意思決定の階層化による強化学習の学習効率の向上

意思決定の階層化による強化学習の学習効率の向上

山森　一人 a)・渡部　将人 b)・相川　勝 c)

Accuracy Improvement for TSP by Multi-Level Perturbed Parallel Island Model

Kunihito YAMAMORI, Masato WATANABE, Masaru AIKAWA

Abstract

Tracking problem is one of the popular benchmark to evaluate reinforcement learning. In the tracking problem, some hunters trace a target and try to catch target in shorter steps. In the paper, we propose to separate decision marking process of reinforcement learning from two points of view; strategy decision and tactical decision. Strategy decision decides the movement policy of the hunters, and tactical decision decides the movement direction of each hunter. Experimental results showed that our method could catch the target with 54% steps by the conventional reinforcement learning.

Keywords: Reinforcement Learning, Multi-layer decision making, Tracking problem

1. はじめに

2014年にDeepMind社の発表したDQN[1]が、ブロック崩しやピンボールなど49種類の古典ゲームで人間を負かすほどのゲームスキルを強化学習により獲得できたことが話題となった。

強化学習（reinforcement learning ： RL）[2]は、試行錯誤的に選択を繰り返すことで環境モデルに適応する教師無し機械学習の1つである。RLは、環境モデルの状態を知覚し、過去の経験から選択すべき行動を学習する学習エージェントの学習手段として用いられる。学習エージェントが選択した行動は環境モデルを遷移させ、遷移した環境モデルは学習エージェントに対して報酬を与える。学習エージェントは、行動の結果得られた報酬をもとに先の選択を評価することで学習を進めていく。RLの目的は、将来の報酬も踏まえた上で、最も多くの報酬を得る行動の選択方針を学習することである。

RLでは、実社会の環境への適応を目指して、複数の学習エージェントを用いて学習を行うマルチエージェント学習にも用いられている。その代表的な問題として、複数のハンターが1匹の獲物をより少ないステップ数で捕獲する追跡問題がよく用いられる。追跡問題を解決する手法として、渡邊らの手法[3]がある。渡邊らは、ハンターが獲物を追跡する動きの学習を、ハンターが次に向かうべき目標位置の選択の学習と、目標位置への最短経路の学習に分割することで、行動の選択を戦略と戦術の視点から２つに分割できることを示した。しかしながら、渡邊らの手法では、環境全体を考慮して目標位置を予測しているものの、獲物を追うための戦略が獲物の追尾しかなく、先回りや獲物の進行妨害といったハンター同士の協力を考慮した行動の選択がなされていなかった。

本研究では、渡邊らの手法に新たな行動戦略を追加し、高い視点から行動戦略を選択する階層型RLを提案する。具体的には、獲物を追う戦略と、先回りして挟み打ちを行う戦略の2つを用いて、追跡問題の学習を行う。本研究の目的は、追跡問題において、複数の戦略を用いることで学習効率が向上することを示すことである。

2. 追跡問題

本研究では、マルチエージェント学習の代表的な問題である追跡問題を扱う。追跡問題では、学習エージェントはハンターとして獲物を追跡し、獲物を少ないステップ数で捕獲できる行動の選択方針を学習する。本研究では以下の条件による追跡問題を用いて提案手法を評価する。

- 15×15 の正方形のフィールドを想定する。
- フィールドの境界は上下・左右をそれぞれ連結したトーラス構造とする。
- 獲物 1 体とハンター4 体を、重複しないランダムな

a) 情報システム工学科教授
b) 情報システム工学科
c) 宮崎大学工学部教育研究支援技術センター技術職員

位置に初期配置する。
- 獲物とハンターは上下左右方向へ1マス移動、その場に停止の5種類の行動を選択できる。
- 獲物とハンターは各時間ステップで同時に行動を選択する。
- ハンターは自身を中心とした周囲5×5の範囲内の状態を知覚する視野を持つ。
- 1体のハンターが獲物を見つけると、他のハンターに獲物の座標が伝達される。

3. 強化学習

本研究では、学習を通して報酬が最大になるような行動の選択方針を RL によって学習エージェントが学習する。そのため、報酬の他に、現在の状態で選択した行動を実行すると、将来的にどれだけの報酬が得られるかを表す値である価値を用いて行動を選択する。

