1L3-OS-24a-3in

自己組織的に環境変化を感知し方策を遷移する緩対称性推論モデル

Loosely Symmetric model that Senses an Environmental Change by Self-Organizing and Changes a Policy

甲野 佑[*1] 高橋 達二[*2]
Yu Kohno Tatsuji Takahashi

[*1]東京電機大学大学院 先端科学技術研究科 [*2]東京電機大学 理工学部
Graduate School of Tokyo Denki University Tokyo Denki University

Loosely Symmetric model (LS) exists as a model with three properties (Relative estimation, Reliability consideration, Reference satisficing) of the cognitive value evaluation of the animals. We focused attention on a value estimation in consideration of the satisfactory degree for the reference, and developed Loosely Symmetric model with Variable Reference (LS-VR) as the model specialized in it. In this study, we handled the N-armed bandit problem as a simple problem to make decision in unknown environment, and showed that LS-VR had significant results in the transient environment.

1. はじめに

本研究は我々が開発した認知的アイディアに基づくヒューリスティクス Loosely Symmetric model with Variable Reference (以下 LS-VR) [1] が意思決定課題において最も優れた既存モデルの一種である $UCB1$-$tuned$ [2] と比較し，価値の優劣に対する規準の学習や過去の情報の減衰，忘却の実装が容易である点と，それらの実装によって成績が向上する点を示す．またそれらの点から過去の研究から認知的価値を有するモデル [3] でありながら工学的にも有用である事を論じる．

2. N 本腕バンディット問題

本研究では工学的な有用性を示す指標として N 本腕バンディット問題を例に，何も情報の無い状態から, trade-off を抱える課題，環境に対し主体的に情報を獲得して行く際の不確実な知識の扱い方や値付けを論じる．ここでの不確実な知識とは観測が不十分で, 正しいか否か断定出来ない曖昧な知識を意味する．これは強化学習課題における初期において学習を促進するためにどのような方策や価値観数を用いるかの問題に対応する [4].

N 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手段 (腕) 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．この課題の難しさは探索と収穫のジレンマという単語で表される．目的を達成するための手段の中から最も効率の良い手段を知るためには, 情報探索のための試行に多くの時間を費やさなければならない．それは結果的に見れば, 非効率的な手段を何度も行った事になり, 探索のための試行を行えば行う程, 最終的に得られる報酬は低くなる．しかし, 探索が不十分だと正確な確率を知る事が出来ず, 不幸にも偶然それまで報酬が多く得られていただけの非効率的な手段を効率的であると判断を誤ってしまう可能性が高くなる．

生き物が効率的に生きるためには, 度々このようなバンディット問題的な課題に直面する．例えば, ある野良猫にとって数カ所の餌場があったとする．それらの餌場を訪れると確率的に餌を得られるが, 時間的制約があり, 全ての餌場を廻る事は出来ない．餌場を腕, 餌を報酬とした時, これは正にバンディット問題に置き換える事が出来る．現実において, バンディット問題における対応すべき環境は複雑であり非定常である．猫にとって餌場の価値は, 餌を出していた住人が突然病で餌を出せなくなったり, 移住で猫好きの住人に変わり, 餌を出す頻度が増したりする等, 突然変化する事が有る．あまつさえギャンブルのスロットマシンでさえ, 時間毎に報酬獲得の確率が変動する場合がある．そのような非定常な環境に対し, 複雑な準備をせず, かつ早く簡便に対応するのは難しい．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．N 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からなる普遍的な “早さ” と “正確さ” の trade-off を端的に表す事が出来る．

連絡先: 東京電機大学 理工学部
〒 350-0394 埼玉県比企郡鳩山町大字石坂
E-mail: yu.kohno.02@gmail.com

3. LS-VR

LS-VR は Loosely Symmetric model (以下 LS) を改良したモデルである [5]．LS は表 1 に表される頻度分布，あるいは表 1 を確率として正規化した表 2 で表される完全結合分布における任意の 2 事象間に対する主観確率のモデルである．例えば $LS(E|C_i)$ であれば，選択肢 C_i を選択した際の結果 E の生起に対する主観確率を意味する．LS は主観確率という一種の価値関数でありながら方策としての性質も有し，価値の評価と共により良い価値を求めた探索行動を自発的に引き起こす事が出来る [1].

