社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

決定木モデルにおける予測アルゴリズムについて

須子 統太† 野村 亮† 松嶋 敏泰† 平澤 茂一†

† 早稲田大学理工学部 経営システム工学科
〒 169-8555 東京都新宿区大久保 3-4-1
E-mail: †suko@matsu.mgmt.waseda.ac.jp

あらまし 従来，決定木モデルを用いた予測（分類，判別）を行う場合，決定木生成アルゴリズムが用いられてきた．これらは，データが与えられたもとでのモデル選択アルゴリズムとみなすことができる．つまり，従来はデータが与えられたもとでモデルを一つ選択し，選択したモデルを用いて予測を行っているといえる．よって，予測誤り率に対して理論的な評価を行うのは非常に困難であった．そこで本研究では，平均予測誤り率を最小にする予測アルゴリズムを示す．そのためにまず，従来の決定木モデルをパラメトリックな確率モデルとして再定式化する．そのモデルを用いて，ベイズ決定理論にもとづく最適な予測アルゴリズムを示す．更に，モデルクラスを制約することで効率的に予測分布を計算するアルゴリズムについて述べる．

キーワード 階層モデル，ベイズ決定理論，決定木

Prediction Algorithm for Decision Tree Model

Tota SUKO†, Ryo NOMURA†, Toshiyasu MATSUSHIMA†, and Shigeichi HIRASAWA†

† School of Science and Engineering, Waseda University
3-4-1 Okubo Shinjyuku-ku, Tokyo 169-8555, Japan.
E-mail: †suko@matsu.mgmt.waseda.ac.jp

Abstract Conventionally, decision tree generation algorithm has been used when performing prediction using the decision tree model. It can be considered that these are the model selection algorithm in the basis to which data was given. And, It predicts using the model chosen by the basis to which data was given. Therefore, it was very difficult to perform theoretical evaluation to the rate of a prediction error. In this work, we shows the prediction algorithm which makes the rate of an average prediction error the minimum. First, we re-formulize a decision tree model as a parametric stochastic model. The optimal prediction algorithm based on Bayes decision theory is shown using the model. Furthermore, the algorithm which calculates a prediction distribution efficiently by restraining a model class is described.

Key words hierarchical model, Bayes decision theory, decision tree

1. はじめに

離散の属性ベクトル $x=(a_1,a_2,...,a_k)\in\mathcal{X}$ を観測して，それが属するカテゴリ $y\in\mathcal{Y}$ (注1) を予測（分類，判別）する問題を考える．この時，予測誤り率を最小にする決定は，

$$\hat{y}=\arg\max_{y} P(y|x), \tag{1}$$

で与えられることは良く知られている．つまり，x が与えられたもとでの y の条件付確率 $P(y|x)$ が既知の場合にはこの問題は解決されていると言える．

一方，$P(y|x)$ が未知である場合については，従来からパターン認識やデータマイニングなどの分野で様々な研究が行われている．この $P(y|x)$ の確率構造を表現するモデルの一つとして決定木モデルがあげられる．これは木を用いて確率構造を表現するモデルで，見た目にも分りやすく実用的にも非常に扱いやすいモデルである．

従来，決定木モデルにおいて $P(y|x)$ が未知である場合の予測法については，与えられたデータから $P(y|x)$ を推定する方法が考えられてきた．これらは決定木生成アルゴリズムと呼ばれ，CART，ID3，C4.5 など数多くのアルゴリズムが研究されている．[1] [2] [3] [4] この決定木生成アルゴリズムは，与えられ

(注1)：本研究では，以下特に断らないかぎり y を離散値として扱うが，連続値としても同様の議論が可能である．

たデータのもとで，考えうる全ての決定木モデルの中から一つの決定木モデルの選択を行う．つまり，データが与えれたもとでのモデル選択アルゴリズムであると言える．しかし，決定木モデルにおけるモデル選択は非常に困難な問題であり，一般に予測誤り率について理論的に評価するのは難しい．そのため，どの決定木生成アルゴリズムを用いた場合，予測誤り率を最小にするのか一概に比較することが出来ない．

そこで本研究では，$P(y|x)$ が未知である場合の決定木モデルにおける予測に対して，平均予測誤りを最小にするアルゴリズムを示す．そのためにまず，決定木モデルをパラメトリックな確率モデルとして再定式化する．そのもとで，再定式化した確率モデルに対しベイズ決定理論に基づく最適な予測アルゴリズムを示す．更に，モデルクラスを制約した場合の効率的な予測アルゴリズムについて述べる．

