D
T_0 をただ 1 つの葉からなる決定ダイアグラムとする.
for $d = 1$ **to** D **do**
1. T_{d-1} の各葉 ℓ をベース仮説 h_ℓ でラベルづけされた内部頂点で置き換え, 決定ダイアグラム T'_d を得る.
2. $L(T'_d)$ のいくつかの葉をマージし, 深さ d の決定ダイアグラム T_d を得る.

end.

葉をマージする条件について解析を行う. ベース仮説 h_ℓ は条件 (1) を満たすので,

$$H_S(Y|T'_d) \leq (1-\gamma)H_S(Y|T_{d-1})$$

が成り立つ. よって, マージ後の決定ダイアグラム T_d について, $\left|L(T_d)\right| \leq k$, かつ, エントロピーの上昇を

$$H_S(Y|T_d) \leq \left(1+\frac{\gamma}{2}\right) H_S(Y|T'_d) \tag{2}$$

のように抑えることができれば,

$$\begin{aligned} H_S(Y|T_d) &\leq (1-\gamma)\left(1+\frac{\gamma}{2}\right) H_S(Y|T_{d-1}) \\ &\leq \left(1-\frac{\gamma}{2}\right) H_S(Y|T_{d-1}) \\ &\leq \left(1-\frac{\gamma}{2}\right)^d \end{aligned}$$

より, 深さ $D = O(\frac{1}{\gamma}\ln\frac{1}{\varepsilon})$ で, $H_S(Y|T_D) \leq \varepsilon$ となるサイズ $O(kD)$ の決定ダイアグラム T_D を得ることができる.

2 つの葉 ℓ と ℓ' の様相が似ている ($\vec{q_\ell} \approx \vec{q_{\ell'}}$) ならば, それらをマージしてもエントロピーはそれほど上昇しない. この考えに基づき, 確率空間 $P^N = \left\{(q_1,\ldots,q_N) \in [0,1]^N \mid \sum_i^N q_i = 1\right\}$ の次のような分割 $\Delta = \{\Delta_1,\ldots,\Delta_k\}$ を考える.

定義 分割 Δ が $(\boldsymbol{\delta},\boldsymbol{\lambda})$ ネットであるとは,

1. 任意の Δ_i, 任意の $\vec{q_1},\vec{q_2} \in \Delta_i$ に対して,
$$G(\vec{q_1}) \leq \max\{\delta, (1+\lambda)G(\vec{q_2})\},$$
2. 各 Δ_i は凸集合

を満たすことである.

このとき, 以下の定理が成り立つ.

定理 $\delta = \gamma H_S(Y|T_{d-1})/6$, $\lambda = \gamma/3$ とする. 分割 Δ を (δ,λ) ネットとし, アルゴリズムDD の 2 のマージ基準として,「$\vec{q}_{\ell_1},\ldots,\vec{q}_{\ell_m} \in \Delta_i$ ならば, $\ell_1,\ldots,\ell_m$ をマージする」を採用すると, 式 (2) が成立する.

これより, δ, λ が与えられたとき, 分割数 $|\Delta|$ の小さい (δ,λ) ネットを構成することが望ましいことがわかる.

4 $N=3$ における (δ,λ) ネットの構成

$N=3$ のときのエントロピー関数 $G(\vec{q})$ を図 1 に示す. 以下では, 図 1 の影で示した空間 P^N の 1/6 の部分の分割 Δ を与える. 残りの領域についても同様に分割する. 図 2 のように点 $\vec{q}_{i,j}$ ($i=1,2,\ldots,I$, $j=1,2,\ldots$) を以下の関係を満たすようにとる.

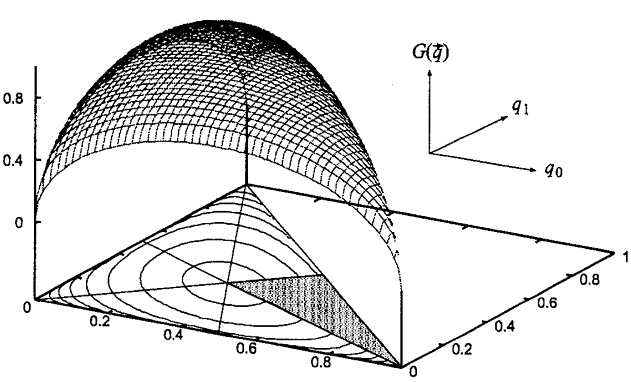

図 1: $N=3$ におけるエントロピー関数 $G(\vec{q})$

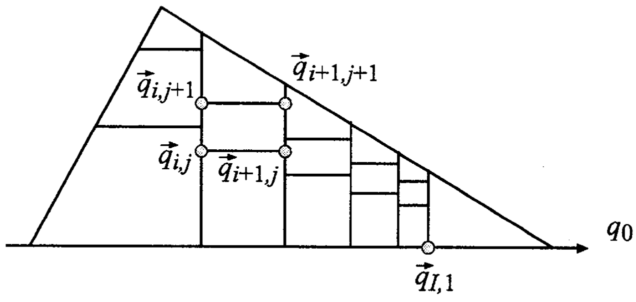

図 2: 分割 Δ

$$\begin{aligned} G(\vec{q}_{i,j}) &= \sqrt{1+\lambda}G(\vec{q}_{i+1,j}) \\ &= \sqrt{1+\lambda}G(\vec{q}_{i,j+1}) \\ &= (1+\lambda)G(\vec{q}_{i+1,j+1}) \end{aligned}$$

ただし, I は $G(\vec{q}_{I,1}) \leq \delta$ を満たす最小の値とする. 各 $1 \leq i < I$, j に対して, 4 つの頂点 $\vec{q}_{i,j}$, $\vec{q}_{i+1,j}$, $\vec{q}_{i,j+1}$, $\vec{q}_{i+1,j+1}$ で囲まれた 4 角形を $\Delta_{i,j}$ とする. また, 点 $\vec{q}_{I,1}$ を通り q_0 に直交する線分を底辺とし, 点 q_0 を頂点とする 3 角形を Δ_I とする. そして, 求める分割を $\Delta = \Delta_I \cup \{\Delta_{i,j} \mid 1 \leq i < I, j \geq 1\}$ とする. このとき分割 Δ は (δ,λ) ネットとなり, 分割数は以下のようになる.

$$|\Delta| \leq \left(1+2\left\lceil\frac{\ln(1/\delta)}{\ln(1+\lambda)}\right\rceil\right)\left(1+2\left\lceil\frac{\ln 1.5}{\ln(1+\lambda)}\right\rceil\right).$$

これより, $|T_D| = O\left(\frac{1}{\gamma^3}\left(\ln\frac{1}{\varepsilon}\right)^2\right)$ で, $H_S(Y|T_D) \leq \varepsilon$ を達成できることがわかる.

参考文献

[1] Y. Mansour and D. McAllester. Boosting using Branching Programs. In *13th COLT*, 220–224, 2000.

[2] E. Takimoto and A. Maruoka. Top-down decision tree learning as information based Boosting. To appear in Theoretical Computer Science.