表 1: 事象 C，E 間の頻度分布

	E	$\bar{E}$
C_1	a_1	b_1
C_2	a_2	b_2
$\vdots$	$\vdots$	$\vdots$
C_n	a_n	b_n

このような LS の性質は，(1) 本来関係ない価値，事象間を結びつけて評価を行ってしまう “相対評価” [6]，(2) 価値判断の材料となるサンプル数の少なさに応じて価値を歪める “信頼性考慮” [7]，(3) ある規準を定めて，規準に対する高低により価値の種類を二値化する “規準充足化” [8] という認知的な 3 種

表 2: 事象 C，E 間の完全結合分布

	E	$\bar{E}$
C_1	$P(C_1, E)$	$P(C_1, \bar{E})$
C_2	$P(C_2, E)$	$P(C_2, \bar{E})$
$\vdots$	$\vdots$	$\vdots$
C_n	$P(C_n, E)$	$P(C_n, \bar{E})$

の性質によって説明できる．

$$LSVR(E|C_i) = \frac{a_1 + s_p}{a_1 + b_1 + \rho_R(s_n + s_p)} \quad (1)$$

$$\text{Positive bias}: s_p = \frac{b_H b_L}{b_H + b_L} \quad (2)$$

$$\text{Negative bias}: s_n = \frac{a_H a_L}{a_H + a_L} \quad (3)$$

$$a_H = \arg\max_{a_j}(a_j + b_j), b_H = \arg\max_{b_j}(a_j + b_j) \quad (4)$$

$$a_L = \arg\min_{a_k}(a_k + b_k), b_L = \arg\min_{b_k}(a_k + b_k) \quad (5)$$

$$\rho_R = \frac{1}{R_t} - 1 \quad (6)$$

$$R_0 = 0.5 \quad (7)$$

$$R_{t+1} = \alpha R_t + (1 - \alpha) r_t \quad (8)$$

$$\rho_R = \frac{1}{R_t} - 1 \quad (9)$$

我々が新たに開発した LS-VR は表 1 で与えられる変数，選択肢 i を試行し報酬獲得頻度 a_i と報酬非獲得頻度 b_i から式 1 で定義される．変数 R_t は t step における価値規準 (参照点 R) の値であり，その詳細な意義は後述にて説明する．機能的には “規準充足化” における規準の動的な学習を可能にした LS の改良モデルであると定義できる．

3.1 参照点の役割

参照点とはある選択肢に対する振る舞いが変化する境界線であり，端的に言えば図 1 のように価値の損得をわける値である．参照点を上回れば利得，下回れば損失として解釈される．参照点の値そのものは損得の中間であり中性的な価値として扱われる．通常の LS の参照点は確率が取りうる値のうち丁度中間の値である 0.5 に固定されており，従来は可変にする方法が見出されていなかった．

LS は環境に 0.5 以上と判断できる価値があるとき，収束を早めて早期に探索を辞めてしまう．逆に 0.5 以下の価値しか無い時，(どこかに 0.5 以上の価値があると考えて) 探索をし続ける．これは信頼性の低い価値を，最も中性的な値である 0.5 に近似するためである．0.5 以上の価値がある場合は信頼性の低い価値は選択せず，0.5 以下の価値しか無い場合は信頼性の低い価値こそが，真に高い価値である可能性を考慮して選択する．このように恵まれた環境でリスク回避し，貧しい環境でリスク追及する傾向を持つ点で，行動経済学における反射効果 (reflection effect) に準えられる．

