

線形回帰モデルの混合を用いた層別木モデルによるベイズ最適な予測法

1X09C102-7 藤原直広
指導教員 後藤正幸

1 研究背景と目的

近年、統計分析やデータマイニングの分野において、様々な確率モデルの学習による予測アルゴリズムの有用性が示されてきた。これらの手法には、決定木モデルや線形回帰モデルに基づく方法などがあるが、決定木モデルは、説明変数と目的変数が線形関係である場合、必要以上に複雑な木構造が構成されてしまう。また、線形回帰モデルは交互作用があるデータに対し、モデル化が困難であるという問題がある。

このような問題に対し、坂口ら [1] は、決定木モデルの各葉ノードに線形回帰モデルを付与した層別木を用いて、ベイズ最適な予測アルゴリズムを提案した。この手法では、データが持つ交互作用の構造と線形構造を共に表現することができる。しかし、この層別木モデルでは、各葉ノードに割り当てられる線形回帰モデルが、層別に用いられない残りの全ての説明変数を含むモデルのみとなっている。実問題を考えた場合、目的変数が必ずしも全ての説明変数を用いた線形回帰モデルに従うとは限らない。

そこで本研究は、考えられる全ての線形回帰モデルについて全てのモデルの混合を行うベイズ最適な予測法を提案する。さらに、人工データによる数値実験を行い、提案手法の有効性を示す。

2 従来手法

2.1 問題設定

q 次元の説明変数ベクトルには、離散変数と連続変数が混在しているものとする。説明変数 x のうち、層別に用いる D 個の離散説明変数を $u = (u_1, u_2, \dots, u_D)^T$ 、残りの説明変数を $v = (v_1, v_2, \dots, v_{q-D})^T$ とし、 $x = (u^T, v^T)^T$ とする。ここで、 T は転置を表す。また、目的変数を y とし、 y は v による線形回帰モデルに従う。 u は回帰係数を変化させる可能性のある離散変数を集めたベクトルである。学習データとして、 $x^n = (u_1, v_1)(u_2, v_2) \dots (u_n, v_n)$ と $y^n = y_1 y_2 \dots y_n$ を用いる。 (u_i, v_i) と y_i の組を $z_i = (u_i, v_i, y_i)$ とし、合わせて $z^n = z_1 z_2 \dots z_n$ と表記する。

本研究では、 z^n が得られているとき、新たに u_{n+1}, v_{n+1} が与えられたもとで y_{n+1} を予測する問題を考える。

2.2 層別木モデル

層別木モデルは、決定木モデルの各葉ノードに線形回帰モデルを割り当てたモデルである。層別に用いる D 個の離散説明変数 u を使い、データは木の根ノードから順に葉ノードに到達するまで $u_i (i = 1, 2, \dots, d, d \leq D)$ のとる値によって層別される。また、 $u^d = u_1 u_2 \dots u_d$ によって決まる葉ノードを s_{u^d} とする。図 1 に $u_i \in \{0, 1\}$ 、最大深さ $D = 2$ の層別木モデルの一例を示す。

2.3 混合モデルの下でのベイズ最適な予測の定式化

坂口らは、最大深さが D の全ての層別木によって定まるモデルクラスを \mathcal{M} とし、 \mathcal{M} に属する全ての層別木 $m \in \mathcal{M}$ の混合を行ったベイズ最適な予測アルゴリズムを提案した。このアルゴリズムでは、 \mathcal{M} に属する全ての層別木を混合し

た混合層別木は深さ D の完全木で表される [2]。1 つの層別木 $m \in \mathcal{M}$ は混合層別木の部分木となる。ここで、 $D = 2$ の混合層別木を図 2 に示す。

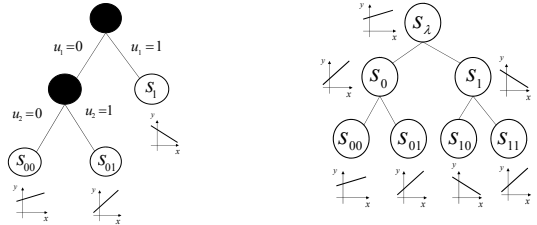


図 1. 1 つの層別木モデル

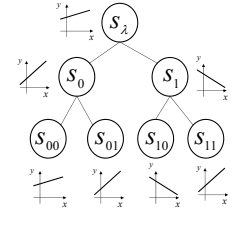


図 2. 混合層別木モデル

混合層別木のノード集合を \mathcal{S} としたとき、各ノード $s \in \mathcal{S}$ に線形回帰モデルが割り当てられている。また、 u が与えられると、根ノードから葉ノードへ 1 つのパスが決まり、このパス上のノード集合を $\mathcal{S}(u)$ とする。

全ての層別木の混合モデルのもとで、ベイズ最適な予測式は以下の式で求めることができる [1]。

$$\hat{y}_{n+1} = \sum_{s \in \mathcal{S}(u_{n+1})} \hat{y}(v_{n+1}, s) P(s | z^n). \quad (1)$$