3.1. 構成要素

RL は以下の要素から成り立つ。
- 報酬関数
  環境モデルの状態が遷移したときに、環境がエージェントに与える報酬を決定する。
- 価値関数
  現在の状態において選択した行動の価値を決定する。
- 環境モデル
  ハンターが適応する環境を示す。本論文では、追跡問題のフィールドを指す。
- 方策
  ハンターが行動を選択する方法を定義する。一般的な手法ではε-greedy 方策[4]が有名であり、時刻tの環境モデルの状態において行動aを選択する価値$Q(s_t, a)$をもとに、式(1)に従って行動aが選択される。

$$a_t = \begin{cases} arg\max_{a \in A} Q(s_t, a), & (1.0 - \varepsilon), \\ random, & otherwise. \end{cases} \quad (1)$$

ここでεは$0 \le \varepsilon \le 1.0$の小さい定数値が設定される。また、A はハンターの選択できる行動の集合である。

3.2. 学習の流れ

RL の流れを、最も多く使用されている手法である Q 学習[5]を例に説明する。時刻tの環境モデルの状態$Q(s_t, a)$において行動aを実行したときの価値関数の更新の手順を以下に示す。

STEP1. 現在の環境モデルの状態s_tを観測し、価値関数を用いて選択できるすべての行動の価値を求める。

STEP2. STEP1 で得た価値を基に、方策に従って行動を選択する。

STEP3. 学習エージェントの行動により環境モデルの状態が遷移し、報酬R_{t+1}が発生する。

STEP4. 得られた報酬R_{t+1}を基に、ハンターは式(2)により選択した行動の価値$Q(s_t, a)$の更新を行う。

$$Q(s_t, a) \leftarrow Q(s_t, a) + \alpha \cdot \mathrm{TDerror},$$

$$\mathrm{TDerror} = \left[R_{t+1} + \gamma \cdot \max_{p \in A} Q(s_t, p) + Q(s_t, a)\right]. \quad (2)$$

ここで、TDerrorは予想していた価値と実際に得られた価値の差を表す。αは学習率と呼ばれ、$0 < \alpha \le 1.0$の定数である。一方、γは割引率と呼ばれ、$0 \le \gamma \le 1.0$の定数である。

学習エージェントが、状態を観測して行動を 1 回選択するまでをサイクル（cycle）と呼ぶ。また、初期状態からサイクルを繰り返し、終了状態に到達するまでをエピソード（episode）と呼ぶ。

4. 提案手法

4.1. 階層型 RL

本研究では、行動の選択を行動戦略と行動戦術の 2 つに分割する手法を提案する。行動戦略の選択では、他のハンターとの協調を考えた上で、ハンターがとるべき行動の選択方針を選択する。一方、行動戦術の選択では、ハンターの視野内の状態を踏まえて、戦略を達成するために取るべき行動を選択する。ハンターの行動戦略として、以下の 2 種類を設定する。
- type1：
  獲物を追いかける行動戦略。
- type2：
  獲物に先回りする行動戦略。

提案する学習モデルを図 1 に示す。上位層は行動の目的、下位層は停止を含む移動すべき方向をそれぞれ学習する。上位層は現在の状態からハンターの行動戦略である type1 か type2 を選択し、下位層に伝える。下位層は目的を実現するためにハンターが実行するべき行動を選択する。また、上位層、下位層とも Q 学習によって学習を行う。Q 学習の方策にはε-greedy 方策を用いる。

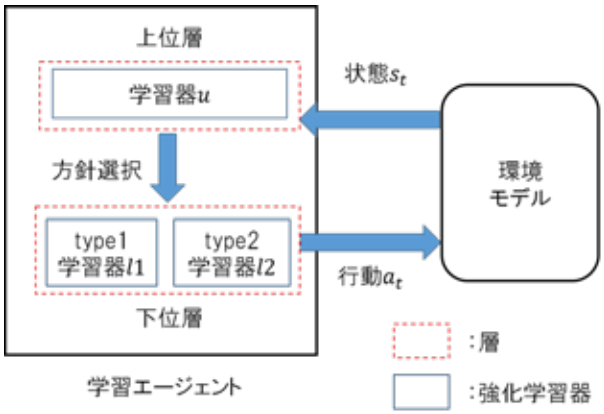

図 1：提案手法の学習モデル

4.2. 上位層の学習

上位層は、ハンターからみた獲物との距離と方向をもとに、選択すべき行動戦略を学習する。上位層が考慮する環境の状態の例を図 2 に示す。黄色のマスはハンターの視野を表し、視野の範囲内の状態のみを考慮する。獲物が視野外の場合は、ランダムウォークによって獲物の探索を行う。ただし、獲物を他のハンターが見つけている場合は、獲物が視野外でも獲物の位置を考慮する。上位層に対する報酬関数の設計内容を式(3)に示す。

$$R_{t+1} = \begin{cases} \lambda^u, & (\text{獲物を捕まえているか}), \\ \mu^u, & otherwise. \end{cases} \quad (3)$$

