多目的タスクにおけるタスク空間分割を利用したマルチエージェント強化学習

橋本　祐馬†　　長行　康男‡

神戸電子専門学校†　神戸情報大学院大学‡

1．はじめに

マルチエージェントシステムへの強化学習[1]の適用は，工学及び認知科学の観点から興味深い研究課題であり，これまでに多くの研究成果[2]が報告されている．ところで，これらの研究で取り扱われているタスクは，エージェントの学習目的（ゴール）が単一のものがほとんどである．しかしながら，現実的な環境において，学習目的が一つだけである場合は特別で，環境内に学習目的が複数存在している場合の方が自然である．

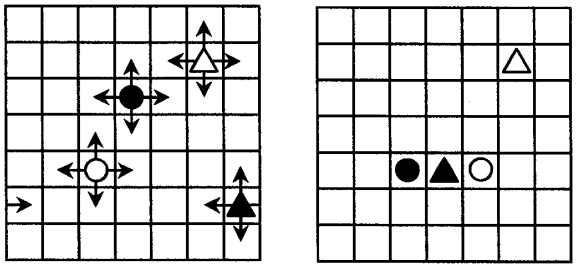
Whitehead ら[3]は，学習目的が複数存在するシングルエージェント環境において，学習目的ごとに状態空間（タスク空間）を分割して学習を行うQ学習[1]に基づいた強化学習法を提案し，その有効性を示した．

本研究では，シングルエージェント環境とは環境の性質が大きく異なる（複数の学習目的が存在する）マルチエージェント環境においても，学習目的ごとに状態空間を分割して学習を行う強化学習法が有効に働くかどうかを調査する．

2．追跡問題

本研究では，実験タスクとして追跡問題を採用する．追跡問題は，ハンターが獲物を追いかけて捕獲する課題である．以下に，本研究における追跡問題の問題設定を示す．

- ２次元（７×７）のトーラス状グリッド空間中に，２体のハンター（hunter_1，hunter_2）と２体の獲物（prey_1，prey_2）が存在する（図１）．
- 本研究では，ハンターを『エージェント』と定義する．
- 各時間ステップごとに，すべてのハンターと獲物は，それぞれ１つの行動を同期して実行する．ここで，ハンター，獲物が実行可能な行動は，隣接する上，下，左，右のグリッドへ移動する（図１（ａ）の矢印），現在位置に留まる，の５通りとする．
- ハンターの目標は，どちらか１体の獲物を捕獲することとする．ここで獲物捕獲の定義は，『２体のハンターが獲物を上下，あるいは左右から挟んだ状態』とする（図１（ｂ））．
- 獲物が各時間ステップで実行する行動は確率的で，prey_1 の行動確率は，現在位置に留まるを 0.25，右へ移動を 0.75，その他の行動を 0 としている．また，prey_2 の行動確率は，上，下，左，右への移動をそれぞれ 0.25，現在位置に留まるを 0 としている．これらの行動確率は時不変とする．ここで，prey_1 の行動の方が prey_2 の行動よりランダムさが小さく，prey_1 の方が捕獲しやすいことに注意されたい．
- 初期配置から，どちらかの獲物が捕獲されるまでを『１エピソード』とする．どちらか一方でも獲物が捕獲されると，すべてのハンター，獲物はグリッド空間中にランダムに初期配置され，新たにエピソードを開始する．



（ａ）グリッド空間　　（ｂ）捕獲状態の例

○：hunter_1，●：hunter_2，△：prey_1，▲：prey_2

図１　追跡問題のグリッド空間

本研究では，上記の追跡問題に対して，個々のエージェント（ハンター）がそれぞれ独立に強化学習を行うことにより，獲物捕獲行動（協調行動）を学習することを考える．

3．マルチエージェント強化学習

3．1　Q学習

強化学習法として多くの研究で広く用いられているQ学習[1]を上記の追跡問題にそのまま適用した場合の学習の流を次に示す．

① ある時間 t (= 0, 1, 2 …) において，hunter_i (i=1, 2) は，現在の環境状態 $s_t \in S$ (S は環境状態の集合) を観測する．ここで，環境状態 s は，自分 (hunter_i) から見た，hunter_j (j≠i)，prey_1，prey_2 のそれぞれの相対位置の組合せとする．例えば，図１（ａ）における hunter_1 から見た環境状態 s は（[1, 2]，[3, 3]，[-3, -1]）である．

