

強化学習を用いたマルチエージェントの 相互通信による協調行動獲得

函館高専 ○佐藤崇正 石若裕子 竹原直美

要 旨

共通のタスクを持つマルチエージェント系においては、エージェント間の協調が重要である。本研究では Q-learning を用い、マルチエージェントの相互通信による協調行動の獲得を試みる。タスクとして連続空間に於ける追跡問題を取り上げ、エージェント間の通信の有無によるタスク達成率の変化を評価する。シミュレーション結果により、相互通信を行う本手法の有効性が示された。

1. はじめに

共通のタスクを持つマルチエージェント系においては、個々のエージェントが単独で問題解決するよりも、複数のエージェントが協調して解決するほうがより効率的である。本研究の目的は、エージェント間の相互通信によって協調行動を獲得することである。強化学習を用いて、マルチエージェント系において典型的な問題である追跡問題を扱い、その結果について議論する。

2. 追跡問題

追跡問題²⁾とは、複数のハンターエージェント(ハンター)がターゲットエージェント(ターゲット)を囲い込み、捕獲することを目的としたものである。この問題は、マルチエージェントの標準問題として広く研究されている。ここでは、Benda らの提案した追跡問題を、二次元格子状の環境から連続空間へと拡張する。

連続的な二次元平面状の環境を設定し、ここに 2 体のハンターと 1 体のターゲットを配置する。各ハンターは、予め決められた視界 Φ 内にあるターゲットの方角と、もう一体のハンターの方角との間の角度 θ を観測可能であり (Fig.1(a)), 距離に関する情報は観測不可能とする。ハンターの選択可能な行動は Fig.1(c) に示されるように、ターゲットの方角を基準とした 8 方向 $\{a_0, a_1, \dots, a_7\}$ とし、選択した方向に移動量 Δ_h だけ移動する。

また、ハンター同士は観測した θ を既定の通信範囲 Ψ

内にいる他のハンターと相互通信する。通信によって得た θ を元に、各ハンターは、ターゲットが視界にない場合でもターゲットの方向を予測し行動する。Fig.1(b) に示されるように、相互通信するハンター間の距離 ϕ ($\Phi < \phi < \Psi$)、およびターゲットを捉えたハンターからターゲットまでの距離 ϕ ($0 < \phi < \Phi$) をランダムに決定し、ターゲットの方向を予測する。

ターゲットは、ハンターから逃避行動をとる。ターゲットは視界にハンターを観測すると (Fig.1(d)), 常にハンターから離れる方向 a_{t_0} を選択し、複数のハンターを観測した場合は、その合成した方向を選択する。視界にハンターがない場合は、ランダムに行動する。ターゲットの移動量は Δ_t とする。

各エージェントは、環境に存在する壁および他のエージェントに衝突する行動はとらないものとする。移動することが出来なくなった場合、その場に静止する。

3. 手法

ハンターの学習には、強化学習手法の一つである Q-Learning を用いる⁴⁾⁵⁾。

追跡問題の強化学習に於ける問題は、以下のものが挙げられる。

- (1) エージェントの観測する状態は、周囲のエージェントの行動により、動的に変化する(状態遷移の不確定性)¹⁾。
- (2) マルチエージェント系であること、環境が連続空間であることから、状態数の爆発的な増加が起きる(状態爆発)³⁾。

(1)の問題により、ハンターは次状態 s' を特定できない。ここでは、ターゲットの行動を推定する手法を用いた。ハンターは、ターゲットが Fig.1(d) に示される 8 方向 $\{a_{t_0}, a_{t_1}, \dots, a_{t_7}\}$ の行動を選択すると推定し、 s' を確率的に観測するものとする。推定し行動した結果得られた報酬を元に、その時推定したターゲットの行動選択確率を高くする。

(2)の問題に対しては、各ハンターの状態を上で述べた角度 θ のみを用いて表すことにより、状態数は問題空間の規模に依存しない。

ハンターの状態を $s = \{\theta_{own}, \theta_{partner} \mid \theta \in \Theta\}$ とする。ここで、 θ_{own} はハンターが観測した角度、 $\theta_{partner}$ はもう一方のハンターから通信によって得た角度であり、 Θ はハンターが観測可能な角度の集合である。ハンターが実行可能な行動集合 $A = \{a_0, a_1, \dots, a_7\}$ において、行動 $a \in A$ を

