マルチエージェントシステムのための信頼度を用いた強化学習 —相補ゲームにおける競合回避行動の獲得—

山田 和明*・高野 慧**

A Reinforcement Learning Approach Using Reliability for Multi-agent Systems —Acquiring Conflict Avoidance Behaviors in Dispersion Games—

Kazuaki YAMADA* and Satoru TAKANO**

This paper proposes a new reinforcement learning approach for acquiring conflict avoidance behaviors in multi-agent systems. Multi-agent systems are able to establish orderly systems autonomously through interaction with autonomous agents. We expect to be able to construct flexible and robust systems for the environmental changes by using multi-agent system approaches. However, it is difficult for designers to preliminarily embed appropriate behaviors to avoid conflict because complex dynamics emerges by interaction between many agents. We apply the proposed method to the narrow road problem that many agents go by each other in a narrow road, and verify the effectivity of the proposed method. In the narrow road problem, it is the optimal strategy that an agent selects going forward and another agent selects giving way. However, it is difficult for agents to decide which strategy to select because they cannot predict other agents' behaviors beforehand. The proposed method can differentiate into agents preferring to go forward and agents preferring to give way, by using Q-learning that can adjust discount rates. We solve conflict problems in multi-agent systems through autonomous functional differentiation of many learning agents. Through experimental results, we showed that agents differentiated into two type of agents, and acquired stable conflict avoidance behaviors with high probability than a conventional Q-learning.

Key Words: reinforcement learning, reliability, multi-agent system, dispersion game

1. 緒 言

マルチエージェントシステム¹⁾は多数の自律エージェントによって構成されており、各エージェントが近傍のエージェントとの相互作用を通して自律的に秩序を形成することができる。そのため、マルチエージェントシステムはシステム内外の環境変化に対して柔軟で頑健なシステムを構築することができる²⁾。しかし、多数のエージェントに限られた空間で相互作用することでエージェント間に競合状態が発生するため、設計者があらかじめ個々のエージェントに適切な行動を組込んでおくことは困難である。

上記の問題を解決する方法として強化学習が注目されている^{3),4)}。強化学習はエージェントが観測した状態において行動を実行したとき、環境から与えられる報酬を頼りに、エー

ジェントの状態行動間の適切な写像関係を学習することができる⁵⁾。われわれは強化学習をマルチエージェントシステムに導入することで、個々のエージェントが状況に適した行動を学習し、エージェント間に発生する競合状態を解消できると考えている。

強化学習をマルチエージェントシステムに適用する場合、少なくともつぎの二点がエージェント間の秩序形成にとって重要な鍵となる。まず、1つは個々のエージェントに与える報酬である。生天目らはエージェントに与えられる報酬を基に、エージェント間の関係を協調ゲーム、相補ゲーム、支配戦略均衡ゲーム、堂々巡りゲームに分類している^{6),7)}。たとえば、狭路において2台のエージェントがすれ違う状況は相補ゲームにあたる。この場合、一方が道を譲り、一方が通過することが必要である。すなわち、相補ゲームでは、エージェントが互いに異なる行動を選択することが最適戦略となる。しかし、一般に通過したエージェントの報酬は、道を譲ったエージェントの報酬より高いと考えられるため、エージェント間の報酬は非対称な関係となる。また、各エージェントは事前に相手の行動を知ることができないため、どちらの行動を選択すべきか判断できない状況に陥る可能性が高い。

もう1つは、エージェントの知覚能力である。マルチエー

* 東洋大学理工学部 川越市鯨井 2100

** 東洋大学大学院工学研究科 川越市鯨井 2100

* Undergraduate School of Sciences and Engineering, Toyo University, 2100 Kujirai, Kawagoe

** Graduate School of Engineering, Toyo University, 2100 Kujirai, Kawagoe

(Received April 11, 2012)

(Revised August 9, 2012)

ジェント環境では、エージェントが他のエージェントの識別や進行方向などを正しく知覚できない場合、異なる状態を同一の状態として誤認する不完全知覚問題^{8),9)}が発生し、適切な入出力関係を学習するのが困難になる。したがって、強化学習をマルチエージェントシステムに適用する場合、上記の課題を解決する必要がある。