2. 従来研究

2.1 決定木

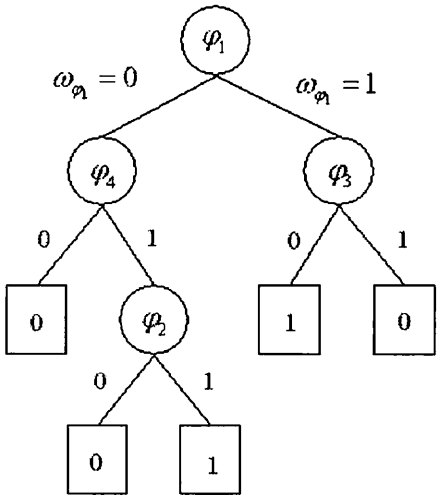

図 1　決定木モデルの例

以降簡単のため $x \in \{0,1\}^k$，$y=\{0,1\}$ とする．今，x に対する質問のインデクスを ψ_j，$j=1,2,....J$ とし，$\omega_{\psi_j}(x) \in \{0,1\}$ は質問 ψ_j に対し x が真か偽を返す関数とする．これは例えば，

$$\omega_{\psi_j}(x)=\begin{cases} 0 & a_1=0 \\ 1 & a_1=1 \end{cases}, \tag{2}$$

のように x の 1 番目の要素が 0 であるか 1 であるかの質問に対して，真か偽かを返すような関数である．他にも，

$$\omega_{\psi_j}(x)=\begin{cases} 0 & (a_1=1)\cap(a_3=0) \\ 1 & otherwise \end{cases}, \tag{3}$$

のように複数の要素に対しての質問も考えられる．

決定木モデルは図 1 のような木で表現される．各ノードには質問 ψ，枝には質問に対する回答が割り付けられ，葉ノードにはカテゴリ y が割り付けられる．これは各ノードに割り付けられた質問 ψ によって，根ノードから順に属性ベクトルの空間 $\mathcal{X}$ を分割していると捉えることができる．葉ノードは最終的に分割された $\mathcal{X}$ の部分空間に対応している．今，$\mathcal{X}$ の部分空間を s とすると，葉ノードには $\hat{y}=\arg\max_y P(y|s)$ となるカテゴリを割り付ける．葉ノードにカテゴリそのものを割り付けるのではなく，カテゴリの分布 $P(y|s)$ を割り付けた場合も本質的にその意味は変わらない．そのため，ある一つ決定木モデルはデータ構造を表す一つの確率モデルを表現していると考えられる．

2.2 決定木生成アルゴリズム

x と y の n 個の組がデータとして与えられた時，決定木生成の基本的なアルゴリズムを以下に示す．

【決定木生成アルゴリズム】

step-1. ノード t において $cost(\psi_j)$ を最小にする ψ_j をノードに割り付ける．

step-2. t から枝を伸ばし子ノードをつくる

step-3. 停止基準を満たす場合は子ノードにカテゴリをつけ step-4 へ，それ以外はどちらかの子ノードに移り step-1 へ．

step-4. 子ノードにカテゴリを割り付ける．全ての葉ノードにカテゴリが割り付けたらアルゴリズムを終了．それ以外はカテゴリの割り付けられていない葉ノードへ移り step-1 へ．

代表的な $cost(\psi_j)$ としては以下の例があげられる．[1] Ψ_t をノード t から根ノードまでの質問の集合，t_0，t_1 を t の子ノードとする．

$$cost(\psi_j)= g(\omega_{\psi_j}=0|\Psi_t)H(Y|\Psi_{t_0})+g(\omega_{\psi_j}=1|\Psi_t)H(Y|\Psi_{t_1}). \tag{4}$$

但し，$g(\omega_{\psi_j}=0|\Psi_t)$，$g(\omega_{\psi_j}=1|\Psi_t)$ をそれぞれノード t における $\omega_{\psi_j}=0$，$\omega_{\psi_j}=1$ となるデータの割合とし，

$$H(Y|\Psi_t)=-\sum_y g(y|\Psi_t)\log g(y|\Psi_t), \tag{5}$$

とする．$g(y|\Psi_t)$ はノード t においてカテゴリが y であるデータの割合である．

これは，質問により分割された $\mathcal{X}$ の部分空間に含まれるカテゴリが，なるべく同一のカテゴリで満たされるように設定されている．その他の従来用いられている $cost$ についても，これと同様の考えに基づき設定されている．

