計測自動制御学会東北支部 第 346 回研究集会 (2023.12.20)
資料番号 346-21

運動特性の向上が強化学習の収束に及ぼす影響の評価

Evaluating the Effects of Improved Motion Control on the Convergence of Reinforcement Learning

○小川祐樹 *，村松鋭一 *

○ Yuuki Ogawa*, Eiichi Muramatsu*

*山形大学

*Yamagata University

キーワード： キーワード (keyword), キーワード (keyword), キーワード (keyword), キーワード (keyword)

連絡先： 〒 992-8510 米沢市城南町 4–3–16 山形大学大学院理工学研究科機械システム工学専攻
村松鋭一，Tel.: (0238)26-3327，Fax.: (0238)26-3327，E-mail: muramatu@yz.yamagata-u.ac.jp

1. 研究の背景と目的

近年，機械学習とダイナミクスの理論の融合に向けた研究の動向がある [1)~3)]．特に動的計画法を通してシステム制御理論との接点を持つ強化学習は，動的システムに対する制御の観点からのアプローチが可能な研究分野と考えられる [3)]．

スタートからゴールへ向かう経路を探索する経路計画は強化学習の典型的な問題である．ここで車輪を持つ移動ロボットを用いる場合，ロボットの運動特性が考えている問題に関わってくる．本研究では，ロボットが未知な環境でゴールへ向かう経路を探索する強化学習において，ロボットに運動特性に関する推定機構を持たせることが強化学習の効率向上につながるのかという問題を考えてみる．

このような経路の探索においては，どこに水たまりがあるか，ゴールはどこかについて情報は報酬として得られる．強化学習では，行動選択に伴う状態遷移で報酬を得ながら，行動価値関数をできるだけ少ないエピソード回数で環境と整合のとれた値になるように学習することが望まれる．

ロボットが報酬を環境から受け取って学習するとき，自身の運動特性を推定しないロボットでも，報酬を環境から受け取るならば環境との整合性を保ちながら行動価値関数を学習できる可能性がある．したがって，運動特性を推定できることが学習効率の向上につながるかどうかは自明な問題ではない．

一方，強化学習は動的計画法との関連をもち，現在から未来へ向けて得られる報酬の総和を大きくしようとする意図が学習則の背景に存在している．学習則に状態遷移が関わり，その状態遷移にロボットの運動特性が関わるならば，運動特性の推定能力が学習効率に関わる可能性もある．

本研究では，運動特性の推定機構を持ち，その推定と強化学習を同時進行するロボットを提案する．そして，そのロボットによる経路計画

問題を考え，運動特性の推定が強化学習の効率に影響を及ぼすかを評価する．ロボットは水たまりがある場所ではスリップによって運動特性が劣化するとする．またロボットは，環境から得られる報酬をもとに水たまりの有無を検知し，その情報を自身の運動特性のパラメータの推定に用いる．このような推定機構を持つロボットに強化学習によるゴール探索を行わせる．そして，推定機構の有無と学習効率の関係をシミュレーションによって評価する．

2. 強化学習

2.1 強化学習の定式化

強化学習問題ではエージェントは行動から状態遷移し, その状態によって環境から報酬が与えられる. この相互作用の関係からエージェントは方策を変更して学習をする. 基本的に強化学習はマルコフ決定過程を対象としている. マルコフ決定過程とは時刻 $t+1$ の状態 s_{t+1} は, その前の状態 s_t と行動 a_t によって決まり, それ以前の値には依存しないことを言う.

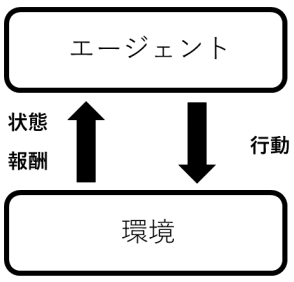

Fig. 1　エージェントと環境の相互作用

強化学習は環境から報酬を与えられることで学習を行うがその時の方策 π での状態価値は（1）式で表される.

$$V^{\pi}(s) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \,|\, s_t = s\right] \quad (1)$$

$V^{\pi}(s)$ は状態価値関数と呼ばれる．ここでの γ は割引率を表している.$\gamma < 1$ より時刻 t が進むことで割引率が大きくなり報酬の重みが小さくなる.(1) 式はさらに

$$V^{\pi}(s_t) = r_{t+1} + \gamma V^{\pi}(s_{t+1}) \quad (2)$$

と表される. ここで,(2) 式の状態価値をもとにどの行動をとるべきかの評価を考える. 状態 s_t で行動 a_t を実行した時の価値は $Q(s_t, a_t)$ を用いて

$$Q^{\pi}(s_t, a_t) = r_{t+1} + \gamma Q^{\pi}(s_{t+1}, a_{t+1}) \quad (3)$$

と表される. これを行動価値関数 Q と呼び, 方策の改善で用いる. ある方策の中で最大の価値を与えるときの価値関数は

$$V^*(s_t) = \max_a V^{\pi}(s_t) \quad (4)$$

$$Q^*(s_t, a_t) = \max_a Q^{\pi}(s_t, a_t) \quad (5)$$

で表現し, 強化学習はこの最適価値関数を推定することが目的である.

