# 連続変数に対応した決定木モデルにおけるベイズ最適な予測アルゴリズム

1G06H049-3　坂口卓也
指導教員　後藤正幸

## 1　研究背景と目的

近年，情報技術の発展により，データマイニングやパターン認識の技術が注目を集めている．これらの技術の中で決定木モデルによる学習と予測の有用性が示されており，CHAID, CART，ID3 など様々な決定木生成アルゴリズムが提案されてきた．これらのアルゴリズムは，学習データが与えられたもとで考えうる全ての決定木モデルの中から 1 つの決定木モデルを選択する方法である．しかし，学習データが与えられたもとで未観測のデータを予測するという問題を考えた場合，必ずしも 1 つのモデルを選択する必要はない．

そこで，須子ら [1] は考えうる全ての決定木モデルの混合をとり，ベイズ基準で平均予測誤り率を最小にしつつ効率的な計算アルゴリズムを提案している．しかし，このアルゴリズムでは予測対象である目的変数を離散値に限定しているが，決定木モデルをより一般的な問題に適用する場合，予測対象として連続目的変数も扱えることが望ましい．

本研究では須子らのアルゴリズムを拡張し，予測対象が連続変数である問題に対応するベイズ最適な予測アルゴリズムを提案する．また，数値実験により提案手法の有効性を示す．

## 2　須子らの手法

須子らの手法 [1] では松嶋らによるベイズ符号のアルゴリズム [2] を応用することで，考えうる全ての決定木モデルの混合モデルを考え，平均予測誤り率を最小にする予測アルゴリズムを示している．

### 2.1　問題設定

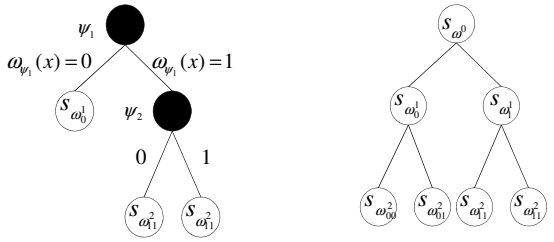
あるデータを $K$ 次元の離散属性ベクトル $\boldsymbol{x} \in \{0,1\}^K$ と，そのデータが属するカテゴリ $y \in \{0,1\}$ のセットで表す．学習データとして $\boldsymbol{x}^n = \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ と $y^n = y_1, y_2, \cdots, y_n$ の長さ $n$ の系列を考え，$\boldsymbol{x}_i$ と $y_i$ の組を $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ とし，合わせて $\boldsymbol{z}^n = \boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n$ と表記する．

本研究で対象とする予測問題は，$\boldsymbol{z}^n$ が得られているもとで，新たに $\boldsymbol{x}_{n+1}$ が与えられたとき，対応するカテゴリ $y_{n+1}$ を逐次的に予測する問題である．

### 2.2　決定木モデルの構成

前述の予測問題を扱うため，決定木モデルのクラスで $\boldsymbol{x}$ に対する質問の内容を $\psi_d (d = 1, 2, \cdots, D)$ とし，質問 $\psi_d$ に対し $\boldsymbol{x}$ が真 (1) か偽 (0) かを返す関数を $\omega_{\psi_d}(\boldsymbol{x}) \in \{0,1\}$ とする．ただし，$D \leq K$ である．質問が $\psi_1, \psi_2, \cdots, \psi_D$ の順番で与えられるとし，質問 $\psi_1, \psi_2, \cdots, \psi_d (d = 1, 2, \cdots, D)$ に対する $\omega_{\psi_d}(\boldsymbol{x})$ の系列を $\omega^d = \omega_{\psi_1}(\boldsymbol{x}), \omega_{\psi_2}(\boldsymbol{x}), \cdots, \omega_{\psi_d}(\boldsymbol{x})$ とする．$\omega^d$ が与えられた時に一意に定まる状態を $s_{\omega^d}$ とし，$s_{\omega^d}$ に基づき予測を行う．

図 1 の (a) は $D = 2$ における 1 つの決定木モデルの例である．予測対象である $y$ の条件付分布パラメータは，葉ノードのみに与えられる．一方，決定木モデルの混合モデルは，最大次数の決定木モデルのクラスに属するため，やはり木の形で描くことができる．そこで，全ての決定木の混合モデルの各ノードを状態 $s$ とし，全ての $s$ の集合を $S$ とする．このとき，状態 $s \in S$ を決定木モデルの葉ノードに対応させた場合，$D = 2$ における全ての決定木の混合モデルは図 1 の (b) で表すことができる．




(a) 1 つの決定木モデル　　(b) 全ての混合モデル

**図 1．決定木モデル**

### 2.3　効率的な計算アルゴリズム

予測対象が離散値なので $0-1$ 損失 を考え，このときベイズ最適な予測は以下で求めることができる [2].

$$\hat{y}_{n+1} = \arg\max_{y_{n+1}} \sum_{m \in M} \int_{\boldsymbol{\theta_m}} P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n, \boldsymbol{\theta_m}, m) P(\boldsymbol{\theta_m}|m, \boldsymbol{z}^n) P(m|\boldsymbol{z}^n) d\boldsymbol{\theta_m}. \quad (1)$$

