ISS-SP-052

動的環境を有する迷路を対象とした強化学習における学習モデルの再利用

寺井　輝†　　小宮山　摂††
† 青山学院大学大学院 理工学研究科　　†† 青山学院大学理工学部

1. はじめに

近年，強化学習の分野では動的な環境変化に対する学習方法が注目を集めている．本稿では迷路を対象とした強化学習において，一度学習した経験を再利用してより効率的に学習を進める解法について検討する．

2. 強化学習手法について

今回は迷路を対象にするため Q 学習と softmax 方策を用いている．

2.1 Q 学習

Q 学習は以下の式(1)で行動価値関数 Q を更新するアルゴリズムであり，Q の更新が方策に依存しないのが特徴である．これにより，行動価値関数の収束を早める効果が見込まれる．

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

ここで，s_t: 時刻 t における状態，a_t: 時刻 t における行動，R_{t+1}: 時刻$(t+1)$における報酬，η: 学習率，γ: 割引率である．

2.2 softmax 方策

softmax 方策は行動価値関数の高い順に行動選択確率が与えられる方策であり，状態 s における行動 a の選択確率 P を以下の式(2)で決定する．

$$P(s,a) = \frac{exp\,(Q(s,a)/\tau)}{\sum_{a' \in \mathcal{A}(s)} exp\,(Q(s,a')/\tau)} \quad (2)$$

ここで，τ: 温度$(\tau > 0)$，
$\mathcal{A}(s)$: 状態 s における選択可能な行動の集合 である．

3. 実験概要

迷路探索課題に対して強化学習を用いる．タスクは迷路のスタート地点からゴール地点まで到達することで，ゴールに到達した際にエージェントは報酬 1 を受け取る．それ以外の報酬は設定しない．今回作成した迷路は図 1 に示す．図中記号は以下の通り．
Si(0≤i≤24): 通行可能
実線: 通行不可
点線: 後に通行可能

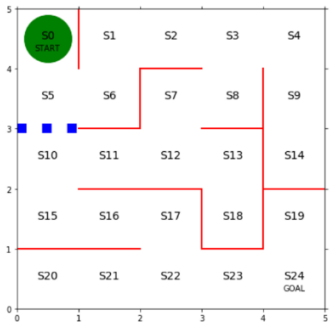

図 1: 作成した迷路

3.1 初期状態の迷路について

まずは初期状態の迷路に対して学習を行う．行動価値関数の更新には Q 学習を用いて 100 エピソード学習させ，初期迷路の学習モデル Q_F を作成する．

3.2 迷路の変更

その後，迷路の形状を変更し，新規に学習を行ったモデルと初期迷路の学習モデル Q_F を再利用して学習を行ったモデルの学習速度の比較を行った．一般に，環境変化のある迷路において，初期迷路の学習モデルをそのまま再利用すると学習が上手く進まないことがわかっている[1]．そこで，本稿では学習モデル Q_F を再利用する際に，Q_F を一定値で底上げをして薄めることにした．変換に利用したのは以下の(3)式である．

$$Q_k(s,a) = kQ_F(s,a) + (1-k)\sum_{a' \in \mathcal{A}(s)} Q_F(s,a') \quad (3)$$

ただし，k: Q_F の再利用率$(0 \le k \le 1)$ である．

4. 実験結果

学習モデルを再利用したものと新しく学習した結果との比較となるため，ゴールまでのステップ数や価値関数の変化を単純に比較することは困難である．そこで，本稿では変更後の迷路の最小手数(図 1 の場合 8 手)でゴールした際に疑似的な報酬として＋1 を設定した．k を変更したときの疑似累積報酬和を比較したものが図 2 である．

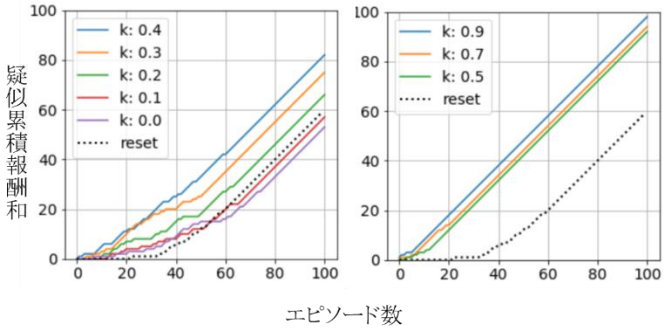

図 2: 再利用率による疑似累積報酬和の変化

0.5 ≤ k のモデルについては概ね初期状態から学習させるよりも速く最小手数に収束する結果が得られた．図1のようなショートカットの場合は一定比率以上再利用することで収束が速くなる．

5. 今後の展望

迷路の形状や変更箇所の影響を調査し，モデルの再利用時の変換式を改良していく予定である．

参考文献

[1] 齋藤智輝，大枝真一，動的環境を対象とした適応的強化学習の提案，情報処理学会第 73 回全国大会(2011)