F-040

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

A Study on Prediction Method based on Mixtured Decision Tree for Continuous variable

坂口卓也† **石田崇**‡ **後藤正幸***
Takuya Sakaguchi Takashi Ishida Masayuki Goto

1 はじめに

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変数も扱えることが望ましい．そこで著者らは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問題において，須子らのアルゴリズムを連続変数に対応したモデルと事前分布のクラスに拡張することにより，予測対象が連続変数である場合のベイズ最適な予測アルゴリズムを提案している [2].

本研究では，[2] で提案した手法に対して，賃貸物件による実データを用いて決定木生成アルゴリズムの一つである CHAID 分析との比較実験を行うことで，提案手法の有用性の検証を行う．

2 連続変数に対する混合決定木モデル

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．そこで著者らは，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している [2]. 以下で，その手法の概要について述べる．

2.1 問題設定

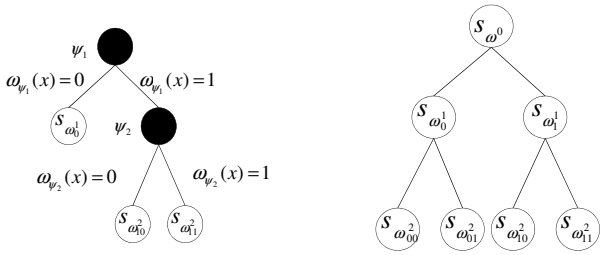
あるデータを K 次元の離散属性ベクトル $\boldsymbol{x} = (a_1, a_2, \cdots, a_K)$ と，そのデータが属するカテゴリ $y \in \mathcal{Y}$ のセットで表す．学習データとして $\boldsymbol{x}^n = \boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_n$ と $y^n = y_1y_2\cdots y_n$ の長さ n の系列を考え，$\boldsymbol{x}_i$ と y_i の組を $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ 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，対応するカテゴリ y_{n+1} を逐次的に予測する問題である．また，POS データなどの大量のデータから顧客の購売特性分析を行うようなケースを考えた場合，連続データの分布はしばしば正規分布に従うと考えられる．よって，本研究では目的変数 y が離散の属性ベクトル $\boldsymbol{x}$ 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する．

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 ψ_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．また，全ての $d \in \{1, 2, \cdots, D\}$ に対し，$\omega^d = \omega_{\psi_1}(\boldsymbol{x}), \omega_{\psi_2}(\boldsymbol{x}), \cdots, \omega_{\psi_d}(\boldsymbol{x})$ 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 2 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ズム

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．

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

モデル m のもとでカテゴリ y の発生する確率を $P(y|m, \boldsymbol{x}, \boldsymbol{\mu}_m, \boldsymbol{\sigma}^2_m)$ とする．このとき，$m \in \mathcal{M}$ は 1 つの決定木モデルを表し，$\boldsymbol{\mu}_m \in \mathcal{U}_m$ 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 (1) は，予測分布の平均値を表している．

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 $|\mathcal{M}|$ は指数的に増大

†早稲田大学大学院創造理工学研究科
‡早稲田大学メディアネットワークセンター
*早稲田大学理工学術院

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 (1) を効率的に計算することができる．

式 (1) を計算するためには，状態 s_{ω^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．著者らの手法における事後予測分布 $p(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d})$ は，以下の式で表すことができる．

$$P(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d}) = \int_{\mu_m(s_{\omega^d})} \int_{\sigma^2_m(s_{\omega^d})} P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n, \mu_m(s_{\omega^d}), \sigma^2_m(s_{\omega^d}), s_{\omega^d}) \cdot P(\mu_m(s_{\omega^d}), \sigma^2_m(s_{\omega^d})|s_{\omega^d}, \boldsymbol{z}^n) d\mu_m(s_{\omega^d}) d\sigma^2_m(s_{\omega^d}) \tag{2}$$

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．

$$P(\sigma^2_m(s)) \sim \chi^{-2}(\nu_0(s), \lambda_0(s)),$$
$$P(\mu_m(s)|\sigma^2_m(s)) \sim N(\mu_0(s), \sigma^2_m(s)/n_0(s)). \tag{3}$$