ここで，$m \in M$ は 1 つの決定木モデル（木の構造）を表し，$\boldsymbol{\theta_m} \in \Theta_m$ はモデル $m$ のパラメータとする．式 (1) は，予測分布のモードを表している．

式 (1) では全ての決定木モデル $m$ を混合しているが，$D$ が大きくなると考慮すべきモデルの数 $|M|$ は指数的に増大してしまう．そこで，松嶋らにより提案されたアルゴリズム [2] を応用することで，図 1 の (b) の全ての決定木の混合モデルのもとで式 (1) を効率的に計算することができる．

$\boldsymbol{z}^n$ が得られたもとでの状態 $s$ の事後確率 $P(s|\boldsymbol{z}^n)$ は，重みパラメータ $q(s|\boldsymbol{z}^n)$ を用いて次式のように計算される．$s'$ は $s$ の祖先ノードとし，これを $s' < s$ と表記する．

$$P(s|\boldsymbol{z}^n) = q(s|\boldsymbol{z}^n) \prod_{s' < s} (1 - q(s'|\boldsymbol{z}^n)). \quad (2)$$

式 (1) で用いられる予測分布 $P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n)$ は式 (2) の重みパラメータを用いることにより，$\boldsymbol{x}_{n+1}$ が与えられたときに定まる状態の列 $s_{\omega^0}, s_{\omega^1}, \cdots, s_{\omega^D}$（混合モデルの木における根から葉までの 1 つのパスを表す）に対する以下の再帰計算で計算される．

$$P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n) = q(y_{n+1}|\boldsymbol{z}^n, s_{\omega^0}), \quad (3)$$

$$q(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d}) = q(s_{\omega^d}|\boldsymbol{z}^n) P(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d}) + (1 - q(s_{\omega^d}|\boldsymbol{z}^n)) q(y_{n+1}|\boldsymbol{z}^n, s_{\omega^{d+1}}). \quad (4)$$

このとき，パラメータの事前分布としてベータ分布を仮定することによって，式 (4) の状態 $s_{\omega^d}$ における事後確率 $P(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d})$ は Laplace 型推定量で計算できる [2].

## 3　提案手法

決定木モデルをマーケティング分析など実問題へ適用することを考えた場合，予測する対象 $y_{n+1}$ が連続値のケースにも対応できることが望ましい．そこで本研究では，連続値に対応した決定木モデルの予測アルゴリズムを提案する．

### 3.1　問題設定

ここでは，予測する対象 $y_{n+1}$ が連続値で正規分布に従う場合を考える．すなわち，$\boldsymbol{z}^n$ が得られている上で，新たに離散の属性ベクトル $\boldsymbol{x_{n+1}}$ が与えられたもとでの条件付正規分布に従う目的変数 $y_{n+1}$ の予測問題を対象とする．

### 3.2　連続値に対応した効率的な計算アルゴリズム

予測対象が連続値なので二乗誤差損失を考える．このときベイズ最適な予測は以下の式で求められる．

$$\hat{y}_{n+1} = \int_{y_{n+1}} y_{n+1} \sum_{m\in M} \int_{\boldsymbol{\mu}_m} \int_{\boldsymbol{\sigma}^2_m} P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n, \boldsymbol{\mu}_m, \boldsymbol{\sigma}^2_m, m) P(\boldsymbol{\mu}_m, \boldsymbol{\sigma}^2_m|m, \boldsymbol{z}^n) P(m|\boldsymbol{z}^n) d\boldsymbol{\mu}_m d\boldsymbol{\sigma}^2_m dy_{n+1}. \quad (5)$$

ここで，$m \in M$ はモデルであり，$\boldsymbol{\mu}_m \in U_m$ と $\boldsymbol{\sigma}^2_m \in \Sigma_m$ はモデル $m$ の未知のパラメータである．式 (5) は，予測分布の平均値を表している．

須子らの手法と同様に，状態 $s_{\omega^d} = s_{\omega^d}(\boldsymbol{x}_{n+1})$ を用いた図 1 の (b) の混合モデルの下で予測を行う．

式 (5) を計算するためには，状態 $s_{\omega^d}$ における $y_{n+1}$ の事後予測分布 $P(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d})$ を計算する必要がある．須子らの手法では予測対象が二項分布であったため，パラメータの事前分布としてベータ分布を仮定していた．これに対し，本研究では予測対象である目的変数 $y$ が $\boldsymbol{x}$ の条件付正規分布に従うことを仮定しているため，正規分布に対して共役な事前分布を仮定する必要がある．そこで，各状態 $s$ における未知のパラメータ $\mu_m(s)$ と $\sigma^2_m(s)$ の事前分布として，以下のような分布を設定する．

$$P(\sigma^2_m(s)) \sim \chi^{-2}(\nu_0(s), \lambda_0(s)),$$
$$P(\mu_m(s)|\sigma^2_m(s)) \sim N(\mu_0(s), \sigma^2_m(s)/n_0(s)). \quad (6)$$

