決定木の信頼上界を活用した文脈付きバンディットアルゴリズムに関する研究

1X20C012-1　大岩将
指導教員　後藤正幸

1. 研究背景・目的

近年，イーコマースを筆頭に様々な領域で推薦システムが活用されている．特に，データが追加されるたびに推薦アルゴリズムを学習するモデルをオンライン推薦システムと呼び，その 1 つに文脈付きバンディットアルゴリズムが挙げられる．文脈付きバンディットアルゴリズムでは，ユーザの属性や過去の購買履歴などの「文脈情報」から，商品の購入有無などを「報酬」として推定し，なるべく報酬が高い商品を推薦する．この手法では，与えられた文脈情報に対して，期待報酬が不明な商品を試しに推薦する「探索」と最も期待報酬の大きいと思われる商品を推薦する「活用」のバランスを取り累積報酬の最大化を図る．

文脈付きバンディットアルゴリズムの中でも，文脈-報酬間に非線形な関係が存在する場合に適した手法として Elmachtoub らの手法 [1] が挙げられる．この TreeBootstrap と呼ばれる手法は，推薦の選択肢 (以下，行動) ごとの期待報酬の推定に決定木を用いることにより，文脈-報酬間の非線形な関係性を捉えている．この手法では，決定木の学習データをブートストラップサンプリングにより作成し推薦行動にランダム性を付与することで，探索と活用のバランスを取っている．しかし，学習データのサイズが小さい場合，学習データは外れ値の影響を受けやすいため，母集団の傾向を反映しないことが多い．この状況でブートストラップサンプリングを行う場合には誤った学習をしてしまう危険があり，ひいては累積報酬の最大化の妨げになり得る．

そこで本研究では，探索と活用のバランスを学習データのブートストラップサンプリングではなく，行動毎の期待報酬の信頼上界により評価する方法を用いる TreeUCB を提案する．具体的には，各決定木から得られる期待報酬の信頼上界に基づき行動を決定し探索と活用のバランスを取ることにより，累積報酬の最大化を図る．人工データを用いた実験により，提案手法の有効性を示す．

2. 関連研究

2.1. 文脈付きバンディットアルゴリズム

文脈付きバンディットアルゴリズムとは，逐次的に推薦と学習を行うオンライン推薦システムの 1 つである．各時刻 $t = 1, 2, ..., T$ において，文脈情報 $\boldsymbol{x}_t \in \mathbb{R}^d$ を用いて行動の集合 A から推薦行動 $a_t \in A$ を 1 つ決定し，得られた報酬 $r_t \in \{0, 1\}$ を用いてモデルの学習を行う．文脈付きバンディットアルゴリズムの概要を図 1 に示す．

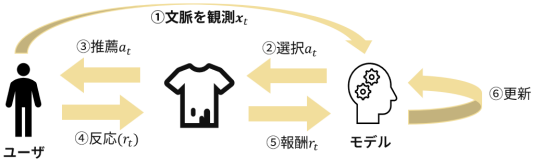

図 1: 文脈付きバンディットアルゴリズムの流れ

2.2. TreeBootstrap

TreeBootstrap は，文脈情報と報酬の関係性を決定木により学習する文脈付きバンディットアルゴリズムである．決定木は各時刻 t の各行動 a における報酬を，文脈情報から推定するように作成される．各決定木の学習は，時刻 t までに得られた文脈情報と報酬のデータ組 $(\boldsymbol{x}_t, r_t)$ の集合 $D_{t,a}$ から $|D_{t,a}|$ 回の復元抽出により作成した $\tilde{D}_{t,a}$ を用いて行われる．ブートストラップサンプリングにより得られたデータの不確実性が学習データに反映され，探索と活用のバランスが調整されることにより累積報酬の最大化を促している．TreeBootstrap のアルゴリズムを以下に示す．

