Actor-Critic によるロボットの強化学習

2005MM029 木下 直人
指導教員：高見 勲

1 はじめに

ロボットをさまざまな環境で適用する要求に対し，ロボット自身に知能をつけるための学習手法である強化学習 [1] の研究がされている．強化学習のなかで一般的な方法として用いられている TD 誤差学習の方法の一つである Actor-Critic 法に注目する．本研究では Actor-Critic 法により迷路に対して最短経路を学習により探索する手法に取組み，ロボットを用いた実験によりその有効性を検証する．

最短経路探索問題に対してはダイクストラ法 [2] があるが，障害物がどこにあるかわからないような迷路に対して試行錯誤し経路を学習していく Actor-Critic 法により最短経路を導く．

2 TD 誤差学習 (Temporal Difference learning)

TD 誤差学習は強化学習の一般的な手法の一つである TD 誤差学習の学習方法は, 自分自身の評価をおこない, それを更新していくことである．評価値の更新には TD 誤差（見積もりと，実際に行動した時に得られる評価値の誤差）を使い，この TD 誤差を 0 に近づけていく学習法である．

TD 誤差は次式となる．

$$\begin{aligned}&\text{(行動して得られた評価値) – (見積もり)}\\ &= r_{t+1} + \gamma V(S_{t+1}) - V(S_t) \qquad (1)\end{aligned}$$

評価値の更新は次式となる．

$$V(S_t) \leftarrow V(S_t) + \alpha(r_{t+1} + \gamma V(S_{t+1}) - V(S_t)) \qquad (2)$$

$V(S_t)$ は時刻 t における状態 S_t の評価値であり，r_t は時刻 t における状態 S_t の報酬である．

$\gamma(0 \leqq \gamma \leqq 1)$ は割引率である．これは未知の状態評価値にノイズや遅れがあることを考慮し，評価値をそのまま用いずに割り引くものである．

$\alpha(0 < \alpha \leqq 1)$ は学習率である．これは現在の評価値と次の状態の評価値をどれだけ近づけるかを調整するものである．

3 Actor-Critic 法

3.1 Actor-Critic 法とは

Actor-Critic 法は TD 誤差学習の手法の一つである．Actor-Critic 法のメリットは，行動選択に最小限の計算量しか必要ないことと，確率的な行動選択を学習できることである．

この手法は行動選択と状態評価が独立したものとして考える．Actor は行動選択の確率分布を持ち，各状態はそれぞれ評価値をもつ．エージェントは Actor に与えられている行動選択確率に従い行動を選択し，その行動に対して Critic が評価をし，評価値を更新する．Critic は評価値から TD 誤差を求める．TD 誤差に応じて Actor の確率分布を更新する．これを繰り返すことにより，より最適な行動を選択するようになる．

3.2 Actor-Critic の演算プロセス

Actor-Critic の演算プロセスを図 1 に基づき説明する．

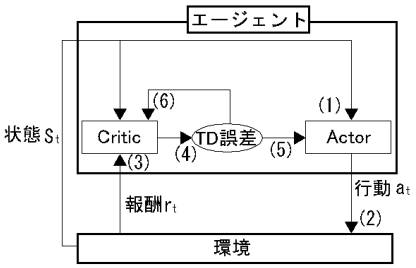

図 1 Actor-Critic の構成

状態観測 Actor が環境から状態 S_t を認識する．

行動 エージェントは Actor に与えられている確率分布に従い行動 a_t をとり, 状態 S_{t+1} へ移動する．

報酬 Critic が環境 S_{t+1} から報酬 r_{t+1} を得る．

強化信号 Critic が TD 誤差を (1) 式より計算する．

行動選択確率の更新 TD 誤差に応じて確率分布を更新する．

状態価値関数の更新 Critic は $V(S_t)$ を TD 誤差が 0 に近づくように (2) 式より更新する．

終了条件 目的の試行回数になった時学習を終了する．そうでない場合　へ戻る．

3.3 行動選択確率の更新

例として図 2 のようにエージェントを設定する．演算プロセスの　行動選択確率の更新において TD 誤差が正の時にのみ確率分布を更新する．エージェントが行動として上を選択した時によい結果が得られたとすると，この時の更新式は (3), (4) となる．ここで c は学習調節パラメータとする．行動選択確率を更新した結果，次回の試行ではエージェントは行動として上を選択する確率が上がる．

a_0 の行動選択確率の更新は次式となる．

$$a_0 \leftarrow \frac{c + a_0}{c + a_0 + a_1 + a_2 + a_3} \qquad (3)$$

a_1，a_2，a_3 については次式となる．i=1，2，3 である．

$$a_i \leftarrow \frac{a_i}{c + a_0 + a_1 + a_2 + a_3} \qquad (4)$$

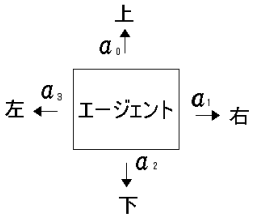

図 2　エージェント

4　シミュレーション

4.1　１０マスの迷路によるシミュレーション

今回のシミュレーションに使用する迷路は図 3 であり，エージェントは図 2 とする．試行回数は 21 回，学習率 0.5，割引率 0.9 である．確率分布の更新式は式 (3)，(4) であり，学習調節パラーメータ c を変化させることで学習の正確さを調整する．