ここで、λ^uは正の報酬、μ^uは負の報酬を表す。

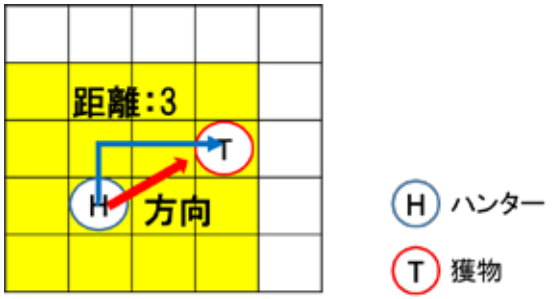

図 2：上位層の考慮する環境の状態

4.3. 下位層の学習

下位層では、ハンターの位置から見た獲物の位置までの距離をもとにハンターの行動を選択する。獲物までの距離は、上下左右のそれぞれの方向に対する座標間の差として定義する。下位層が考慮する環境の状態の例を図 3 に示す。図 3 の場合、ハンターと獲物の上下左右方向への距離は、それぞれ上 1、下 4、左 3、右 2 となる。

行動戦略による行動の違いは報酬関数の設計の仕方で制御することができる。それぞれの指針での報酬関数の設計内容を式(4)及び式(5)に示す。

- Type1

$$r^1_{t+1} = \begin{cases} \lambda^l_1, & dist_{H_t,T_t} > dist_{H_{t+1},T_t}, \\ \mu^l_1, & otherwise. \end{cases} \quad (4)$$

- Type2

$$r^2_{t+1} = \begin{cases} \lambda^l_2, & dist_{H_t,T_{t+1}} > dist_{H_{t+1},T_{t+1}} \\ \mu^l_2, & otherwise. \end{cases} \quad (5)$$

ここで、$dist_{a,b}$は位置aとbの最短距離を表す。H_tは移動前のハンターの位置、T_tは移動前の獲物の位置を表す。H_{t+1}は移動後のハンターの位置、T_{t+1}は移動後の獲物の位置を表す。また、λ^l_1とλ^l_2は正の報酬、μ^l_1とμ^l_2は負の報酬をそれぞれ表す。

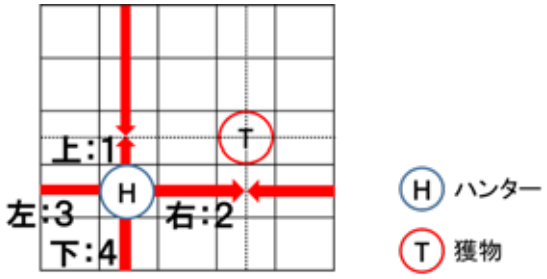

図 3：下位層の考慮する環境の状態

5. 提案手法の評価

5.1. 学習環境

評価実験は、第 2 章で説明した環境を用いた追跡問題で行う。表 1 は藤田ら[6]の評価実験を参考に、獲物が行動を選択する確率を示している。実験時の学習パラメータは、追跡問題でよく用いられる学習率αとして 0.10、割引率γは 0.99、εは 0.85 にそれぞれ設定する。また、各階層の学習で用いる報酬の値を表 2 に示す。報酬の値は Q 学習でよく用いられる値としている。

エピソード開始時に獲物とハンターをランダムに配置し、ハンターが獲物と同じ位置に到着するか、300 サイクル経過するまで試行を続ける。1 エピソード 300 サイクルの学習を 3,000 エピソード繰り返す。また、学習した行動の価値はエピソード終了時にリセットされず、次のエピソードに持ち越される。

表 1：獲物の行動選択の確率

行動	確率(%)
上方向へ移動	40%
右方向へ移動	40%
左方向へ移動	0%
下方向へ移動	0%
停止	20%

表 2：実験時の各報酬値

報酬	正の報酬	負の報酬
r^1_{t+1}	$\lambda^l_1 = 10.0$	$\mu^l_1 = -10.0$
r^2_{t+1}	$\lambda^l_2 = 10.0$	$\mu^l_2 = -10.0$
R_{t+1}	$\lambda^u = 10.0$	$\mu^u = -10.0$

