目的変数がポアソン分布に従う決定木モデルにおけるベイズ最適予測アルゴリズム

1X08C120-1　峯苫和史
指導教員　後藤正幸

1　研究背景と目的

従来，交通事故件数などのリスク発生頻度の予測としてポアソン回帰分析 [1] が適用されている. このモデルは目的変数がポアソン分布に従うと仮定し，説明変数との線形関係で表されたモデルである.

しかし，データが得られたもとで，説明変数と目的変数の関係に線形性が仮定できなかったり，説明変数が離散で交互作用があるケースにおいてポアソン回帰モデルでは適用が困難である. そこで，本研究ではこのようなデータに対してデータマイニングやパターン認識技術の中で学習と予測の有用性が示されている決定木モデルの適用を考える.

さらに，この決定木モデルのもとで予測を行う場合に，考えうる全ての決定木モデルの混合モデルを考えることでベイズ最適な予測分布を構成することができる [2][3].

本研究では，予測対象がポアソン分布に従う場合の決定木モデルについて，効率的にベイズ最適な予測値を計算する予測アルゴリズムを提案し，シミュレーション実験を通じて，提案手法の有効性を示す.

2　従来手法

2.1　問題設定

あるデータとして K 次元離散属性ベクトル$\boldsymbol{x} \in \{0, 1\}^K$ とそのデータが属するカテゴリ y の組を考える. いま，$\boldsymbol{x_i}$, y_i をそれぞれ i 番目のデータとし，その n 個のデータ列を $\boldsymbol{x^n}=\boldsymbol{x_1x_2}\ldots\boldsymbol{x_n}$，$y^n = y_1y_2\ldots y_n$ と表す.

また i 番目の$\boldsymbol{x}$と y の組を $\boldsymbol{z_i} = (\boldsymbol{x_i}, y_i)$ とし，その n 個の集合を $\boldsymbol{z^n} = \boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_n}$ と表記する．予測問題として$\boldsymbol{z^n}$が得られたもとで，$\boldsymbol{x_{n+1}}$に対応するカテゴリ y_{n+1} を逐次的に予測する問題を考える.

2.2　ポアソン回帰モデル

ポアソン回帰モデルはカテゴリ y を式 (1) のポアソン分布に従うと仮定した回帰モデルであり，パラメータ $\boldsymbol{a}_i$ を用いて平均値 λ を式 (2) のように表す.

$$P(y_i|\lambda) = \frac{\lambda^{y_i}}{y_i!}e^{-\lambda} \tag{1}$$

$$\lambda = e^{\sum_i \boldsymbol{a}_i\boldsymbol{x}_i} \tag{2}$$

2.3　決定木モデル

ポアソン回帰モデルに対し，決定木モデルでは学習データを属性値によって階層的に部分集合に分割し，そのもとで葉ノードの分布のパラメータを学習する．そのため，決定木モデルでは説明変数と目的変数の関係に線形性が仮定できないデータに対しても適用することが可能である.

2.4　混合決定木モデル

須子らの手法 [2] や坂口らの手法 [3] では，この決定木モデルによる学習と予測に対して, 松嶋らによるベイズ符号アルゴリズム [4] 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，ベイズ最適な予測アルゴリズムを示している．前述の予測問題を扱う上で，決定木モデルクラスの$\boldsymbol{x}$に対する質問の内容を $\psi_d(d = 1,\ldots,D)$ とし，$\omega_{\psi_d}(\boldsymbol{x}) \in \{0, 1\}$ を質問 ψ_d に対して$\boldsymbol{x}$が真 (1) か偽 (0) を返す関数とする. いま，質問の順番が $\psi_1,\ldots,\psi_D$ として既に与えられているものと仮定し，質問 $\psi_1,\ldots,\psi_D$ に対する $\omega_{\psi_d}(\boldsymbol{x})$ の系列を ω^d=$\omega_{\psi_1}(\boldsymbol{x}),\ldots,\omega_{\psi_D}(\boldsymbol{x})$ とする. ω^d と x により一意に定まる状態を s_{ω^d} として，状態 s_{ω^d} にもとづき予測を行う.

図 1 の左図は深さ D=2 の決定木モデルの例である. 予測対象である y の分布パラメータは，葉ノードのみに与えられる. ここで，このような全ての決定木モデルの混合をとるために K 個の属性と y の関係性を考慮し，図 1 の左図の全ての部分木がモデルの候補となる.

混合決定木モデルを最も深い深さ D の木で表現する. また混合決定木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，同じ位置に葉を持つ決定木モデルの葉ノードに対応させる. 例として深さ D=2 における混合決定木モデルは図 1 の右図で表現することができる.