ここで、 $\hat{y}(v_{n+1}, s)$ は v_{n+1} が与えられたもとで、ノード $s \in \mathcal{S}(u_{n+1})$ における線形回帰モデルによる予測値であり、 $P(s | z^n)$ は、 z^n が与えられたもとでのノード $s \in \mathcal{S}(u_{n+1})$ の事後確率である。

坂口らは、ノード $s \in \mathcal{S}$ において割り当てられている線形回帰モデルを、層別に用いていない残りの全ての説明変数を含んだ以下のモデルに限定している。

$$y = \beta_0^{(s)} + \beta_1^{(s)} v_1 + \dots + \beta_{q-D}^{(s)} v_{q-D} + \varepsilon. \quad (2)$$

ここで、 $\beta_i^{(s)} \in \mathcal{R}$ は、ノード $s \in \mathcal{S}$ における回帰パラメータ ($i = 0, 1, 2, \dots, q - D$)、 ε は正規分布に従う残差項である。

3 提案手法

坂口らの手法では、各ノード $s \in \mathcal{S}$ に割り当てられた線形回帰モデルが式 (2) に限定されていたが、 v についても説明変数の選択により複数のモデル候補を考慮することができる。そこで、本研究では、各ノード $s \in \mathcal{S}$ ごとに、考えられる全ての線形回帰モデルを含んだモデルクラスを仮定し、全てのモデルの混合を用いたベイズ最適な予測法を提案する。提案手法は坂口らの手法よりも広いモデルクラスを設定し、それに対するベイズ最適な予測法を構成するものである。

3.1 提案手法の問題設定

$P = \{p_1, p_2, \dots, p_{2^{q-D}}\}$ を考えられる全ての線形回帰モデル集合とし、以下のように与える。

$$\begin{cases} y = \beta_0^{(p_1)} + \varepsilon. \\ y = \beta_0^{(p_2)} + \beta_1^{(p_2)} v_1^{(p_2)} + \varepsilon. \\ \vdots \\ y = \beta_0^{(p_{2^{q-D}})} + \beta_1^{(p_{2^{q-D}})} v_1^{(p_{2^{q-D}})} \\ + \dots + \beta_{q-D}^{(p_{2^{q-D}})} v_{q-D}^{(p_{2^{q-D}})} + \varepsilon. \end{cases}$$

このとき、本研究で仮定する層別木モデルのクラスは、各層別木モデル $m \in \mathcal{M}$ の全ての葉ノード $s \in \mathcal{S}(m)$ にそれぞれ1つの線形回帰モデル $p \in \mathcal{P}$ を付与した $\{\mathcal{S}(m) \times \mathcal{P} \mid m \in \mathcal{M}\}$ となる。ただし、 $\mathcal{S}(m)$ はモデル m の葉ノード集合とする。

3.2 混合モデルのもとでのベイズ最適な予測の定式化

考えられる全ての線形回帰モデルを含んだモデルクラス上で、全てのモデルの混合を用いたベイズ最適な予測式は以下の式で表される。

$$\hat{y}_{n+1} = \sum_{s \in \mathcal{S}(u_{n+1})} \sum_{p \in \mathcal{P}} \hat{y}(v_{n+1}, p, s) P(p \mid z^n, s) P(s \mid z^n). \quad (3)$$

ここで、

$$\hat{y}(v_{n+1}, s) = \sum_{p \in \mathcal{P}} \hat{y}(v_{n+1}, p, s) P(p \mid z^n, s),$$

である。 $\hat{y}(v_{n+1}, p, s)$ は v_{n+1} が与えられたもとで、ノード $s \in \mathcal{S}$ における線形回帰モデル $p \in \mathcal{P}$ による予測値であり、 $P(p \mid z^n, s)$ は、 z^n が与えられたもとでのノード $s \in \mathcal{S}$ における線形回帰モデル $p \in \mathcal{P}$ の事後確率である。

ここで、 v_{n+1} が与えられたもとで、 u_{n+1} によって決定されるパス上のノードにおける予測値 $\hat{y}(v_{n+1}, s)$, $s \in \mathcal{S}(u_{n+1})$ が与えられれば、式 (3) は混合層別木の葉ノードから根ノードへ以下の式で再帰的に計算することができる [1]。

$$\hat{y}_{n+1} = \hat{y}_{n+1}(v_{n+1}, s_{u^0}), \quad (4)$$

$$\begin{aligned} \hat{y}_{n+1}(v_{n+1}, s_{u^d}) &= (1 - g(s_{u^d} \mid z^n)) \hat{y}(v_{n+1}, s_{u^d}) \\ &\quad + g(s_{u^d} \mid z^n) \hat{y}_{n+1}(v_{n+1}, s_{u^{d+1}}). \end{aligned} \quad (5)$$

3.3 線形回帰モデルの混合

各ノード $s \in \mathcal{S}$ において、考えられる全ての線形回帰モデルの混合を行い、以下の式で予測値を計算する。

$$\begin{aligned} \hat{y}(v_{n+1}, s) &= \sum_{p \in \mathcal{P}} \hat{y}(v_{n+1}, p, s) P(p \mid z^n, s) \\ &= \sum_{p \in \mathcal{P}} P(p \mid z^n, s) \left\{ \hat{\beta}_0^{(p,s)} + \left(\hat{\beta}^{(p,s)} \right)^T v_{n+1}^{(p)} \right\}. \end{aligned} \quad (6)$$