相補ゲームにおいてエージェント群に秩序を形成させる研究として、柴田ら¹⁰⁾は、強化学習を電車乗降問題に適用し、乗車グループと降車グループを識別できるエージェント群が学習により整列乗車ルールを獲得できることを示している。また、環境の非対称性や身体性の差により獲得されるルールに違いが生じることを示している。しかし、この研究では、乗降の順番を譲ったエージェントが不利益を被るといった報酬の非対称性は扱っていない。一方、Moriyama ら¹¹⁾は、狭い通路を多数のエージェントがすれ違う狭路問題において、道を譲ったエージェントが不利益を被る報酬の非対称性を扱っている。そして、道を譲ったエージェントが周囲のエージェントから報酬を得ることで、道を譲ったエージェントが著しく不利益になることを軽減する近隣報酬を導入している。しかし、各エージェントは、事前に相手の行動を知ることができないため、どちらの行動を選択するべきか判断できない状況に陥る可能性がある。上田ら¹²⁾は、各エージェントが難易度の異なるタスクを選択・処理するタスク割当問題において、人の同情や嫉妬といった感情を模擬した憐憫悲パラメータを導入し、他のエージェントが負の報酬を受けたとき自身の報酬を下げる操作を行なうことで、難度の高いタスクを選択する利他的なエージェントが発現することを示している。しかし、設計者が問題ごとに憐憫悲パラメータを適切に設定する必要がある。

本研究では、マルチエージェントシステムにおける競合問題の1つとして狭路問題を取り上げ、道を譲ったエージェントが先に通過したエージェントより不利益を被る報酬の非対称性と、エージェントの知覚能力が学習性能に与える影響について取り扱う。そして、均質なエージェント群が、強化学習により道を譲るエージェントと先に通過するエージェントに分化することで、エージェント間の競合を回避できることを示す。また、信頼度を用いて強化学習の割引率を自律的に調整することで、先行研究のようにエージェント間で報酬を分配・調整することなく、従来の強化学習より高い確率で競合回避行動を獲得できることを示す。そして、競合回避行動の獲得プロセスを分析し、信頼度が競合回避行動の獲得に対し有効に作用するメカニズムを解明する。次章では本研究で取り扱う狭路問題について説明し、3章では提案手法について詳述する。4章ではシミュレーション設定と実験結果について説明し、最後にまとめと今後の課題について述べる。

2. 狭路問題

本研究では、提案手法の有効性を検証するためのテスト問題として、狭い通路において多数エージェントがすれ違う狭

Fig. 1 A narrow road problem

Table 1 The reward matrix of a narrow road problem

		B	
		passing way	giving way
A	passing way	-2	-1
	giving way	1	-1

Table 2 The reward matrix that translated the diagonal element into zero by local shift

		B	
		passing way	giving way
A	passing way	0	1
	giving way	2	0

路問題を取り扱う。狭路問題では、Fig. 1に示すように、狭い通路を多数のエージェントがすれ違うときエージェント間に競合が発生する。エージェントが互いに通過を選択したとき、エージェントが互いに通路を塞ぐため通過できない。一方、エージェントが互いに道を譲ったとき、エージェントは互いに通路を通過できない。狭路問題では、エージェントが互いに異なる行動を選択するように分化する必要がある。

本研究では、狭路問題におけるエージェントの報酬をTable 1に示すように与える。エージェントが互いに通過を選択したとき、エージェントは互いに衝突するため報酬-2が与えられ、エージェントが互いに道を譲ったとき、エージェントは互いに通過できないため報酬-1が与えられる。エージェントが互いに異なる行動を選択したとき、通過を選択したエージェントは報酬1が与えられ、道を譲ったエージェントは報酬-1が与えられる。

行動と報酬の関係をわかりやすくするために、Table 1に局所シフト⁷⁾を施し、Table 2のように変換する^(注1)。Table 2を見ると、エージェントが互いに異なる行動を選択したときの報酬が同じ行動を選択したときの報酬より高いことがわかる。すなわち、狭路問題では、エージェントは互いに異なる行動を選択することが最適戦略となる。しかし、エージェントが互いに異なる行動を選択したときの報酬が非対称であるため、どちらか一方のエージェントが報酬の低い道を譲る行動を選択する必要がある^(注2)。強化学習は、得られる報酬の期

(注1) 局所シフトを施した利得行列は、元の利得行列と合理的な行動の分岐点が同じとなる⁷⁾。

(注2) 柴田ら¹⁰⁾は、報酬を基に競合状態を、弱い衝突、中くらいの衝突、強い衝突の3つに分類している。本問題設定は強い衝突に分類され、従来の強化学習では学習困難な問題にあたる。

期待値を最大化する行動を学習するため、従来の強化学習をそのまま狭路問題に適用しただけでは、エージェント間に秩序を形成することは困難である。また、各エージェントは、事前に相手の行動を知ることができないため、どちらの行動を選択すべきか判断できない状況に陥る。さらに、エージェントは、周回方向の異なるグループの違いやエージェントの進行方向を識別できない場合、不完全知覚状態が発生する。

本研究では、従来の強化学習を適用しただけでは学習困難な上記の狭路問題に対し、信頼度を用いた強化学習により個々のエージェントの行動を分化させることで競合状態を回避するアプローチを採用する。