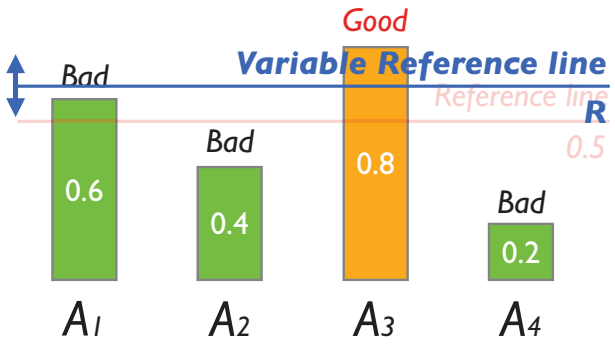

図 1: 可変する参照点

LS-VR は新たに設けたパラメータ (式 6) により参照点を任意に変更しオンラインに更新する事が可能になった．これにより恵まれた環境で早期に探索を辞めてしまう事や，貧しい環境で永遠に探索し続ける事を回避する．また参照点によって参照点以上の価値 (良い選択肢)，参照点以上の価値 (悪い選択肢) を分離する．参照点が学習により上昇すれば良価値の数を絞る事ができ，選択肢が多い場合にも効率よく探索を行う事ができる．

3.2 忘却と割引率

新たな環境に適応するためには有る程度は過去の情報を忘れなければならない．そのために本研究ではモデルに対して割引率 γ の適用を試みた．割引率とは試行回数 (報酬獲得回数 a_i + 報酬非獲得回数 b_i) を更新する際に過去の情報を減衰させて行く，この形式はマルコフ決定過程における Bellman 方程式の仕組みを応用したものであり $0.0 \leq \gamma \leq 1.0$ の値域を持つ [4]．試行した選択肢を C_k としたとき，具体的な更新式は以下の通りである．

$$\begin{bmatrix} a_i \\ b_i \end{bmatrix} = \begin{cases} \gamma \begin{bmatrix} a_i \\ b_i \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} & (C_i = C_k) \wedge E \\ \gamma \begin{bmatrix} a_i \\ b_i \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} & (C_i = C_k) \wedge \bar{E} \\ \gamma \begin{bmatrix} a_i \\ b_i \end{bmatrix} & otherwise \end{cases} \quad (10)$$

割引率 γ の振る舞いを端的に言えば，古い情報に対する新しい情報の影響度を決定するパラメータである．本研究で用いた $\gamma = 0.999$ を例にすれば，新しい情報が全体に対して最低でも $1 - \gamma = 0.1\%$ の影響度を有し続ける事を意味する．この割引率パラメータを用いて頻度情報を更新する事により，適度に忘れる事で成績を向上させる事が可能になると考えられる．

4. シミュレーション

本項では N 本腕バンディット問題のシミュレーションによって LS-VR の有用性を示す．この課題では 1 step 毎に一度 N 個ある選択肢の中から一つの選択肢を方策に従って選択して試行する．そして選択肢に設定された未知の報酬確率に基づき報酬を獲得する．これらの情報から各選択肢の価値を推定して行き，最終的に得られる価値を最大化する事を目的としている．

この課題における評価基準は，正しい選択肢を選べているかと，なるべく良い選択を行い，探索中でも損をし過ぎていないかである．前者を正解率，後者を理想的な選択を行っていた場合との差を表す期待損失似よって評価する．また環境の変化への対応力と，$LS\text{-}VR$ と割引率の相性を見るため，途中で環境が再度一様乱数から設定されるようにした．

本研究では選択肢が 16 の場合を扱った．10,000 step までは最初に一様分布から決定された真の報酬確率は変化せず，10,000 step 毎に真の報酬確率が再設定される．即ち本課題は最初の 10,000 step までにどれだけ正解率を上昇できるかと，報酬確率の再設定という環境変化の度に 10,000 step でどれほど正解率を回復できるかを検証する課題である．以上の設定で 4 回の環境変化を伴う 50,000 step のシミュレーションを 1,000 回行い，正解率と期待損失の平均をとった．