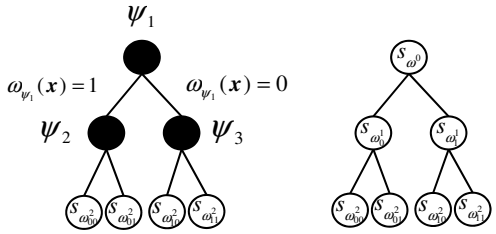

図 1. 混合決定木モデル

この混合決定木モデルを用いて須子らは予測対象に二項分布を仮定し，坂口らは正規分布を仮定したもとで効率的に予測分布を計算するアルゴリズムを示している. 本研究では予測対象にポアソン分布を仮定したもとで効率的に予測分布を計算するアルゴリズムを提案する.

3　提案手法

3.1　問題設定

$\boldsymbol{x}$と y の n 個の組 $\boldsymbol{z^n}$ が得られたとき $\boldsymbol{x_{n+1}}$ に対応するポアソン分布に従うカテゴリ y_{n+1} を逐次的に予測する問題を考える.

3.2　効率的予測値計算アルゴリズム

予測対象が可算無限の離散値をとるため二乗誤差損失 $Loss_1$ を考える.

$$Loss_1 = (y_{n+1} - \hat{y}_{n+1})^2. \tag{3}$$

このとき，y_{n+1} のベイズ最適な予測値 $\hat{y}_{n+1}$ は以下の式で求められる.

$$\begin{aligned}\hat{y}_{n+1} &= \sum_{y_{n+1}} y_{n+1} \sum_{m\epsilon M} \int_{\lambda_m} P(y_{n+1}|\boldsymbol{x_{n+1}}, \boldsymbol{z^n}, \boldsymbol{\lambda_m}, m) \\ &\qquad P(\lambda_m|m, \boldsymbol{z^n})P(m|\boldsymbol{z^n})d\boldsymbol{\lambda_m}dy_{n+1} \\ &= \sum_{m\epsilon M} \bar{y}_{n+1}(\boldsymbol{x_{n+1}}, \boldsymbol{z^n}, m)P(m|\boldsymbol{z^n}) \end{aligned} \tag{4}$$

ここで，$m \in M$ は 1 つの決定木モデルを示しており，$\boldsymbol{\lambda_m} \in \Lambda_m$ はモデル m の未知のパラメータである．

式 (4) は予測分布の平均値を表しており，考えうる全ての決定木モデル m を混合しているが，最大深さ D が大きくなると考慮すべきモデル数 $|M|$ は指数的に増加する．そこで，松嶋らにより提案された効率的計算アルゴリズムを応用することで，図 1 の混合決定木モデルのもとで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_{ω^d} の事後確率 $P(s_{\omega^d}|\boldsymbol{z^n})$ は混合決定木モデルの各状態が持っている重みパラメータ $q(s_{\omega^d}|\boldsymbol{z^n})$ を用いて式 (5) で求めることができる．

$$P(s_{\omega^d}|\boldsymbol{z^n}) = q(s_{\omega^d}|\boldsymbol{z^n}) \prod_{i=0}^{d} (1 - q(s_{\omega^i}|\boldsymbol{z^n})) \tag{5}$$

式 (4) の右辺の予測分布 $P(y_{n+1}|\boldsymbol{x_{n+1}}, \boldsymbol{z^n})$ は式 (5) 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根から葉までの 1 つのパス上の状態列 $s_{\omega^0}, s_{\omega^1}, \ldots, s_{\omega^D}$ に対して再帰計算として次式で求めることができる．

$$P(y_{n+1}|\boldsymbol{x_{n+1}}, \boldsymbol{z^n}) = q(y_{n+1}|\boldsymbol{z^n}, s_{\omega^0}) \tag{6}$$

$$q(y_{n+1}|\boldsymbol{z^n}, s_{\omega^d}) = q(s_{\omega^d}|\boldsymbol{z^n}) P(y_{n+1}|\boldsymbol{z^n}, s_{\omega^d}) + (1 - q(s_{\omega^d}|\boldsymbol{z^n})) q(y_{n+1}|\boldsymbol{z^n}, s_{\omega^{d+1}}) \tag{7}$$

本研究では，予測対象である目的変数 y が $\boldsymbol{x}$ のポアソン分布に従うことを仮定するため，ポアソン分布に対して自然共役事前分布である以下のガンマ分布 $Ga(\alpha, \beta)$ を各状態 s におけるパラメータ $\lambda_m(s)$ の事前分布として設定する．

$$P(\lambda_m(s)) \sim Ga(\alpha_0(s), \beta_0(s)) \tag{8}$$

ここで $\alpha_0(s)$ と $\beta_0(s)$ は状態 s における事前分布のパラメータを表している．式 (8) の事前分布をもとにベイズの定理を用いて推測を行うことで事後予測分布 $P(y_{n+1}|\boldsymbol{z^n}, s_{\omega^d})$ を, 次の式で与えられるポアソンガンマ分布 $Pg(\alpha,\beta)$ として求めることができる．

$$P(y_{n+1}|\boldsymbol{z^n}, s_{\omega^d}) \sim Pg(y_{n+1}|\alpha'_{s_{\omega^d}}, \beta'_{s_{\omega^d}}) \tag{9}$$