② hunter_i は，環境状態 s_t におけるQ関数値 $Q(s_t, a)$ を基に，時間 t で実行すべき行動 $a_t \in A$ (A は行動の集合) を決定する．この行動決定（行動選択）は確率的で，確率 ε でQ関数値が最大となる行動を，残りの確率 (1-ε) で実行可能な全ての行動の中からランダムに行動を選択する．

③ hunter_i は，手続き②で選択した行動 a_t を実行する．このとき，hunter_j，prey_1，prey_2 も同期して行動を実行する．これらの行動により，環境状態は s_t から s_{t+1} へ遷移し，hunter_i は環境から報酬 r_{t+1} を受け取る．本研究では，この報酬を，獲物捕獲時に r=1（prey_1，prey_2 のどちらを捕獲しても同じ報酬 1 とする），それ以外のときに r=0 としている．

Multi-agent Reinforcement Learning Using the Decomposition of the Task Spaces for Multiple Tasks Environments

†Yuma HASHIMOTO: Kobe Institute of Computing - College of Computing

‡Yasuo NAGAYUKI: Kobe Institute of Computing - Graduate School of Information Technology

④ hunter_i は，環境状態 s_t，行動 a_t に対するQ関数値を式（1）に従って更新（学習）する．

$$Q(s_t, a_t) \leftarrow (1-\alpha)\, Q(s_t, a_t) + \alpha\,(r_{t+1} + \gamma \max_{a \in A} Q(s_{t+1}, a)) \quad (1)$$

ここで，$\alpha \in (0, 1]$，$\gamma \in [0, 1]$ はそれぞれ学習率，割引率と呼ばれるパラメータである．

⑤ 学習の終了条件を満たしていれば学習終了．そうでなければ，tに1を加えて，手続き①に戻る．

3．2　タスク空間分割を利用したQ学習

複数の学習目的が存在する環境において，前節のようにQ学習をそのまま適用した場合，手続き①のように，すべての学習目的（獲物）をまとめて一つの環境状態として扱うことになってしまう．これは，二兎を追うようなタスク設定になってしまうことを表す．しかしながら，我々人間が上記の追跡問題のハンターの立場に置かれた場合，ハンター，獲物の位置状況にもよるが，ほとんどの状況において二兎は追わず，報酬が得やすい（捕獲しやすい）獲物（prey_1）のみを追うはずである．なぜなら，その方が効率的だからである．

本研究では，そのような，我々人間が実際に行うであろう行動の学習が可能となるよう，複数の学習目的を一つの状態空間（タスク空間）としては扱わず，学習目的ごとに状態空間を分割することを考える．上記の追跡問題では，自分（hunter_i）から見た環境状態を，hunter_j，prey_1のそれぞれの相対位置の組合せ（s^1とする）と，hunter_j，prey_2 のそれぞれの相対位置の組合せ（s^2 とする）の二つに分割する．例えば，図1（a）における hunter_1 から見た環境状態s^1は([1, 2], [3, 3])，s^2は([1, 2], [-3, -1])となる．そして，それぞれの状態空間に対応したQ関数 $Q^1(s^1, a)$，$Q^2(s^2, a)$ を用意し，それぞれのQ関数で別々にQ学習を行う．学習の流れは，以下の3点を除いて，前節の①～⑤と同様である．

- ①で観測する環境状態を s から s^1，s^2 に変更する．
- ②の行動選択では，$Q^1(s^1, a)$，$Q^2(s^2, a)$ の両方のQ関数値（10通り）の中から最も大きいQ関数値である行動を確率 ε で選択し，残りの確率（$1-\varepsilon$）で，実行可能な全ての行動の中からランダムに選択する．
- ④の行動学習では，$Q^1(s^1, a)$，$Q^2(s^2, a)$の両方のQ関数を式（1）と同様の更新式で更新する．