4.1 結果および考察

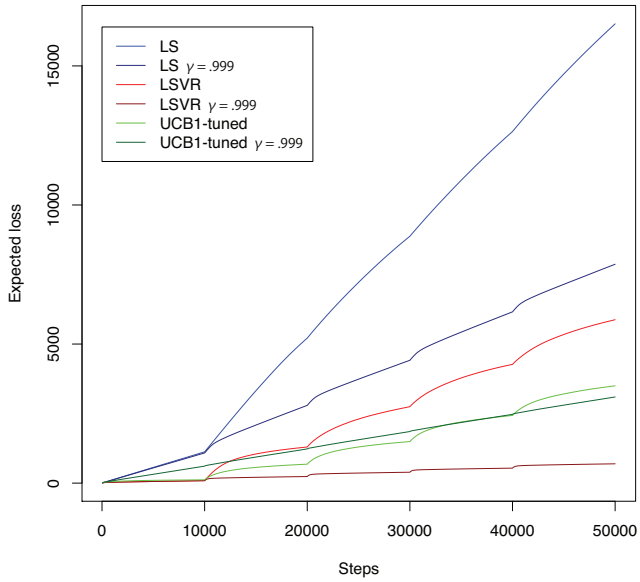

図 2: 16 本腕バンディット問題における step 経過に対する正解率の推移

図 2 は 16 本腕バンディット問題のシミュレーション結果であり，1,000 回のシミュレーションにおいて正しい選択肢を選んだ割合 (正解率) である．この結果から $LS\text{-}VR$ は通常の LS と比較して成績が向上しているだけでなく，現在もっとも優れたモデルの一つであるとされる $UCB1\text{-}tuned$ と比較しても高い正解率を有し，正確に良い選択肢を選んでいる事がわかる．これは $LS\text{-}VR$ が環境に応じて参照点を変化し，あらゆる環境において着目すべき選択肢の数を絞って行く事が出来るためである．また図 2 に示す通り，$UCB1\text{-}tuned$ に割引率を適用しても成績が下がってしまっているが，$LS\text{-}VR$ に割引率を適用した場合は環境変化前 (10,000 step 以前)，変化後 (10,000 step 以後) においても通常の $UCB1\text{-}tuned$ を越える成績を有している．他のモデルが環境変化の度に正解率が低下している中で $LS\text{-}VR$ のみ環境変化の前の成績を保ち続けている．これは $LS\text{-}VR$ が割引や忘却といった概念と相性の良い事を表し，現実的な非定常環境における学習，価値判断において優れたモデルである事を示している．

