適格度トレースと報酬予測により温度変化を行う Softmax を用いた環境変化に適応する強化学習

Adaptive reinforcement learning in dynamic environment using eligibility traces and softmax controlled by reward prediction

亀井 圭史 (PY)†, 石川 眞澄 ‡
Keiji Kamei(PY) and Masumi Ishikawa
† 西日本工業大学 工学部 電気電子情報工学科
‡ 九州工業大学 大学院 生命体工学研究科
kamei@nishitech.ac.jp

Abstract— Reinforcement learning is considered to be suitable for navigation of mobile robots in mainly static environment. In this paper, we propose adaptive reinforcement learning in dynamic environment, and succeed in improving the learning performance significantly.

Keywords—**Reinforcement Learning, Eligibility Trace, Dynamic Environment**

1 はじめに

強化学習 [1] は移動ロボットの最適経路学習に適している。我々はこれまでにセンサ情報を導入した強化学習により学習を加速させる手法や学習パラメータ最適化について研究 [2] してきた。しかしながら，これまでは環境変化の伴わない静的な環境のみを扱ってきた。実環境での応用を考慮すると，環境が変化する場合が殆どであり，環境変化に対応する学習法が必要である。

本研究では，学習過程が進んだ段階で急激に環境を変化しする場合に，それに適した学習法を提案し，シミュレーション実験によりその有効性を示した。

2 強化学習

2.1 適格度トレース: Q(λ) 学習

適格度トレース [1] は，観測された全ての系列から学習するモンテカルロ法と観測された 1 エピソードのみから学習する Q 学習の中間的な学習法である。これは，1 つのみの報酬より多く，1 エピソード終了までの報酬よりは少ない報酬で学習する。本研究では式 (1) で表される Watkins の Q(λ) 学習を用いた。

$$
\begin{aligned}
&Q_{t+1}(s,a) = Q_t(s,a) + \alpha\delta_t e_t(s,a) \\
&\delta_t = r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s_t,a_t) \\
&e_t = I_{ss_t}\cdot I_{aa_t} + \begin{cases} \gamma\lambda e_{t-1}(s,a) \; ; Q_{t-1}(s_t,a_t) \\ \qquad\qquad = \max_a Q_{t-1}(s_t,a) \\ 0 \; ; otherwise \end{cases}
\end{aligned}
\tag{1}
$$

ここで，I_{xy} は一致関数で，$x = y$ なら 1，それ以外では 0 である。$\lambda = 0$ なら従来の Q 学習と同じとなる。

2.2 報酬予測から温度変化させる Softmax 行動選択

強化学習の行動選択方法の 1 つに Softmax [1] がある。Softmax の各行動の選択確率は式 (2) で表される。

$$
\Pr(a) = \frac{e^{Q_t(a)/\tau}}{\sum_{a=1}^{n} e^{Q_t(a)/\tau}} \tag{2}
$$

ここで，$Q_t(a)$ は Q 値，τ は温度パラメータであり，温度パラメータ τ が大きければ行動のランダム性が大きくなり，逆に小さければ $greedy$ に行動する。τ は式 (3) で減少させた。

$$
\tau = \frac{0.1}{1 + 0.02 \cdot S \cdot R_c} \tag{3}
$$

ここで，S は移動ロボットが移動に成功した履歴，R_c は予測報酬と実報酬の差を表し，それぞれ式 (4)，式 (5) で設定する。

$$
S = \begin{cases} 0 & ; Collide\,to\,obstacles \\ S+1 & ; otherwise \end{cases} \tag{4}
$$

$$
R_c = \begin{cases} 1/||r_e| - |r|| & ; r_e \neq r \\ 1 & ; otherwise \end{cases} \tag{5}
$$

ここで，r_e は予測報酬，r は実際に得られた報酬である。予測報酬とは，移動した結果得られる報酬をエージェントが予測したものであり，報酬は表 1 に示す様に，我々が設定している。S 及び R_c の値が大きくなれば式 (3) の温度 τ が小さくなる。

一般に，Softmax では学習エピソード数に応じて温度係数を変化させるが，これは環境が変化しても温度が変化しない。本研究では，環境変化に対応するために，**式 (4)：“衝突せずに行動できた回数”，式 (5)：“予測報酬と実報酬の差”** の 2 点から温度を設定する。これにより，学習エピソード数に関係なく探索環境に対する学習の程度により，探索的な行動と greedy な行動を適切に選択するようになる。

3 実験

3.1 シミュレーション設定

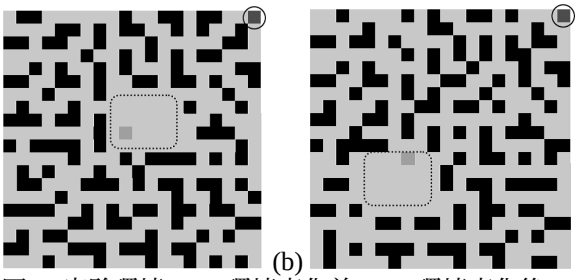
本研究では，仮想移動ロボットを定義してシミュレータにより実験した。シミュレーションの強化学習パラ