ただし，$\nu_0(s)$, $\lambda_0(s)$, $\mu_0(s)$, $n_0(s)$ は状態 $s$ における事前分布のパラメータ，$\chi^{-2}(\nu_0(s), \lambda_0(s))$ は逆カイ二乗分布である．

式 (6) をもとにベイズの定理を用いて推測を行うと，事後予測分布 $P(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d})$ は以下に示す一般化 $t$ 分布に従うことがわかる．

$$P(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d}) \sim t\left[\bar{y}_{s_{\omega^d}}, \left(1 + \frac{1}{n_{s_{\omega^d}}}\right) b^2_{s_{\omega^d}}, \nu_{s_{\omega^d}}\right]. \quad (7)$$

ただし，$\bar{y}_{s_{\omega^d}}, b^2_{s_{\omega^d}}, \nu_{s_{\omega^d}}$ は，それぞれ状態 $s_{\omega^d}$ における $y$ の平均，不偏分散，自由度であり，$(1 + 1/n_{s_{\omega^d}})b^2_{s_{\omega^d}}$ は，データ数 $n_{s_{\omega^d}}$ によって変化する $b^2_{s_{\omega^d}}$ のパラメータである．

式 (7) を用いて式 (5) の予測分布の平均値を変形することにより，$\hat{y}_{n+1}$ は $\boldsymbol{x}_{n+1}$ が与えられたときに定まる状態の列 $s_{\omega^0}, s_{\omega^1} \cdots, s_{\omega^D}$ における平均値 $\bar{y}_{s_{\omega^0}}, \bar{y}_{s_{\omega^1}}, \cdots, \bar{y}_{s_{\omega^D}}$ を用いて以下の再帰計算で計算される．

$$\hat{y}_{n+1} = \bar{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^0}), \quad (8)$$

$$\bar{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^d}) = q(s_{\omega^d}|\boldsymbol{z}^n)\bar{y}_{s_{\omega^d}} + (1 - q(s_{\omega^d}|\boldsymbol{z}^n))\bar{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^{d+1}}). \quad (9)$$

## 4　数値実験と結果

提案手法の有効性を検討するために，数値実験を行った．比較対象として，Minimum Description Length (MDL) 基準 [3] によってモデル選択する方法を扱う．

### 4.1　実験条件

木の深さ $D = 3$ とする．データ長 $n = 200$ までの逐次予測の実験を 1 セットとして，繰り返し 500 セット実験する．その際，データを発生させる真の決定木モデルは，1 セット毎に考えられる全ての決定木モデルの中からランダムに 1 つ選択することとした．ただし，真の決定木モデルの各ノードの正規分布パラメータは，予め設定した値を用いて実験を行った．

### 4.2　実験結果及び考察

図 3 に実験結果を示す．横軸はデータ長，縦軸は予測値 $\hat{y}_{n+1}$ と観測値 $y_{n+1}$ の平均二乗誤差とする．$(\sigma^*)^2$ は，全ての決定木モデルの各ノードにおける $\sigma^2_m(s_{\omega^d})$ の重み付け和であり，平均予測誤差の理論下限値である．

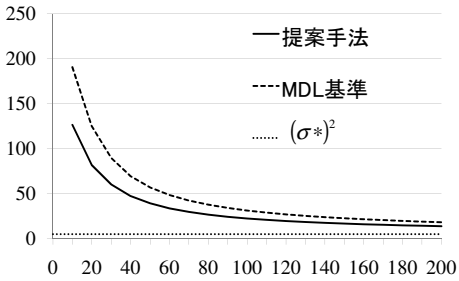



**図 3．提案手法と MDL 基準の比較**

図 3 より，提案手法の方が MDL 基準による決定木モデルよりも早く誤差が減少することがわかる．これは，決定木モデルを 1 つ選択するよりも全ての決定木モデルを混合する提案方法の方が，データ長 $n$ が有限のときの予測精度が高いことを示している．

この提案手法により，POS (販売時点情報) データから顧客属性と売上高というデータセットを得た場合，新たな顧客属性を得たときの売上高を効率的に予測できるなど，マーケティング分析への応用が可能となったと考えられる．

## 5　まとめと今後の課題

本論文では，予測対象が連続値である場合に対し，決定木の混合モデルを用いた予測値の効率的な計算アルゴリズムを示し，数値実験によりその有効性を示した．また，MDL 基準による方法よりも提案手法である混合モデルの方が予測精度の面で優れていることが示されている．今後の課題は，実問題への適用と評価である．

## 参考文献

[1] 須子統太, 野村亮, 松嶋敏泰, 平澤茂一, “決定木モデルにおける予測アルゴリズムについて,” 電子情報通信学会技術研究報告, COMP, コンピュテーション, Vol. 103, pp. 93–98, July 2003.

[2] 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, pp. 1288–1293, 1991.

[3] J. Rissanen, “Modeling by shortest data description,” *Automatica*, Vol. 46, pp. 465–471, 1978.