1L3-OS-24a-5

対称性推論と運動学習の分節化: LS モデルを応用した Q 学習による大車輪ロボットの実現

Symmetric Reasoning and Segmentation of Motor Learning: Realization of Giant-Swing Robot by using Q-leaning with LS-model

浦上 大輔[*1] Daisuke Uragami　高橋 達二[*2] Tatsuji Takahashi　アルスビヒーン ヒシャム[*1] Hisham Alsubeheen

アルアルワン アリー[*1] Ali Alalwan　関口 暁宣[*1] Akinori Sekiguchi　松尾 芳樹[*1] Yoshiki Matsuo

[*1] 東京工科大学コンピュータサイエンス学部 School of Computer Science, Tokyo University of Technology

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

This study proposes an application of symmetric reasoning to Q-learning method, and the proposed method is applied to motor learning of the Giant-swing robot. The simulation shows the proposed method efficiently performs without segmentation of learning. This research argues how symmetry reasoning enhances motor learning and sign generation.

1. はじめに

ロボット設計者があらかじめ行動規則を定めることができないような環境においては，ロボットが自律的に適切な行動を獲得する必要がある．その手法として，強化学習をロボットの行動獲得に適用した研究は多い．しかし，これらの研究では，複数のコントローラの選択に強化学習を適用したものや，内部モデルの学習と組み合わせて強化学習を行うなど，他の手法の補助的な手段として強化学習を適用したものがほとんどであり，設計者の予想の範疇を超えた行動の獲得に成功した例は少ない．また，シミュレーションによる研究がほとんどで，実ロボットにシミュレーション結果を適用した例は多くない．

これに対して，坂井らは強化学習の代表的な手法である Q 学習をシンプルに実ロボットに適用し，その際の問題点を整理している[坂井 2010]．坂井らの研究では，大車輪ロボットの運動学習を例として，サンプリングや状態分割の粗さに起因する非マルコフ性のため，学習が段階的にならざるを得ないという問題を明らかにしている．この問題を解消する手法として，坂井らの研究では，Q 学習の報酬として2種類の報酬を組み合わせて使用する手法を提案している．一方，筆者らは同様の問題に対して，人間の推論傾向をモデル化した LS モデル[篠原 2007] を Q 学習の方策に応用する手法を提案している[Uragami 2011]．本研究では，この提案手法の有効性を，特に学習率と学習時間に着目して評価した．

2. LS モデルと提案アルゴリズム

2.1 LS モデル

人間には「q ならば p」から「p ならば q」を推論する傾向がある．LS (Loosely Symmetric) モデルは，この傾向性を定量的に表現したものである[篠原 2007]．

表 1 の a は，イベント p とイベント q の両方が発生した回数である．このとき，p が発生したという条件の下で q が発生する条件付き確率P$(q|p)$ は，P$(q|p) = a/(a+b)$ となる．

連絡先：浦上大輔，東京工科大学コンピュータサイエンス学部，東京都八王子市片倉町 1404-2，uragami@cs.teu.ac.jp

表 1：共起頻度

	p	not p
q	a	c
not q	b	d

一方，LS モデルでは，同様の条件付き確率に対する人間の直観的な見積りを次式で算出する．

$$\mathrm{LS}(q|p) = \frac{a + bd/(b+d)}{a + bd/(b+d) + b + ac/(a+c)} \quad (1)$$

式 (1) は，実際の人間の推論傾向と高い精度で一致することが知られている．LS モデルで表現される推論は，論理的には必ずしも正しくないが，実世界においてはしばしば有用である．著者の一人は，不確実性の下で学習と意思決定における LS モデルの有効性を実証しつつ内部観測理論 [松野 2000] との関係を論じている[高橋 2013]．

