F_040

Modular Profit Sharingによるマルチエージェント強化学習

Modular Profit Sharing for Multi Agent Learning System

相良 幸範† 延澤 志保‡ 太原 育夫‡
Yukinori Sagara Shiho Nobesawa Ikuo Tahara

1. はじめに

マルチエージェント強化学習において，エージェントは自分以外のエージェントも環境の一部として観測する．そのためエージェント数の増加に伴い状態空間が指数的に増加し，状態空間の爆発により学習速度が著しく低下するという問題が生じる．この問題を解決する手法の1つとして，Q-learningを拡張したModular Q-learningという手法が提案されている．本研究では，同様の方法を経験強化型の強化学習アルゴリズムであるProfit Sharingに適用したModular Profit Sharingが，状態空間の爆発を低減し，マルチエージェント環境での学習を効果的に行なうことを実験により示す．

2. Modular Q-learning

Modular Q-learningは状態空間の爆発を防ぐ手法としてOnoらによって提案された[1]．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ので，状態空間の大きさは，エージェントの数に関わらず常に$|s|^2$となり，状態空間の爆発を防ぐことが可能となる．

エージェントがN体存在する場合，部分状態空間はN-1個存在し，それぞれに対して一つずつ学習器が割り当てられる．図1にエージェントが4体の場合のエージェント1の構成を示す．ここで，エージェント1は4体の内の1体であり，三つの部分状態空間$< s_1, s_2 >, < s_1, s_3 >, < s_1, s_4 >$を学習に用いるエージェントである．各学習器はQ-learningを行い，Q値を行動選択器へと渡す．Q-learningでは，状態s_tで行動a_tを選択して状態s_{t+1}へと遷移した場合，選択したルールのQ値$Q(s_t, a_t)$を次の式(1)によって更新する．

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_a Q(s_{t+1}, a)) \quad (1)$$

ここで，αは学習率$(0 < \alpha < 1)$，γは割引率$(0 < \gamma < 1)$，r_tは報酬を表している．行動選択器は，式(2)に示すように受け取ったQ値の合計値がより大きな値となる行動を優先して選択する．

$$\arg\max_{a \in A} \sum_{i=1}^{i=N-1} Q^{k,i}(S_{k,i}, a) \quad (2)$$

ここで，aは行動，Aは行動の集合，$\arg\max_{a \in A} Q$はQを最大とするような引数aを意味する．また，$S_{k,i}$はエージェントkとそれ以外のエージェントiから構成される部分状態を表している．

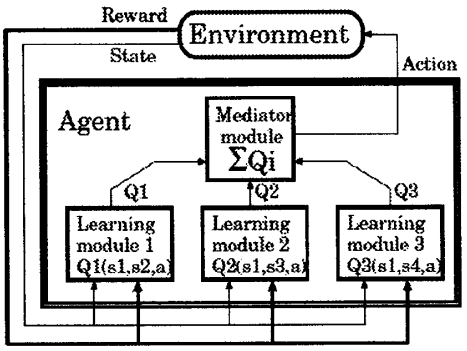

図1 Modular Q-learningの構成

Modular学習では部分状態のみを観測するため，不完全知覚が生じ学習性能が低下してしまうという欠点がある．藤田らはこの部分状態による学習性能の低下に対し，状態空間の部分的高次元化法を用いることによる解決法を提示している[2]．

3. Modular Profit Sharing

Modular Q-learningは各エージェントの学習器の学習にQ学習を用いている．しかし，マルチエージェントシステムでは，Q学習を用いるエージェント集団よりもProfit Sharingを用いるエージェント集団の方が不完全知覚問題を解消し，エージェント間の協調行動を獲得できることが知られている[2]．そこでModular Q-learningの学習器をProfit Sharingに換え，それに伴い行動選択手法もボルツマン選択からルーレット選択に換えた方法を検討する．

4. 実験

Modular Profit Sharingの有効性を検証するため，マルチエージェント強化学習における標準問題である追跡問題で評価する．エピソード数と目標達成までのステップ数の関係について，Modular Profit Sharing，Modular Q-learning，Profit Sharing，Q-learningの4つを比較する．

4.1. 実験環境

実験環境は，参考文献[2]と同様に，

- ハンターエージェントは3体，獲物エージェントは1体．
- 各エージェントは同一マスに進入できない．
- 獲物エージェントはその場に停止を40%，上方向に移動を20%，右方向に移動を20%の確率でランダムに行動する．

