アンサンブル学習を用いた近似ベイズ予測アルゴリズムに関する研究

1X12C004-8　荒井 琢充
指導教員　後藤 正幸

1　研究背景と目的

近年の高度情報化社会においても，離散カテゴリの予測問題は広い適用場面を持つ重要な問題である．この種の予測問題に対し，従来から CHAID, CART, ID3 など様々な決定木作成アルゴリズムが提案されている．これらのアルゴリズムは，考えられ得る決定木モデルの中から何らかのモデル選択基準を最小にする決定木モデルの選択を行う．しかし，有限の学習データから予測を行う問題を考えた場合，この様な唯一の決定木モデルを選択する方法が最適とは限らない．そこで須子らは，与えられた説明変数の並びによって一意に定められる完全木の部分木で与えられるモデルクラスを仮定したもとで，考えられ得る全ての部分木モデルの混合をとり，平均予測誤り率を最小にしたベイズ予測アルゴリズムを提案している [1]．この手法では，与えられた完全木の部分木で表される決定木モデルについては，全モデルが考慮されているので，このような前提がうまく当てはまる対象問題については良い予測性能を発揮する．

しかしながら，一般の予測問題では木を分岐させる説明変数の順番は分析前に決めることができず，全ての並び方の可能性を考慮して予測モデルを構築しなければならない場合の方が多い．また，全ての説明変数を用いて木を構築するため，説明変数が多いデータに対しては，必要となるメモリ量が膨大となりモデルの構築が困難となってしまう．そこで本研究では，全ての説明変数を用いることなく，比較的少ない説明変数を用いた従来のベイズ予測アルゴリズムによる混合モデルを複数構築し，それらをアンサンブルすることで予測を行う新たなアルゴリズムを提案する．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．

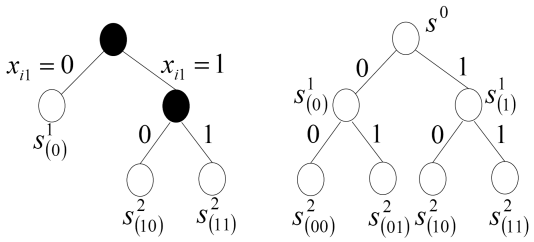
2　従来研究

2.1　問題設定

以下では，説明変数ベクトル $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})$, $(x_{id} \in \{0,1\}, d = 1,2,...,K)$，そのデータが属する 2 値の目的変数 $y_i \in \{0,1\}$ の組を考える．ただし，n 個の学習データを $\boldsymbol{x}^n = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$，$y^n = (y_1, y_2, ..., y_n)$ と表す．また，i 番目の説明変数 $\boldsymbol{x}_i$ と目的変数 y_i の組を $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ とし，その n 個のデータ集合を $\boldsymbol{z}^n = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n)$ と表記する．このとき，$\boldsymbol{z}^n$ が与えられたもとで，$n+1$ 番目の説明変数 $\boldsymbol{x}_{n+1}$ に対応する目的変数 y_{n+1} を予測する問題を扱う．

2.2　決定木モデルの構成

前述の予測問題を扱うため，各枝には説明変数 $x_{id} \in \{0,1\}$，葉ノードには目的変数 y_i が割り付けられた決定木モデルを考える．ここで，説明変数が $x_{i1}, x_{i2}, ..., x_{id}$ の順番で必ず与えられるとし，$\boldsymbol{x}_i = x_{i1}, x_{i2}, ..., x_{id}$ とする．また，$\boldsymbol{x}_i$ が与えられた時，一意に定まる状態を s^d とし，s^d に基づき予測を行う．決定木モデルは木の最大深さ K の部分木で表される．図 1(a) に深さ $K = 2$ の決定木モデルの例を示す．一方，決定木モデルの混合モデルは最大深さの決定木モデルのクラスに属する．いま，混合モデルの各ノードの状態を s とし，全ての s の集合を $\mathcal{S}$ と定義する．このとき，同じ位置にノードを持つ決定木モデルにおいて，それらモデルのノードを状態 s に集約ことにより混合モデルを構成する．図 1(b) に，深さ 2 の混合モデルを示す．