2.2 提案アルゴリズム

LS モデルを Q 学習の行動方策に応用する．Q 学習の行動価値関数（Q 値）は次式で定義される．

$$\mathrm{Q}(s_t, a_t) \leftarrow (1-\alpha)\mathrm{Q}(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a \mathrm{Q}(s_{t+1}, a)\right] \quad (2)$$

s_t と a_t は時刻 t における状態と行動，r_{t+1} は報酬，α は学習率，γ は割引率である．ある状態において，もっとも Q 値が大きくなる行動を greedy な行動と呼び，学習過程では greedy な行動かそれ以外の行動をある確率で選択する．

提案アルゴリズムでは，状態ごとに，greedy な行動を選択した回数とそれ以外の行動を選択した回数を表 2 のように記録する．2種類の行動 A と B があり，ある状態において行動 A を選択し，そのとき行動 A が greedy な行動であった回数が a である．表 2 より条件付き確率P(greedy|A) とP(greedy|B) を計算して比較し，値が大きいほうの行動を NS-greedy な行動とする．同様に，

表 2 と式(1)よりLS(greedy|A) とLS(greedy|B) を計算して比較し，値が大きいほうの行動を LS-greedy な行動とする．

学習過程においては，ε-greedy 行動選択によって，それぞれの greedy な行動とそれ以外の行動を定められた確率で選択する．行動選択，状態遷移，Q 値の更新，表2の更新を学習の1ステップとして繰り返し学習を行う．

表 2：行動選択の履歴（各状態）

	行動 A	行動 B
greedy	*a*	*c*
not greedy	*b*	*d*

3. 大車輪ロボットへの適用

3.1 大車輪ロボット

大車輪ロボットは，アクロボットとも呼ばれ，強化学習のテスト課題としてよく知られている．鉄棒とロボットの接続部分はフリージョイントになっており，腰部の関節のみが能動的に稼働することが特徴である（図 1）．本研究では，研究の第1段階として，ODE (Open Dynamics Engine) を用いて大車輪ロボットのシミュレータを構築した．

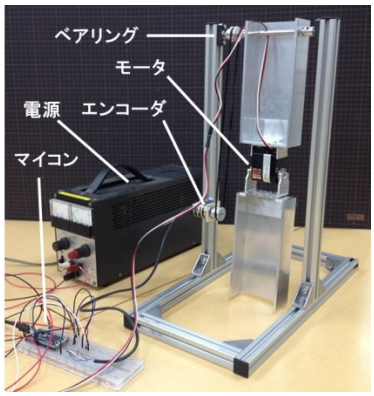

図 1：大車輪ロボット（製作中）

3.2 シミュレーション

提案アルゴリズムを大車輪ロボットの運動学習に適用し，シミュレーションによってその学習能力を評価した．学習の条件は次のとおりである．状態は，位置，速度，姿勢を分割して離散的に定義した．行動は，腰部の関節を曲げる，伸ばす，停止の 3 通りである．報酬は，ロボットの足部の先端の位置によって定義し，その位置が鉄棒の真上にあるときに最大値になるように設定した．詳細は文献[Uragami 2011] による．

図 2 は，横軸を学習時間[/ 1000 step]，縦軸を 1000 step 毎の獲得報酬の合計としたシミュレーション結果である．1000 step 毎にロボットの状態は初期状態（鉄棒の真下に速度 0 でぶら下がった状態）に戻している．学習の開始時はランダムに行動を選択し，2000 step ごとに greedy な行動を選択する確率を 0.02 ずつ増やした．このとき，通常の Q 学習における greedy な行動を選択した場合を GQ，NS-greedy な行動を選択した場合を NS，LS-greedy な行動を選択した場合を LS と表記し，それぞれ 100 試行の平均である．

図 2 において，GQ は学習の終了時付近で獲得報酬が減少している．この現象は，坂井らの研究[坂井 2010]でも発見されている．その原因は，次節で詳しく議論するが，ロボットのスイング角が小さい状態からうまく抜け出せないことによる．NS では同様の傾向が見られ，greedy な行動を選択する確率が 1.00 になる付近で獲得報酬が大きく落ち込んでいる．一方，LS ではこのような現象を回避しており，最終的な獲得報酬も3方式の中で最も大きい．この結果は提案アルゴリズムの有効性を示している．