† 東京理科大学大学院 理工学研究科 情報科学専攻
‡ 東京理科大学 理工学部 情報科学科
Tokyo University of Science

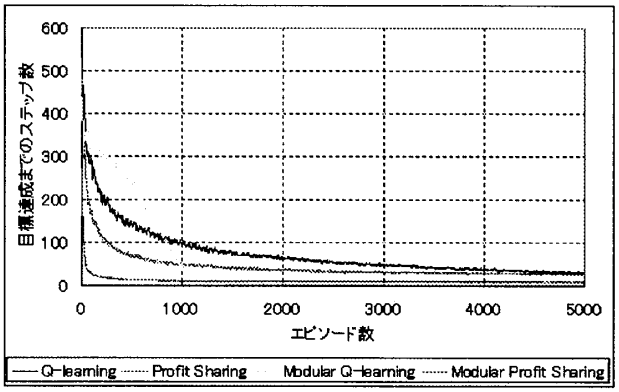

図2　各学習法のエピソード数とステップ数 5x5

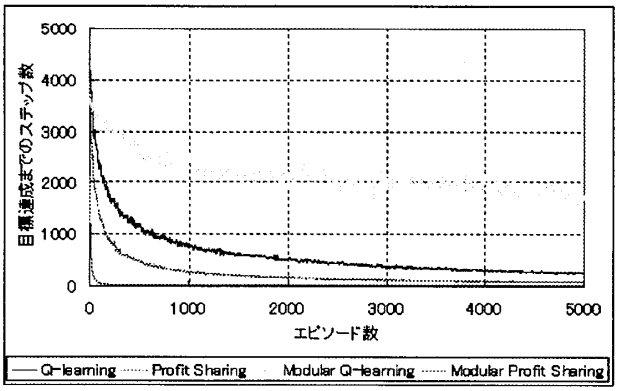

図3　各学習法のエピソード数とステップ数 7x7

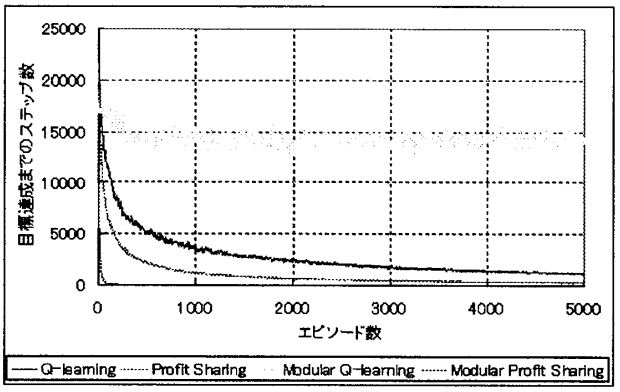

図4　各学習法のエピソード数とステップ数 9x9

- 目標状態はハンターが異なる 3 方向から獲物に隣接した状態.

とした．また，各学習法のパラメータは学習率 α は 0.3，割引率 γ は 0.8 とした．

4.2. 実験結果

図 2，3，4 に，Modular Profit Sharing，Modular Q-learning，Profit Sharing，Q-learning を用いた実験結果を示す．横軸はエピソード数，縦軸は獲物捕獲に要したステップ数を表している．各学習法の結果は共に 100 試行の平均値である．各学習法の 5000 エピソード後の目標達成までのステップ数は表 1 のようになった．

表1　5000 エピソード後の目標達成までのステップ数

	5x5	7x7	9x9
Q-learning	31	240	1160
Profit Sharing	28	78	269
Modular Q-learning	4	1669	13741
Modular Profit Sharing	8	11	14

また，図 2，3，4 及び表 1 の値は 10 エピソード毎の移動平均である．

5. 考察

図 2，3，4，および表 1 から，Modular Profit Sharing は状態空間の増加に関わらず学習が非常に高速に収束し，一方，Modular Q-learning は状態空間の増加に対して収束が著しく遅くなることがわかる．各状態空間での Q-learning，Profit Sharing は状態空間の増加によらず学習の収束は安定的であることから，Modular Q-learning は状態空間の増加に対して不安定であるといえる．更に，Q-learning は環境が MDPs であれば、学習結果の最適性が保証されるが，Modular 学習においては，学習器の構成上全エージェントの状態を同時に観測できないため，不完全知覚環境下での学習となり，そのため学習の収束が不安定になると考えられる．一方 Modular Profit Sharing では，Profit Sharing エージェント集団は Q-learning エージェント集団よりも不完全知覚問題を解消し，エージェント間の協調行動を獲得できるため[3]，不完全知覚環境においても適切に学習することができ，それに加え部分状態のみを観測することにより状態空間が減少したため，学習の収束が速くなったと考えられる．

6. まとめ

本研究では，Modular Q-learning の考え方を Profit Sharing に適用した Modular Profit Sharing が，状態空間の爆発を低減し，マルチエージェント環境での学習を効果的に行なうこと，また Modular Q-learning が，状態空間の増加に対して収束が著しく遅くなるのに対し，Modular Profit Sharing では状態空間の増加の影響を受けにくく，学習が非常に高速に収束することを実験により示した．

今後の課題として，Modular Q-learning，Modular Profit Sharing の状態空間の増加に対する学習性能の低下の差に関する理論的考察，また Modular Profit Sharing の学習器の更なる改良などが考えられる．

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, Vol. J88-D1, No.4, pp. 864-872 , 2005.

[3] 荒井幸代, "マルチエージェント強化学習－実用化に向けての課題・理論・書技術との融合－," 人工知能学会誌, Vol.16, No.4, pp.476-481, 2001.