$cost$ が最小になるまで木を完全に成長させ続けると，一般に，データが過剰に適合するという問題が生じる．そのため，交差点検証法や MDL (minimum description length) 基準などを用いた様々な停止基準が考えられている．[3] また，一度木を生成した後に，冗長な部分の枝の刈り込みを行うことも考えられる．

上記のアルゴリズムの場合，根ノードからトップダウン式に木を生成していくアルゴリズムであるので，大域的な意味で $cost$ を最小化するアルゴリズムにはなっていない．そのため，計算量は増大するものの大域的な $cost$ を設定し，それを最小

化するアルゴリズムについても研究がなされている.[5]

これら決定木生成アルゴリズムは，考えうる全ての決定木モデルの中からデータを上手く説明するモデルを一つ選択するアルゴリズムであると捉えることができる．そのため，予測誤り率を評価基準としたとき，どのアルゴリズムを用いるべきか理論的な評価をするのは困難である．

3. モデル化と最適予測アルゴリズム

従来の決定木生成アルゴリズムは，モデル選択アルゴリズムであった．しかし，データが与えられたもとで，予測そのものを目的とした場合，必ずしもモデルを選択する必要は無い．そこで，本研究では平均予測誤り率を最小にする予測アルゴリズムを示す．

3.1 データからの予測

まず，予測問題を以下で示す．

x_i, y_i をそれぞれ i 番目の x と y とし，その n 個の集合を $x^n = \{x_1, x_2, ..., x_n\}$, $y^n = \{y_1, y_2, ..., y_n\}$ とする．また，i 番目の x と y の組を $z_i = (x_i, y_i)$ とし，その n 個の集合を $z^n = \{z_1, z_2, ..., z_n\}$ とする．今，x と y の n 個の組 z^n が得られたとする．このとき x_{n+1} が与えられたもとで x_{n+1} に対応するカテゴリ y_{n+1} を逐次的に予測する問題を考える．

3.2 モデル化

決定木モデルは質問 ψ をノードとした木で表現されていた．これを確率モデルとして捉えた場合，同じ確率構造をもつ決定木モデルが複数存在することとなり扱いにくい．そこで，新たに決定木モデルをパラメトリックな確率モデルとして再定式化する．決定木モデルでは前述の通り，$\mathcal{X}$ の空間を質問によって分割し，分割された部分空間に対しカテゴリ y を割り当てていた．そこで，本研究では $\mathcal{X}$ の部分空間に対し，カテゴリ y の発生する確率分布を割り当てることでモデル化を行う．

カテゴリ y の出現確率を $P(y|x, \theta_m, m)$ とする．この時，$m \in M$ はモデル，$\theta_m \in \Theta_m$ はモデル m におけるパラメータである．次に，$\mathcal{X}$ の部分集合を状態 s と定義する．全ての s の集合を S，状態 s におけるパラメータを $\theta_s \in \Theta_s$ とする．更に，モデル m における状態の集合を $S(m)$ とした時，モデル m を以下で定義する．

$$m = \{(s, \theta_s) | s \subseteqq \mathcal{X}\}. \tag{6}$$

但し，$\forall s, s' \in S(m)$, $s \neq s'$, $s \cap s' = \phi$, $s = \phi$, $\bigcup_{s \in S(m)} s = \mathcal{X}$ とする．この時，

$$\theta_m = \{\theta_s | s \in S(m)\}, \tag{7}$$

$$P(y|x, \theta_m, m) = P(y|x, \theta_{s(m,x)}, s(m, x)), \tag{8}$$

となる．但し，$s(m, x)$ は $S(m)$ の中で $x \in s$ となる s とする．

また，定義を満たすモデル m の全ての集合をモデルクラス M と定義する．

3.3 ベイズ決定理論に基づく最適予測

前節で定義したモデル及びモデルクラスに対して，ベイズ基準に基づく最適な予測法（以下ベイズ最適な予測と呼ぶ）を示す．

まず，予測に対する損失を $Loss$ と定義する．今，離散のカテゴリに対する平均予測誤り率を最小にしたいので，損失を以下の 0-1 損失で与える．但し，$\hat{y}$ は y の予測値とする．

$$Loss_1 = \begin{cases} 0 & y_{n+1} = \hat{y}_{n+1} \\ 1 & y_{n+1} \neq \hat{y}_{n+1} \end{cases}. \tag{9}$$