3. 提案手法

本研究では、狭路問題におけるジレンマを解消するために強化学習の一手法である Q 学習⁵⁾を用いる。しかし、マルチエージェント環境では、複数のエージェントが相互作用するため複雑なダイナミクスが発生する。そのため提案手法では、不確実性の高い環境において Q 学習の性能を高めるために、尾川ら¹³⁾が提案している信頼度による割引率の自律的な調整方法を導入する。また、Q 学習では、割引率の値によりエージェントの意思決定に違いが生じることが知られている^(注 3)。そのため、割引率を自律的に調整することで、先に通過するエージェントと相手に道を譲るエージェントへの分化を促進させることが期待できる。

3.1 提案手法の学習手順

提案手法の学習手順は以下のとおりである。

- ① エージェントは、観測した状態 s における行動 a の行動価値の期待値 $Q(s, a)$ を基に、(1) 式によりボルツマン選択により行動を決定する。
- ② 報酬 r は、実行した行動に対し環境から与えられる。
- ③ 割引率 γ は、信頼度 $R(s, a)$ に基づいて (3) 式により計算される。
- ④ 実行した行動 a の Q 値は、新しい割引率 γ に基づいて (2) 式により更新される。
- ⑤ 信頼度 $R(s, a)$ は、誤差 δ に基づいて (4) 式により更新される。
- ⑥ 終了条件を満たさない場合、①に戻る。

以下に、行動選択および Q 値の更新、信頼度を用いた割引率の自律的な調整について説明する。

3.2 行動選択および Q 値の更新

状態 s における行動 a の選択確率 $p(a|s)$ を次式により計算する。

$$p(a|s) = \frac{\exp(Q(s, a)/T)}{\sum_{b \in A} \exp(Q(s, b)/T)} \quad (1)$$

ただし、 T は温度パラメータ、 b は行動集合 A の要素である。

(注 3) Q 学習では、割引率の値が低い場合、エージェントは即時的報酬を重視し、割引率の値が高い場合、エージェントは長期的報酬を重視して意思決定を行なうことが知られている⁵⁾。

期待値の更新を次式により計算する。

$$\delta = r + \gamma \max_b Q(s', b) - Q(s, a) \quad (2-a)$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha \delta \quad (2-b)$$

ただし、 $\max_b Q(s', b)$ は次状態 s' における Q 値の最大値である。また、 α は学習率であり、 γ は割引率である。

3.3 信頼度を用いた割引率の自律的な調整

提案手法は、各エージェントの行動を分化させるために、尾川らが提案している信頼度を用いた割引率の自律的な調整方法を用いる¹³⁾。割引率 γ は、信頼度 $R(s, a)$ に基づいて次式により計算される。

$$\gamma_t = \gamma_0 \min \left(1, \frac{1}{R(s, a)} \right) \quad (3)$$

信頼度 $R(s, a)$ は次式により更新される。

$$\zeta = \delta^2 + \gamma_R R^2(s', b) - R^2(s, a) \quad (4-a)$$

$$R^2(s, a) \leftarrow R^2(s, a) + \alpha_R \zeta \quad (4-b)$$

ただし、 α_R 、 γ_R はそれぞれ信頼度の学習率、割引率である。なお、尾川らは、Actor-Critic 法を用いているため、 $R(s', b)$ として次状態 s' において実行した行動の信頼度を用いている。しかし、本研究は Q 学習を用いるため、 $R(s', b)$ として次状態 s' において最も Q 値の高い行動の信頼度を用いる。

信頼度 $R(s, a)$ は、エージェントが状態 s において行動 a を実行したとき、得られた報酬と Q 値との二乗誤差 δ^2 の累積を表わす。そのため、報酬と Q 値の誤差が大きいつき、信頼度 $R(s, a)$ の値は大きくなり、割引率 γ の値は小さくなる。一方、報酬と期待値の誤差が小さいとき、信頼度 $R(s, a)$ の値は小さくなり、割引率 γ の値は大きくなる。

4. シミュレーション

4.1 実験設定

Fig. 2 (a) に示すシミュレーション環境には、時計回り/反時計回りに周回するエージェントがそれぞれ 3 台ずつ配置されており、各エージェントには 0 から 5 までの番号が割り振られている^(注 4)。実験では、エージェントが非同期に意思決定する状況を模擬するために、時刻 t において 0 番目のエージェントから順番に意思決定するのではなく、ランダムな順番で行なう。そのため、Fig. 2 (b) のように 2 台のエージェントが狭路で対峙する場合、左右どちらのエージェントが先に意思決定するかわからないため学習がより困難になる。

エージェントの目的は、壁や他のエージェントに衝突しないで時計回りあるいは反時計回りに安定して周回し続けることである。エージェントは、進行方向に対し、Fig. 2 (c) に示す 11 近傍の状態と、通路の内側または外側の違いを識別できる。11 近傍の入力は、マスが空白のときゼロ、壁のとき 1、周回方向が同じグループのエージェントのとき 2、周回方向が

