1L3-OS-24a-4in

認知的な強化学習モデルに対する基準学習の応用と考察

Application and consideration of learning of reference to a cognitive reinforcement learning model

高橋 優太 Yuta Takahashi　甲野 佑 Yu Kono　高橋 達二 Tatsuji Takahashi

東京電機大学理工学部
School of Science and Technology, Tokyo Denki University

LS(loosely symmetric) model is effective in N armed bandit problem. LS model is subjective probability model derived from human cognition. Uragami devised a method for Giant-Swing that takes advantage of LS properties. In this research, the refined model of LS which Kono devised (LS-VR) was implemented in the above-mentioned method, and aim to improve the performance.

1. はじめに

単純な強化学習課題である N 本腕バンディット問題において，価値関数でありながら方策としても機能する緩対称性推論モデル(LS)が存在する[1]. LS は人間認知由来の主観確率モデルでもあり，環境の曖昧さからの推論に長けている．浦上はその性質を活かした LS 強化学習手法を考案し，複雑な運動制御課題で優れた成績を示した．本研究では，前述の手法に甲野の LS 改良モデル(LS-VR)を実装し，更なる成績の向上を行なった．

2. 強化学習

強化学習はエージェントが環境から与えられる数値化された報酬を最大にする事を目的とし，どのような行動をとればよいかを学習する機械学習の一種である[2]. 同じく機械学習の種類である教師あり学習，教師なし学習と異なり，エージェントが主体的に行動してそれに伴う報酬から選択すべき行動を学習しなければならない．エージェントは環境の“状態”と取りうる“行動“を得る事ができ，実際には以下に記述されるような状態行動価値関数を報酬から学習してその時点での状態に対する最良の行動を学習する．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)] \quad (1)$$

上式は最も単純な強化学習手法のひとつである Q 学習の状態価値関数である．ここで S_t, a_tは時刻 t における状態，行動をそれぞれ表し，r_{t+1} は環境から得られた報酬を表す．αは新しい学習結果をどの程度考慮するかの学習率を意味し，γは未来の状態行動価値をどの程度考慮するかを意味する割引率である．エージェントはこの状態価値関数の値と，それをどのように扱うかを記述した方策を用いて実際に取る行動を学習する．また，このとき最も Q 値が大きくなる行動を greedy な行動という．

以下では浦上の論文で使用された方策について説明する．

2.1 GQ policy

GQ policy とは学習された Q 値に基づき最も Q 値の高い行動を選択する方策である．

連絡先：高橋優太，東京電機大学理工学部
10rd137@ms.dendai.ac.jp

2.2 NS policy

表 1　greedy な行動頻度分布

	greedy	not greedy
A_0	l	u
A_1	m	v
A_2	n	w

表 1 は状態ごとに記録される，greedy な行動をした頻度，greedy でない行動をした頻度の表である．NS policy は表 1 の頻度を用い，例えば行動 A_0 であれば以下のような評価式を計算する．そしてこの値の最も高い行動を選択する方策である[3].

$$NS(greedy|A_0) = \frac{l}{l+u} \quad (2)$$

2.3 LS policy

LS policy は篠原[1]が考案した緩い対称性モデル(LS)を用いた評価式を使用し，行動を決定する方策である．例えば行動 A_0 であれば以下のような評価式を計算する．LS は対称性推論を用いたモデルであり，「p ならば q」が真であるという前提があるとき，逆命題である「q ならば p」も真であると捉えてしまう傾向を有している．

$$LS(greedy|A_0) = \frac{l + \frac{u(v+w)}{u+v+w}}{l + \frac{u(v+w)}{u+v+w} + u + \frac{l(m+n)}{l+m+n}} \quad (3)$$

考案した浦上[3]によって LS policy は非常に複雑なダイナミクスを持つ運動制御の強化学習課題において NS や GQ と比較しても優位に学習可能である持つ事が示されている[3].

3. 提案手法

浦上は推論モデルでありながら意思決定課題の一種である 2 本腕バンディット問題においても優れた結果を持つ LS を強化学習の方策として扱えるような形式を考案し，優れた成績を持つ事示した．しかしながら LS は 2 種類の行動に対応するのみで，強化学習等の取りうる行動が複数の一般的な課題には対応していない．そこで本論文では甲野が考案した N 本腕バ

ンディット問題における 2 種の一般化 LS 式を用いて LS policy の拡張を行った.