もし，y が量的データである場合には，次式の二乗誤差損失などで損失を与えることもできる．

$$Loss_2 = (y_{n+1} - \hat{y}_{n+1})^2. \tag{10}$$

その他にも，y を予測するのではなく y の分布 $P(y|x)$ を予測したい場合には，次式の対数損失などを与えることも考えられる．但し，$\hat{P}(y|x)$ は $P(y|x)$ の予測値とする．

$$Loss_3 = \log P(y_{n+1}|x_{n+1}) - \log \hat{P}(y_{n+1}|x_{n+1}). \tag{11}$$

次に，データによる損失の期待値である危険関数を以下の式で定義する．

$$Risk = \sum_{y^n} \sum_{x^n} Loss \times P(y^n|x^n, \theta_m, m) P(x^n|\nu). \tag{12}$$

但し，$P(x^n|\nu)$ は x^n が従う確率分布，$\nu \in \mathcal{N}$ はその確率分布のパラメータとする．

更に，事前分布で期待値をとったベイズ危険関数を以下の式で定義する．

$$BR = \sum_M \int_{\mathcal{N}} \int_{\Theta_m} Risk P(m) P(\theta_m|m) P(\nu) d\theta_m d\nu. \tag{13}$$

$Loss_1$ を仮定した場合，これは平均予測誤り率となる．ベイズ最適な予測はこのベイズ危険関数を最小にする予測である．このとき，$Loss_1$ に対するベイズ最適な最適な予測は以下で求めることができる．

$$\hat{y}^*_{n+1} = \sum_Y y_{n+1} \sum_M \int_{\Theta_m} P(y_{n+1}|x_{n+1}, \theta_m, m) P(\theta_m|z^n) P(m|z^n) d\theta_m. \tag{14}$$

つまり，上式を計算することで平均予測誤り率を最小にする予測が可能であることがわかる．

また，$Loss_2$ に対する最適な予測は以下で与えられる．

$$\hat{y}^*_{n+1} = \arg\max_{y_{n+1}} \sum_M \int_{\Theta_m} P(y_{n+1}|x_{n+1}, \theta_m, m) P(\theta_m|z^n) P(m|z^n) d\theta_m. \tag{15}$$

同様に $Loss_3$ に対する最適な予測は以下で与えられる．

$$\hat{P}^*(y_{n+1}|x_{n+1}, z^n) = \sum_M \int_{\Theta_m} P(y_{n+1}|x_{n+1}, \theta_m, m) P(\theta_m|z^n) P(m|z^n) d\theta_m. \tag{16}$$

ここで，各損失に対する最適な予測に注目すると，全てに

(16) 式の右辺が入ってきていることが分る．つまり，どのような損失に対しても (16) 式を計算することによりベイズ最適な予測は可能となり，予測の本質はこの (16) 式の計算にあると言える．以降 (16) 式を予測分布と呼び，この予測分布の計算について述べる．

このように予測分布を計算することでベイズ最適な予測を行う方法は，従来から様々な分野で扱われてきた．例えば，ユニバーサル無歪データ圧縮の分野では，予測分布を符号化確率とすることで冗長度を最小にする符号化法が考案されている．[6][7][8][9]

3.4 計 算 量

予測分布の計算にかかる計算量について述べる．パラメータの事前分布に自然共役な事前分布を仮定すると，パラメータ空間における積分計算にかかる計算量はデータ数 n の線形オーダとすることができる．他方，モデルに対する期待値を求める計算に必要な計算量は $O(|M|)$ となり，モデル数が多い場合この部分の計算量が全体の計算量の主要項となる．前節で定義したモデルクラス M におけるモデルの数は，$\mathcal{X}$ の空間の任意の分割の数だけあるので，$\mathcal{X}$ の大きさ $|\mathcal{X}|$ に関して，

$$|M| = \sum_{v=1}^{|\mathcal{X}|} \sum_{w=0}^{v-1} \begin{pmatrix} v \\ w \end{pmatrix} \frac{(-1)^w (v-w)^{|\mathcal{X}|}}{v!}, \tag{17}$$