2.2 強化学習の手法

状態価値関数または行動価値関数において (1) 式のように期待値の計算過程で報酬関数と状態遷移確率が既知の場合は動的計画法で解くことができるが, 通常の強化学習ではどちらかが未知, または両者ともに未知のものとして扱っている. そのため期待値の計算ができないため, 今起きた事象から推定を行う.

$$\begin{aligned} Q(s_t, a_t) \leftarrow & (1-\alpha) Q(s_t, a_t) \\ & + [r + \max_{a'} Q(s_{t+1}, a_{t+1})] \end{aligned} \quad (6)$$

この手法は, 実際に得られる報酬 r_{t+1} を含んでいるため左辺の Q 値よりも正しい推定値を得ることができる. また, 状態遷移後の状態 s_{t+1} では最大値の Q 値を用いるこの手法を Q 学習と呼ぶ.

3. 運動特性モデルの推定方法

本研究で扱う移動ロボットの制御量を示す.

$$a_t = (v, \omega)^T \quad (7)$$

v は速度, ω は角速度を表している.

次に, 2 輪型移動ロボットのモデルを示す.

$$\begin{bmatrix}\dot{x}_t\\ \dot{y}_t\\ \dot{\theta}_t\end{bmatrix} = \begin{bmatrix}\dot{x}_{t-1}\\ \dot{y}_{t-1}\\ \dot{\theta}_{t-1}\end{bmatrix} + \begin{bmatrix}v_t\cos(\theta_{t-1})\Delta t\\ v_t\sin(\theta_{t-1})\Delta t\\ \omega_t\Delta t\end{bmatrix} \quad (\omega_t = 0), \tag{8}$$

$$\begin{aligned}\begin{bmatrix}\dot{x}_t\\ \dot{y}_t\\ \dot{\theta}_t\end{bmatrix} &= \begin{bmatrix}\dot{x}_{t-1}\\ \dot{y}_{t-1}\\ \dot{\theta}_{t-1}\end{bmatrix}\\ &+ \begin{bmatrix}v_t\omega_t^{-1}\left[\sin(\theta_{t-1}+\omega_t\Delta t)-\sin(\theta_{t-1})\right]\\ v_t\omega_t^{-1}\left[\cos(\theta_{t-1})-\cos(\theta_{t-1}-\omega_t\Delta t)\right]\\ \omega_t\Delta t\end{bmatrix}\\ &(\omega_t \neq 0).\end{aligned} \tag{9}$$

この運動モデルに影響を及ぼす外乱を推定することを考える.

$$a_t = (cv, \omega)^T \tag{10}$$

$$\begin{cases} c = 1 & (\text{水たまりなし})\\ c = 0.5 & (\text{水たまりあり}) \end{cases} \tag{11}$$

制御量 a_t に水たまりによるスリップ現象を再現させる. このスリップ係数 c を (11) 式のように設定し, 速度 v に影響を与えることとする.

エージェントは水たまり中では負の報酬 r を基に判定をし, また報酬とスリップ係数の関係がモデル化されていることを前提として推定を行う.

$$\begin{cases} g_r = 1 & (r < 0)\\ g_r = 0 & (r = 0) \end{cases} \tag{12}$$

g_r は報酬から設定される環境を示すパラメータである. ここでエージェントは環境パラメータ g_r から自身がもつ環境パラメータ $\hat{g}$ を (13) 式で更新し, スリップ係数 c のモデル式（14）に代入することで推定を行う.

$$\hat{g} = \hat{g} + \beta(g_r - \hat{g}) \tag{13}$$

$$\hat{c} = -0.5\hat{g} + 1 \tag{14}$$

この推定機構を持たせた学習のシステムを Fig2 に示す

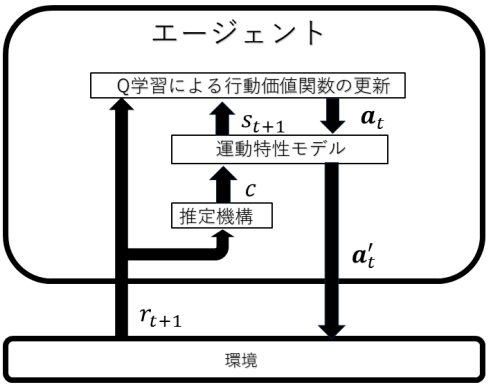

Fig. 2　推定機構を備えた学習

4.　シミュレーション

今回のシミュレーションでは, 連続状態空間を離散空間に変換することで Q 学習を行う. 離散化による状態空間は $S = (x, y, \theta)$ とし, ロボットは状態遷移後の状態 S_{t+1} と報酬 r_{t+1} を与えられる.