図 3 は学習率および学習時間の影響を比較したものである（各 10 試行の平均）．図 3-右上のグラフを見ると，学習率が大きい場合（α=0.9），学習時間を長くしても（図 2 の 2 倍），GQ は図 2 と同様に学習の終了時付近に獲得報酬が減少している．つまり，この現象は，単に学習時間にのみに依存した問題ではないということがわかる．図 3-右上のグラフを見ると，学習率を小さくすると（α=0.5），この現象は回避できるようみえる．しかし，図 3-左上と右上を比較すると，学習時間が短い場合は学習率が大きい方が良いことがわかる．

図 4 は学習の終了時付近の 20,000 step について獲得報酬を比較したものである（1000 step 毎，各 10 試行の平均）．線で繋がっている点は学習時間が同じで，学習率が左から 0.1, 0.3, 0.5, 0.7, 0.9 である．特筆すべき点は，LS の学習率 0,9 の場合が，他の条件・学習方法と比較して常により良いということである．一方，GQ は学習時間に依存して最適な学習率が異なる．一般的に学習率の設定は常に課題であり，経験的に決定される場合が多い．したがって，LS の学習率についての普遍性は，提案アルゴリズムの有用な利点の1つと言える．

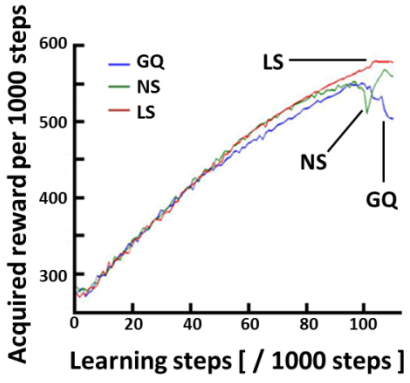

図 2：学習曲線（獲得報酬）

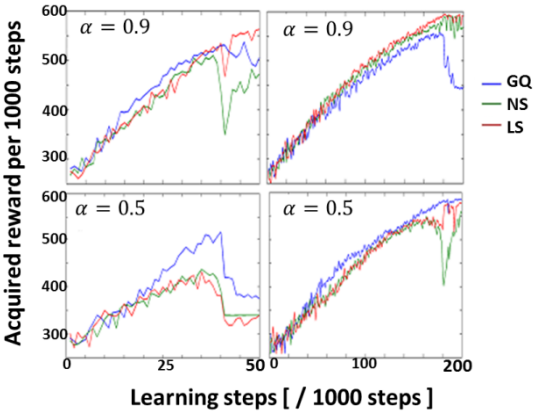

図 3：学習曲線（獲得報酬）の学習率および学習時間による比較

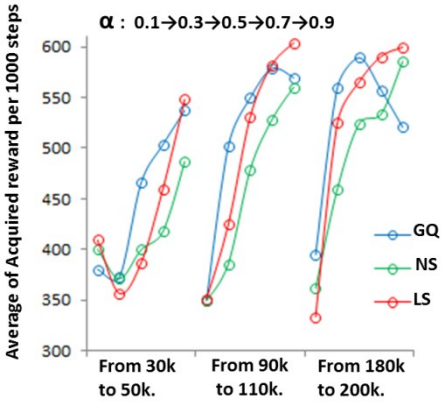

図 4：学習終了時付近の獲得報酬の比較

4. 考察

シミュレーション結果を運動学習の分節化という視点から考察する.

ロボットの運動学習に強化学習を適用するときは，内部モデルの学習と組み合わせる手法が一般的である. 大車輪ロボットの場合は，1つの内部モデルでは上手く学習できないことが指摘されており，複数の内部モデルの切り替えのタイミングが学習課題となる. 言い換えると，運動をどのように分節するかが問題となる.