となる．そのため，$|\mathcal{X}|$ が大きくなる場合には予測分布の計算は非常に困難な問題となる．

4. 効 率 化

従来，ユニバーサル無歪データ圧縮の分野で，予測分布を効率的に計算するアルゴリズムが提案されている．[6][7][9] 決定木モデルにおける予測と，このユニバーサル無歪データ圧縮は本質的に等価な問題である．そこで本研究では，[9] で提案されたアルゴリズムを応用することで，決定木モデルにおける効率的な予測分布計算アルゴリズムを示す．

4.1 制約モデルクラス

前述の通り，予測分布の計算の計算はモデルクラス M 上では非常に難しい．そこで，制約を加えた新たなモデルクラス $\tilde{M}$ を考え，そのもとで効率的に予測分布を計算するアルゴリズムを示す．

今，質問が $\psi_1, \psi_2, ..., \psi_D$ の順番で必ず与えられるとする．真のモデルは質問 $\psi_1, ..., \psi_d$, $d = 1, ..., D$ によって分割された状態により表現可能であると仮定する．この時，質問 $\psi_1, ..., \psi_d$ に対する ω の系列を $\omega^{\psi_d} = \omega_{\psi_1}\omega_{\psi_2}...\omega_{\psi_d}$ とする．以降表記を簡単にするため ω^{ψ_d} を ω^d と書く．ω^d と x により一意に定まる状態を s_{ω^d} とすると，s_{ω^d} は質問 ψ_{d+1} によって，$s_{\omega^d 0}$ と $s_{\omega^d 1}$ の二つの状態に分割される．つまり，$\mathcal{X}$ は質問 $\psi_1, \psi_2, ..., \psi_D$ によって 2^D 個の状態に分割される．このとき，$s_{\omega^0} = \mathcal{X}$ である．

この時，モデルクラス $\tilde{M}$ は状態 s をノードとする木 $Tree(\tilde{M})$ で表現することができる．(図 2) $Tree(\tilde{M})$ はモデルクラスを表現している木で，単一のモデル表現する決定木とは全く違う表記である．$\tilde{M}$ におけるモデル m は，$Tree(\tilde{M})$ の完全部分木 $tree(m)$ の葉ノードに対応する s とそのパラメータ θ_s で定義される．

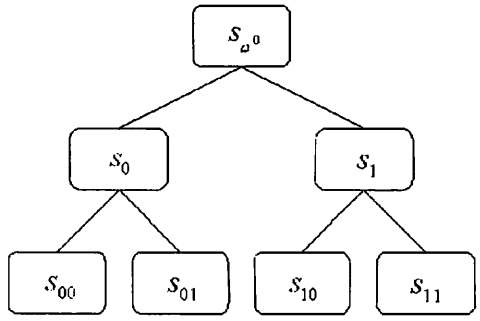

図 2 $Tree(\tilde{M})$ の例

4.2 効率的予測分布計算アルゴリズム

$\tilde{M}$ 対する，効率的な予測分布計算アルゴリズムを示す．

状態 s の事前分布を以下で定義する．

$$P(s) = \sum_{\{m | s \in S(m)\}} P(m). \tag{18}$$

このとき，$Tree(\tilde{M})$ における，ある葉ノードから根ノードまでの一本のパス上の s に対して，

$$\sum_{d=0}^{D} P(s_{\omega^d}) = 1. \tag{19}$$

が成立する．但し，s_{ω^0} は根ノードを表し，$s_{\omega^0} = \mathcal{X}$ である．

また，複数のモデルに同一の s が含まれてくるが，s のパラメータ θ_s の事前分布 $P(\theta_s)$ は同一の s であれば全て等しいと仮定する．

$P(s_{t_i}|z^n)$ を次式のように表現する．

$$P(s_{\omega^d}|z^n) = q(s_{\omega^d}|z^n) \prod_{l=0}^{d} (1 - q(s_{\omega^l}|z^n)). \tag{20}$$

【効率的予測分布計算アルゴリズム】

step-1. x_{n+1} を受け取る．

step-2. $x_{n+1} \in s$ を満たす $Tree(\tilde{M})$ 上の葉ノード s_{ω^D} を求める．

step-3. s_{ω^D} から根ノードへ向かい以下の再起計算を行う．

$$q(y_{n+1}|x_{n+1}, z^n, s_{\omega^d}) = \begin{cases} P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^d}) & d = D \\ (*) & \text{otherwise} \end{cases}. \tag{21}$$

$$(*) = q(s_{\omega^d}|z^n) P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^d}) + (1 - q(s_{\omega^d}|z^n)) q(y_{n+1}|x_{n+1}, z^n, s_{\omega^{d+1}}). \tag{22}$$