ここで, $\alpha'_{s_{\omega^d}}$ と $\beta'_{s_{\omega^d}}$ は状態 s_{ω^d} ごとにもつパラメータであり，各状態 s におけるカテゴリの和 $\sum y_{s_{\omega^d}}$ とカテゴリの出現回数 $n_{s_{\omega^d}}$ によって次式で与えられる．

$$\alpha'_{s_{\omega^d}} = \alpha_{s_{\omega^d}} + \sum y_{s_{\omega^d}}, \quad \beta'_{s_{\omega^d}} = \beta_{s_{\omega^d}} + n_{s_{\omega^d}} \tag{10}$$

式 (9) を用いて式 (7) の平均値を変形することで $\hat{y}_{n+1}$ は $\boldsymbol{x_{n+1}}$によって一意に定まる状態列 $s_{\omega^0}, s_{\omega^1}, \ldots, s_{\omega^D}$ の平均値 $\bar{y}_{s_{\omega^0}}, \bar{y}_{s_{\omega^1}}, \ldots, \bar{y}_{s_{\omega^D}}$ を用いて以下の再帰計算で求めることができる．

$$\hat{y}_{n+1} = \bar{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^0}) \tag{11}$$

$$\bar{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^d}) = q(s_{\omega^d}|\boldsymbol{z}^n) \bar{y}_{s_{\omega^d}} + (1 - q(s_{\omega^d}|\boldsymbol{z}^n)) \bar{y}_{n+1}(\boldsymbol{z}^n, s_{\omega^{d+1}}) \tag{12}$$

4 数値実験と結果

提案手法の有効性を検討するために，数値実験を行なった．比較対象として，一般化線形モデルによるポアソン回帰分析を扱う．

4.1 実験条件

木の最大深さを $D = 2$ と仮定する．データ長 $n = 150$ までの逐次予測の実験を 1 セットとし，繰り返し 10 セット実験を行う．

比較手法として一般化線形モデルを用いたポアソン回帰式を実験データセット毎に算出し，テストデータ 1000 件に対しての予測を行う．

また，真のモデルの構造は最大深さ $D = 2$，分岐数 $L = 2$ とする．その構造のもとで葉ノードの出現確率を等確率とし，各葉ノードにおけるポアソン分布のパラメータは，平均予測誤差理論値が $\lambda = 3.0$ となるように $\lambda = 2.0$,$\lambda = 4.0$ のいずれかを与えて実験を行った．

4.2 実験結果及び考察

図 2 に実験結果を示す．横軸はデータ長 n，縦軸は平均二乗誤差損失 $\bar{L}oss_1$ を示した．ここで，平均予測誤差理論値 $\lambda = 3.0$ までの収束過程を示している．

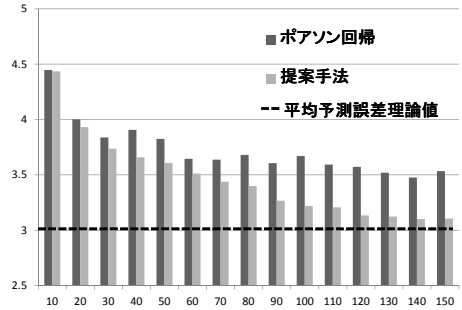

図 2．実験結果

図 2 より提案手法の方が一般化線形モデルを用いたポアソン回帰分析よりも早く誤差が減少することがわかる．これは，ポアソン回帰モデルとして 1 つのモデルを選択するよりも交互作用を含むデータに対して決定木モデルの混合をとる提案手法の方が，予測精度が高いことを示している．

5 まとめ

本論文では，予測対象としてポアソン分布に従う可算無限の離散値データを扱う場合を想定し，混合決定木モデルのもとで予測値の効率的計算アルゴリズムを考え, 数値実験によりその有効性を示した．また，一般化線形モデルを用いたポアソン回帰よりも混合決定木モデルの方が予測精度が優れていることを示した. 今後の課題としては実問題の適用と評価を考えていくことがあげられる．

参考文献

[1] 州浜源一,“ 計数データと回帰分析：中国地域の交通事故発生モデルの展開, ”尾道大学経済情報論集 3(2),pp. 1–9, Dec., 2003.

[2] 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, ”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.

[3] 坂口卓也, 石田崇, 後藤正幸, 寺本賢一, “ 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, ”経営情報学会 秋季全国研究発表大会, Nov., 2010

[4] T.Matsushima and S.Hirasawa, “ Universal coding algorithms FSMX sources based on bayes coding, ” IEEE IT. ISIT., 1994