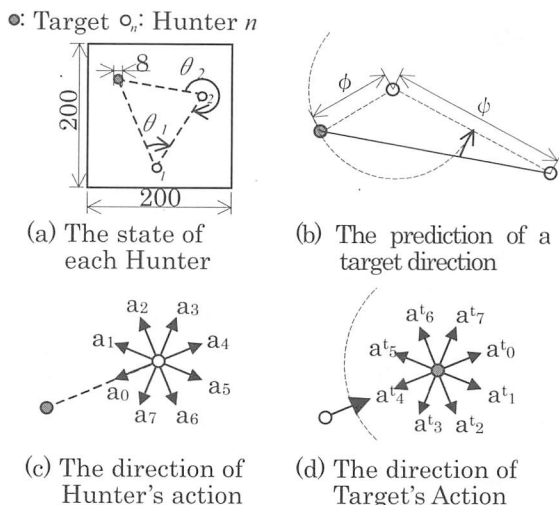


Fig.1 Hunting problem space

Table 1 The comparison of communication (Average and Hunting rates)

		Average	Hunting rate
Without Communication	$\Phi=60$	1140.5	35.6%
	$\Phi=\infty$	225.7	99.4%
With Communication	$\Phi=60$	161.6	98.2%
	$\Phi=\infty$	127.86	98.8%

選択したとき、状態は s' に遷移し、価値関数 $Q(s, a)$ は次式で更新される。

$$Q^{agent_i}(s, a) \leftarrow (1-\alpha)Q^{agent_i}(s, a) + \alpha [\gamma \max_a Q^{agent_i}(s', a) + r] \quad \dots(1)$$

ここで、 α は学習率、 γ は割引率、 r は報酬である。

Q 値の更新は、環境が連続空間であることを考慮し、各状態、行動に対しラジアル基底関数 (RBF) を用いる。RBF には次式に示すガウス分布を用いる。

$$Gaussian(\theta_m, \theta_n, a) = \frac{(\theta_m - \theta_m c)^2 + (\theta_n - \theta_n c)^2}{\sigma_\theta^2} + \frac{(a - ac)^2}{\sigma_a^2} \quad \dots(2)$$

ここで、 $m=own, n=partner$ 、 σ_θ^2 は角度 θ に対する分散、 σ_a^2 は行動 a に対する分散、 $\theta_m c, \theta_n c, ac$ はそれぞれ、現在のハンターの状態 $\theta_{own}, \theta_{partner}$ 、行動 a である。

次状態 s' で観測する $Q(s', a)$ は、 $\theta'_{own}, \theta'_{partner}$ における状態を $s'(\theta'_{own}, \theta'_{partner})$ としたとき、次式で求められる。

$$Q(s', a) = \sum_{\theta_{own} = -\rho - \theta'_{own}}^{\rho - \theta'_{own}} \sum_{\theta_{partner} = -\sqrt{\rho^2 - x^2} - \theta'_{partner}}^{\sqrt{\rho^2 - x^2} - \theta'_{partner}} Q(s, a) / \pi \rho^2 \quad \dots(3)$$