ここで報酬関数を定義する.

$$\begin{aligned} r_{time}(s, a, s') &= -\Delta t\\ r_{puddle}(s, a, s') &= -10\Delta t \end{aligned} \tag{15}$$

(15) 式において Δt は 1 ステップを示し, 時間にして 0.1[s] とする. またゴールについた場合 $r = +10$ を与え, 1 エピソードでステップ数が 220 回を越えた場合 $r = -10$ を与えて次のエピソードを実行する. また, 初めに何も規則性のない方策を与えることは学習に時間がかかってしまうため, 初期位置固定と初期位置からゴールへまっすぐ進むような方策を与えることとする.

まず初めに, 運動モデルを用いない通常の Q 学習を示す (Fig2). エピソード数は 100 回で各エピソードごとに得られる総報酬を縦軸に示している.

Fig3 より通常の Q 学習では, エピソード数が 80 回目程度で収束し始めていることがわかる.

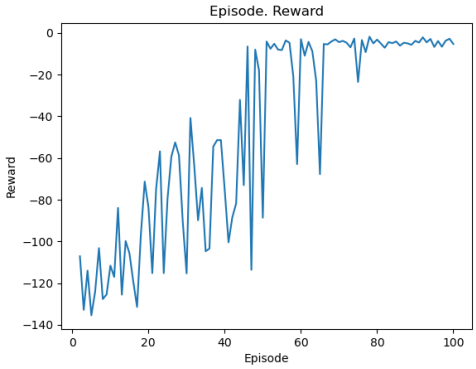

Fig. 3　Q 学習の様子

次は, エージェントに不完全な運動特性モデル持たせてスリップ係数を推定しない場合と推定させる場合に分けてシュミレーションを行う. この運動特性モデルは (6) 式における右辺の $\max_{a'} Q(s_{t+1}, a_{t+1})$ を推定するときに用いることとし, 行動価値関数 Q 値を更新する.

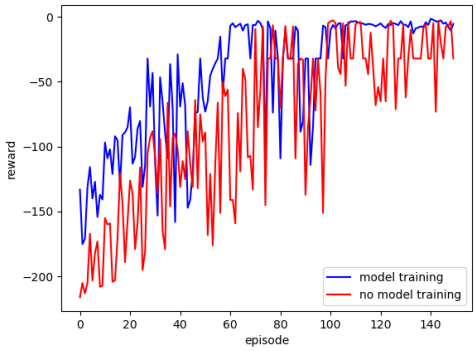

Fig. 4　運動特性モデルを用いた学習の様子

スリップ係数を考慮する場合としない場合で比較する (Fig4).　全体的に水たまりによるスリップ係数を推定しながら学習する方が各エピソードにおいて総報酬が高い傾向が見られた. これは,(6) 式の $\max_{a'} Q(s_{t+1}, a_{t+1})$ の値がエージェントが持つ運動特性モデルの推定と実際の運動特性と比較して誤差が小さくなっていくから $Q(s_t, a_t)$ の値は適切に更新される.

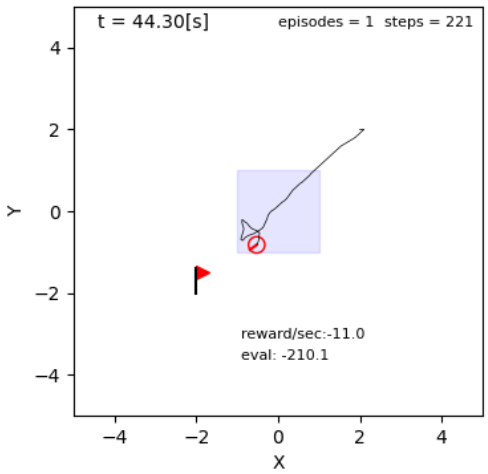

Fig. 5　推定なし学習開始直後の様子

また, 水たまりによるスリップ係数をずっと推定しない場合, 水たまりの中にいるときに制御量がスリップによって変更されていることをエージェントが知らないため水たまりなしの運動制御をもとに状態遷移が考えられてしまう. そのため初期の学習段階において, 実際には移動ロボットはゴールまで着かず, 環境からゴール情報をもらえずに迷っている動きになっている (Fig5).

5.　まとめ

今回のシミュレーション結果から通常の強化学習では, 遷移後の状態観測が正しく行われていることを前提とした学習であるため, 環境から与えられる情報に誤差がないことが総報酬と学習の収束において良い影響を与える結果となった. しかし, 状態観測が不確かさを伴う場合や観測不可能な場合, いかに状態を正しく観測するか, また推定できるようになることが重要である. そのため移動ロボットに環境による運動特性を持たせることや推定する機構の有無では獲得できる総報酬や学習の収束に影響を及ぼすことが明らかにすることができた.

参考文献

1) 加嶋健司:機械学習と調和する制御理論を模索して，計測と制御，vol.58, no.3, pp.153-155 (2019)

2) 大川，佐々木，岩根:強化学習とモデルベースト制御を並列した制御アプローチ，第 61 回自動制御連合講演会資料，pp.152–159 (2018)

3) 加嶋:制御工学者のための強化学習入門，計測と制御，vol.58, no.3, pp.182-188 (2019)

4) R. S. Sutton and A. Gl. Barto : 強化学習，森北出版 (2022)

5) 上田隆一 : 詳解確率ロボティクス，講談社 (2019)