(注 4) 実験では、シミュレーション環境の拡張性（障害物やバイパスの追加など）を考慮して周回環境を用いている。

Fig. 2 Simulation setting

異なるグループのエージェントのとき3となる。エージェントは、前進、後進、左旋回、右旋回、停止から行動を選択する。

エージェントは、通路に対して前に進んだとき報酬 $r = 1$ が与えられ、他のエージェントや壁に衝突したとき報酬 $r = -2$ が与えられる。エージェントは前に進めなかったとき報酬 $r = -1$ が与えられる。ただし、エージェントがコーナにいるときは、周回方向に旋回したとき、つぎの通路に入ったと判断し、その後は新しい通路に沿って上記ルールを適用する。

Q 値および信頼度の初期値はすべてゼロとする。Q 学習および信頼度の学習率を $\alpha = 0.1$ とし、割引率の初期値を $\gamma_0 = 0.9$ とする。ボルツマン選択における温度パラメータを $T = 0.1$ とする。実験では、エージェントが500 ステップ行動したときエピソードを更新し、エピソードを更新するとエージェントは初期位置から再スタートする。3000 エピソードを1 トライアルとし、25 トライアル行なう。

シミュレーションでは、提案手法の有効性を検証するために、従来の Q 学習と、周回方向の異なるグループの違いを識別できる場合/できない場合、信頼度の割引率が $\gamma_R = 0.0$ の場合/ $\gamma_R = 0.9$ の場合の計6つの実験設定における学習性能を比較する。なお、信頼度の割引率が $\gamma_R = 0.0$ のとき、(4) 式より時刻 t の二乗誤差 δ^2 のみが信頼度 R_t に影響を与え、 $\gamma_R = 0.9$ のとき、時刻 t の二乗誤差 δ^2 と次状態の信頼度 R_{t+1} が信頼度 R_t に影響する。すなわち、信頼度の割引率 γ_R の値が大きい場合、将来の期待値の予測誤差が現在の信頼度に強く影響する。

4.2 獲得した競合回避行動の評価指標

エージェントが獲得した競合回避行動の安定性を評価する

ために以下の4つの指標を用いる。

- (a) 正の報酬回数 (the number of positive rewards) 各エピソードにおいて、エージェントが正の報酬を得た回数。
- (b) 観測状態数 (the number of observed states) 各エピソードにおいてエージェントが観測した状態数。
- (c) 条件付エントロピー (conditional entropy) 各エピソードにおいて、エージェントが観測した状態において実行した行動の条件付エントロピー。
- (d) 平均割引率 (average discount rate) 各エピソードにおけるエージェントの割引率 γ の平均。

エージェント群が競合回避行動を獲得すると、各エージェントが正の報酬を得る回数が増加し、観測する状態数が減少する。また、各エージェントがある状態 s において特定の行動 a を選択するようになると、条件付エントロピー $H(a|s)$ は減少し、得られた報酬と Q 値との二乗誤差が減少するため割引率 γ の値は0.9に収束する。そのため、各エージェントの正の報酬回数が増加し、観測状態数および条件付エントロピーが減少し、平均割引率が0.9に収束したとき、エージェント群は競合回避行動を獲得したと判断することができる。なお、条件付エントロピーを次式により計算する。

$$H(a|s) = - \sum_j^N p(s_j) \sum_i^M p(a_i|s_j) \log p(a_i|s_j) \quad (5)$$

ただし、 N はエージェントが各エピソードで観測した状態数であり、 M はエージェントの行動数である。 $p(s_j)$ は状態 s_j の出現確率であり、 $p(a_i|s_j)$ は状態 s_j において行動 a_i が実行される確率である。

4.3 獲得した競合回避行動の分類

エージェント群が競合回避行動を獲得するのに要する時間はトライアルごとに異なる。そのため本研究では、全トライアルにおける各評価指標の平均値を用いて6つの実験設定の学習性能を比較するのではなく、エージェント群が獲得した競合回避行動を以下の3つに分類して学習性能の比較を行なう。

安定した競合回避行動 (stable)

Fig. 3 に示すように、正の報酬回数が増加し、観測状態数および条件付エントロピーが減少して一定になった場合、エージェント群は安定した競合回避行動を獲得していると分類する。

周期的かつ安定した競合回避行動 (periodic)

Fig. 4 に示すように、正の報酬回数が増加し、観測状態数が減少しているが、条件付エントロピーが一定の周期で増減している場合、エージェント群は周期的かつ安定した競合回避行動を獲得していると分類する。

不安定な競合回避行動 (unstable)

Fig. 5 に示すように、正の報酬回数、観測状態数、条件付エントロピーがともに非周期的に増減している場合、エージェント群は不安定な競合回避行動を獲得していると分類する。