(a) 決定木モデル　　(b) 混合モデル

図 1. 決定木モデル

2.3　効率的なベイズ予測アルゴリズム

須子らは，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，考えられ得る全ての決定木モデルを混合モデルへ集約し，平均誤り率を最小にした効率的なベイズ予測アルゴリズムを提案している [1]．ここで，最大深さが K の決定木モデルにおける $\boldsymbol{x}_i$ に対応する葉ノードを s_i^K, 根ノードを s_i^0, s_i^K と s_i^0 を結ぶパス上のノード集合を $\mathcal{S}_i = \{s_i^0, s_i^1, ..., s_i^K\}$, s に保持されている y_i の生起確率ベクトルを $\boldsymbol{\theta}(s)$ とする．全ての決定木の混合モデルの下で，ベイズ最適な y_i の予測確率は以下のように計算することができる．

$$
\begin{aligned}
&P(y_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}) \\
&= \sum_{s_i \in \mathcal{S}_i} \int_{\boldsymbol{\theta}(s_i)} P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}(s_i), s_i) \\
&\qquad \times\ P(\boldsymbol{\theta}(s_i)|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}, s_i) P(s_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}) d\boldsymbol{\theta}(s_i) \\
&= P_C(y_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}, s = s_i^0) \qquad (1)
\end{aligned}
$$

ただし，$P(y_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1})$ は，式 (2) により葉ノードと根ノードを結ぶパスからの確率 $P_C(y_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}, s_i)$ を再帰的に計算することにより与えられる．s_i' は s_i の子ノードを指すものとする

$$
\begin{aligned}
&P_C(y_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}, s_i) \\
&= \begin{cases}
P(y_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}, s_i^K) & (s = s_i^K) \\
q(s_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}) P(y_i|\boldsymbol{x}_i, s_i) & \\
\quad +(1 - q(s_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1})) P_C(y_i|s_i') & (otherwise)
\end{cases} \qquad (2)
\end{aligned}
$$

ただし s_i の事前確率 $q(s_i|\boldsymbol{z}^i)$ を式 (3) により更新する．

$$
q(s_i|\boldsymbol{z}^i) = \frac{q(s_i|\boldsymbol{z}^{i-1}) P_C(y_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}, s_i)}{P(y_i|\boldsymbol{x}_i, \boldsymbol{z}^{i-1}, s_i)} \qquad (3)
$$

3　提案手法

3.1　概要

須子らの手法では，説明変数 $x_{i1}, x_{i2}, ..., x_{id}$ $(d = 1,2,...,K)$ の並びによって一意に定められるモデルクラスを仮定した下で，ベイズ最適な混合モデルを構築しているため，考慮できていないモデルクラスが存在する可能性がある．例えば，従来手法では上記のような説明変数の並びが与えられた際に，$\{x_{i1}, x_{iK}\}$ のような組み合わせからなる決定木モデルを考慮できていない．加えて，説明変数が多いデータに対してはメモリ量が膨大となり，モデルの構築が困難となってしまう．そこで本研究では，アンサンブル学習の枠組みを援用し，上記の問題を解決する．すなわち，ランダムに比較

的少数の説明変数の組み合わせを複数抽出し，それらを用いて混合モデルを構築する．そのうち，広いモデルクラスを表現できるような混合モデルを予め定めた数だけ残したもとでそれらを組み合わせることにより，モデル集合全体を近似的に網羅したベイズ予測アルゴリズムを提案する．これにより従来手法と比較して，より多くのモデルクラスを考慮することができ，予測精度の観点から優れた新たな手法を示す．

3.2 複数のベイズ決定木の生成

提案手法では，深さごとに割り付ける説明変数をランダムに選択し，木の削減・複製を繰り返しながら，木を任意の深さまで成長させる．まず，全ての説明変数から深さ 1 の決定木を K 個構築し，不純度を基準に木を J 個残す．ただし不純度は，各ノードに含まれる学習データの目的変数の割合により定義する．そして，新たな深さでの説明変数をランダムに割り当てながら $I(>J)$ 個に複製する．このように木の削減と複製を繰り返しながら木を成長させることで，任意の深さの決定木を複数作成していく．ただし，残された J 個の木を I 個に複製する際には，不純度が低い木から順に新たな深さでの説明変数をランダムに割り当てるものとする．すなわち，最終的に作成する木の深さを $D(\leq K)$ と定めると，深さ D の決定木を J 個作成し，それらの木に従来手法の学習アルゴリズムを適用することで目的変数の生起確率を求める．ただし，より広いモデルクラスを考慮するため，新たな深さでの説明変数を選択する際には，複数の木で説明変数の組み合わせの重複を避ける．

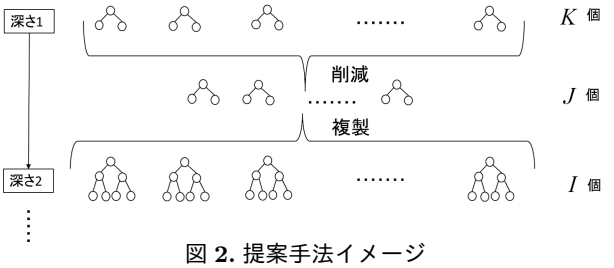


図 2. 提案手法イメージ

3.3 アルゴリズム

いま，l 番目の木の深さ d に割り当てられた説明変数を $a_{l,d}$ $(l = 1, 2, ..., J,\ d = 1, 2, ..., D)$ とし，全ての木における 2 つの説明変数の組み合わせ $\{a_{l,d}, a_{l,d-1}\}$ を成分とする集合を $\mathcal{N}$ と定義する．提案アルゴリズムを以下に示す．

