層別回帰モデルの Tree 表現によるベイズ最適な予測法

情報数理応用研究

5210C015-8　坂口卓也
指導教員　後藤正幸

Bayes Optimal Prediction Method for Tree Structures of Hierarchical Regression Models

SAKAGUCHI Takuya

1　はじめに

近年，情報技術の発展により，大量のデータを解析するデータマイニングや多変量解析の技術が注目を集めている．これらの技術の中で，様々な予測モデルが提案されてきており，有用性が示されてきた．その中で，説明変数を入力として，目的変数を予測するための代表的なモデルとして，決定木モデルや線形回帰モデルがある．

決定木モデルとは，データの説明変数をもとに木構造を用いて目的変数を予測するモデルであり，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 [1] ．決定木モデルでは，モデルの性質上，目的変数は離散変数，連続変数は問わないが，説明変数は離散変数であるか，あるいは離散化されて木構造に取り込まれる．実問題を考えたとき，説明変数に離散変数と連続変数が混在している場合を考えるのが望ましい．また，決定木モデルでは説明変数と目的変数間に線形の関係があっても，これを細かく離散化して複雑な木構造を作ってしまう．

一方，線形回帰モデルとは，データの説明変数をもとに線形関数を用いて目的変数を予測するモデルであり，重回帰分析，数量化 I 類など様々な分析手法が提案されてきた [2] ．線形回帰モデルでは，説明変数には特に制約もなく扱うデータにおいて柔軟に対応できる特徴を持っている．しかし，一般的に線形回帰モデルの特徴として交互作用があるデータに対しては，1 つの回帰式を用いて直接モデル化することが困難である．このような層別因子をもつ回帰分析では，層別して推定した回帰パラメータに有意差がある場合には，層別して複数の回帰式を当てはめた方が良いことも知られている．層別因子をもつ場合，層別される回帰式を木構造を用いて表現するモデルがある [3],[4]．さらに，この層別因子を交互作用基準で決定することで．より効果的な層別を行う研究も行われている [5].

一方，これまでの代表的な予測モデルは，学習データが与えられたもとで考えられる全てのモデルの中から適切な 1 つのモデルを選択する方法がほとんどであった．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，考えられる全ての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するベイズ最適な予測アルゴリズムの研究がされており，これを効率的に計算する予測アルゴリズムが提案されている [6].

そこで本研究では，より実問題に近いデータを想定して，離散変数と連続変数の混在する説明変数により考えられる連続の目的変数を予測する問題を対象とする．そして，線形回帰モデルをベースとし，説明変数間に交互作用が考えられる部分において決定木モデルを用いて階層的層別化を表現することで，交互作用のあるデータに対して適用できるモデルを考える [3]-[5]．本研究では，このようなモデルクラス上で効率的なベイズ最適な予測アルゴリズムを提案する．さらに，提案モデルの有効性を示すために，人工データによる検証と実データによる検証を行う．

2　関連研究

層別因子をもつ回帰分析による研究として，Quinlan や Karalic は層別された複数の回帰式を決定木を用いて表現するモデルを提案している [3],[4]．さらにこのモデルを拡張した研究として，関らは線形回帰モデルを交互作用効果の大きさを基準とした階層的な層別を行い，樹形モデルの中間ノードにも回帰項を割り当てることを許容したモデルを提案している [5]．関らの研究では，層別の分岐基準としてモデル選択手法の一つである MDL 基準 [7] を用いている．これらの研究では，考えられるモデルクラス上から MDL 基準などを用いてモデルを 1 つ選択している．これらのモデル選択手法に対して，本研究では，考えられる提案モデルクラス上で，効率的にベイズ最適な予測法を提案する．

ベイズ最適な予測アルゴリズムの研究はいくつか存在する．例えば，須子らは目的変数が多項分布に従う離散変数を予測対象と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を提案した [8]．これに対し，著者らは目的変数が正規分布に従う連続変数を予測対象とし，決定木の混合モデルによるベイズ最適な予測アルゴリズムを提案している [9]．これら 2 つの研究は，松嶋らによるベイズ符号アルゴリズム [6] を決定木モデルに応用したものであり，ベイズ最適な予測を効率的に計算するアルゴリズムを提案している．

一方で，鈴木らは線形回帰モデルにおけるベイズ最適な予測アルゴリズムを提案している [10]．この研究では，線形回帰モデル上で成り立つ事後確率に漸近正規性を用いて，混合事後分布を漸近近似的に求めることでベイズ最適な予測を計算している．