なお、Fig. 3 から 5 は信頼度の割引率 $\gamma_R = 0.0$ のときの結果の一例であり、Fig. 6 は信頼度の割引率が $\gamma_R = 0.9$ の

Fig. 3 Learning results when agents acquired a stable conflict avoidance behavior ($\gamma_R = 0.0$)

Fig. 4 Learning results when agents acquired a periodic conflict avoidance behavior ($\gamma_R = 0.0$)

Fig. 5 Learning results when agents acquired an unstable conflict avoidance behavior ($\gamma_R = 0.0$)

Fig. 6 Learning results when agents acquired a periodic conflict avoidance behavior ($\gamma_R = 0.9$)

Table 3 The number of acquired conflict avoidance behaviors

	Functions of an agent	stable	periodic	unstable
1	Q-learning	2 (8%)	0 (0%)	23 (92%)
2	Q-learning +reliability ($\gamma_R = 0.0$)	2 (8%)	0 (0%)	23 (92%)
3	Q-learning +reliability ($\gamma_R = 0.9$)	0 (0%)	0 (0%)	25 (100%)
4	Q-learning +identification	2 (8%)	2 (8%)	21 (84%)
5	Q-learning +identification +reliability ($\gamma_R = 0.0$)	7 (28%)	8 (32%)	10 (40%)
6	Q-learning +identification +reliability ($\gamma_R = 0.9$)	3 (12%)	8 (32%)	15 (56%)

ときの結果の一例である。

4.4 実験結果

各実験設定においてエージェントが競合回避行動を獲得した回数を **Table 3** に示す。なお、表の stable は安定した競合回避行動を獲得した回数, periodic は周期的かつ安定した競合回避行動を獲得した回数, unstable は不安定な競合回避行動を獲得した回数である。

各エージェントが、周回方向の異なるグループの違いを識別できない場合、エージェント群は競合回避行動を 25 トライアル中 2 回 (8%) しか獲得できていない。一方、各エージェントが、周回方向のグループの違いを識別できる場合、従来の Q 学習を用いたエージェントは、安定した競合回避行動および周期的かつ安定した競合回避行動を合わせて 4 回 (16%) 獲得している。さらに、各エージェントが、周回方向のグループの違いを識別でき、かつ、信頼度の割引率が $\gamma_R = 0.0$ の場合、エージェントは競合回避行動を 15 回 (60%) 獲得し、 $\gamma_R = 0.9$ の場合、エージェントは競合回避行動を 11 回 (44%) 獲得している。上記の実験結果から、周回方向の異なるグループの違いを識別でき、かつ、信頼度を用いた Q 学習のほうが、従来の Q 学習と比較して高い確率で競合回避行動を獲得していることがわかる。

以下では、上記のように信頼度の割引率が競合回避行動の獲得に影響するメカニズムを解明するために、信頼度を用いたときの競合回避行動の獲得プロセスを、信頼度の割引率が $\gamma_R = 0.0$ と $\gamma_R = 0.9$ の場合に分けて分析する。

4.4.1 信頼度の割引率が $\gamma_R = 0.0$ の場合

4.4.1.1 安定した競合回避行動

信頼度の割引率が $\gamma_R = 0.0$ のとき、エージェント群が獲得した競合回避行動の一例を **Fig. 7** に示す。エピソード初期において、Fig. 7 (a) のようにエージェント no.0 と no.2 が道を譲り、エージェント no.1 は前進行動を選択している。しかし、周回を重ねると、Fig. 7 (b) のようにエージェント no.1 が道を譲り、エージェント no.0 と no.2 が前進行動を選択するように行動が変化している。エージェント no.4 は、Fig. 7 (c) と Fig. 7 (d) のように他のエージェントと遭遇すると道を譲る行動を獲得している。

Fig. 7 An acquired stable conflict avoidance behavior

このように、各エージェントが状況に合わせて道を通過する/道を譲るという行動を獲得して競合を回避することで、Fig. 3 に示すように正の報酬数が増加し、各エージェントが観測する状態数は減少している。また、各エージェントの平均割引率は、学習初期においてエージェントが壁や他のエージェントと衝突するため低下するが、エージェントが競合回避行動を獲得すると、得られる報酬と Q 値との二乗誤差が減少するため平均割引率は約 0.9 に収束している。

各エージェントの条件付エントロピーは、各状態において特定の行動の選択確率が高いと、小さくなる。そのため、安定した競合回避行動を獲得すると、Fig. 3 (c) のように各エージェントの条件付エントロピーは減少し、収束する。また、道を譲るより通過するほうが高い報酬を得られるため、道を通過する行動の選択確率のほうが高くなる。そのため、道を譲るエージェントより通過を好むエージェントのほうが、条件付エントロピーが小さくなる。Fig. 3 (c) を見ると、エピソード初期にしか道を譲らないエージェント no.1 の条件付エントロピーは小さく、他のエージェントに遭遇すると道を譲るエージェント no.4 の条件付エントロピーは大きいことがわかる。

