

動的環境を対象とした適応的強化学習の提案

齋藤智輝†

大枝真一‡

木更津工業高等専門学校 情報工学科本科†

木更津工業高等専門学校 情報工学科‡

1. まえがき

学習は生物だけでなく人工システムにとっても、環境に適応するための重要な機能であると考えられる。従来の多くのシステムでは時刻によって動的に変化する環境（以下、動的環境）に適応させるために、設計者が状況に依存しない環境として、起こり得る変化をすべて事前に想定する必要があった[1]。この場合、環境ごとにシステムを用意し、個々に特化した学習を行い、環境が変化したときに切り替えさせることも考えられるが、この手法では各環境に対して独立に学習を行っており、他の環境で学習した経験を次の学習に再利用していない。

本研究では強化学習において、動的環境下における効率的な解法について検討を行う。計算機実験としては強化学習の代表的な問題である「迷路の経路探索」を対象とする。

2. 強化学習

学習と意思決定を行う個体をエージェントと呼び、そのエージェントの学習方法として強化学習が広く用いられている。強化学習の目標は、未知の環境に置かれた知的エージェントが環境との間の相互作用を通して、環境において報酬が最も多く得られるような方策を学習することである。本研究では強化学習として有名な手法の1つである Q 学習法 (Q-Learning) を用いることとした。

2.1 Q 学習法

Q 学習法[2]は行動価値関数 Q を使用している行動選択方策とは独立に、最適行動価値関数 Q^* を近似する方法である。

本研究で用いた Q 学習の1ステップにおける更新式を式(1)に示す。この更新式を用いて行動価値関数 Q を近似する。具体的に、ある時刻 t において行動を決定した際の Q 値の更新の様子を図1に示す。

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

s_t : 時刻 t における状態(statement)

a_t : 時刻 t における行動(action)

r_{t+1} : 時刻 $(t+1)$ における報酬(reward)

α : 学習率 γ : 割引率 a : 行動

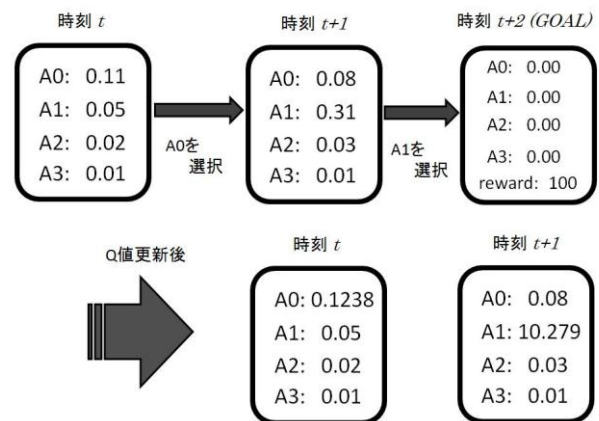


図1: Q 値の更新の様子

2.2 行動選択方策

エージェントと環境は離散的な時間ステップの各々において相互作用を行う。各時間ステップ t においてエージェントは何らかの環境の状態の表現 s_t を受け取り、これに基づいて行動 a_t を選択する。各時間ステップ t においての行動選択方策によって学習効率が大きく変化していくが、本研究では ϵ -greedy 方策[2]を用いる。

2.3 ϵ -greedy 方策

ϵ -greedy 方策では確率 ϵ でランダムに行動を選択するが、主に時間 t における最大価値をもつ行動を選択する。

ϵ を適切に設定すると常に時刻 t における最大価値を持つ行動を取るときとは異なる環境を作れるため、解空間が広がっていく。すなわち、局所解に陥りにくくなるという利点がある。

デメリットとしては、 ϵ が大き過ぎると等確率ランダム方策に近いものになってしまう、逆に ϵ が小さ過ぎると局所解に陥る可能性が高くなることが挙げられる。

A proposal of adaptive reinforcement learning for dynamic environments

†Tomoki Saito, Dept. of Information and Computer Engineering course, Kisarazu National College of Technology

‡Shinichi Oeda, Dept. of Information and Computer Engineering, Kisarazu National College of Technology

3. 計算機実験

「迷路の経路探索」として強化学習を用いる。動的環境下での観測を行うために Goal までの複数経路を持つ迷路に一定の周期で通行することができない場所を作成していくこととした。作成した迷路の例を図2に示す。