Algorithm 1 TreeBootstrap

for $t = 1, 2, ..., T$ **do**
　文脈情報 $\boldsymbol{x}_t$ を観測する
　for $a = 1, ..., |A|$ **do**
　　$D_{t,a}$ から $\tilde{D}_{t,a}$ を得る
　　$\tilde{D}_{t,a}$ から決定木 $\hat{\theta}_{t,a}$ を得る
　end for
　行動 $a_t = \arg\max_a p(\hat{\theta}_{t,a}, \boldsymbol{x}_t)$ を選択する
　得られたデータ組 $(\boldsymbol{x}_t, r_{t,a_t})$ を D_{t,a_t} に追加する
end for

ただし，$p(\hat{\theta}_{t,a}, \boldsymbol{x}_t)$ は決定木の出力 (期待報酬) で，時刻 t における文脈情報 $\boldsymbol{x}_t$ が該当する葉ノードに蓄積しているデータの平均報酬である．なお，各決定木は CART アルゴリズムを用いて学習される．

3. 提案モデル (TreeUCB)

3.1. 着想

TreeBootstrap はブートストラップサンプリングを行うことで探索と活用のバランスを取るモデルであるが，学習データ数が小さい時に外れ値の影響を受けやすく，誤った学習を行い活用のフェーズに入ってしまうことで累積報酬の最大化を妨げる原因の 1 つになっている．一方，UCB[2] は期待報酬の信頼上界を利用しており，外れ値が存在していても信頼上界により探索を続けやすい．このことから，TreeBootstrap における探索と活用のバランスを取る機構をブートストラップサンプリングから提案手法の信頼上界を用いた手法に変更することで，累積報酬の更なる最大化が期待される．

3.2. 提案手法

提案手法 (以下，TreeUCB) では，TreeBootstrap におけるブートストラップサンプリング処理を省略し，代わりに各決定木の出力 $\hat{p} = p(\hat{\theta}_{t,a}, \boldsymbol{x}_t)$ の信頼上界を新たに $\bar{p}(\hat{\theta}_{t,a}, \boldsymbol{x}_t)$ として，各時刻における推薦行動の選択に利用する．

信頼上界は，ある決定木の葉ノード s におけるデータ数を

N_s，そのうち報酬が 1 であるデータ数を K_s とした時，K_s が母比率 p，試行回数 N_s の二項分布 $B(N_s, p)$ に従うと仮定して計算する．特に N_s がある程度大きい場合に中心極限定理を用いて二項分布 $B(N_s, p)$ が正規分布 $N(Np, Np(1-p))$ に近似できることを利用すると，母比率 p の信頼水準 $\alpha/2$ の信頼上界は式 (1),(2) で表すことが出来る．ここで，$z_{\alpha/2}$ は標準正規分布における上側確率が $\alpha/2$ となる値である．

$$\hat{p} = K_s/N_s \tag{1}$$

$$p \leq \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/N_s} = \bar{p}(\hat{\theta}_{t,a}, \boldsymbol{x}_t) \tag{2}$$

また，TreeUCB のアルゴリズムを以下に示す．

Algorithm 2 TreeUCB(提案手法)

for $t = 1, 2, ..., T$ **do**
  文脈情報 $\boldsymbol{x}_t$ を観測する
  for $a = 1, ..., |A|$ **do**
    $D_{t,a}$ から決定木 $\hat{\theta}_{t,a}$ を得る
  end for
  行動 $a_t = \arg\max_a \bar{p}(\hat{\theta}_{t,a}, \boldsymbol{x}_t)$ を選択する
  得られたデータ組 $(\boldsymbol{x}_t, r_{t,a_t})$ を D_{t,a_t} に追加する
end for

4. 人工データによる実験

提案手法の有効性を検証するため，人工データを用いたオフライン評価により提案手法と各種バンディットアルゴリズムの性能比較を行う．

4.1. 実験条件

本実験では，3 次元の文脈情報 $x_1 \in \{0, 1, 2, ..., 9\}, x_2 \in \{0, 1, 2, 3, 4\}, x_3 \in \{0, 1, 2, ..., 9\}$ と 3 種類の行動を離散一様分布から生成する．その後，予め行動ごとに作成した決定木から得られる期待報酬に従って報酬を生成し，人工データ (ログデータ) を作成する．実験は，1 セットをラウンド数 $T = 1,000$ として，手法ごとに 10 セット実行する．評価指標には，各手法の平均報酬±信頼区間 (標準偏差) 及び平均実行時間を用いる．実験に用いる各手法の概要，及びハイパーパラメータの設定値を表 1 に示す．

