ISSN 2186-5647

P-16

人の認知特性に触発された強化学習アーキテクチャの鉄棒ロボットへの応用

日大生産工　○柳原　勇希　　日大生産工　　柳　凜之介
日大生産工　　浦上　大輔

1 はじめに

自律的に活動するロボットの実現に貢献する技術の一つとして，強化学習が注目されている。人間と同じ環境で活動するロボットは，個別的な環境に素早く適応するだけでなく，環境の変化や人間の行動に応じてロボット自身の行動も変化させる必要がある。その際，既に蓄積された知識に基づいた行動を選択するか，あるいはより良い行動を追及するために別の行動を選択するかの判断は難しい。この判断の難さは「探索と知識利用のジレンマ」と呼ばれ，強化学習エージェントが直面する普遍的な課題として知られている。このような課題は，人間もまた実環境において直面する課題である。人間がそれにどのように適応しているかは人間情報学的に興味深いだけでなく，工学的な示唆に富むと考えられる。

我々の研究グループでは，人間の認知特性を模倣して応用する強化学習アーキテクチャ「LS-Q」を提案し，LS-Qの有効性を鉄棒ロボットの運動獲得をテスト課題として検証している[1]。これまでの研究成果として，LS-Qは状態分割の精度が粗くても適応的に学習できるということがシミュレーションによって明らかになっている。本研究では，実ロボットによる実験によって，シミュレーションによる結果が実環境においても妥当であることを検証する。また，主要な学習パラメータの一つである割引率の影響を調査する。

2 方法

2．1　LSモデル

人間の認知特性の一つに「対称性バイアス」と呼ばれるものがある。対称性バイアスは，「p ならばq」より「q ならばp」を推論する傾向性である。このような推論は論理的には必ずしも正しくないが，経験的にはしばしば有用であり，人間の知能の柔軟さに関係していると考えられる。対称性バイアスの強度は事象の頻度や多様性に依存すると考えられるが，それを数理モデルとして表現したものが「LSモデル」である[2]。事象 p と事象 q の共起頻度が表 1 のように与えられているとき，LSモデルは「 p ならば q 」に対する信念の程度を次式よって見積もる。

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

LSモデルは認知実験のデータと高い精度で一致する一方で，LSモデルに基づいて意思決定を行うエージェントは，探索と知識利用を巧妙に調整することが知られている[篠原 2007]。

表 1：共起頻度

	p	not p
q	a	c
not q	b	d

2．2　LS-Q

LS-Qは強化学習法の一つであるQ学習にLSモデルを応用したものである。Q学習は，行動の価値を表すQ値を次式によって更新することによって学習を行う。

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

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。

Human Cognition Inspired Reinforcement Learning Architecture and its Application to Giant-swing Robot

Yuki YANAGIHARA, Rinnosuke YANAGI, Daisuke URAGAMI

LS-Q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1のように記録し，それを「C-Table」と呼ぶ。例えば，表１の事象 p がある行動Aで，その行動がGreedyであることが事象 q であり，行動Aを選択してその行動がGreedyな行動であった回数が a である。C-Tableに対して式（1）を適応してLS-Greedyな行動を決定する。例えば，ある状態において行動Aと行動Bが可能である場合， LS(Greedy|A) と LS(Greedy|B) を計算して値の大きい方がLS-Greedyな行動となる。学習の過程ではLS-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。

Q学習は，実装が容易で多くの対象に適用可能であるが，学習に要する試行回数が膨大であることや環境の不確実性の影響が大きいなどの課題がある．これに対して，LS-Qは学習速度を改善し且つ不確実性の大きい環境において適応的に探索をおこなうということが明らかになりつつある