3.1 一般化 LS policy (LSN policy)

LS を複数の要因に対応させるため, 最も観測した割合の高い事象, 最も観測していない事象を LS の式に組み込んだ.

$$LS(greedy|A_0) = \frac{l + \frac{x_{max}x_{min}}{x_{max}+x_{min}}}{l + \frac{x_{max}x_{min}}{x_{max}+x_{min}} + u + \frac{k_{max}k_{min}}{k_{max}+k_{min}}} \quad (4)$$

$$LS(greedy|A_1) = \frac{m + \frac{x_{max}x_{min}}{x_{max}+x_{min}}}{m + \frac{x_{max}x_{min}}{x_{max}+x_{min}} + v + \frac{k_{max}k_{min}}{k_{max}+k_{min}}} \quad (5)$$

$$LS(greedy|A_2) = \frac{n + \frac{x_{max}x_{min}}{x_{max}+x_{min}}}{n + \frac{x_{max}x_{min}}{x_{max}+x_{min}} + w + \frac{k_{max}k_{min}}{k_{max}+k_{min}}} \quad (6)$$

ここで x_{max} は最も試行した行動, x_{min}は試行していない行動に対する not greedy な頻度を意味する. 同じくk_{max} , k_{min}は greedy な頻度を意味する. このようにx_{max}, x_{min}, k_{max} , k_{min} を使用したバイアス項を共通化させる事でする事で各 LS 式を簡略化し, かつ本来 LS が有する地の不変性という各選択肢に対するバイアス項(視覚における非着目事象)が等しくなる認知的性質を付加する事が出来る[4].

3.2 LS-VR policy

LS はある価値に対してエージェントの振る舞いが変化する境界線を考慮した価値評価が可能であり, 通常の LS はその境界線(参照点)は 0.5 に固定され, 変える事ができない. それに対して甲野が考案した LS-VR はパラメータ ρを導入する事によって参照点の値 R を任意に変えることを可能とした LS の改良モデルである. LS-VR は認知的性質に由来するモデルでありながら, N 本腕バンディット問題において最も優れたモデルの一種であるとされる UCB1-tuned と同等の成績を有しており[5], 強化学習課題に置いても有用だと考えられる.

$$\rho = \frac{1}{R} - 1 \quad (7)$$

$$LS(greedy|A_0) = \frac{l + \frac{x_{max}x_{min}}{x_{max}+x_{min}}}{l + u + \rho\left(\frac{x_{max}x_{min}}{x_{max}+x_{min}} + \frac{k_{max}k_{min}}{k_{max}+k_{min}}\right)} \quad (8)$$

4. 大車輪問題シミュレーション

大車輪問題は, 強化学習の中でも特に状態数が多く, 複雑なダイナミクスを有する問題として知られている[6]. ロボットの目的は自分自身を回転させることであり, ロボットは腰のアクチュエータのみを稼働させることができる. 本研究では Open Dynamics Engine を用いてシミュレータを構築し, シミュレーションを行った[3].

取りうる行動は, “A0 : 左の方向に動く”, “A1 : 右の方向に動く”, “A2 : 動かない”の三種がある. それぞれの行動について価値関数と比較して”greedy”, ”not-greedy”を記録し, その値から各ポリシーを用いて行動を選択する. 状態は“上半身の角度”, “下半身の角度” , “上半身の角速度”で決まり, 報酬はエージェントの角度から 0～1 の範囲で与えられる. 状態はそれぞれ 24, 5, 6 分割されており, 計 720 の状態を持つ. 学習方法として Q 学習, NS, LS , 一般化 LS, LS-VR を用い, 1000step 毎の獲得報酬の合計を記録した. それぞれについて 50 回のシミュレーションを行い獲得報酬の平均を比較した.

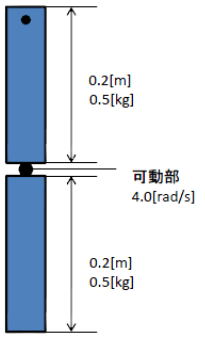

図 1：ロボットの構造

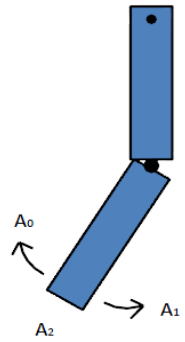

図 2：行動

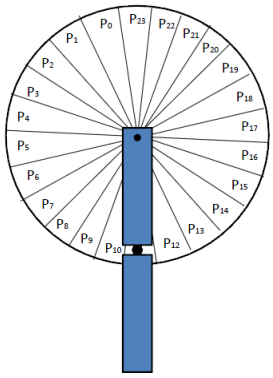

図 3：状態 1(上半身の角度)

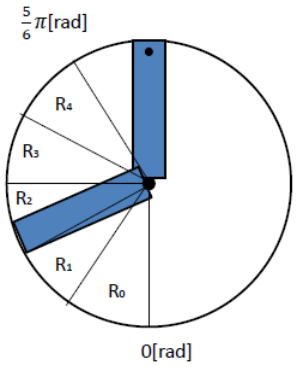

図 4：状態 2(下半身の角度)