表 1: 各手法の概要とハイパーパラメータ設定

名称	文脈の利用	概要
ε -greedy($\epsilon = 0.05$)	×	確率 ε で探索，1- ε で活用を行う．
UCB($\alpha = 0.05$)	×	文脈情報を使わない．信頼上界を用いる．
LinUCB($\alpha = 0.05$)	○	文脈-報酬間に線形性を仮定する．
TreeBootstrap	○	報酬の推定に決定木を用いる．
TreeUCB($\alpha = 0.05$)	○	提案手法

人工データを用いて，複数モデルを同一環境下で比較するオフライン評価を行う．

4.2. 実験結果

各手法の平均報酬±標準偏差と平均実行時間を表 2 に，実験 1 セット目の各ラウンドにおける各手法の平均報酬の推移を図 2 に示す．

表 2: 各手法の平均報酬と平均実行時間

名称	平均報酬±標準偏差	平均実行時間 (s)
ε -greedy($\epsilon = 0.05$)	0.4505 ± 0.0574	0.0969
UCB($\alpha = 0.05$)	0.4681 ± 0.0049	0.135
LinUCB($\alpha = 0.05$)	0.4235 ± 0.0491	0.286
TreeBootstrap	0.5613 ± 0.0090	**25.21**
TreeUCB($\alpha = 0.05$)	**0.6068** ± 0.0093	**5.55**

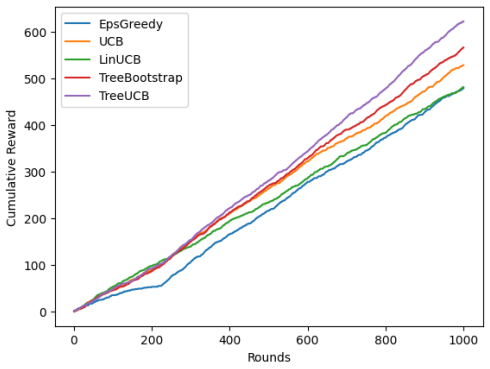

図 2: 各手法の累積報酬の推移

結果より，提案手法がベース手法の TreeBootstrap を含めたすべての従来手法よりも高い報酬を得ており，最も優れた性能を示していることが分かる．また提案手法とベース手法の標準偏差を比較すると，提案手法がベース手法とほとんど変わらず結果のばらつきを抑えていることが分かる．加えて，実行時間に関してもベース手法の 2 割程度に削減できていることが分かる．

5. 考察

図 2 において，提案手法は最も優れた累積報酬を記録している．これは提案手法の決定木の学習において，母比率の信頼上限がより適切に探索と活用のバランスを調整していることを示唆している．一方で，ラウンド数が小さい場合の提案手法の累積報酬は他の一部手法と比べて劣っている．これは母比率の信頼区間の計算において，データ数 N_s がある程度大きい場合でないと二項分布から正規分布への近似が不安定になることに起因していると考えられる．

6. まとめと今後の課題

本研究では，決定木を用いた文脈付きバンディットアルゴリズムに，信頼上界の考え方を導入した TreeUCB を提案した．人工データを用いた実験では，精度・計算量の両面で，TreeBootstrap を含めた他手法に対する提案手法の有効性が確認できた．

今後の課題としては，実データによる実験や，達成しうる最大の報酬と提案手法の期待報酬との差を解析する後悔分析などが挙げられる．

参考文献

[1] Sechan Oh Marek Petrik Adam N. Elmachtoub, Ryan MeNellis. A practical method for solving contextual bandit problems using decision trees. *Proceedings of the 33rd Conference on Uncertainity in Artificial Intelligence*, 2017.

[2] Paul Fischer Peter Auer, Nicolò Cesa-bianchi. Finite-time analysis of the multiarmed bandit problem. *15th International Conference on Machine Learning*, pp. 235–256, 2002.