5J－4

強化学習型マルチエージェント系における職能の分担の学習

吉田 将志　中西 正和　斎藤 博昭
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

現在, 強化学習型マルチエージェント系について数多くの研究がなされている. その中で, Tan[1], 岩下ら [2], 荒井ら [3] は, 追跡問題における協調に関する研究を行い, ハンターエージェントに協調行動が創発されることを確認している. そこでは, ある特定の場合に複数のエージェントが個々に異なる役割を演じるような特別な協調行動が見られることを報告している.

本稿では, この役割分担的な協調行動を職能の分担と呼び, これに焦点を当て, 強化学習型マルチエージェント系において, 職能の分担が見られるかどうかを観察し, 分担に関する評価方法を定めた上で, それを用いて職能の分担が学習されているかどうかを実験により確認する.

変化していく環境において自律的に振舞うエージェントの集団が職能の分担を獲得すれば, 効率的に知的作業を行うロボット知能の実現も期待され, 工学的にも有意義であると考えられる.

2. 先行研究

岩下ら [2] や荒井ら [3] は, 追跡問題を用いて, 複数のハンターが獲物を捕獲する過程において協調行動が創発されていることを実験により確認している. ここで, 特に獲物が逃避行動をとる場合において, 追い詰め役と待ち伏せ役を演じるという職能の分担が観察されている.

2.1 岩下らの研究

問題設定は, 10 × 10 のグリッドワールドにおいて, 制限された視覚領域を持つ 2 人のハンターと 1 匹の獲物が存在する. 捕獲は 2 人のハンターが獲物に同時に隣接したときとする. ハンターの学習モデルとして Profit Sharing[4][5] を用いている.

以上のような問題設定において, 獲物の動き方や世界を変えたり, 学習途中で獲物の初期位置や動き方を変化させたりしてシミュレーションを行い, ハンターの協調行動が得られたかどうかを観察している. この研究において, 特に獲物が回避行動をとるときに, 追い詰め役と待ち伏せ役を演じるという職能の分担が見られた.

2.2 荒井らの研究

この研究も同様な追跡問題を用いており, グリッドワールドは 7 × 7 格子型とし, 四方が壁に囲まれている場合とトーラス状の場合を対象としている. ハンターの数を 2 人または 3 人, 獲物の数を 1 匹または 2 匹とし, それらの各組合せを考えている. この研究では, ある環境の下で学習させた後, 環境を変化させると以前に獲得された行動ルールがどのような影響を及ぼすかについても調べている. ここでもハンターの学習モデルとして Profit Sharing [4][5] を用いている. この研究においても, 追い詰め役と待ち伏せ役を演じるという職能の分担が見られた.

3. 本研究の手法

本稿における問題設定, 学習のアルゴリズム, および職能の分担に対する評価方法について以下に述べる.

3.1 問題設定

本稿では, 上の先行研究 [2][3] と同様に追跡問題を用いる. 世界を四方が壁に囲まれた 9 × 9 のグリッドワールドとし, ハンター 2 人, 獲物 1 匹の場合を考える. 各エージェントは, 各タイムステップにおいて上下左右いずれかへ 1 マスだけ移動あるいはその場に停止することができる. ここで, 獲物は, いずれかのハンターが自分の視覚領域に入った場合にのみ逃避行動を取るものとする. 逃避行動選択アルゴリズムについては, 視覚領域内にいるすべてのハンターとの距離 (マンハッタン距離) の和が最大となるような行動を選択することにする. 捕獲条件は獲物の周囲 2 箇所以上をハンターのみあるいはハンターと壁によって占有することとする. 各エージェントはそれぞれ視覚

Learning of Performing part of all functions each agent has among Reinforcement-Learning Agents
Masashi YOSHIDA, Masakazu NAKANISHI, Hiroaki SAITO
Department of Computer Science, Graduate School of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223, Japan

領域を持ち, 自分を中心に 5×5 マスとする. また, 学習途中において各試行におけるそれぞれのエージェントの初期位置はランダムとする.