4.4.1.2 周期的かつ安定した競合回避行動

つぎに、エージェントが獲得した周期的かつ安定した競合回避行動の一例を **Fig. 8** に示す。実験では、各エージェントは時刻 t において 0 番目のエージェントから順番に意思決定するのではなく、ランダムな順番で行なう。そのため、Fig. 8 (a) のつぎのステップで、先にエージェント no.0 が意思決定する場合 (Fig. 8 (b)) と、先にエージェント no.5 が意思決定する場合 (Fig. 8 (c)) が生じる。同様のことが、エージェント no.1 と no.4 においても観測されている。Fig. 4 (c) において、エージェント no.1, no.4, no.5 の条件付エントロピーが

Fig. 8 An acquired periodic conflict avoidance behavior

Fig. 9 The detailed process of the conflict avoidance behavior that was shown in Fig. 7 (a)

上下に振れているのは、そのためである。ただし、エージェント no.0 は、異なる行動を選択したのは上記の 1 箇所のため、条件付エントロピーはほぼ一定になっている。このように、各エージェントは状況にあった数種類の競合回避行動を獲得することで、周期的かつ安定した競合回避行動を獲得している。

Table 4 The conflict avoidance behavior constructed by six rules

rule	agent no.0	rule	agent no.1
r_{01}	if s_{01} then right turn	r_{11}	if s_{11} then go forward
r_{02}	if s_{02} then left turn	r_{12}	if s_{12} then go forward
r_{03}	if s_{03} then go forward	r_{13}	if s_{13} then stop
r_{04}	if s_{04} then go forward	r_{14}	if s_{14} then go forward
r_{05}	if s_{05} then left turn	r_{15}	if s_{15} then go forward
r_{06}	if s_{06} then right turn	r_{16}	if s_{16} then right turn

(a) Roles of discount rate γ in updating Q values (b) Roles of discount rate γ_R in updating reliability

Fig. 10 Roles of two kinds of discount rates in reward propagation

4.4.1.3 競合回避行動獲得プロセスの分析

Fig. 7 (a) に示した競合回避行動の各ステップを Fig. 9 に示す。エージェント no.0 は、12 ステップ目でエージェント no.1 と遭遇したとき道を譲り、エージェント no.1 は直進することで競合を回避している。このとき各エージェントは、Table 4 に示す 6 つのルールを獲得することで回避行動を実現している。各エージェントは同時に 6 つのルールを獲得したのではなく、まず、つぎの 2 つのルールを獲得している。

エージェント no.0 は、エージェント no.1 を前方に観測した状態 s_{01} でルール r_{01} により右旋回し、エージェント no.1 を左後方に観測した状態 s_{02} でルール r_{02} により左旋回し、道を譲っている。そして、エージェント no.1 は、エージェント no.0 を前方に観測した状態 s_{11} でルール r_{11} により前進し、エージェント no.0 を左前方に観測した状態 s_{12} でルール r_{12} により前進し、衝突を回避している。

本来、ルール r_{01} は、Fig. 10 (a) に示すように、道を譲る行動のため負の報酬 $r = -1$ が与えられ、状態 s_{01} で右旋回行動は選択されなくなる。しかし、ルール r_{02} は、左旋回して前進するため正の報酬 $r = 1$ が与えられ、 $Q(s_{02})$ の値は高くなっている。このとき、もし割引率 γ の値が高い場合、ルール r_{01} は $\gamma \max Q(s_{02})$ を得て、状態 s_{01} で右旋回行動が強化される。そのため、報酬に非対称性がある場合でも道を譲る行動を学習できる。この 2 つのルールによる競合回避行動は、エピソード No.9 に初めて出現している。

Fig. 11 A situation that the agent no.1 confuses with the situation in Fig. 9 (b)

Fig. 12 The relationship between reliability discount rates and the number of conflict avoidance behaviors

しかし、競合回避行動を構成するルール、たとえば、ルール r_{02} は Fig. 9 (b) と類似した Fig. 11 の状態でも選択される。このときルール r_{02} は、類似状態において他のエージェントと衝突して負の報酬を受ける、あるいは、次状態の Q 値が低いために現在の Q 値も下がり、選択されなくなる可能性がある。そのため、2 エージェントは上記 2 つのルールを獲得するまでにエピソード No.46 までかかり、6 つのルールを獲得するまでに、エージェント no.0 はエピソード No.65 までかかり、エージェント no.1 はエピソード No.109 までかかっている。