3　準備

3.1　線形回帰モデルの構成

本研究で取り扱う線形回帰モデルについて説明する．p 個の離散変数と連続変数が混在している説明変数ベクトル $\boldsymbol{x} = (a_1, a_2, \cdots, a_p)^{\mathrm{T}}$ を用いて，連続型の目的変数 y を予測する問題を考える．ただし，T は転置を表わす．こ

のとき線形回帰モデルは以下のように定義する．

$$\hat{y} = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} + \epsilon. \tag{1}$$

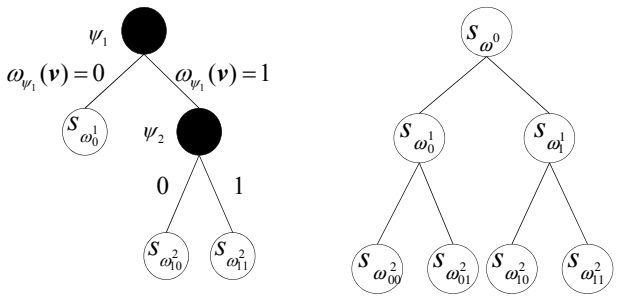
このとき，$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^{\mathrm{T}}$ を p 個の偏回帰係数ベクトル，ϵ を残差項とし，$\epsilon \sim N(0, \sigma^2)$ に従うものと仮定する．

3.2 決定木モデルの構成

本節では，決定木モデルとベイズ最適な混合モデルの構成法について述べる．あるデータを K 次元の離散説明変数ベクトル $\boldsymbol{v} = (c_1, c_2, \cdots, c_K)$ と，そのデータが属する目的変数 y のセットで表す．

決定木モデルのクラスで $\boldsymbol{v}$ に対する質問の内容を $\psi_d (d = 1, 2, \cdots, D)$ とし，質問 ψ_d に対し $\boldsymbol{v}$ が真 (1) か偽 (0) かを返す関数を $\omega_{\psi_d}(\boldsymbol{v}) \in \{0, 1\}$ とする．ただし，$D \leq K$ である．また，全ての $d \in \{1, 2, \cdots, D\}$ に対し，$\omega^d = \omega_{\psi_1}(\boldsymbol{v}), \omega_{\psi_2}(\boldsymbol{v}), \cdots, \omega_{\psi_d}(\boldsymbol{v})$ とする．

ω^d が与えられた時に一意に定まる状態を s_{ω^d} とし，s_{ω^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 $\mathcal{S}$ とする．このとき，状態 $s \in \mathcal{S}$ 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

4 提案手法

4.1 問題設定

p 個の離散変数と連続変数が混在している説明変数ベクトル $\boldsymbol{x} = (a_1, a_2, \cdots, a_p)^{\mathrm{T}}$ を用いて，連続型の目的変数 y を予測する問題を考える．このとき，p 個の説明変数ベクトルのうち，離散変数が r 個，連続変数が $(p - r)$ 個が混在しているものとする．r 個の離散説明変数のうち，交互作用のある D 個を層別因子とし，これを質問に対応させた深さ D の木を生成する．

学習データとして $\boldsymbol{x}^n = \boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_n$ と $y^n = y_1y_2\cdots y_n$ 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 y_{n+1} を逐次的に予測する問題である．また，目的変数 y は層別因子が与えられたもとでの，残りの説明変数による条件付き線形回帰モデルに従うものとする．

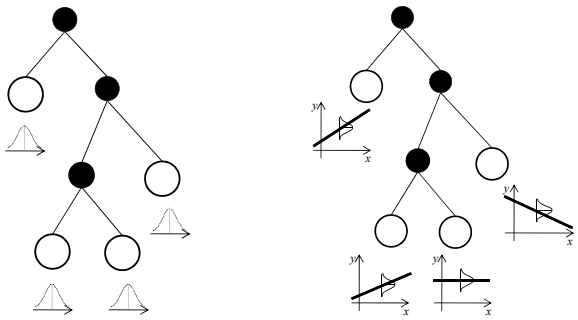
4.2 概要

図 2 の (a) は，回帰木のイメージ図であり，図 2 の (b) は層別木のイメージ図である．

回帰木とは，決定木のなかでも目的変数が連続変数の場合を木構造で表したモデルである．著者らの従来研究 [9] では，目的変数 y を離散説明変数ベクトル $\boldsymbol{x}$ が与えられたもとでの条件付正規分布に従うモデルの予測問題を対象としていたため，回帰木の葉ノードに正規分布を仮定したもとで，効率的なベイズ最適な予測法を提案した．