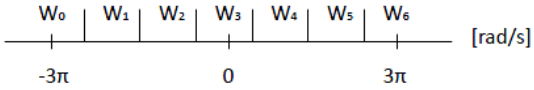

図 5：状態 3(上半身の角速度)

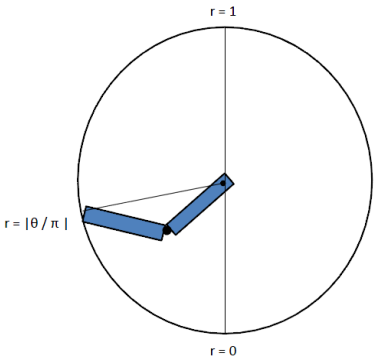

図 6：報酬

今回の設定では，1000step 毎にロボットの状態を初期状態に戻した．また，1000step 毎に greedy な行動を選択する確率を 0.05 ずつ増やし，20000step 以降から必ず greedy な行動を選択するように設定した．その後 50000step までシミュレーションを行い獲得報酬の合計の推移を観察した．

4.1 結果及び考察

図7は大車輪シミュレーションの結果である．横軸は step 数，で縦軸は 1000stepg 毎の報酬の平均を表している．LS-VR0.2 は LSN にとって損得の規準が 0.5 である参照点 R の値をパラメータ ρ を用いて 0.2 に変更したモデルであり，LS-VR0.8 は同じく 0.8 に変更したモデルである．

必ず greedy な行動を選択するようになる 20,000 step になった直後にどの policy においても獲得報酬が落ち込む現象がみられるが，LS-VR はその減少の値が小さいことが確認できる．その後の推移を確認すると LS-VR0.8 が最も良い成績を示している．この大車輪課題における参照点とは，エージェントの行った行動が Q 値に照らし合わせて greedy だった割合の基準値である．基準値が低ければあまり greedy でない行動でもその選択に執着したり、高ければ規準を満たす行動を見付け出さない限り探索をし続ける．このことから，LS は基準を満たすまで探索を続け，更に有る程度高い基準を設ける事でより良い選択を見つけ出せる事が出来ると考えられる．

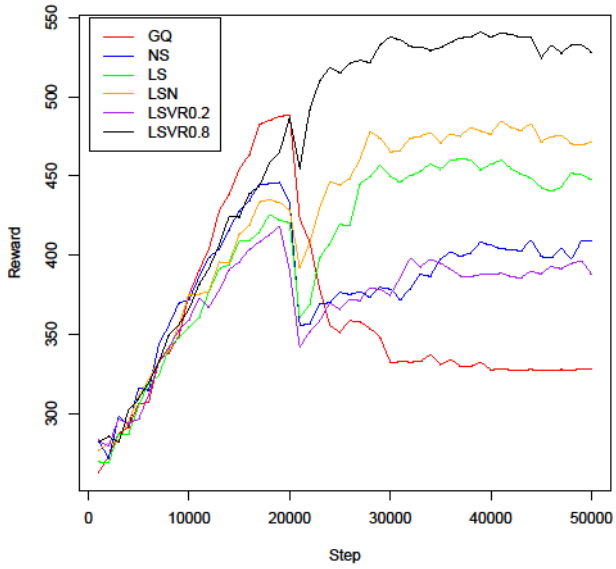

図 7：獲得報酬の学習率

5. 結論

シミュレーションの結果から，大車輪問題における学習に対し LS-VR は他の学習方法に比べ，高い効果があることがわかる．

LS-VR は N 本腕バンディット問題において参照点を動的に学習し高い成績を残している．本研究では参照点の動的な学習方法の手法が考案されていないため，参照点を固定しシミュレーションを行った．その結果，参照点 R=0.8 において高い成績を残すことができたが，大車輪問題の各状態における最適な参照点を動的に学習する手法を考案することで，更なる成績の向上を目指すことができると考えられる．

6. 謝辞

本研究のシミュレーションプログラムを提供と解説いただいた東京工科大学コンピュータサイエンス学部浦上大輔氏に深謝する．

参考文献

[1] 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.

[2] Richard S.Sutton, Andrew G.Barto 三上貞芳 皆川雅章 共訳: 強化学習.

[3] Daisuke Uragami, Tatsuji Takahashi, Hisham Alsubeheen, Akinori Sekiguchi, Yoshiki Matsuo: The Efficacy of Symmetric Cognitive Biases in Robotic Motion Learning.

[4] 甲野佑 高橋達二:緩い対称性推論を用いた強化学習アルゴリズム Reinforcement Learning Algorithm using Loosely Symmetric Reasoning.

[5] Kohno, Y., Takahashi, T. (2012), Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off, SCIS-ISIS 2012.

[6] 原正之 川辺直人 久嶋肇 黄健 藪田哲郎(横国大): 強化学習を用いた人型ロボットよる大車輪運動の獲得.