但し，

$$P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^d})$$

$$= \int_{\Theta_s} P(y_{n+1}|x_{n+1}, \theta_{s_{\omega^d}}, s_{\omega^d}) P(\theta_{s_{\omega^d}}|z^n) d\theta_{s_{\omega^d}}. \quad (23)$$

step-4. $q(y_{n+1}|x_{n+1}, z^n, s_{\omega^0})$ を予測分布として出力.

step-5. $s_{\omega^0}, ..., s_{\omega^D}$ に対して，$q(s_{\omega^d}|z^n)$ を以下の式で更新する.

$$q(s_{\omega^d}|z^{n+1}) = \frac{q(s_{\omega^d}|z^n) P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^d})}{(**)}. \quad (24)$$

$$(**) = q(s_{\omega^d}|z^n) P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^d}) + (1 - q(s_{\omega^d}|z^n)) q(y_{n+1}|x_{n+1}, z^n, s_{\omega^{d+1}}). \quad (25)$$

step-6. step-1 へ戻る.

定理 効率的予測分布計算アルゴリズムによる出力は，以下の式を満たす.

$$q(y_{n+1}|x_{n+1}, z^n, s_{\omega^0}) = \hat{P}^*(y_{n+1}|x_{n+1}, z^n). \quad (26)$$

証明は付録参照.

4.3 計 算 量

モデルクラス $\tilde{M}$ においてモデルの期待値計算にかかる計算量は，効率的アルゴリズムを用いない場合 $O(|\tilde{M}|)$ となる．$\tilde{M}$ には $Tree(\tilde{M})$ の完全部分木の数だけのモデルが含まれので，$|\tilde{M}|$ は質問の数 D によってきまる．今，$\tilde{M}_D$ を質問の数 D の時のモデルクラスとすると，

$$|M_D| = 1 + |M_{D-1}|^2 > 2^D, \quad (27)$$

となる．それに対し効率的アルゴリズムを用いた場合，アルゴリズムは $D+1$ 回の再起計算をおこなっているので，モデルの期待値計算にかかる計算量は $O(D)$ で済むことになる.

5. ま と め

本研究では決定木モデルにおける予測に対し，ベイズ基準のもとで最適な予測アルゴリズムを示した．予測分布を計算するには，モデルクラスに含まれる全てのモデルにおける重み付け和を計算することになり，モデルの数が多い場合この計算は非常に困難になる．それに対し，モデルクラスを制約することで，予測分布を効率的に計算するアルゴリズムを示した.

$\tilde{M}$ は強い制約のもとでのモデルクラスであり，効率的予測分布計算アルゴリズムを実問題へそのまま適用することは難しい．そのため，より広いモデルクラスに対しこのアルゴリズムを拡張する必要がある．これについては今後の課題としたい.

謝 辞

本研究を行うにあたり，数多くの御助言，御支援を賜りました，早稲田大学平澤研究室および松嶋研究室の各氏に感謝致します．なお，本研究の一部は文部省科学研究費基盤研究 (C)(No.15560338) の助成による.

文 献

[1] L. Breiman, J, Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[2] J. R. Quinlan, "Induction of decision trees," *Machine Learn.*, vol. 1, pp. 81-106, 1986.

[3] J. R. Quinlan and R. L. Rivest, "Inferring decision trees using minimum description length principle," *Inform. Comput.*, vol. 80, pp. 227-248, 1989.

[4] J. R. Quinlan, "Improved use of contimuous attributes in C4.5," *Journal of Artificial Intelligence*, 4: 77-90, 1996.

[5] Donald Geman and Bruno Jedynak," "Model-Based Classification Trees," *IEEE Trans. Inf. Theory*, vol. 47, No. 3, page 1075, 2001.

[6] Frans M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, "The Context-Tree Weighting Method: Basic Propreties," *IEEE Trans. Inf. Theory*, vol.41, No.3, page 653, 1995.

[7] Frans M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, "Context Weighting for General Finite-Context Sources," *IEEE Trans. Inf. Theory*, vol.42, No.5, page 1514, 1996.

[8] T. Matsushima, H. Inazumi and S. Hirasawa, "A Class of Distortionless Codes Designed by Bayes Decision Theory," *IEEE Trans. Inf. Theory*, vol.37, No.5, page 1288, 1991.