それに対し，図 2 の (b) のように，葉ノードに対して線形回帰式を付与させたモデルを層別木と呼ぶことにする．これは，根ノードから階層的に層別して回帰モデルを構成していることを意味する．層別木を考えている研究は，2 章で述べたようにこれまでにいくつか存在している [3]-[5]．本研究でも目的変数 y の分布が離散と連続が混在している説明変数ベクトル $\boldsymbol{x}$ が与えられたもとでの条件付線形回帰モデルで表される予測問題を対象としているので，層別木を考えている．

従来の層別木による研究がモデル選択手法であるのに対し，本研究では，層別木を用いて階層化した全ての部分木モデルの混合を効率的に計算する予測法を提案する．



(a) 回帰木　　(b) 層別木

図 2．回帰木と層別木イメージ

4.3 ベイズ最適な予測の定式化

予測対象が連続値なので，二乗誤差損失で考え，そのベイズ最適な予測は以下の式で求めることができる．ただし，$\hat{y}$ は y の予測値とする．

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

$$= \sum_{m \in \mathcal{M}} \hat{y}_m P(m|\boldsymbol{z}^n). \tag{3}$$

モデル m のもとでカテゴリ y の発生する確率を $P(y|m, \boldsymbol{x}, \boldsymbol{\beta}_m, \boldsymbol{\sigma}^2_m)$ とする．このとき，$m \in \mathcal{M}$ は 1 つの決定木モデルを表し，$\boldsymbol{\beta}_m \in \mathcal{B}_m$ 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．また，$\hat{y}_m$ はモデル m に含まれる回帰式の予測値である．

式 (2) は，全ての考えられるモデルの混合事後予測分布の平均値を表している．本研究では，式 (2) を式 (3) のように変形することにより，混合事後予測分布の平均値をとるのではなく，各モデル m の事後予測分布の平均値の混合をとることでベイズ最適な予測値を算出する．各モデルの事後予測分布の平均値は，各モデルに含まれる回帰式の予測値である．

4.4 混合モデルのもとでのベイズ最適な予測

式 (3) 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 $|\mathcal{M}|$ は指数的に増大してしまう．そこで，図 2 の (b) の全ての決定木の混合モデルのもとで式 (3) を効率的に計算することができる．式 (3) を計算するためには，$\boldsymbol{x}_{n+1}$ が与えられた時に定まる各状態 s_{ω^d} における y_{n+1} の事後予測分布 $P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n, s_{\omega^d})$ を計算する必要がある．事後予測分布を計算するために，各状態 s_{ω^d} における未知のパラメータ $\beta_{s_{\omega^d}}$ と $\sigma^2_{s_{\omega^d}}$ の事前分布として，以下の式で表される局所一様事前分布を採用する．

$$P(\beta_{s_{\omega^d}}, \sigma^2_{s_{\omega^d}}) \propto (\sigma^2)^{-1}. \tag{4}$$

式 (4) 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 $P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n, s_{\omega^d})$ は以下に示す多変量 t 分布に従うことがわかる [11].

$$\begin{aligned} &P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n, s_{\omega^d}) \\ &\sim t\left[\hat{y}_{s_{\omega^d}}, \{1 + \boldsymbol{x}^t_{n+1}\left(\boldsymbol{X}^t_{s_{\omega^d}} \boldsymbol{X}_{s_{\omega^d}}{}^{-1}\right)\boldsymbol{x}_{n+1}\} b^2_{s_{\omega^d}}, \nu_{s_{\omega^d}}\right]. \end{aligned} \tag{5}$$

ただし，$\hat{y}_{s_{\omega^d}}, b^2_{s_{\omega^d}}, \nu_{s_{\omega^d}}$ は，それぞれ状態 s_{ω^d} における予測値，残差平方和，多変量 t 分布の自由度であり，$\boldsymbol{X}_{s_{\omega^d}}$ をデータ数 $n_{s_{\omega^d}} + 1$ 次元の説明変数行列とする．

式 (5) で示すように，各状態 s_{ω^d} の事後予測分布の平均値は，各状態 s_{ω^d} に含まれる回帰式の予測値である．よって，式 (3) を混合モデルのもとに置き換えると以下の式で示すことができる．

$$\hat{y}_{n+1} = \sum_{s_{\omega^d}} \hat{y}_{s_{\omega^d}} P(s_{\omega^d}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n). \tag{6}$$

式 (6) の通り，ベイズ最適な予測値は各状態の予測値の混合（期待値）を取ればよい．

4.5 効率的な計算アルゴリズム

状態 s_{ω^d} の事後確率を以下の式で定義する重みパラメータで表すことで，学習のパラメータ更新を効率的に行うことが出来る．

$$P(s_{\omega^d}|\boldsymbol{z}^n) = q(s_{\omega^d}|\boldsymbol{z}^n) \prod_{l=0}^{d} (1 - q(s_{\omega^l}|\boldsymbol{z}^n)). \tag{7}$$

