3T－5

# 通信手段を持つマルチエージェント系における強化学習

山本 真也　山口 文彦　中西 正和
慶應義塾大学大学院 理工学研究科 計算機科学専攻

## 1. はじめに

マルチエージェント系強化学習において、協調行動の創発の原理に関して、各エージェントが他とは独立に自己の目的を追求して、利己的に振舞うにもかかわらず、集団内に協調的行動が創発される場合がありその原動は環境からの報酬に求められるという仮説があり、いくつかの研究がなされている [1] [5]。

また、エージェント間で情報を共有することに関する研究は多々あるが、いずれも何らかの形でヒューリスティックを与えている。

そこで、本研究では複数のハンターが獲物を捕える問題について、各エージェントに通信手段を持たせることにより強化学習法に基づくマルチエージェント系において通信手段をどのような状態で使用するようになるのかを調べる。また、それにより、目的達成までのステップ数を減少させることができるかどうかを考察する。

## 2. 強化学習

強化学習法 [2] は2つの側面から分類することができ、1つめは環境のクラスがマルコフ的か非マルコフ的かということによる分類である。

2つめは接近の指向性が環境同定型か経験強化型かということによる分類である。マルコフ的な環境における代表的な環境同定型学習として **Q-学習** [3]、経験強化型学習として **Profit Sharing** [4] などがある。

### 2.1 Q学習

状態と行動の組に対する評価を見積もった値を **Q値** と呼ぶ。Q値は、以下の式で更新され、特定の条件のもとで収束する。しかし、環境の同定に正確を期するためには、非常に多くの試行を要する。また、一般に環境が非マルコフ的のときQ値は振動し収束しないことがある。

$$\Delta Q(x_t, a_t) = \alpha(r_t + \gamma \max_b Q(x_{t+1}, b) - Q(x_t, a_t))$$

### 2.2 Profit Sharing

Profit Sharing では報酬を獲得したときにそれに貢献したルールの強度を強化する。強化値の分配を決定する関数を**強化関数**と呼ぶ。Q学習に比べて学習に要する試行の回数は非常に少ないが、一般に最適性の保証はない。強化関数としては等比減少数列が最適とされている [4]。

### Tan の研究

Tan は、Q学習を用いて以下のような実験を行った [5]。タスクは $10 \times 10$ のグリッドワールドにおいてランダムに動く獲物とハンターからなる。ハンターは視覚フィールドを持ち、上下左右の4方向に動く。この実験において次の3つの場合において目的の達成までのステップ数が減少することが確認されている。

1. 入力情報を共有する
2. 経験 (エピソード) を共有する
3. 行動戦略を共有する

この実験では、上記のように情報を共有することで、環境同定型であるQ学習においても動的な環境に対応できるということが確認された。

### 岩下の研究

岩下ら [1] は、Tan の研究と同様の環境において各エージェントが独立に自己の目的のために行動するにもかかわらず、協調行動が創発されるという仮定のもとで Profit Sharing を用いて実験を行った。この実験では、被食者の行動パターンを次の3つの場合に分けている。

---

Reinforcement Learning on Multi Agents System with Signals
Shinya YAMAMOTO Fumihiko YAMAGUCHI Masakazu NAKANISHI
Department of Computer Science, Faculty of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

1. 被食者が動かない
2. 被食者が軌道上を動く
3. 捕食者が近づくと被食者が逃げる

いずれの場合も、エージェント間で情報の交換がなくても協調行動が創発されることが確認された。

## 3. 実験環境

本研究では、トーラス上のグリッドワールドを用い、グリッドワールド上には、二人のハンターと一匹の獲物が存在し、グリッドワールドの大きさは実験ごとに変化させる。ハンターは一回の行動で上下左右のいずれか一方向に移動可能で、二人のハンターが同時に獲物に隣接するか、または、セルを共有した場合に獲物を捕まえることができ、報酬を得られる。獲物の動きは常にランダムである。ハンターは視覚範囲内にいる他のハンターと獲物を知覚できるものとする。

## 4. エージェント間通信

本研究では、ハンターの行動として上下左右の一方向に移動するという四つの行動のほかに、シグナルを発信してから上下左右の一方向に移動するという四つの計八行動選択可能とする。ハンターは自分の視覚範囲内にいる仲間のハンターがシグナルを発信している状態と発信していない状態を異なる状態として認識できる。

学習前には、シグナルをいつ、どのように発信するかということに関する一切の制約やヒューリスティックを与えない。そのため、学習の初期においてはシグナルの発信は完全にランダムとなるが、シグナルを発信するという行動を学習し、エージェント間でシグナルの発信に関する何らかのルールが生まれる可能性も考えられる。

### 4.1 シグナルのメモリの導入

ハンターに一回前の行動のときにシグナルを発信したかどうかのメモリを持たせる。これにより、エージェント間でシグナルの連携による意味付けのようなものが見られる可能性があると考えられる。

## 5. 実験結果及び考察

ハンターの学習法としてQ学習を用いた場合には、シグナルを用いない方が目的達成までのステップ数が少なく、また、Q値を解析した結果からも特殊な状態におけるシグナルの発信を学習しているように見えなかった。Q学習では、動的な環境と状態数の爆発に対応するためには何らかのヒューリスティックを与える必要があると考えられる。

Profit Sharingを用いた場合には、獲物を捕まえるまでのステップ数は表1のようになった。

表 1: Profit Sharingの実験結果

| グリッドワールド | 視界 | Profit Sharing | シグナル | メモリ |
|---|---|---|---|---|
| 7 × 7 | 3 × 3 | 21.10 | 20.45 | 19.12 |
| | 5 × 5 | 12.10 | 10.27 | 10.11 |
| 9 × 9 | 3 × 3 | 45.03 | 41.05 | 34.58 |
| | 5 × 5 | 19.67 | 18.92 | 19.66 |

表1より、シグナルを発信する能力をハンターに持たせた場合に獲物を捕まえるまでのステップ数が減少していることが分かる。このことより、ハンターがシグナルを有効に使っていることが推測できる。

## 6. 今後の課題

Q学習に関しては、既にその効果の確認されている状態数削減手法などに対して適用することによりシグナルの効果を確認することが考えられる。

Profit Sharingに関しても既存の手法との組合せにより更なる効果が現れるかどうかを調べることなどが挙げられる。

## 参考文献


[1] 岩下健久, 山村雅幸, 小林重信 : 強化学習に基づくマルチエージェント系の協調の創発, 第 *21* 回 知能システムシンポジウム予稿集, pp. 37-42, 1995.

[2] 山村雅幸, 宮崎和光, 小林重信 : エージェントの学習, 人工知能学会誌, *Vol. 10, No. 5,* pp. 683-689, 1995.

[3] Watkins, C. J. H and Dayan, P. : Technical Note:Q-Learning,*Machine Learning, Vol.8,* pp. 55-68, 1992.

[4] 宮崎和光, 山村雅幸, 小林重信 : 強化学習における報酬割当ての理論的考察, 人工知能学会誌, *Vol. 9, No. 4,* pp. 580-587, 1994.

[5] Tan, M.: Multi-Agent Reinforcement Learning : independent vs. Cooperative Agents, *Proceedings of the 10th International Conference on Machine Learning,* pp. 330-337, 1993.