２．３　鉄棒ロボット

本研究では，オリジナルな鉄棒ロボットを設計・自作した（図１）。鉄棒とロボットの接続（第１関節）はフリーで，腰部の関節（第２関節）のみが能動的に稼働する。腰部の間接をタイミング良く伸縮することによってロボットの状態を制御することが課題となる。比較的シンプルであるにもかかわらず非線形で動的なシステムであるため，状態分割の精度を変えることにより環境の不確実さの程度を調整できる点が，学習対象として鉄棒ロボットを採用した理由である。鉄棒ロボットは理論的な研究対象と実用的な研究対象の中間に位置し，両者の橋渡しをするための研究対象として適している。

２．４　LS-Qの鉄棒ロボットへの応用

本研究では，鉄棒ロボットの運動獲得をテスト課題としてLS-Qの学習能力を検証する。学習アルゴリズムの概要は図２のとおりである。状態は第１関節の角度と速度，第２関節の角度をそれぞれ分割して離散的に定義した。状態数は １２×７×５＝４２０ である。行動は第２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。C-Tableが表１とは異なり２×３のテーブルであるが，a＝l , b＝u , c＝m＋n , d＝v＋w とおくことによりLS値を算出する。

C-TableからLS値を算出して行動を選択，状態遷移と報酬の獲得，Q値の更新，C-Tableの更新，を繰り返すことにより学習は進行する。行動選択と学習は０．１秒に１回おこない，１０秒（１０００回）を学習の１セットとする。１セットごとにロボットの状態を初期状態に戻した。学習の１セット目はランダムに行動を選択し，徐々にLS-Greedyな行動を選択する確率を増加させ、２００セット目以降はLS-Greedyな行動のみを選択するようにした。学習回数はトータルで２５０回である。学習はシミュレーションによっておこない，獲得された行動パターンを実ロボットに移植して実験をおこなった。

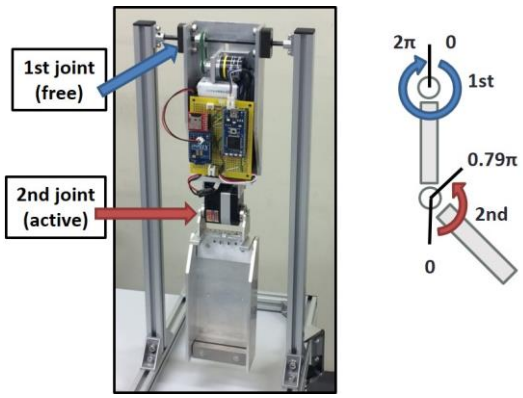

図１　鉄棒ロボット

図２　学習アルゴリズム

3 結果

図３は，学習によって獲得された運動である。上図（ａ）は LS-Q による結果で，下図（ｂ）はQ学習による結果である。それぞれ実ロボットによる実験とシミュレーションの両方のグラフを重ねてプロットしている。LS-Qの場合は，第２関節をタイミングよく稼働せることにより第１関節の回転角が徐々に増加していることがわかる。実ロボットとシミュレーションでわずかな差があるが，両者とも約８秒後に回転運動に至っている。一方，Q学習の場合は，第２関節は振り上げた状態で固定されているため，第１関節の回転角は小さい状態に留まっている。