ここで、 $s = \{\theta_{own}, \theta_{partner}\}$ 、 ρ は Q 値の観測範囲である。

終端状態に達したとき、すべてのハンターに対して一定の報酬が与えられる。また、一方のハンターのみがターゲットに接触した場合、そのハンターに対し即時報酬を与える。

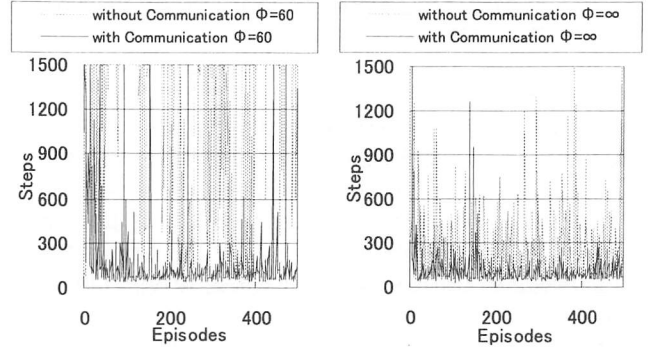
4. 実験

問題空間の大きさを 200×200 、各エージェントの形状を半径 4 の円形とし、ターゲットを空間の中心に、ハンターをエピソードごとにランダムに配置した。 Q 値の初期値は 0、学習率 $\alpha=0.1$ 、割引率 $\gamma=0.8$ 、報酬を 1、即時報酬を 0.0003、方策は ϵ -greedy を用い、 $\epsilon=0.1$ とした。各エージェントの移動量 $\Delta_h = \Delta_t = 4$ 、ガウス分布による更新に於ける $\sigma_\theta = 4$ 、 $\sigma_a = 16$ と設定した。 Q 値の観測範囲 $\rho = 30$ 、ハンターの通信範囲 $\Psi = 500$ とし、最大ステップ数は 1500 ステップ、エピソード数は 500 回とした。ハンターの視界が $\Phi = 60$ の場合と $\Phi = \infty$ の場合、及び通信を行わないハンター(状態として θ_{own} のみを持つ)について実験をそれぞれ行った。

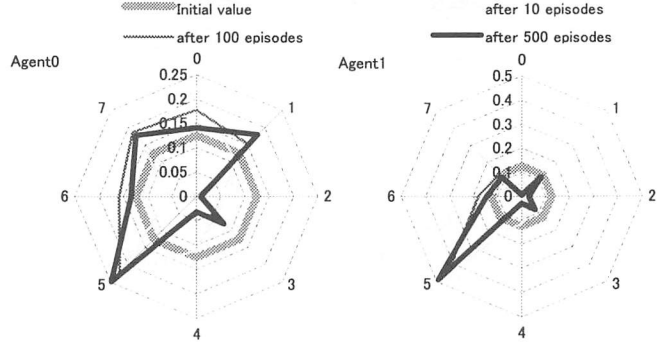
各実験におけるターゲット捕獲率、平均ステップ数を Table 1 に、各実験の学習曲線を Fig.2(a) に、Fig.2(b) には各ハンターが獲得したターゲットの行動推定確率を示す。 $\Phi = 60, \infty$ どちらの場合においても、通信するハンターのほうがよい性能を示している。 $\Phi = 60$ に於ける通信なしの場合は収束していない。また、環境全体を見渡すことが出来る通信なしのハンターよりも、視界の限られた通信するハンターの方がよい性能を示していることは、ハンター間の通信が追跡問題の解決により効果的であることを示している。

5. 終わりに

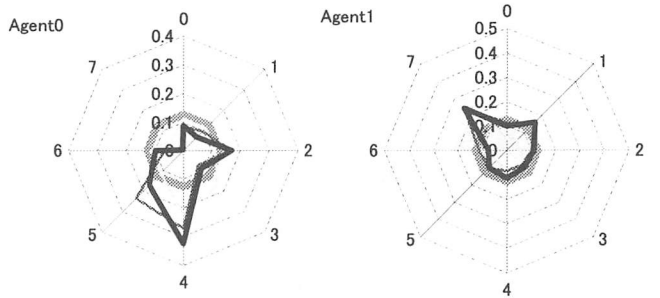
マルチエージェント系に於ける、エージェント間の通信の有無による学習の変化を示した。本研究では追跡問題を扱い、実験の結果、ハンター間の相互通信によってより効率的にターゲットを捕獲できることが示された。



(a) The learning curves (Left hand side: $\Phi=60$ Right hand side: $\Phi=\infty$)



(i) with communication ($\Phi=60$)



(ii) with communication ($\Phi=\infty$)

(b) Action estimation probabilities for a Target of each Hunter

Fig.2 Results of Experiments

参考文献

- 1) 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.5, pp.609-618, 1998
- 2) Benda, M., Jagannathan, V., and Dodhiawalla, R.: On Optimal Cooperation of Knowledge Sources, Technical Report, BCS-G2010-28, Boeing Advanced Technology Center, Boeing Computer Services, Seattle, Washington, 1986
- 3) 三上貞芳: 強化学習のマルチエージェント系への応用, 人工知能学会誌, Vol.12, No.6, pp.845-849, 1997
- 4) Sutton, R. S. and Barto, A. G.: Reinforcement Learning —An Introduction—, The MIT Press, 1998
- 5) Watkins, C.J.H., and Dayan, P.: Technical note: Q -learning, Machine Learning, Vol.8, pp.55-68, 1992