Step1 全ての説明変数を深さ 1 の変数に割り当て，K 個の木を作成し，各木で不純度を計算する．そのうち，不純度の低い木を J 個残す．

Step2 不純度が低い木から順に深さ 2 での説明変数をランダムに割り当て，I 個の木を作成する．

Step3 各木で不純度を算出し，不純度が低い木を J 個残す．

Step4 選ばれた J 個の木における説明変数の組み合わせ $\{a_{l,d}, a_{l,d-1}\}$ を $\mathcal{N}$ に追加する．

Step5 $d < D$ の時，**STEP6** へ．
$d = D$ の時，**STEP7** へ．

Step6 J 個の中で不純度が低い木から順に，$d+1$ での説明変数 $a_{l,d+1}$ をランダムに割り当て，I 個の木を作成し，$d = d + 1$ として，**STEP3** へ戻る．ただし，$a_{l,d+1}$ を選択する際，深さ d で選ばれる説明変数 $a_{l,d}$ との組み合わせが $\mathcal{N}$ に含まれる説明変数を除くものとする．

Step7 不純度が低い木を J 個残し，これらの木構造に対し従来手法のベイズ学習アルゴリズム (式 (1)〜式 (3)) により生起確率 $P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n)$ を算出する．

Step8 各木からの出力を $\hat{y}_l = \arg\max_{y_{n+1}} P(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{z}^n)$ とし，J 個の木の出力から多数決により最終予測を算出する．

4 実験

4.1 実験条件

2 値の 15 個の説明変数とそれに対する 2 値の目的変数からなる人工データを用い実験を行う．ここで，15 個の説明変数のうち，5 つの説明変数で真のモデルを構築し，それに従って目的変数を定める．また，残りの 10 個の説明変数は理論誤差が 0.2 となるようにランダムに値を割り当てる．そして，真のモデル構造が未知であるようにするため，これらの説明変数の並び順をランダムに入れ替えを行うことで人工データを作成した．また，学習データ数を 10 件から 100 件の 10 件刻みとし，目的変数が未知であるテストデータ 1000 件に対して予測を行う．従来手法は深さ $K = 15$ の木を 1 つ作成し，提案手法に用いるパラメータは $D = 5$, $I = 6, 7$, $J = 1, 2$ とする．また，各データセットを 100 セット用意したもとで，予測誤り率を式 (4) で算出し，その平均をモデルの評価指標とする．

$$\text{予測誤り率} = 1 - \frac{\text{正しく予測されたテストデータ数}}{\text{テストデータ数}} \quad (4)$$

4.2 実験結果と考察

図 3 に実験結果を示す．

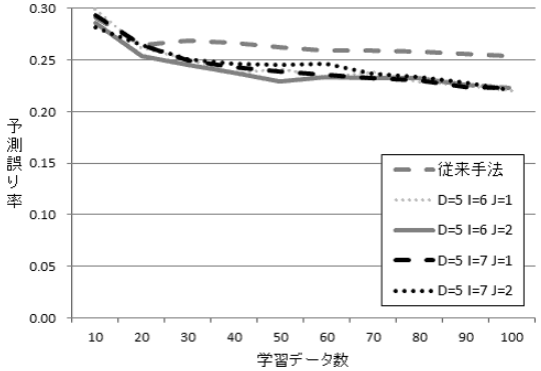


図 3. 実験結果

図 3 より，全ての学習データ数において提案手法が従来手法より低い予測誤り率を示した．提案手法では，多様な説明変数の組み合わせを用いた混合木を作成できたため，精度が向上したと考えられる．また，従来手法では $2^K - 1$ 個のノードが作成されるのに対し，提案手法では $J(2^D - 1)$ 個のノードが作成されるため，$D \ll K$ のとき，メモリ量の観点からも優れているといえる．

5 まとめと今後の課題

本研究では，浅い木をアンサンブルすることで，モデル集合を近似的に網羅したベイズ予測アルゴリズムを提案し，人工データを用いた実験により本手法の有効性を示した．今後の課題としては，実データに対しての評価実験，近似性の理論解析などが挙げられる．

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,” 信学技報, COMP, コンピュテーション, Vol. 103, pp. 93 − 98, 2003.

[2] 松嶋敏泰, 平澤茂一, “定常有限記憶情報源に対するベイズ符号化アルゴリズム,” 信学技報, IT, Vol. 95, No. 79, pp. 1 − 6, 1995.