4.4.2 信頼度の割引率が $\gamma_R = 0.9$ の場合

信頼度の割引率が $\gamma_R = 0.9$ のとき、エージェント群が獲得した学習結果の一例を Fig. 6 に示す。Fig. 6 (d) を見るとエピソード初期に割引率 γ の値は 0.5 まで下がり、その後、上昇していないことがわかる。これは、信頼度の割引率が $\gamma_R = 0.9$ の場合、Fig. 10 (b) に示すように、エージェントは次状態の信頼度 $R(s)$ も考慮して現在の状態の信頼度を更新する。そのため、遠い未来の誤差 δ が現在の信頼度に影響し、全状態の誤差が小さくならないと割引率 γ の値が高くなることを意味する。Fig. 10 (a) で説明したように、割引率 γ の値が小さいと、道を譲る行動が強化されにくいため、競合回避行動を獲得する確率が低くなる。そのため、エージェント群は、信頼度の割引率が $\gamma_R = 0.0$ と比べて、 $\gamma_R = 0.9$ の場合、競合回避行動の獲得回数が少なかったものと考えられる。

実際、Fig. 12 に示すように、信頼度の割引率 γ_R と競合回避行動の獲得回数を見ると、信頼度の割引率 γ_R が高くなるほど、競合回避行動の獲得回数が少なくなっていることがわかる。

Fig. 13 A perceptual aliasing problem that results when an agent cannot observe the other agents' direction

4.5 考察

4.5.1 信頼度の役割

本項では、信頼度を用いて Q 学習の割引率を自律的に調整した場合、従来の Q 学習に比べて競合回避行動の獲得確率が高い理由について考察する。学習初期、各エージェントはランダムに行動を選択するため、あるエージェントが競合回避行動を獲得しても他のエージェントが未学習の場合、エージェント同士が衝突して行動系列が崩れる可能性がある。また、競合回避行動を構成するルール群は、類似状態でも選択されるため、次状態の Q' 値のほうが小さい場合、(2-a) 式の $\delta = r + \gamma \max Q' - Q$ により値を下げる。このように、各エージェントは、競合回避行動を構成する有用なルール群を崩壊させようとする圧力を常に受ける。

しかし、信頼度を用いて Q 学習の割引率を自律的に調整する場合、もし、得られた報酬や次状態の Q' 値が、選択されたルールの Q 値より小さいと、誤差 δ が大きくなり、(4) 式により信頼度 R の値が変化し、割引率 γ の値が小さくなる。そのため、有用なルールが不完全知覚状態において誤って選択され、負の報酬を受けるといった場合でも、割引率 γ の値が次第に小さくなるため、有用なルールの Q 値が激減するとといった危険性を低減することができる。

このように、信頼度を用いて Q 学習の割引率を自律的に調整することで、一度獲得した有用なルール群を維持しつつ、新しいルール群を獲得することができる。そのため、従来の Q 学習より高い確率で競合回避行動を獲得できたものと考えられる。

4.5.2 不完全知覚状態における行動学習

本項では、 Q 学習が不完全知覚状態の発生する環境下において、競合回避行動を獲得できた理由について考察する。本稿の狭路問題の設定では、エージェントは、他のエージェントの個体の違いや進行方向などの情報を観測できないため、異なる状態を同一状態と誤認する不完全知覚状態が発生する^(注5)。特に学習初期は、Fig. 13 (a) に示すように、エージェントがランダムに行動するため不完全知覚状態が発生しやすい。しかし、Fig. 13 (b) に示すように、各エージェントが学習により、ある状態において特定の動作を選択するようになると、

(注5) 不完全知覚状態の発生を抑えるためには、エージェントの識別能力を高める方法が考えられる。しかし、学習に必要な状態数が増大し、学習の収束速度が遅くなる、あるいは、学習できない可能性が高くなるといった別の問題が発生する。

エージェントは、状態が一意に特定できるようになるため、エージェントが観測する状態数は減少し (Fig. 3 (b)), 不完全知覚状態が発生する確率が減少する。エージェントは、学習初期より少ない状態において適切な行動を学習すれば良いため、エージェントが探索しなければならない状態空間は狭くなり、学習が容易になる。そのため、不完全知覚状態が発生する状況でも競合回避行動が獲得できたものと考えられる。

5. 結 言

マルチエージェントシステムは、システム内外の環境変化に対して柔軟で頑健なシステムを構築できるアプローチとして注目されている。しかし、多数のエージェントが限られた空間で複雑に相互作用するため、エージェント間に競合状態が発生する。本研究では、マルチエージェントシステムにおいて発生する競合状態を解消するために、信頼度を用いた強化学習により競合回避行動を獲得する方法を提案した。