$v^{(p)}$ は、線形回帰モデル $p \in \mathcal{P}$ に用いられる説明変数である。 $\left(\hat{\beta}^{(p,s)} \right)^T$ は、ノード $s \in \mathcal{S}$ における線形回帰モデル $p \in \mathcal{P}$ の回帰パラメータベクトルである。事後確率 $P(p \mid z^n, s)$ は各モデルの重みと解釈でき、ベイズの定理を用いて以下の式で求められる。

$$P(p \mid z^{n+1}, s) = \frac{P(p \mid z^n, s) f(y_{n+1} \mid v^{n+1}, y^n, p, s)}{\sum_{p \in \mathcal{P}} P(p \mid z^n, s) f(y_{n+1} \mid v^{n+1}, y^n, p, s)}. \quad (7)$$

ここで、 $f(y_{n+1} \mid v^{n+1}, y^n, p, s)$ はノード $s \in \mathcal{S}$ 、線形回帰モデル $p \in \mathcal{P}$ のもとで、 y_{n+1} の発生する事後予測分布の確率密度であり、以下に示す一般化 t 分布に従う [3]。

$$f(y_{n+1} \mid v^{n+1}, y^n, p, s) \sim t \left(\hat{y}(v_{n+1}, p, s), \phi, \nu^{(p,s)} \right). \quad (8)$$

$$\phi = \left\{ 1 + \left(v_{n+1}^{(p)} \right)^T \left(\left(V^{(p,s)} \right)^T V^{(p,s)} \right)^{-1} v_{n+1}^{(p)} \right\} (b_s^p)^2.$$

ただし、各ノード $s \in \mathcal{S}$ に属するデータ数を n_s 、線形回帰モデル $p \in \mathcal{P}$ の説明変数の数を N_p とするとき、 $V^{(p,s)}$ は $(n_s + 1) \times N_p$ の説明変数行列、 $(b_s^p)^2$ を残差平方和、 $\nu^{(p,s)} (= n_s - N_p - 1)$ を一般化 t 分布の自由度とする。

4 実験

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。

4.1 実験条件

離散変数を2つ、連続変数を6つの8次元の説明変数を用意する。2つの離散変数は、木の枝に対応させ層別木を構成する。最大深さ $D = 2$ の木をランダムに選び、目的変数をそのモデルのもとで交互作用があるように発生させる。目的変数の回帰パラメータはあらかじめランダムに定めた。

比較手法として、坂口らの手法 [1] と、各ノード $s \in \mathcal{S}$ において、全ての線形回帰モデルから MDL 基準でモデルを1つ選択する手法を用いる。学習データは100件から500件の100件刻みとし、テストデータ1000件に対して逐次的に予測を行い、平均二乗誤差を計算する。これを1セットとして100セット繰り返し、それらの平均を実験結果とする。

4.2 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を表している。

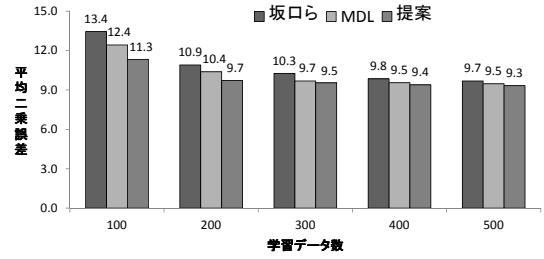


図3. 実験結果

図3より、全ての学習データ数において、提案手法の方が比較手法より低い平均二乗誤差を示した。また、少ない学習データ数において、提案手法が有効であることがわかる。

学習データが十分にある場合は、ノード $s \in \mathcal{S}$ や線形回帰モデル $p \in \mathcal{P}$ の事後確率が1または0に収束し、混合をとる必要がなくなるため、提案手法と比較手法の予測精度の差は小さい。これに対して、学習データが少ない場合は、これらの事後確率が収束していないため、1つのモデルのみ用いている比較手法と、全てのモデルの混合を行った提案手法の予測精度の差が大きくなったと考えられる。

5 まとめと今後の課題

本研究では、考えられる全ての線形回帰モデルを含んだモデルクラスに対し、線形回帰モデルの混合を行いベイズ最適な予測法を提案した。また、人工データを用いた数値実験により提案手法の有効性を示した。今後の課題は、層別に用いる変数の順序の決定方法などが考えられる。

参考文献

- [1] 坂口卓也, “層別回帰モデルの Tree 表現によるベイズ最適な予測法,” 早稲田大学修士論文, 2012.
- [2] T. Matsushima, S. Hirasawa, “Reducing the Space Complexity of a Bayes Coding Algorithm using an Expanded Context Tree,” Proc. ISIT.2009, Seoul Korea, pp.719–723, 2009.
- [3] 繁樹算男, “ベイズ統計入門,” 東京大学出版会, pp.169–180, 2003.