5G-04

バンディット問題による主観的リターンの観測可能性について

仲里　慎司†　　下川　哲矢‡

東京理科大学　経営学研究科†　　東京理科大学　経営学部ビジネスエコノミクス学科‡

1, 本研究の位置づけ

日々の人間活動は,購入する商品の選択や取引先の選択など，不確実性下における選択問題の連続である．この意思決定問題は一種の探索問題とみなすことができ，どのようにして環境の情報探索と知識利用を行うかの戦略を適切に決定することは重要である． このバランスの問題は the exploration-exploitation dilemma(trade-off)として知られており， Multi-Armed Bandit(MAB)問題としてその枠組みが一般化されている。近年，これらの研究成果を人間行動モデリングに応用する潮流がある．

最近では Wu et al.(2018)[1]や Schulz et al.(2018)[2]に代表されるように，より現実的な環境設定として， 広大な選択肢空間を扱った MAB 問題の研究が増えている．これら研究結果の中で特筆すべきは,Gaussian Process(GP)と Upper Confidence Bound(UCB)方策を採用したモデルが人間行動モデルとしても有効であることが実証された点である．特に Wu らは,これと Softmax 型型選択関数を組み合わせたモデルが， 27 の有力な行動モデルのなかで最も記述精度が良いことを報告している． また Schulz らは， このモデルをフードデリバリーサービスのビッグデータに応用し，その有効性を実証している．これらの結果は，エピソード記憶による学習のモデルとしてだけでなく，認知科学的観点からも，GP に基づいた類推過程のモデルとして意味を持ち，商品選択など選択肢が非常に多い様々な経済活動に応用できると期待される．

しかしながら，同モデルを幅広い問題に応用するためには 1 つのボトルネックが存在する．それは， GP による期待のアップデートのために,主体が選択から得られたリターンを分析者が観測する必要がある点である． Schulz et al.(2018)では，このリターンに事後的な評価のフィードバックを用いており，それが得られなかった選択に関しては分析から除外されている．だが一般的には,主体が選択から得たリターンは主観的なものであり，観測者の手に入らないことがほとんどである． そこで本研究では， MAB 問題で定式化される広域探索問題において，主体が得たリターンが観測できない場合に選択の極限分布を用いて顕示する手法を提案する．また，提案手法の有効性を,Wu らの実験を参考にして行った独自実験のデータと，現実のマーケティングデータを用いて示す.

2, 問題・モデル設定

主体が選択肢集合Dの中から任意の選択肢$x \in D$を選択する計T回の通時的な意思決定問題を考える． 主体は,毎期自身の選択からリターン$y = f(x) + \varepsilon$ を観測する． ここで$f: D \to \mathbb{R}$は未知の利得関数であり，ε は分散σ^2 をパラメータとするガウスノイズ$\varepsilon \sim N(0, \sigma^2)$である． この時，意思決定主体の目的関数は以下で与えられる．

$$\max_{\{x_1,\cdots,x_T\}} \sum_{t=1}^{T} f(x_t)$$

続いて，先行研究にならい，リターンに関する期待更新を GP で行うものとする． したがって，利得関数fが平均μ, 共分散行列Kの GP に従うとし，このことを

$$f \sim GP(\mu, K)$$

と表記する．ただし，共分散行列の各要素はカーネル関数$k(x, x')$から与えられる． また， 任意のt期における過去の選択ベクトルとリターンベクトルをそれぞれ$x_{1:t-1} = [x_1, \cdots, x_{t-1}]^\top$， $y_{1:t-1} = [y_1, \cdots, y_{t-1}]^\top$とする． この時， t期における任意の選択肢xの利得は

$$f(x) \sim N(\mu_t, \sigma_t^2), \quad where$$
$$\mu_t(x) = k(x, x_{1:t-1})[k(x_{1:t-1}, x_{1:t-1}) + \sigma_n^2 I]^{-1} y_{1:t-1}$$
$$\sigma_t^2(x) = k(x, x)$$
$$-k(x, x_{1:t-1})[k(x_{1:t-1}, x_{1:t-1}) + \sigma_n^2 I]^{-1} k(x, x_{1:t-1})^\top .$$

のガウス分布で与えられる．ただし，$k(x, x_{1:t-1})$はxと$x_{1:t-1}$間のカーネル値を計算した列ベクトルであり，同様に$k(x_{1:t-1}, x_{1:t-1})$は$x_{1:t-1}$の各要素同士のカーネル値の行列である．ここでは一般化を失わず$\mu_0(x) = 0, \sigma_0^2(x) = k(x, x) > 0$とする．