[9] T. Matsushima, and S. Hirasawa, "A bayes coding using context tree," *In Proc. Int. Symp. on. Inf. Theory*, page 386, 1994.

[10] 松嶋敏泰, "帰納・演繹推論と予測-決定理論による学習モデル-," *1998* 年情報論的学習理論ワークショップ,page 1, 1998.

[11] 韓太舜, 小林欣吾, 情報と符号化の数理, 培風館, 1999.

[12] R. O. Duda, P. H. Hart and D. J. Stork, *Pattern Classification*, John Wiley & Sons, 2001.

付 録

付録では定理の証明を行う.

はじめに，証明に用いるノーテーションをいくつか定義する．状態 s の子ノードの集合を S_c，S_c のうち $S(x_n)$ に含まれるノードを $s_{c'}$，$S(y_n)$ に含まれる s の子孫ノードの集合を $S_{p'}$ とする．$Tree\tilde{M}$ の完全部分木 $treem$ おける中間ノードを $s_i^I(m) \in S^I(m)$，但し $|S^I(m)| = I$，葉ノードを $s_l^L(m) \in S^L(m)$，但し $|S^L(m)| = L$ とする.

まず，(24) 式による $q(s|z^n)$ の更新により，モデルの事後分布 $P(m|y^n)$ が正しく更新されていることを示す．(20) 式の仮定から，

$$P(m) = \prod_{i=1}^{I} \{1 - q(s_i^I(m))\} \prod_{l=1}^{L} q(s_l^L(m)), \quad (28)$$

が成り立つ．今，(24) 式により更新された $q(s|z^n)$ を用いて，

$$P(m|z^n) = \prod_{i=1}^{I} \{1 - q(s_i^I(m)|z^n)\} \prod_{l=1}^{L} q(s_l^L(m)|z^n), \quad (29)$$

が成り立てば，事後確率が正しく更新されているといえる．まず，ベイズの定理より $P(m)$ の事後分布は，

$$P(m|z^n) = \frac{P^m(y^n|x^n, m) P(m)}{\sum_{m' \in M} P^m(y^n|x^n, m') P(m')}. \quad (30)$$

但し，

$$P^m(y^n|x^n, m) = \int_{\Theta_m} P(y^n|x^n, \theta_m, m) P(\theta_m|m) d\theta_m. \quad (31)$$

この時，(30) 式は (8)，(29) 式より，

$$P(m|z^n) = \frac{\prod_{s \in S(m)} P^S(y^{(n,s)}|x^{(n,s)}, s) \prod_{i=1}^{I} \{1 - q(s_i^I(m))\} \prod_{l=1}^{L} q(s_l^L(m))}{\sum_{m' \in M} P^m(y^n|m') P(m')}.$$

(32)

但し，$x^{(n,s)}$ は x^n の中で $s \in S(x)$ を満たす x の集合，$y^{(n,s)}$ は $x^{(n,s)}$ に対するカテゴリの集合とする．また，

$$P^S(y^{(n,s)}|x^{(n,s)}, s) = \int_{\Theta_s} P(y^{(n,s)}|^{(n,s)}, \theta_s, s)P(\theta_s|s), \quad (33)$$

とする．次に，(24) 式より，

$$\begin{aligned}
&q(s|z^n)\\
&= \{q(s)P^S(y_n|x_n, z^{(n-1,s)}, s)\}\\
&\quad /\{q(s|z^{n-1})P^S(y_n|x_n, z^{(n-1,s)}, s)\\
&\quad +\{1-q(s|z^{n-1})\}\textstyle\prod_{s'\in S_c} Q(y_n|x_n, z^{(n-1,s)}, s')\}\\
&= \frac{q(s)P^S(y^{(n,s)}|x^{(n,s)},s)}{q(s)P^S(y^{(n,s)}|x^{(n,s)},s)+\{1-q(s)\}\prod_{s'\in S_c} Q(y^{(n,s)}|x^{(n,s)},s')}\\
&= \{\textstyle\prod_{s\in S(m)} P^S(y^{(n,s)}|x^{(n,s)}, s)\\
&\quad \prod_{i=1}^{I}\{1-q(s_i^I(m))\}\prod_{l=1}^{L} q(s_l^L(m))\}\\
&\quad /\{\textstyle\sum_{s'\in S_{p'}}\prod_{s'\in S(m)} P^S(y^{(n,s)}|x^{(n,s)}, s')\\
&\quad \prod_{i=1}^{I}\{1-q(s_i^I(m))\}\prod_{l=1}^{L} q(s_l^L(m))\}. \qquad (34)
\end{aligned}$$

