ISSN 2186-5647

P-25

Modular Q-Learningにおける多段高次元化法

日大生産工　○岩間　達希　日大生産工　山内 ゆかり

1 まえがき

強化学習とは、思考錯誤を通じて環境に適応する学習制御のことである。教師付き学習とは異なり、報酬という環境から与えられる情報を手掛かりに学習を行う。

強化学習を用いた研究として、マルチエージェントシステムがある。マルチエージェント強化学習には次元の呪いがある。

次元の呪いを防ぐ方法として Ono らによって提案された Modular Q-learning[1]がある。この手法では、自分と他のエージェントから構成される部分状態空間を用いるため、状態空間の大きさがエージェント数に影響されない。しかし、部分状態空間のみを観測することにより、不完全知覚による学習性能の低下が起きてしまう。

この問題に対して、藤田らは学習性能低下の要因となる状態を高次元化し全状態を見ることで不完全知覚を取り除き、Modular Q-learning の学習性能を改善する、HMQL (Hybrid Modular Q-Learning)[2]を提案している。その結果、Q-learning、Modular Q-learning と比較して、獲物捕獲に要したステップ数の性能を向上させることができた。しかし、この手法は高次元化を行う判断が難しく、学習初期では学習精度と効率が改善されない。

そこで本研究では、高次元化を行うかの判断をエージェントと獲物との距離で判断し、段階的に高次元化を行う多段高次元化法を提案する。

2 従来研究

2.1 追跡問題とは

追跡問題とは、マルチエージェント環境下での協調問題として扱われる問題である。従来研究では、3体のハンターが1体の獲物を追跡し捕獲したときに報酬が与えられる。この試行を繰り返し学習していく。図1に獲物の捕獲状態を示す。○はハンター、▲は獲物を表す。

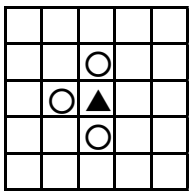

図 1. 捕獲状態

2.2 Modular Q-learning

エージェントが複数いる環境下では、エージェントの数分だけ状態数が指数的に増大してしまう。そのような次元の呪いの問題を改善するためにOnoらによって提案された手法がModular Q-learning[1]である。Modular Q-learningでは、自分と他の1体のエージェントから構成される部分状態空間を用いるため、状態数の指数的な増大を防ぐことができる。

2.3 状態空間の部分的高次元化法

藤田らは、ゴール状態に近く、Modular Q-learningでは不完全知覚が生じ学習性能の低下が予想される状態にある時に、部分状態$\{s_1, s_2\}$や$\{s_1, s_3\}$ではなく高次元化して全エージェントの状態$\{s_1, s_2, s_3\}$を見るということを行っている[2]。図2に不完全知覚が生じる例を示す。

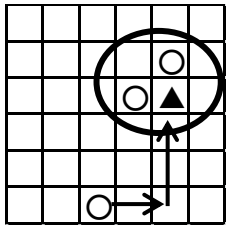

図2. 不完全知覚の例

ここで、獲物に隣接しているエージェントはゴール状態を満たしているが、もう1体のエージェントが獲物に到達するまで停滞していると判断し、状態価値が下がるという問題が起きる。具体的には、獲物付近の2体のエージェントの部分状態のQ値が減少してしまう。

そこで、部分状態の価値を表す状態価値関数V_iを用いて状態の識別を行う手法を提案している。自分の状態s_{self}と他エージェントiの状

Multistage High-Dimensional State Space Method for Modular Q-Learning

Tatsuki IWAMA and Yukari YAMAUCHI

態s_iから構成される部分状態の価値$V_i(s_{self}, s_i)$の学習式を式(1)で与える。

$$V_i\left(s_{self}^{t-1}, s_i^{t-1}\right) \leftarrow V_i\left(s_{self}^{t-1}, s_i^{t-1}\right) + \beta\left(max\left(r_t, \gamma V_i\left(s_{self}^{t}, s_i^{t}\right)\right) - V_i\left(s_{self}^{t-1}, s_i^{t-1}\right)\right) \quad (1)$$