以上により，一兎のみを追うような行動の学習が可能となる．

4．実験結果

前章で述べた2つの強化学習法を追跡問題に適用し，コンピュータシミュレーション実験を行った．ここで，行動選択時，行動学習時に使用したパラメータの値は，$\varepsilon = 0.2 \times 0.999931^{num\_ep}$（num_ep は学習エピソード数），$\alpha = 0.2$，$\gamma = 0.9$ である．また，Q関数の初期値は，すべての状態，行動において 0.0 とした．

実験結果を図2，図3に示す．図2の横軸は学習エピソード数，縦軸は1エピソード中で獲物捕獲までに費やした平均時間ステップ数を表す．図2の結果は，10 学習エピソード毎に，そのときまでの学習性能を評価するため，初期配置を変えた100評価エピソード（このエピソードでは学習を行わない）の実験を行い，その平均時間ステップ数を表示したものである．図2より，タスク空間を分割して学習を行った方が，タスク空間を分割していない通常のQ学習より学習が高速になるっていることがわかる．

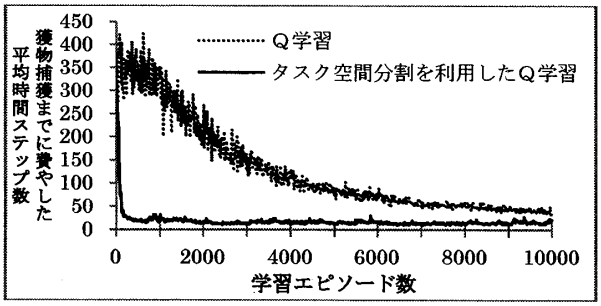


図2　獲物捕獲までに費やした平均時間ステップ数

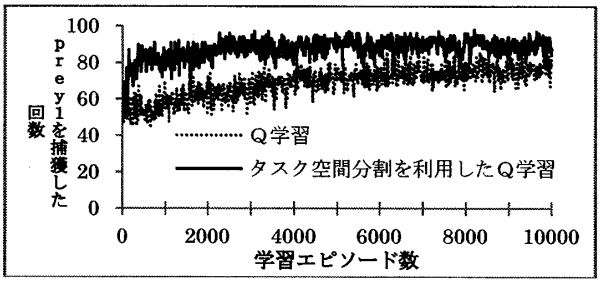


図3　100評価エピソード中で prey_1 を捕獲した回数

図3は，上述の100評価エピソード中で捕獲しやすい方の獲物（prey_1）を捕獲した回数を示したものである．図の横軸が学習エピソード数，縦軸が捕獲回数（100 回中）を表す．図3より，タスク空間を分割して学習を行った場合，8割以上の確率で捕獲しやすい獲物を捕獲するように学習していることがわかる．それに対して，タスク空間を分割していない通常のQ学習では，二兎を追ってしまっているためか，捕獲しやすい獲物を捕獲するという，この追跡問題における効果的な解を6，7割程度しか学習できていないことが見てとれる．

5．おわりに

本稿では，複数の学習目的が存在するマルチエージェント環境において，学習目的ごとに環境状態を分割して学習を行う強化学習法が有効かどうかを調査した．実験により，学習目的ごとに状態空間を分割すると，学習が高速になり，更に我々人間が行うであろう行動（達成しやすい目的に向かっていく行動）を獲得できるようになることがわかった．

参考文献

[1] R.S. Sutton and A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 1998.

[2] L. Busoniu, R. Babuska and B.D. Schutter, "A Comprehensive Survey of Multi-Agent Reinforcement Learning", IEEE Transactions on Systems, Man, and, Cybernetics, Part C, vol. 38, no. 2, pp. 156-172, 2008.

[3] S. Whitehead, J. Karlsson and J. Tenenberg, "Learning Multiple Goal Behavior via Task Decomposition and Dynamic Policy Merging", In Robot Learning, Kluwer Academic Publishers, pp. 45-78, 1993.