5.2. 実験結果

図 4 に、エピソードの経過に対する、ハンターが獲物を捕獲するまでのサイクル数の推移を示す。図 4 から、200 エピソードくらいまでは学習が急激に進み、500 エピソード付近でほぼ学習が完了していることが分かる。また、提案手法を用いた学習の方が、階層化を行わない従来の学習に比べ、捕獲に必要なサイクル数が約 4 サイクル分減少している。ランダムウォークは、学習を行わなかった場合での捕獲にかかるサイクル数を示している。以上から、提案手法を用いてハンターの行動戦略を増やした結果、従来の手法より早く獲物を捕らえることができ、学習効率を向上できたと言える。

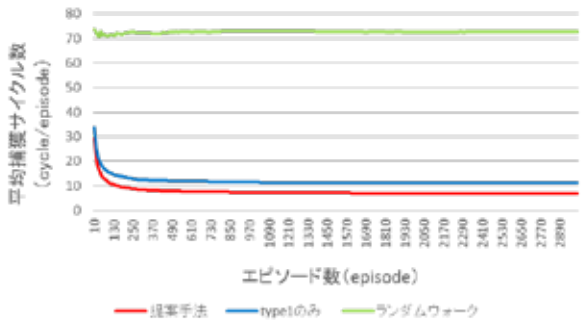

図 4：捕獲にかかるサイクル数の推移

図 5 の 100%積み上げグラフは、学習終了時の上位層の行動戦略の選択傾向を示している。横軸は獲物とハンターの距離を表し、縦軸は行動戦略が選択された割合を表す。図 5 から、獲物とハンターの距離が小さい場合、単純に追いかける戦略である type1 を優先的に選ぶようになることがわかる。一方、獲物とハンターの距離が大きくなると迎撃を優先する戦略である type2 も選択されるようになる。また、視野内である 2 マスの範囲に獲物がいる場合、80%以上の確率で type1 が選ばれている。

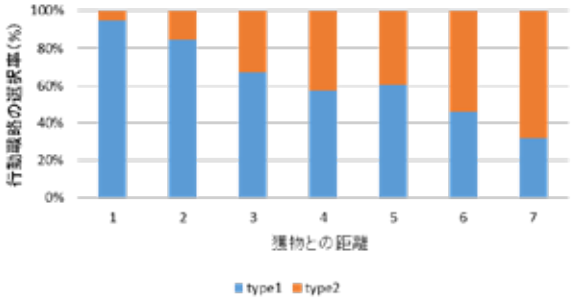

図 5：学習後の上位層の選択の傾向

6. おわりに

RL の研究の代表的な問題として追跡問題がある。渡邊らは、追跡問題における行動の選択方針を、行動戦略と行動戦術に分割する手法を提案した。しかし、行動戦略が 1 つしかないため、ハンター同士の協力を考慮した行動の選択が実現できていなかった。

本研究では、より効率的に獲物を捕獲するため行動戦略を複数用意し、ハンター同士の協調行動の視点から行動戦略を選択する階層型 RL を提案した。

実験の結果、従来の手法より獲物を約 4 サイクル早く捕まえられることを示した。ハンターと獲物の距離が離れているときに先回りを行う行動戦略が選択されていることから、ハンター同士の協力を考慮した行動の選択が行われていることを実験により示した。

獲物が視野外のときの探索手段がランダムウォークのみであるので、効率的に探索するアルゴリズムの導入が今後の課題である。

参考文献

1) V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. A. Riedmiller, “Playing atari with deep reinforcement learning”, Computer Research Repository (CoRR), Vol. abs/1312.5602, (2013)

2) R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction, MIT press (1998).

3) 渡邊俊彦, 和田竜也, “マルチエージェント追跡問題のための相対座標系に基づく階層型モジュ

ラー強化学習”, バイオメディカル・ファジィ・システム学会誌, Vol. 12, No. 2, pp. 65‐74 (2010).

4) 高玉圭樹, マルチエージェント学習: 相互作用の謎に迫る, コロナ社(2003).

5) 荒井幸代, 宮崎和光, 小林重信, “マルチエージェント強化学習の方法論: Q-learning とprofit sharing による接近”, 人工知能学会誌, Vol. 13, No. 4, pp. 609–618 (1998).

6) 藤田和幸, 松尾啓志, “状態空間の部分的高次元化法によるマルチエージェント強化学習(分散協調とエージェント)”, 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, Vol. 88, No. 4, pp. 864–872 (2005).