ここで、βは学習率$(0 < \beta < 1)$、γは割引率$(0 < \gamma < 1)$である。状態価値関数V_iの値は、報酬が与えられるか、より価値の高い状態へと遷移した場合に増加する。それを行うために閾値を設け、状態を構成する部分状態の価値がともに閾値を超える状態の時に高次元化を行う。閾値ηは式(2)で与える。

$$\eta = \gamma^{\lambda} \cdot Reward \quad (2)$$

λはパラメータ、γは状態価値関数Vの学習式に用いた割引率、$Reward$は環境から与えられる報酬を表す。

しかし、図2のような状態に初めて遭遇した時に、状態空間の高次元化を行うことができない。状態価値の更新は複数の試行により経験的に学習されるため、初期の学習速度の低下や不完全知覚の問題を解決することはできない。

3　提案手法

本研究では、高次元化を行う状態をハンターと獲物との距離で判断し、段階的に高次元化を行う多段高次元化法を提案する。

具体的には、図3のようにハンターの上下左右斜め1マス以内に獲物が隣接している場合に、高次元化を行い、ゴール状態に近い場合の不完全知覚を回避する。

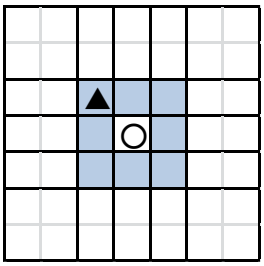

図3.高次元化を行う状態

4　実験環境

本研究の実験環境は、5×5、9×9 の二次元トーラス状グリッドに 3 体のハンターと 1 体の獲物を置いて実験を行う。ハンターと獲物の初期配置はランダムで決定し、各ハンターは長行らによって提案された政策推定を用いたQ-learning[3]を行う。ハンターの状態は、図 4 のように獲物との相対位置で表す。ハンターが選択可能な行動は上下左右とその場に停止の 5 通りで、獲物の選択可能な行動は上(40%)、右(40%)、その場に停止(20%)の 3 通りである。獲物の行動選択確率は一定であり、獲物とハンターが同じグリッドに移動してもよいこととする。ハンターが 3 方向から獲物に隣接したとき捕獲できたとし、初期状態から獲物を捕獲するまでを 1 エピソードとする。

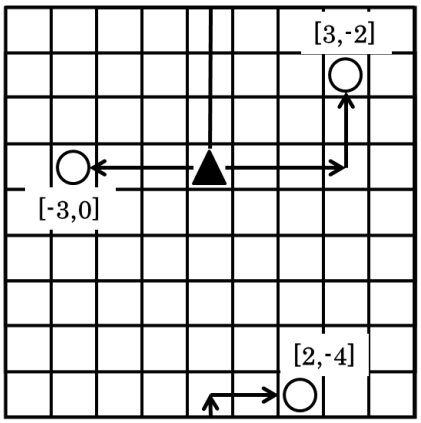

図 4. ハンターの位置情報

5　まとめ

本研究では、Modular Q-learningの高次元化を状態価値関数に依存せず、ハンターと獲物との距離で判断し、段階的に高次元化を行う多段高次元化法を提案した。これにより、従来研究の問題点である、学習初期での学習精度と効率が改善されると考える。

「参考文献」

[1]N.Ono and K.Fukumoto, “Multi-agent reinforcement learning: A modular approach,” Proc. 2nd International Conference on Multi-agent Systems (ICMAS-96), pp.252-258, AAAI Press, 1996.

[2]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, 電子情報通信学会論文誌, (D-1), vol.J88-D-I, no.4, pp.864-872, Nov. 2005

[3]長行康男, 伊藤　実, 2体エージェント確率ゲームにおける他エージェントの政策推定を利用した強化学習法, 電子情報通信学会論文誌, (D-1), vol.J86-D-I, no.11, pp.821-829, Nov. 2003