GP-UCB 方策では， 上記の GP 回帰の結果を用いて通常の UCB 探索を行う．すなわち選択肢xの価値評価を，α_1, α_2を正の定数として

$$UCB_t(x) = \alpha_1 \mu_t(x) + \alpha_2 \sigma_t(x)$$

と与える．選択関数は，実証的な当てはまりの良

The observability of subjective evaluation by the Multi-Armed Bandit problem

†Shinji Nakazato, Graduates School of Management, Tokyo University of Science

‡Tetsuya Shimokawa, School of Management, Tokyo University of Science

さから認知科学分野で通常採用される，以下のSoftmax型とする．

$$p_t\ (x) = \frac{exp\{UCB_t(x)\}}{\sum_{x' \in D} exp\{UCB_t(x')\}}$$

3，提案手法

この，GP-UCB-Softmax型モデルでは，t期における任意の選択肢$x_i, x_j \in D$間での選択確率のオッズ比r_t^{UCB}が以下で与えられる．

$$r_t^{UCB}(x_i, x_j) = \frac{exp\{UCB_t(x_i)\}}{exp\{UCB_t(x_j)\}}$$

ここで，ベンチマークとして，選択肢xの評価を真の利得を用いて$\alpha_1 f(x)$とした場合の選択オッズ比r_t^*を考える．

$$r_t^*(x_i, x_j) = \frac{exp\{\alpha_1 f(x_i)\}}{exp\{\alpha_1 f(x_j)\}}$$

この時，以下の命題が成立する．

命題

主体の期待形成がGPに従っているとし，GP-UCB-Softmax型モデルが採用されているとする．この時，任意の正の定数$\varepsilon > 0$に対してあるτが存在し，任意の$t > \tau$と，任意の選択肢$x_i, x_j \in D$について

$$P\left\{\left|log\frac{r_t^{UCB}(x_i, x_j)}{r_t^*(x_i, x_j)}\right| > \varepsilon\right\} = 0$$

となる．

すなわち，十分な選択後，GP-UCB-Softmax探索を採用する主体の選択分布のオッズ比r_t^{UCB}は，r_t^*に従うことになる．このことは，上記r_t^*の定義式と合わせれば，主体の選択のオッズ比さえわかれば、主体の主観的なリターンを以下の近似式を用いて顕示できることを意味する．

$$f(x_i) = f(x_j) + \frac{log\left(r_t^{UCB}(x_i, x_j)\right)}{\alpha_1}$$

提案手法

(Step1) ベンチマークとなる選択肢$x_B \in D$を決め，そのリターンを$f(x_B) = 0$とする．

(Step2) 任意の選択肢$x_i \in D$について，オッズ比$r^*(x_i, x_B)$を計算し，顕示リターン$f^R(x_i)$を

$$f^R(x_i) = f(x_B) + \frac{log\left(r^*(x_i, x_B)\right)}{\alpha}$$

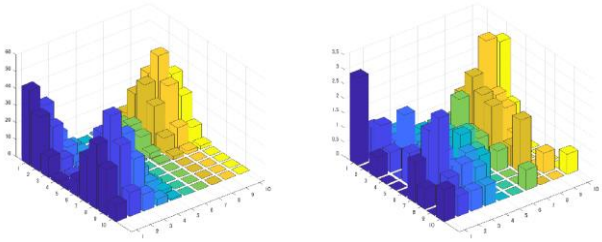
であるとする．ただし$\alpha > 0$の定数とする．

(Step3) $x_i \in D$を選択した場合のリターン$y(x_i)$を顕示リターン$f^R(x_i)$として，通常のGP-UCB-Softmaxモデルを推定する．

4，実証

私たちは提案手法の有効性の検証を，Wuら[1]の実験を発展させておこなった独自実験のデータと，現実のマーケティングデータを用いて行った．

我々が行った実験は一般的なMAB問題であり，10×10の計100本のアームに滑らかなリターン分布を設定し，計30回の選択内に得るリターンの総和を最大化させることを目的としたものである．ただし，Wuらの実験と異なり，eyetrackerとfunctional Near-Infrared Spectroscopyを用いて，同時に生体情報の取得も行った．以下は分析結果の一部である．



図：真のリターン(左)と推定されたリターン(右)

左が我々によって設定された真のリターン平面，右が被験者の選択履歴に提案手法を用いて推測されたリターン平面である．30回という比較的短い期間の選択問題ではあるものの，真のリターン平面の大まかな形状は表現できている．

また提案手法のより現実的な応用として，IDR情報学研究レポジトリを通して提供されるインテージ社の飲料水の購買に関するパネルデータを利用し，現実データにおける提案手法の有効性を検証した．この検証から得られた結果として特筆すべき点は，一部のデータプールからでも，全データを利用した際のリターン平面を比較的高い精度で予測できる可能性が示唆されたことである．具体的には，対象となる消費者のうち，利用頻度の高いユーザー半数の中からランダムに15%のユーザーを抽出し，そのユーザーの購買履歴の直近10%ほどのデータという僅かな情報の獲得で，リターン平面を比較的高い精度で予測できる可能性が得られた．

[主要参考文献]

[1] Wu, Charley M., et al. "Generalization guides human exploration in vast decision spaces." Nature human behavior 2.12 (2018): 915-924.

[2] Schulz, Eric, et al. "Structured, uncertainty-driven exploration in real-world consumer choice." Proceedings of the National Academy of Sciences 116.28 (2019): 13903-13908.