本稿では、マルチエージェントシステムにおける競合問題の1つとして、狭い通路において多数のエージェントがすれ違う狭路問題を扱い、均質なエージェント群が強化学習により道を譲るエージェントと先に通過するエージェントに分化することで、エージェント間の競合を回避できることを示した。また、信頼度を用いて強化学習の一手法であるQ学習の割引率を自律的に調整することで、先行研究のようにエージェント間で報酬を分配・調整することなく、従来の強化学習より高い確率で競合回避行動を獲得できることを確認した。そして、競合回避行動の獲得プロセスを分析し、信頼度が競合回避行動の獲得に対し有効に作用するメカニズムを解明した。

今後、エージェントの規模を数台から十数台の間で増減しても、エージェント群が競合回避行動を獲得できるように、逐次報酬に加えて遅延報酬の導入や、エージェントの知覚能力の向上による不完全知覚状態の低減といった拡張を予定している。また、動的な環境におけるQ学習の性能を高めるために、学習率やボルツマン選択における温度パラメータの調整方法^{14)~16)}の導入を検討している。最後に、将来の展望として、提案手法を次世代生産システムにおける自動搬送車や社会シミュレーション^{17), 18)}などへの応用を検討していく。

参 考 文 献

- 1) K.P. Sycara: Multi-Agent Systems, AI Magazine, **19**-2, 79/92 (1998)
- 2) P. Stone and M. Veloso: Multiagent Systems: A Survey from a Machine Learning Perspective, Autonomous Robots, **8**-3, 345/383 (2000)
- 3) 宮崎, 荒井, 小林: Profit Sharing を用いたマルチエージェント強化学習における報酬分配の理論的考察, 人工知能学会誌, **14**-6, 1156/1164 (1999)
- 4) L. Busoniu, R. Babuska and B. De Schutter: A Comprehensive Survey of Multi-Agent Reinforcement Learning, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, **38**-2, 156/172 (2008)
- 5) R.S. Sutton and A. Barto: *Reinforcement Learning: An Introduction*, The MIT Press (1998)
- 6) 生天目, 岩永, 佐藤: ミクロ-マクロ・ループによる自己組織性とその評価, システム/制御/情報, **46**-9, 561/568 (2002)
- 7) 生天目章: ゲーム理論と進化ダイナミクス, 森北出版 (2004)
- 8) 宮崎, 小林: Profit Sharing の不完全知覚環境下への拡張: PS-r* の提案と評価, 人工知能学会論文誌, **18**-5, 286/296 (2003)
- 9) 半田久志: 強化学習エージェントのための知覚変化予測に基づいた状態分割, 第5回 AI 若手の集い MYCOM2004 資料, 9/15 (2004)
- 10) 柴田, 上田, 伊藤: 強化学習による個性・社会性の発現・分化モデル, 計測自動制御学会論文集, **39**-5, 494/502 (2003)
- 11) K. Moriyama and M. Numao: Self-Evaluated Learning Agent in Multiple State Games, Proc. 14th European Conference on Machine Learning, ECML-2003, 289/300 (2003)
- 12) 上田, 谷澤, 高橋, 宮原: マルチエージェントシステムにおける利他的な行動規則の獲得, 電子情報通信学会論文誌, **J88-D-I-9**, 1278/1286 (2005)
- 13) 尾川, 並木, 石川: 学習進度を反映した割引率の調整, 電子情報通信学会ニューロコンピューティング研究会, 電子情報通信学会技術研究報告, NC2002-129, 73/78 (2003)
- 14) 野田五十樹: マルチエージェント環境下における強化学習のステップサイズパラメータの適応, 人工知能学会全国大会予稿集, CD-ROM, 1A3-1 (2010)
- 15) 野田, H.-T. Kim: マルチエージェント学習下における温度パラメータの調節手法, 人工知能学会全国大会予稿集, CD-ROM, 1G1-1 (2011)
- 16) 石井 信: 強化学習におけるランダムさの自己調節, 日本神経回路学会誌, **9**-4, 254/262 (2002)
- 17) 石田, 鳥居, 村上, 寺野: 社会シミュレーションと参加型デザイン, 情報処理, **48**-3, 271/277 (2007)
- 18) 小川, 山本, 岡田, 諏訪, 太田: エージェントベースシミュレーションによる知識共有コミュニティの報酬制度設計, 電子情報通信学会論文誌, **J94-D-6**, 945/956 (2011)

[著 者 紹 介]

山 田 和 明 (正会員)

2002年神戸大学大学院自然科学研究科修士。同年東京大学先端科学技術研究センター特任助手。2005年東京大学大学院情報学環特任講師。2008年東洋大学工学部機能ロボティクス学科講師。2009年東洋大学理工学部機械工学科講師。現在に至る。機械学習, マルチエージェントシステムの研究に従事 (博士 (工学))。

高 野 慧

2011年東洋大学工学部機能ロボティクス学科卒業。現在, 東洋大学大学院工学研究科在籍中。機械学習, マルチエージェントシステムの研究に従事。