本研究では，あえて内部モデルを使わずにシンプルに Q 学習を大車輪ロボットに適用した. その結果，図 2 の GQ のグラフにみられるように，学習の終了段階で獲得報酬が減少するという現象が確認された. このとき，ロボットの振れ幅が小さい状態で状態遷移がループしている. 学習の初期段階では，ランダムンネスの効果でループから脱出できるが，学習の終了段階ではランダムな行動選択をする確率は小さくなるためループの影響が顕わになる. 理論上は，Q 学習はこのようなループを回避できるはずであるが，実ロボットに適用する際は，サンプリングや状態分割の粗さにくわえて学習時間の制約により，このような現象が起こりえる.

坂井らの研究では，2 種類の報酬を組み合わせることにより，同様の現象を解消している[坂井 2010]. しかし，この 2 種類の報酬は，1つはロボットの振れ幅が小さいときに有効なものであり，もう1つはその段階を脱した後に有効なものである. つまり，実質的に運動状態を2つに分節しており，複数の内部モデルを使用することと本質的には違いがない. 一方，本研究の提案手法では，このような分節を行わずに学習を成功さている.

谷口らの研究では，内部モデルの分割自体をロボットに自律的に学習させている[谷口 2010]. 彼らの研究では，ロボットの身体能力が中程度のときに内部モデルの分割数が多くなるという結果を得られている. また，内部モデルの分割を記号あるいは言語の生成と結びつけて考察している. このような研究を参考に，提案手法と内部モデルの学習を組み合わることにより，対称性推論が内部モデルの分割すなわち記号生成にどのような影響を与えるか調べることができる. また，提案アルゴリズムでは状態を離散化しており，その粗さ，すなわち状態数は学習能力に大きく影響する. 状態分割の粗さはロボットの知覚能力に依存するものであり，状態の離散化を記号生成のプリミティブな形態とみなすこともできる. したがって，本研究のシミュレーション結果は「身体能力を補うものとしての記号生成と対称性推論」として解釈することが可能であり，そのような方向で本研究を谷口らの研究と繋いでいくことができる.

5. おわりに

本研究では，LS モデルを Q 学習に応用する手法を提案し，その手法を大車輪ロボットの運動学習に適用した. シミュレーションによる実験の結果，通常の Q 学習では学習を 2 段階にするなどのアドホックな仕組みが必要な課題を，提案手法はそのような仕組みなしで効果的に学習すること示された. また，提案手法は学習率や学習時間に依存せずに普遍的に効果的であることが示された. サンプリングや状態分割の粗さと提案手法の関係を明らかにすることや，シミュレーションによる学習結果を実ロボットに適用することなどが今後の課題である.

謝辞

本研究の一部は平成 24 年度東北大学電気通信研究所 共同プロジェクト研究 H22/B08「生命にとっての情報・推論・計算の解明と工学的応用の検討」による.

参考文献

[坂井 2010] 坂井直樹，川辺直人，原正之，豊田希，薮田哲郎: 強化学習を用いたスポーツロボットの大車輪運動の獲得とその行動形態の考察，計測自動制御学会論文集 Vol.46，No.3，pp.178-187，2010.

[篠原 2007] 篠原修二，田口亮，桂田浩一，新田恒雄: 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応用，人工知能学会論文誌 22 巻 1 号 G，pp.58-68，2007.

[Uragami 2011] Daisuke Uragami, Tatsuji Takahashi, Hisham Alsubeheen, Akinori Sekiguchi, Yoshiki Matsuo: The Efficacy of Symmetric Cognitive Biases in Robotic Motion Learning, Proceedings of ICMA 2011, 410-415, 2011.

[高橋 2013] 高橋達二: 混同という方法論：内部観測の研究プログラム化，第 24 回計測自動制御学会 SI 部門共創システム部会研究会・第 7 回内部観測研究会（共同開催），2013.3.2 口頭講演.

[松野 2000] 松野孝一郎: 内部観測とは何か，青土社，2000.

[谷口 2010] 谷口忠大: コミュニケーションするロボットは創れるか―記号創発システムへの構成論的アプローチ，NTT 出版，2010.