7L-04

認知特性を実装した価値関数による非定常環境への適応

甲野　佑 *1　　高橋　達二 *2

*1東京電機大学大学院　　*2東京電機大学理工学部

1. はじめに

捉えきれない形で変化する環境で実際に生きる人間は，上手く柔軟に情報収集とそのコストのバランスを取りつつ意思決定していると考えられる．そのような人間の傾向に習った Loosely symmetric (LS) 価値関数が存在する．LS は選択肢の評価に用いる事で，内的な基準とコストの釣り合いと選択肢に関する情報の信頼度を考慮して探索の打ち切りと再開を行う傾向 (満足化) を意思決定に反映できる．本研究では基準を動的に変更して探索を柔軟に行える拡張型 LS(LSX) が，非定常多本腕バンディット問題を通して，非定常環境下で特殊な意思決定構造がなくとも LSX が既存アルゴリズムより良く振舞う事を示す.

2. 多本腕バンディット問題と環境の非定常さ

多本腕バンディット問題とは目的となる報酬を確率的に得る事の出来る幾つかの選択肢集合 $A = \{a_1, a_2, ..., a_n\}$ から最適な手段を探索し, 得られる報酬を最大化させる事を目的とする問題である．最も効率の良い選択肢を知るために, 環境を探索するため多くの試行を費やさなければならない．高い報酬を得るためにはどこかで探索を辞めるべきである．しかし充分に探索しなければ高い報酬を得る事はできない．多本腕バンディット問題はこのような知識の獲得とその利用についての普遍的な “速さ” と “正確さ” のトレードオフを端的に表す課題である．さらに現実的には対応すべき環境は非定常であり，探索の再開を判断する必要が有る．これに複雑な準備をせず, かつ早く簡便に対応するのは難しい.

2.1 メタバンディットアルゴリズム

非定常環境での多本腕バンディット問題には一般的に メタバンディットアルゴリズムと呼ばれる特殊な構造を持つアルゴリズムが用いられる [Hartland 06]. 環境の変化は Page-Hinkley 統計量を用いて検出し，初期化した新規エージェントを新たに生成して，任意の期間，旧エージェントと比較してどちらが多くの報酬を得られるかというバンディット問題を行い，劣っていた方を破棄して通常の意思決定に戻る．メタバンディットは一般に特殊な前提を置かずに定常環境で最も優れた成績を有するアルゴリズムの一種である UCB1-tuned [Wang 05] を評価式としており，以後は Meta UCB1-tuned と呼称する.

本来，検出すべき多本腕バンディット問題の環境の変化には，(1) 最適な選択肢の報酬確率が変化し他の選択肢の報酬確率以下に減少する，(2) 他の選択肢の報酬確率が変化し最適な選択肢の報酬確率以上に上昇する，(3) 前述の変化が両方起こり逆転する，がある．メタバンディットは (1) は検出できるが，その他二種の変化は検出できないという問題が存在する [Hartland 06].

3. 満足化方策と基準値

我々は非定常環境での意思決定では実際に対応している人間の選択傾向が役立つのではないかと考えた．人間の意思決定の特徴として，ある行動系列がある基準を満たす成果を得られた時，その行動系列に執着してあまり探索をしなくなる．このような傾向は満足化と呼ばれ [Simon 56]，最適化とは区別される．しかし前述のトレードオフを考慮する場合，満足化には探索を開始と停止の条件を明確に規定できるという利点が存在する．そこで我々はその評価に基づいて選択するのみで満足化する事ができる，既存の LS 価値関数に着目し，満足化方策の実装形式の一案として LS の拡張モデル LSX を考案した [甲野 14].

3.1 EXtended Loosely Symmetric model

EXtended Loosely Symmetric model (LSX) は客観的な価値を歪めて表現する価値関数である．その最も重要な評価の性質は，各選択肢の試行回数に応じて観測された情報から曖昧な評価 (基準値 R) に近づける事にある．価値関数 LSX の評価値は以下の式により，互いに独立な選択肢 ($a_i \in A$) と，その選択肢を試行した際に観測された目的事象 (報酬，$e \in \{e, \bar{e}\}$) の発生割合 ($\overline{X_{a_i}}$)，またその試行回数 (n_{a_i}) によって定義される.

$$a_{MT} = \arg\max_{a_k}(n_{a_k}),\ a_{LT} = \arg\min_{a_k}(n_{a_k}) \quad (1)$$

$$V_e = \frac{n_{a_{MT}}\overline{X_{a_{MT}}}\, n_{a_{LT}}\overline{X_{a_{LT}}}}{n_{a_{MT}}\overline{X_{a_{MT}}} + n_{a_{LT}}\overline{X_{a_{LT}}}} \quad (2)$$

Adaptation to Non-stationary Environment with Value Function Implementing Cognitive Traits.

Yu Kohno, Graduate School of Tokyo Denki University.