式 (6) と式 (7) を用いて，$\hat{y}_{n+1}$ は $\boldsymbol{x}_{n+1}$ が与えられたときに定まる状態の列 $s_{\omega^0}, s_{\omega^1} \cdots, s_{\omega^D}$ における予測値 $\hat{y}_{s_{\omega^0}}, \hat{y}_{s_{\omega^1}}, \cdots, \hat{y}_{s_{\omega^D}}$ を用いて以下の再帰計算で計算される．

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

$$\begin{aligned} \hat{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^d}) =& q(s_{\omega^d}|\boldsymbol{z}^n)\hat{y}_{s_{\omega^d}} \\ &+ (1 - q(s_{\omega^d}|\boldsymbol{z}^n))\hat{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^{d+1}}). \end{aligned} \tag{9}$$

5 人工データを用いた検証

提案手法の有効性を検証するために，人工データによる数値実験を行う．

5.1 実験条件

離散変数と連続変数が混在した 5 次元の説明変数を用意する．このとき，離散変数を 3 つ，連続変数を 2 つ用意した．

D 次元の交互作用のある離散変数は，回帰の層別因子として木の枝に対応させ，層別木を構成する．目的変数は深さ $D = 2$ の完全層別木のもとで交互作用のあるようにデータを発生させる．ここで完全層別木とは最大深さ D まで全ての葉ノードが開いている状態であり，その深さのモデルクラスの中で最も表現能力の高いモデルとして考えられる．データを発生させるパラメータ偏回帰係数の値は適当に定めた．

このとき，比較手法として 5 つの説明変数全てを使用した線形回帰分析（数量化 I 類）で推定されたモデルでデータを予測する方法を考える．

学習データを 100 件から 1000 件までの 100 件刻みとし，それぞれの学習データで生成されたモデルのもとで，テストデータ 1000 件に対しての平均二乗誤差を計算する．これをを 1 セットとし，繰り返し 100 セットを行うものとする．

5.2 実験結果及び考察

図 3 に実験結果を示す．横軸は，学習データ数，縦軸は予測値と観測値の平均二乗誤差とする．

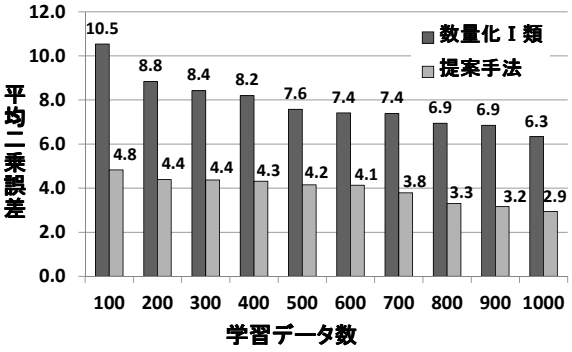


図 3．人工データの結果

図 3 より，学習データ数が 100 件から 1000 件までの全ての場合で，提案手法の方が通常の線形回帰モデル (数量化 I 類) による平均二乗誤差よりも低く抑えることが出来た．

今回の実験では，説明変数間で交互作用のあるように人工データを発生させたため，比較手法である通常の線形回帰モデル (数量化 I 類) では，この構造を表現することができず，交互作用を考慮した提案手法の方が予測精度が高くなったと考えられる．この結果より，交互作用のあるデータに対して提案手法の有効性を示すことができた．

6 実データを用いた検証

提案手法を実データに適用し予測性能の検証を行う．今回扱う実データとして賃貸物件サイト「ＣＨＩＮＴＡＩ」[12] を利用し，そのデータをもとに家賃の価格予測を行う．

6.1 実験対象データ

実験対象データは山手線沿線の賃貸物件データ 12,695 件のデータ (2011 年 6 月 10 日時点) とし，専有面積，築年数など全部で 16 項目の変数を抽出した．抽出してきた変数項目を表 1 に示す．

表 1．変数項目

目的変数	家賃
説明変数	駅までの距離 , 南向き
	バス・トイレ別 , ペット相談可
	2 階以上 , 専有面積
	室内洗濯機置場 , 建物
	駐車場付 , ベランダ
	エアコン , 階層
	フローリング , 築年数
	オートロック , 階建