4.2　シミュレーション結果と考察

式 (3)，(4) について，$c = 0.1$ とすると図 4 から学習の速度が遅いことがわかる．$c = 0.5$ とすると図 5，図 6 より学習で得られら経路が $0 \to 1 \to 2$ を通り最短経路とはならない．これは学習が早すぎたことが原因と予測される．以上のことから c の値を試行錯誤した結果 $c = 0.225$ とした．この時の各マスの確率分布の変化のグラフは図 7 から図 10 であり，学習後の各マスの確率の値は図 11 である．図 11 は各マスにおいて a_0，a_1，a_2，a_3 を左から表したものである．図 11 より学習に得られた経路は $0 \to 3 \to 6 \to 7 \to 9$ となり最短経路となった．この迷路には最短経路が 3 つ存在するが，100 回シミュレーション行った結果 $0 \to 3 \to 6 \to 7 \to 9$ は 32 回，$0 \to 3 \to 4 \to 7 \to 9$ は 33 回，$0 \to 1 \to 4 \to 7 \to 9$ は 35 回となった．

図 3　迷路

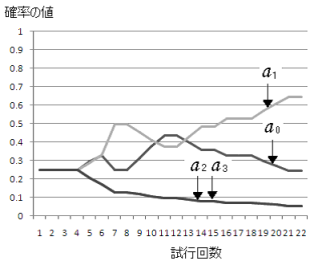

図 4　$c = 0.1$ のマス 0 の確率分布

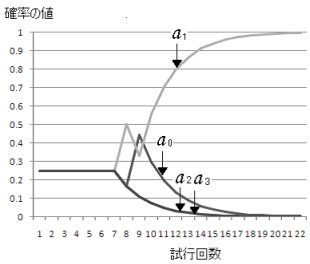

図 5　$c = 0.5$ のマス 0 の確率分布

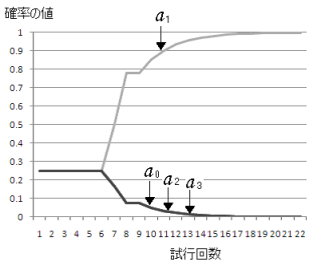

図 6　$c = 0.5$ のマス 1 の確率分布

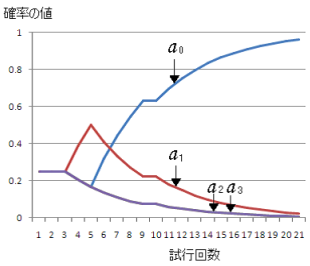

図 7　マス 0 の確率分布

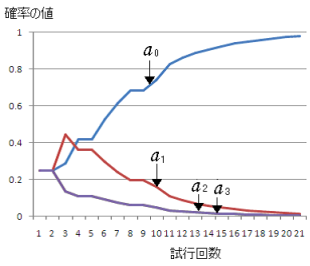

図 8　マス 3 の確率分布

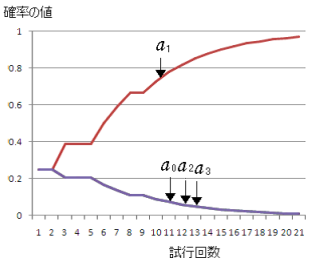

図 9　マス 6 の確率分布

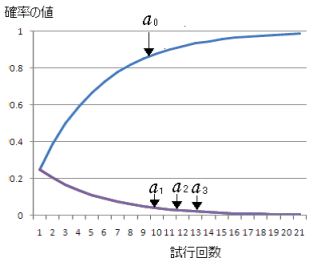

図 10　マス 7 の確率分布

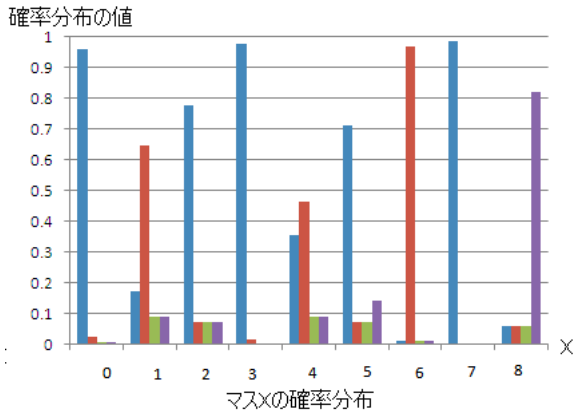

図 11　学習後の確率分布

5　おわり

本研究で得られた成果を以下に示す．

- Actor-Critic 法に離散分布を用いて学習を行った．
- シミュレーションにより Actor-Critic 法による学習から，最短経路探索問題に対して最短経路を得ることができた．
- シミュレーションにより行動選択確率の更新式である式 (3)，(4) において，学習調節パラメータ c の値により学習の速さや正確さを調整することが可能であることが確認できた．

参考文献

[1] 森北 肇：『学習とそのアルゴリズム』．森北出版，東京，2002.

[2] 福島 雅夫：『数理計画入門』．朝倉書店，東京，1996.