以上の結果により，LS-Qの有効性が実ロボットにおいて確認された。

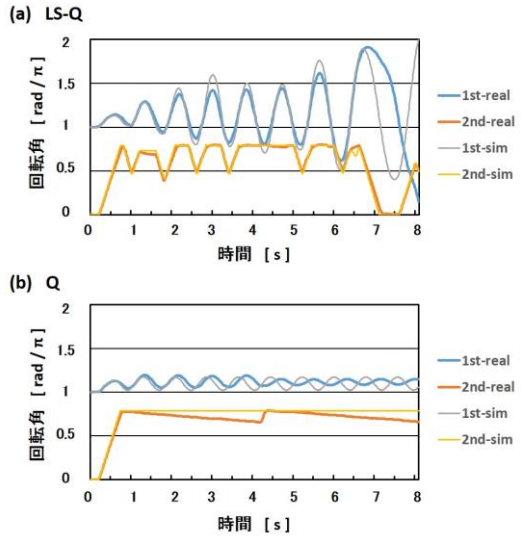

図３　学習によって獲得された運動

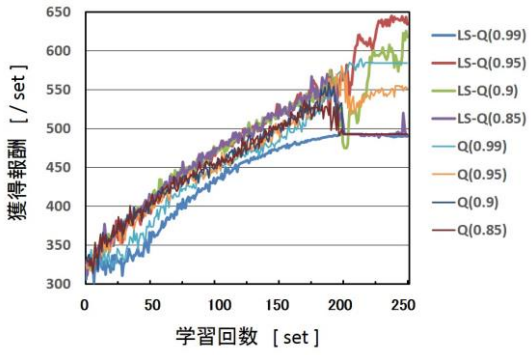

図４　割引率による学習曲線の比較

尚，図３（ｂ）において実ロボットの第２関節が少しずつ降下している原因は，サーボモータの角度で維持するためのフィードバック信号にタイムラグがあるせいだと考えられる。また，実ロボットの第１関節の回転角の振動が徐々に減衰しているのは，空気抵抗が主な要因と考えられる。

図４はシミュレーションによる結果で，割引率の影響を学習曲線の比較によって調査したものである。LS-QおよびQ学習について，割引率 $\gamma = 0.8, 0.85, 0.9, 0.95, 0.99$ とした。それぞれ５試行の平均値をプロットした。獲得報酬の最終的な値によると，LS-Qで$\gamma = 0.95$の場合が最も良く，LS-Qで$\gamma = 0.9$の場合が次に良いということが確認できる。

割引率は，将来的に得られる報酬を現在の行動選択にどの程度考慮するかの指標であると，直感的には説明できる。したがって，適当な割引率は報酬の設定や学習時間に依存する。一般的には，割引率は設計者が学習環境に応じて個別的に設定するか，あるいは割引率自体も学習によってチューニングする。前者は設計者の経験と直感に依拠し，後者は技巧的で複雑なアルゴリズムが必要である。LS-Qのような比較的簡素なアーキテクチャで割引率のチューニングが必要ないのであれば，高い汎用性が期待できる。

4 おわりに

ロボットが活動する実環境の状態量は一般的には連続値であるため，状態の離散化の精度や価値関数の関数近似能力が学習の成否に大きく関係する。近年注目を集めているDeep Q-Network（DQN） は関数近似器を用いた精度の高い状態表現の中での学習アルゴリズムである。一方，本研究では，あえて関数近似器を用いずに，粗い状態分割のもとでロボットの行動獲得を学習課題として取り上げた。粗い状態分割のもとでの学習は，アルゴリズムの簡素化や計算の高速化という点において依然として重要な研究テーマである。

本研究では鉄棒ロボットの運動獲得を学習課題として，提案手法であるLS-Qの，粗い状態分割のもとで有効性を検証した。図３に示され実験結果で確認できるように，本研究の学習環境ではQ学習は適切な運動を獲得できていない。これは粗い状態分割に起因する非マルコフ性の影響であると考えられる。一方，LS-Qでは適切な運動が獲得できている。LS-Qは，一般的には内部モデルの学習や状態空間の関数近似を用いなければ上手く学習できないような環

境において, それらを用いずに学習に成功している。このようなLS-Qの探索・学習能力には, 鉄棒ロボットの非線形なダイナミクスとLSモデルのヒューリスティクスが複雑に関係していると考えられるが, そのメカニズムの解明は今後の課題である。

「謝辞」

本研究の一部は, 平成27年度東北大学電気通信研究所 共同プロジェクト研究H25/A12「不定な環境における適応能の階層横断的解明と工学的応用」による。ロボットの製作は, 東京工科大学の松尾芳樹教授と大学院生のアルアルワン・アリー氏の協力による。

「参考文献」

1) Uragami, D., Takahashi, T., Matsuo, Y.: Cognitively inspired reinforcement learning architecture and its application to giant-swing motion control, BioSystems, 116 (2014) p. 1-9.
2) 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 (2007) p. 58-68.