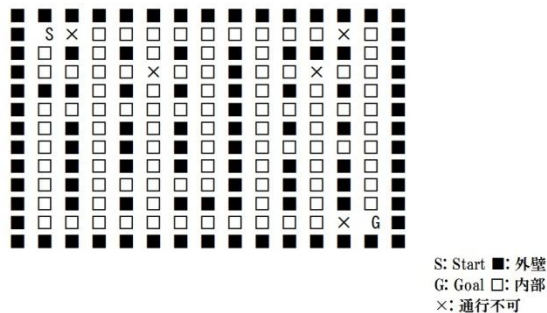


図2：作成した迷路（13×15）

3.1 極端な周期による Q 値の初期化

動的環境下において環境が変化すると Q 値を初期化する場合と、環境が変化しても Q 値を初期化しない場合の2通りについて実験した。実験した結果を図3,4に示す。なお、この実験では150試行毎に環境を変動させた。

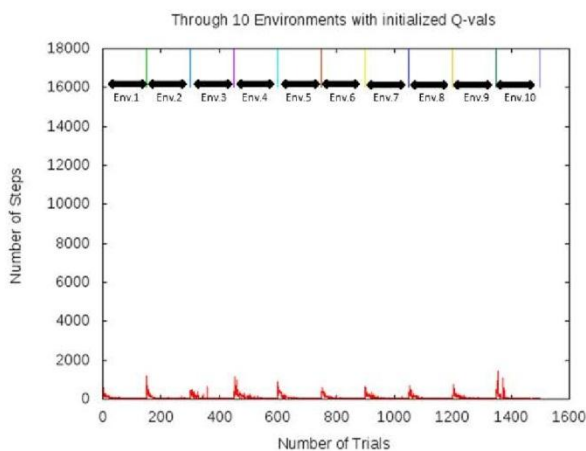


図3：環境が変化する毎に Q 値を初期化した場合

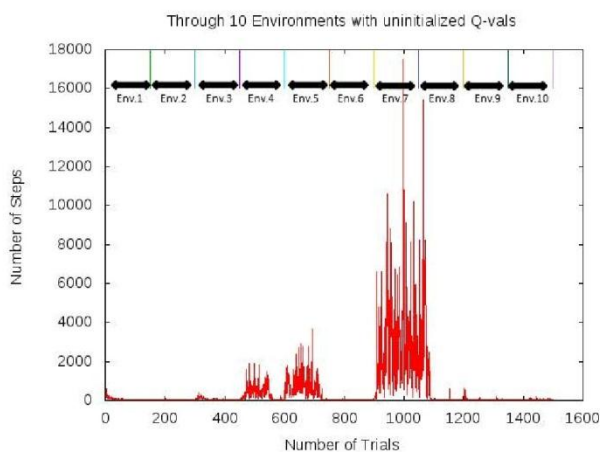


図4：環境が変化しても Q 値を初期化しない場合

3.2 閾値を用いた Q 値の初期化

本実験では、他の環境での学習結果を再利用することにより性能の向上を期待していたが、実験結果から、環境変化毎に Q 値を初期化しない方法では Q 値を初期化する方法と比べて性能の向上が見られなかった。そのため、閾値を用いて Q 値を初期化するかしないかということを決定することとした。閾値は初期環境についてランダム探索によるステップ数とした。擬似乱数関数を用いたため Seed を5通りに変化させて観測を行った。閾値を変化させたときの実行時間を表1に示す。

表1：閾値の倍率による実行時間の比較

	実行時間[s]			
閾値(Step)	0.5 倍	1.0 倍	1.5 倍	2.0 倍
Seed1 (875)	2.88	0.26	0.32	0.71
Seed2 (925)	15.25	1.35	0.49	0.58
Seed3 (865)	8.68	0.96	0.63	0.92
Seed4 (519)	9.50	4.42	1.62	0.56
Seed5 (625)	5.82	2.08	0.31	0.41
Avg.	8.426	1.814	0.674	0.636

表1から閾値の倍率を0.5倍にしてみると実行時間が大幅に長くなってしまったことがわかる。これは Q 値を初期化する回数が多くなってしまると同時に Q 値を初期化後のランダム探索で、その閾値以下のステップ数で探索を終える可能性が低いからだと考えられる。

閾値の倍率を1.5倍、2.0倍にした場合は実行時間を約1/3程度にすることができた。

4. まとめ

本研究では動的環境下における効率的な解法について検討を行った。実験より、動的環境に適用させるためには Q 値を常に初期化する方法や逆に Q 値を全く初期化しない方法などの極端な方法ではなく、常時 Q 値を更新するか否かを判断する対策が必要であると考えられる。

しかしながら、対策をすることによって環境に依存はするものの前の学習結果を受け継ぐことができるため、単に初期化された Q 値を用いた学習よりも早い時間で学習を完了することができ、動的環境に適応することができる。

参考文献

- [1] 港 隆史, 浅田 稔, “環境の変化に適応する移動ロボットの行動獲得”, 日本ロボット学会誌, Vol.18, No.5, pp.706-712, 2000.
- [2] 三上 貞芳, 皆川 雅章, “強化学習”, 森北出版, 2000.