6.2 実験条件

抽出したデータのうち，学習データを 100 件から 500 件までの 100 件刻みとし，残りの件数からランダムに 10,000 件抽出したものをテストデータとし，これを繰り返し 100 回行う．

表 1 からの変数選択方法として変数増加法を用いた．また，学習データごとに変数選択された説明変数らを分散分析を行い交互作用を抽出した．有意水準は 5%有意とした．分散分析を行った結果，交互作用があると判断された場合には，提案手法において，その変数らを質問とみなして層別木モデルを生成し，その層別木モデルのもとでベイズ最適な予測を構成する．

比較手法として，通常の線形回帰分析 (数量化 I 類) と層別因子で階層化した完全層別木のもとでデータを予測していくものとする．

6.3 実験結果及び考察

図 4 に実験結果を示す．横軸は学習データ数，縦軸は予測値と観測値の平均二乗誤差である．

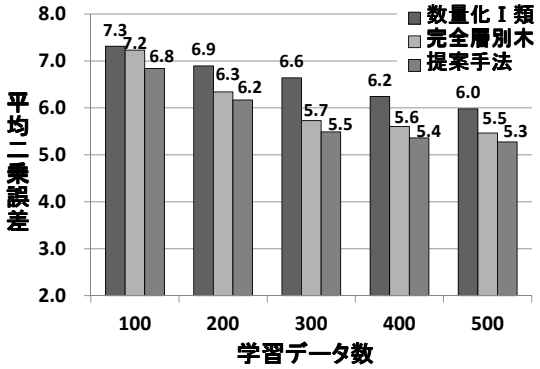


図 4．実データでの検証

図 4 より，学習データ数が 100 件から 500 件の全ての場合で提案手法の方が平均二乗誤差を低く抑えることができた．

提案手法が比較手法である通常の線形回帰モデル (数量化 I 類) による平均二乗誤差より低く抑えたことで，実データに対しても交互作用のある部分を Tree 表現で階層化した提案モデルの有効性を示せた．

また，提案手法が比較手法である完全層別木による平均二乗誤差より低く抑えたことで，交互作用がある部分を全て開いた完全層別木のモデルよりも，完全木を含んだ考えられる全てのモデルを混合したベイズ最適な予測モデルである提案手法の方の有効性を示せたといえる．このことは，未知データの予測という問題においては，モデルを一つ選択するよりもベイズ最適にモデルを混合した予測の方が有用であることを示している．

7 今後の課題とまとめ

本研究では，層別回帰モデルの Tree 表現によるベイズ最適な予測法を提案した．提案モデルでは，複数の層別因子による階層的層別によって考えられる層別回帰モデルのクラス上で，効率的にベイズ最適な混合モデルを導く方法が与えられているモデルと考えられる．

実際に提案手法の有効性を示すために交互作用のある人工データでの検証において通常の線形回帰モデル (数量化 I 類) を比較とした実験を行い，有効性を示した．更に，実データでの提案手法の有効性を示すために，賃貸物件の家賃データでの検証を行った．結果として，通常の線形回帰モデル (数量化 I 類)，交互作用の部分を全て開いた完全木モデルに比べて有効性を示した．

今後の課題は，線形回帰モデルだけでなく様々な確率モデルへのモデル拡張である．

参考文献

[1] J. R. Quinlan, “Induction of decision trees,” *Machine-learn.*, Vol. 1, pp. 81–106，1986.

[2] 永田靖, 棟近雅彦 “多変量解析入門,” サイエンス社, pp.1-10, 2007.

[3] J. R. Quinlan, “Linear regression in regression tree leaves,” *In 5th Australian Joint Conference on Artificial Intelligence*, pp. 343–348, 1992.

[4] A. Karalic, “Employing Learning with Continuous Classes,” *Proceedings of IECAI ’92*, pp. 440–441, 1992.

[5] 関庸一, 野島勇 “交互作用基準による再帰分割線形モデル,” 応用統計学会, Vol. 33，pp. 111–130, Dec. 2004.

[6] 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.

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

[8] 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.

[9] 坂口卓也, 石田 崇, 後藤正幸, “混合決定木モデルによる連続変数の予測法に関する一考察,” 第 10 回情報科学技術フォーラム, pp.503-504, Sep. 2011.

[10] 鈴木友彦, 後藤正幸, 石田崇, 後藤正幸, 俵信彦, “線形回帰モデルのベイズ最適な予測法に関する研究,” 日本経営工学会論文誌 51(1), Vol. 46 , pp. 59–69, 2000.

[11] 繁桝算男, “ベイズ統計入門,” 東京大学出版会, pp.169-180, 2003.

[12] CHINTAI : http://www.chintai.net/