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 (3) 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．

$$P(y_{n+1}|\boldsymbol{z}^n, s_{\omega^d}) \sim t\left[\bar{y}_{s_{\omega^d}}, \left(1 + \frac{1}{n_{s_{\omega^d}}}\right) b^2_{s_{\omega^d}}, \nu_{s_{\omega^d}}\right]. \tag{4}$$

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_{ω^d} に含まれる y_i の平均，不偏分散，t 分布の自由度であり，$\nu_{s_{\omega^d}} = n_{s_{\omega^d}} - 1$ かつ $b^2_{s_{\omega^d}} = \frac{1}{\nu_{s_{\omega^d}}} \sum_{i=1}^{n_{s_{\omega^d}}} (y_i - \bar{y}_{s_{\omega^d}})^2$ で求めることが出来る．

式 (4) を用いて式 (1) 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．

$$\hat{y}_{n+1} = y_{n+1}(\boldsymbol{z}^n, s_{\omega^0}), \tag{5}$$

$$\bar{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^d}) = q(s_{\omega^d}|\boldsymbol{z}^n)\bar{y}_{s_{\omega^d}} + (1 - q(s_{\omega^d}|\boldsymbol{z}^n))\bar{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^{d+1}}). \tag{6}$$

ここで，$q(s_{\omega^d}|\boldsymbol{z}^n)$ は状態 s_{ω^d} における重みパラメータであり，これを用いることで効率的に予測値を計算することができる [2].

3　実データを用いた検証

2 節で述べた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」[3] を利用し，代表的なそのデータを基に家賃の価格予測を行う．比較手法として，決定木生成アルゴリズムの一つである CHAID 分析を用いる．

3.1　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 13635 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 21 項目の変数を抽出した．

3.2　実験条件

抽出したデータの内，学習データを 500 件，1000 件，1500 件，2000 件とし，テストデータを残りの件数として実験を行う．このとき，提案手法の質問（説明変数）の順番は CHAID 分析によって得られたものと同様の変数を扱うものとする．

3.3　実験結果

図 1 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．

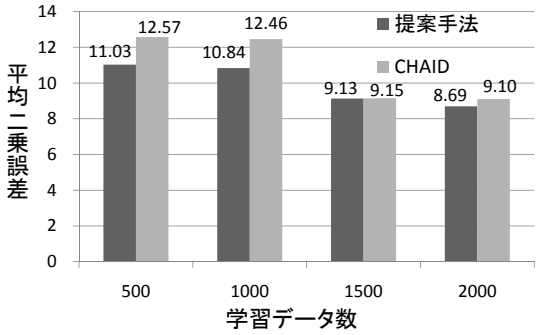


図 2．提案手法と CHAID 分析の比較

3.4　実データ検証における考察

図 5 より，学習データ数が 500 件から 2000 件の全ての場合で平均二乗誤差を低く抑えることができた．これは，学習データから CHAID 分析により得られた一つの決定木モデルが必ずしも未観測のデータに対して全て適用できるとは言えず，考えられる全ての決定木モデルを考慮している提案手法の方が，未観測データに対する予測精度は高いということがわかる．特に学習データが少ない場合，未知のデータに対して十分な予測モデルが生成されていない分，考えられる全てのモデルの混合をとる提案手法の方が予測誤差をより小さく抑えることができたと考えられる．

4　今後の課題とまとめ

本研究では，[2] で著者らが提案した手法の性能を検証を行うため，代表的な決定木生成アルゴリズムの 1 つである CHAID 分析との比較実験を実データに対して行った．その結果，従来の決定木生成アルゴリズムより予測誤差を抑えることができ，特に学習データ数が少ない所で提案手法の方が有効であることを示した．

今後の課題は，質問の順番の考慮が挙げられる．今回，実データでの実験では，比較手法である CHAID 分析によって得られた決定木をもとに質問の順番を考慮した．今後は，提案手法に適した質問の順番の決定法を理論的に検討することが求められる．また，決定木モデルに捉われず，予測モデルという幅広い視野から回帰モデルとの比較，検証をしていくとともに，それらを考慮したアルゴリズムの拡張も今後の課題である．

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] 坂口卓也, 寺本賢一, 石田崇, 後藤正幸, “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,” 営情報学会　全国研究発表大会要旨集, Vol. 2010f , pp. 61–64, Nov. 2010.

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