Tatsuji Takahashi, School of Science and Technology, Tokyo Denki University.

$$V_{\bar{e}} = \frac{n_{a_{MT}}(1-\overline{X_{a_{MT}}})n_{a_{LT}}(1-\overline{X_{a_{LT}}})}{n_{a_{MT}}(1-\overline{X_{a_{MT}}})+n_{a_{LT}}(1-\overline{X_{a_{LT}}})} \quad (3)$$

$$n_V = V_e + V_{\bar{e}} \quad (4)$$

$$\omega_{n_i} = \frac{n_i}{n_i + n_V} \quad (5)$$

$$LSX(e;a_i) = \omega_{n_i}\overline{X_{a_i}} + (1-\omega_{n_i})(2R - \frac{V_e}{n_V}) \quad (6)$$

4. 非定常環境シミュレーション

一回の選択と試行を 1 step として各エージェントそれぞれ 100,000 steps 行い，そのシミュレーション 1,000 回分を平均して各種指標を算出した．選択肢は 20 通りあり，全ての選択肢の真の報酬生起確率は，毎回シミュレーション開始時に一様分布から独立に決定される．非定常環境を表すため，シミュレーション開始から 10,000 step 後に必ず全ての選択肢の真の報酬生起確率が一様分布から独立に再設定される．そのため本シミュレーション課題はメタバンディットが対応し難いとされる (3) 前述の変化が両方起こり逆転する，という非定常環境に分類される．各エージェントは選択肢の変化を直接的に観測する事は出来ず，またそのタイミングを学習する能力も持たない．比較するアルゴリズムには，全身的に基準値を獲得する LSX，理想的な基準を常に与えられている LSX_{OPT}，UCB1-tuned を用いた．また，step 毎に 試行回数を $n_i \leftarrow \gamma n_i$ として圧縮 (それに付随して期待値 $\overline{X_{a_i}}$ も移動平均になる) するエージェントも用いる．忘却率 $\gamma = 0.999$ を用いて，以下の結果ではモデル名 γ と表記する．更にメタバンディット (Meta UCB1-tuned) も比較に用いる．

4.1 結果および考察

図 1 は理想の選択肢を取り続けた際の獲得報酬の期待値との差を示す後悔の度合い，図 2 は一つ前の選択から選択肢を切り替えた入替率の推移である．ほぼ全てのアルゴリズムに対して 10,000 step 毎に後悔の度合いの上昇が見られた．その中でも過去の情報を圧縮して更新している場合の $LSX\gamma$ と最適基準を用いた $LSX_{OPT}\gamma$ が最も低い水準を保っている．環境の変化前と変化後の正解率がほぼ変わらない点であれば忘却率の $UCB1\text{-}tuned\gamma$ も同様であるが，前述した 2 つと比べてかなり成績が低い．図 2 に示す入替率においても，選択肢の入替率から見られる環境の変化によって発生する探索行動の割合が，前述した 2 つアルゴリズムにおいて環境変化の度に高くなるだけでなく，その後急速に減少していた．これは環境の変化に反応し，適宜再探索が行えている事を意味する．

5. 結論

本研究は定常・非定常かもわからない環境の中で意思決定で，実際に複雑な環境下で選択している人間の傾向が有効に働く事を示した．非定常環境において，情報を圧縮して新しい情報の反映度合いを最低限保ち続けるという情報の保存形式が重要になる．LSX は割引率 γ のような情報の圧縮と相性がよく，また圧縮しなくとも基準値が最適であるならば非定常環境によく対応することが本研究のシミュレーションに表れている．それにより，LSX がメタバンディットでは対応し辛い非定常環境にも対応できるという新たな知見が得られた．

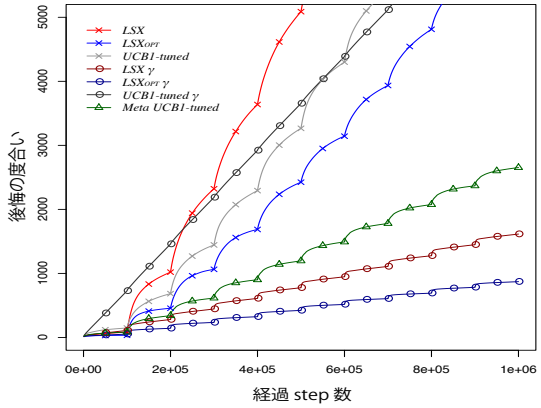

図 1: 非定常 20 本腕バンディット問題：後悔の度合い

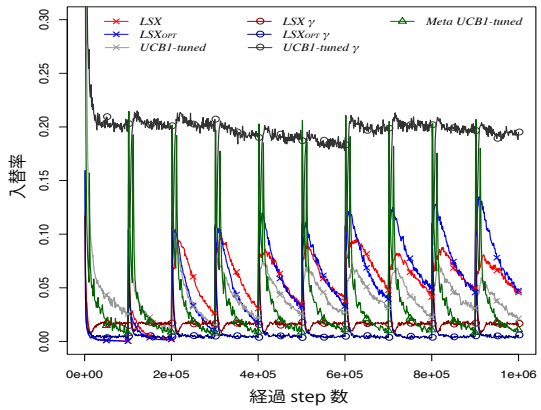

図 2: 非定常 20 本腕バンディット問題：入替率

参考文献

[Simon 56] H. A. Simon, Rational choice and the structure of the environment, *Psychological Review, 63*, 261–273 (1956).

[Wang 05] S. Gelly, Y. Wang., R. Munos. and O. Teytaud., Modification of UCT with Patterns in Monte-Carlo Go, *Technical Report*, No.6062, INRIA (2005).

[甲野 14] 甲野佑, 高橋達二, 柔軟な意思決定機能のための認知特性の応用と検証, JSAI 2014(2014 年度人工知能学会全国大会 (第 29 回)) 予稿集, 2N5-OS-03b-2 (2014).

[Sutton 00] R. S. Sutton. and A. G. Barto., 強化学習, **森北出版**, (三上, 皆川 訳) (2000).

[Hartland 06] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag, Multi-armed bandit, dynamic environments and meta-bandits, In Advances in Neural Information Processing Systems(NIPS-2006) Workshop, Online Trading Exploration Exploitation, 2006.