図 3 は始めから最も良い選択肢を選び続けていた場合に得られた報酬と，実際の選択で得た報酬との差を示す，期待損失

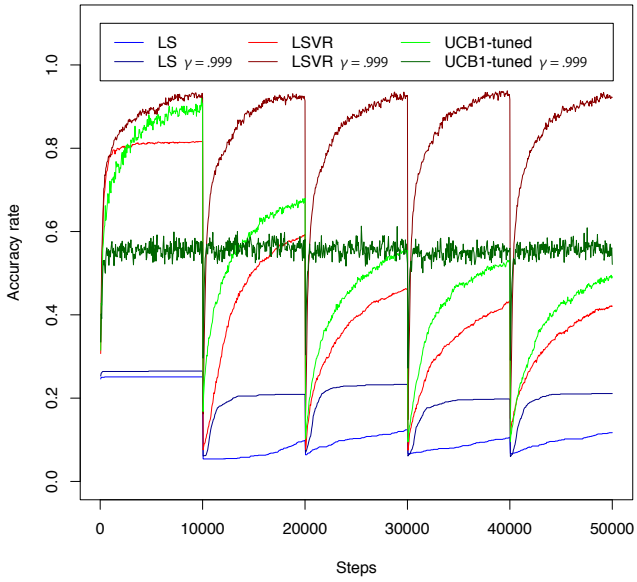

図 3: 16 本腕バンディット問題における step 経過に対する期待損失の推移

のグラフである．環境変化前までは割引率を適用していない $LS\text{-}VR$，$UCB1\text{-}tuned$，割引率を適用した $LS\text{-}VR$ において大きな差は無いが，環境変化後では割引率を適用した $LS\text{-}VR$ は急激に正解率を回復するため，少ない損失に抑えられるのに対し，特に割引率を適用していない $LS\text{-}VR$ は損失の上昇を抑える事が出来ていない．

以上の正解率，期待損失という二つの指標から認知的なアイディアに基づいた $LS\text{-}VR$ は，むしろ割引率を用いたような自然な情報の保存形式において，かえって優秀な成績を持つ事がわかった．これは $UCB1\text{-}tuned$ が正確さを重視するあまり多くの情報を必要とするのに対し，LS および $LS\text{-}VR$ は少ない情報で効率的に価値推定が出来るためだと考えられる．

5. 結論

本研究により $LS\text{-}VR$ は (1) 可変参照点による規準の学習と，(2) 割引率による過去の情報の忘却を効率的に活用できるヒューリスティクスである事がわかった．現時点では飽くまで簡便的な実装であるが様々な認知的アイディアを簡便的な実装可能な点で，工学的に利用できる優秀な認知モデルであると言える．また環境の変化に対する対応性を保ち続けられる点から，$\epsilon\text{-}greedy$ 等の常に一定の割合以上は探索行動をし続ける性質や $UCB1\text{-}tuned$ を始めとした数学的規範性を起源とするモデルに見られる正確性を重視する性質の中間の性質を持ち，それでいて人間の認知に由来するという，それらのモデルとは異なる起源を有する新たな種類のモデルであると言える．更に言えば，本研究の結果は人間の認知が進化の中で獲得してきた，現実の環境への簡便かつ適応的性質の妥当性を端的に示しているものと考えられる．

また，本論文で使用した参照点 R の学習手法や割引率 γ の調整は $LS\text{-}VR$ そのものとは独立した問題であり，これらの問題を解決する事で更なる成績の向上を望む事が出来る．更に強化学習全般への一般化も試行されており [9]，今後，より広い分野での認知研究の成果の利用を簡便にし，かつ工学的な発展にも寄与できると考えられる．

参考文献

[1] Kohno, Y., Takahashi, T. (2012), "Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off", *SCIS-ISIS 2012*, Kobe Convention Center (Kobe Portopia Hotel), pp.1166–1171.

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

[3] Takahashi, T., Nakano, M., Shinohara, S. (2010), "Cognitive symmetry: Illogical but rational biases", *Symmetry: Culture and Science*, Vol.21, No.1-3, pp.275–294.

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

[5] 篠原修二, 田口亮, 桂田浩一, 新田恒雄 (2007), "因果性に基づく信念形成モデルと N 本腕バンディット問題への適用", **人工知能学会論文誌**, Vol.22, No.1, pp.58–68.

[6] Tversky, A., Kahneman, D. (1974). "Judgment under uncertainty: Heuristics and biases". *Science 185 (4157)*, 1124–1131.

[7] Kahneman, D.; Tversky, A. (1984). "Choices, values and frames". *American Psychologist 39 (4)*, 341–350.

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

[9] Uragami, U., Takahashi, T., Alsubeheen, H., Sekiguchi, A., and Matsuo, Y. (2011), "The Efficacy of Symmetric Cognitive Biases in Robotic Motion Learning". *Proceedings of the 2011 IEEE International Conference on Mechatronics and Automation*, August 7–10, Beijing, China , pp. 410-415.