但し，

$$\prod_{s'\in S_c} Q(y_n|x_n, z^{(n-1,s')}, s') = q(y_n|x_n, z^{(n-1,s)}, s_{c'}) \quad (35)$$

$$Q(y^{(n,s)}|x^{(n,s)}, s) = \begin{cases} P^S(y^{(n,s)}|x^{(n,s)}, s) & (\text{s が葉ノード})\\ q(s)P^S(y^{(n,s)}|x^{(n,s)}, s) & \\ +\{1-q(s)\}\prod_{s'\in S_c} Q(y^{(n,s)}|x^{(n,s)}, s') & (\text{その他}) \end{cases}, \quad (36)$$

$$Q(y_n|x_n, z^{(n-1,s)}, s) = \begin{cases} P^S(y_n|x_n, z^{(n-1,s)}, s) & (\text{s が葉ノード})\\ q(s|z^{n-1})P^S(y_n|x_n, z^{(n-1,s)}, s) & (\text{その他}),\\ +\{1-q(s|z^{n-1})\}\prod_{s'\in S_c} Q(y_n|x_n, z^{(n-1,s)}, s') & \end{cases} \quad (37)$$

よって (34) 式を (29) 式右辺へ代入すると．

$$\prod_{i=1}^{I}\{1-q(s_i^I(m)|z^n)\}\prod_{l=1}^{L} q(s_l^L(m)|z^n) = \frac{\prod_{s\in S(m)} P^S(y^{(n,s)}|x^{(n,s)},s)\prod_{i=1}^{I}\{1-q(s_i^I(m))\}\prod_{l=1}^{L} q(s_l^L(m))}{\sum_{m'\in M} P^m(y^n|m')P(m')}. \quad (38)$$

よって，(32) 式より，(24) 式による $q(s|z^n)$ の更新が，$P(m|z^n)$ を正しく更新していることが示された．

次に，(21) の再起計算が (16) 式を計算していることを示す．

(21) 式を展開すると，

$$\begin{aligned}
&q(y_{n+1}|x_{n+1}, z^n, s_{\omega^0})\\
&\quad = P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^0})q(s_{\omega^0}|z^n)\\
&\quad +\cdots\\
&\quad +P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^d})q(s_{\omega^d}|x_{n+1}, z^n)\\
&\qquad \textstyle\prod_{d'<d}(1-q(s_{\omega^{d'}}|z^n))\\
&\quad +\cdots\\
&\quad +P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^D})\textstyle\prod_{d'<D}(1-q(s_{\omega^{d'}}|y^n)). \quad (39)
\end{aligned}$$

(18)，(20) 式より，

$$\begin{aligned}
&q(y_{n+1}|y^n, s_{\omega^0})\\
&= \textstyle\sum_{d=0}^{D} P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^d})q(s_{\omega^d}|z^n)\\
&\qquad \textstyle\prod_{d'<d}(1-q(s_{\omega^{d'}}|z^n))\\
&= \textstyle\sum_{d=0}^{D} P(s_{\omega^d}|z^n)P^S(y_{n+1}|x_{n+1}, z^n, s_{\omega^d})\\
&= \textstyle\sum_{s\in S(x_{n+1})}\\
&\quad \textstyle\int_{\Theta_s} P(y_{n+1}|x_{n+1}, z^n, \theta_s, s)P(\theta_s|z^n, s)d\theta_s P(s|z^n).\\
&= \textstyle\sum_{s\in S(x_{n+1})}\int_{\Theta_s} P(y_{n+1}|x_{n+1}, z^n, \theta_s, s)P(\theta_s|z^n, s)d\theta_s\\
&\qquad \textstyle\sum_{\{m|s\in S(m)\}} P(m|z^n)\\
&= \textstyle\sum_{m\in M}\int_{\Theta_m} P(y_{n+1}|x_{n+1}, z^n, \theta_m, m)P(\theta_m|z^n, m)\\
&\qquad P(m|z^n)d\theta_m\\
&= \hat{P}^*(y_{n+1}|x_{n+1}, z^n) \qquad (40)
\end{aligned}$$

□