3.2 強化学習アルゴリズム

本稿では強化学習アルゴリズムとして, 学習途中でも強化値が十分な意味をもち, 環境変化に対しても速やかに追従できる柔軟性をも合わせもつ Profit Sharing[4] [5] を採用する.

● Profit Sharing

報酬が与えられたときに, それまでの行動系列を一括に強化する方法である. 報酬を行動にどのように分配するかが重要な問題であり, この分配方法を強化関数という. 強化関数は等比減少関数がよいとされている [4].

3.3 評価方法

職能の分担に対する評価方法として, 各ハンターエージェントの獲得した行動パターンに着目し, 分担をしているならば両者の行動パターンに差異が生じると考え, その考えに基づいて以下の手法を提案する.

ハンターエージェントが遭遇し得る場面は, $10 \times 10 \times (10 \times 10 - 1) \times (10 \times 10 - 2) = 970200$ 通りあり, それぞれの場面を $S_i(i = 0, 1, 2, \cdots, 970199)$ とする. ハンター $j(j = 0, 1)$ が場面 x に遭遇したときに取る (強化値最大の) 行動の番号を, $a_j(x)$ で表すことにする. すなわち, $a_j(x)$ は 0 (上に移動), 1 (右に移動), 2 (下に移動), 3 (左に移動), 4 (その場に停止) のうちいずれかの値をとる.

そこで, $a_j(S_i)$ を各成分とする 970200 次元ベクトル

$$\boldsymbol{a}_j = \left(a_j(S_i)\right)$$

を考え, $\boldsymbol{a}_0$ と $\boldsymbol{a}_1$ のなす角度を計算する. これによって両ハンターの行動パターンの違いがわかり, これを職能の分担に対する評価関数とする.

学習過程において, この評価関数が増加していれば, 各ハンターエージェント間で職能の分担が形成されていると考えることができる.

4. 今後の課題

現在, 獲物が逃避行動を取る場合について, 各エージェントの行動をシミュレーションにより観察してあり, その実験結果 (ハンター 2 人, 獲物 1 匹の場合) を図 1 に示す. ここで, 図中の各エージェントの移動経路に添付した数字は, その位置における初期状態 (タイムステップ 0) から経過したタイムステップ数を表す. この図から, 各ハンターエージェントは職能の分担を学習したと考えられる.

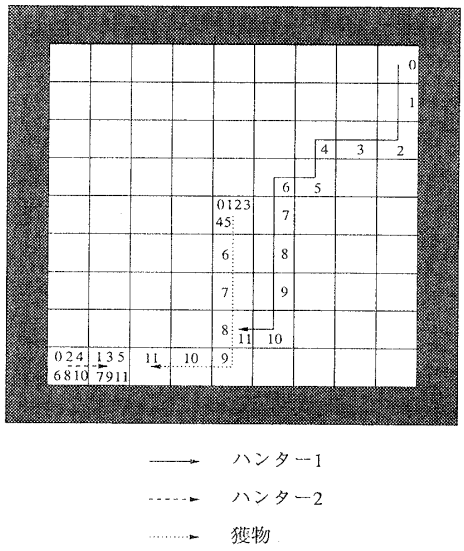

図 1: ハンター 2 人, 獲物 1 匹の場合において強化された行動

今後は, 以上に述べた評価方法を用いて実験を行い, 学習過程において職能の分担が学習されているかどうかを確認する予定である.

なお, 実験結果については発表時に併せて報告する.

参考文献

[1] Tan. M.: Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents, Proc, Machine Learning 10, pp. 330-337, 1993.

[2] 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資料, pp. 37-42, 1995.

[3] 荒井 幸代, 山村 雅幸, 小林 重信: 動的環境における強化学習型マルチエージェント系の協調, 第 9 回 人工知能学会全国大会論文集, pp. 139-142, 1995.

[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, 1994.

[5] J. J. Grefenstette: Credit Assignment in Rule Discovery Systems Based on Generic Algorithms, *Machine Learning*, Vol. 3, pp. 225-245, 1988.