メータは表 1 である。学習環境は学習開始後 3500 エピソード経つと変化する。環変化前は図 1(a)，変化後は図 1(b) である。 移動ロボットは，上下左右の行動を選択

表 1: 強化学習設定

学習率	割引率	報酬		
		衝突	移動	ゴール
0.43	0.999996	-114.29	-1.43	1.00
λ	ϵ			
0.55	0.001			

し，スタート地点はエピソード毎にランダムに設定される。



(a) (b)

図 1: 実験環境，(a) 環境変化前，(b) 環境変化後

図 1 において，右上端部分の円内がゴールで，中央付近の点線内には 1 ステップ毎に，“上下左右にランダムに動く”，もしくは“そのまま停止する” 移動障害物がある。

3.2 シミュレーション実験結果

(1)$\epsilon-greedy$ 行動選択と従来型 Q 学習，(2)Softmax 行動選択と従来型 Q 学習，(3)$\epsilon-greedy$ 行動選択と Q(λ)，(4)Softmax 行動選択と Q(λ) の 4 種類について比較実験を行った。図 2 にゴール回数の比較図，表 2 にゴール，衝突回数，そして学習定常状態でゴールに到達できたエピソードの平均行動数比較を示す。

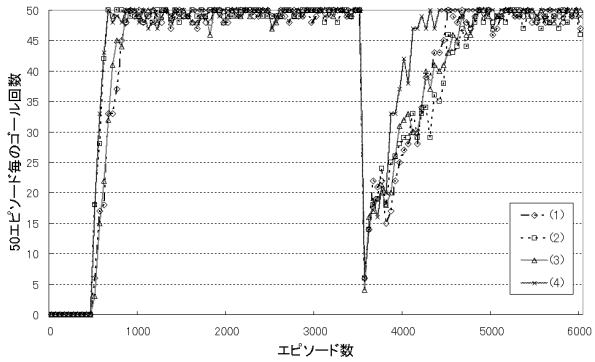


図 2: 実験結果，ゴール回数比較

ゴール回数について，学習初期では ϵ-greedy，Softmax 共に適格度トレースを用いた方が良い結果となった。複数のステップを考慮するため，より速い学習となったためである。環境変化後は，本研究の提案手法がゴール到達数増加率と総ゴール数の両者について最良の結果となっている。衝突回数について，学習初期では適格度トレースを用いた両者は同じとなり，その他より良い結果

表 2: ゴール，衝突回数，平均行動数の比較

	エピソード	(1)	(2)	(3)	(4)
ゴール	[1, 1000]	386	482	401	487
	[3500, 4500]	619	592	618	766
衝突	[1, 1000]	491	468	471	468
	[3500, 4500]	334	318	318	212
行動数	[5000, 6000]	31.86	38.10	31.33	40.43

となっている。環境変化後は，本研究の提案手法が他よりも 100 エピソード程度減少しており，減少率も最良である。定常状態のゴールに到達した平均行動数では，提案手法が悪くなっているように見えるが，これはゴールに到達できなかったエピソードについて除外しているためであり，ゴールに到達できた本研究の提案手法がゴールに多く到達しているため増加していると考えられる。

次に，本研究で提案した手法での Softmax の温度変化を図 3 に示す。図より，学習開始直後は温度が高く学

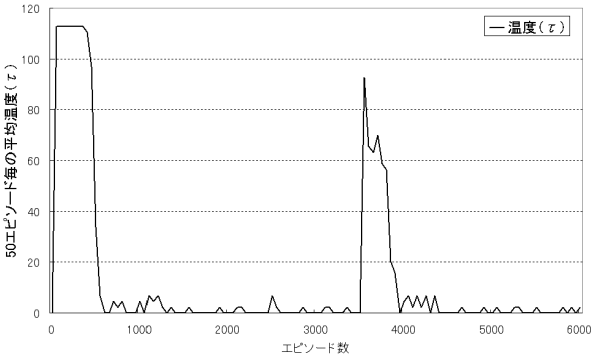


図 3: Softmax 温度 τ の変化

習エピソードが増えるに従って温度が下がるが，環境が劇的に変化した 3500 エピソード後では再び大きく上昇していることがわかる。

4 おわりに

本研究では，移動ロボットの動作する環境の劇的な変化に対応するための強化学習について，報酬予測により温度変化を行う Softmax を用いた Q(λ) 学習を用いることを提案した。シミュレーション実験の結果から，従来よりも柔軟に環境変化へ対応できる方法であることが証明された。今後は，パラメータ，特に λ の値について問題依存性の検証を行い，実機移動ロボットへ適用したいと考えている。

参考文献

[1] Richard S. Sutton and Andrew G. Barto (1998) “Reinforcement Learning,” MIT Press

[2] Keiji Kamei and Masumi Ishikawa (2006) “Dependency of values of parameters in reinforcement learning for navigation of a mobile robot on the environment.” Neural Information Processing - Letters and Reviews, Volume 10 Numbers 7-9 July - Sept.