環境状況の変化に応じて 自己の報酬を操作する学習エージェントの構築

森山 甲一　　沼尾 正行

Abstract

The authors aim at constructing an autonomous agent learning appropriate actions in several situaions of a Multi-Agent environment. Under this aim, the authors constructed an agent having abilities that can handle the agent's own rewards and distinguish the situations, and we saw that the agent can act well in each of the situations of the environment composed of homogeneous agents. In this paper, on the other hand, we check whether the agent can also act well in the environment in which the situations switch over.

1　はじめに

自律エージェントの学習とは可能な選択肢の選択基準を獲得することであり、通常は何らかの外部評価に基づいてその評価値を最大化する選択肢を選ぶようにする。しかし、例えば複数の評価基準が存在してそれらの並立が難しい環境ではどの評価基準に従うかの選択の問題があり、またエージェントが複数存在するマルチエージェント環境では社会的ジレンマの問題がある。

本研究では、マルチエージェント環境における社会的ジレンマの問題に着目する。社会的ジレンマの下では、個々が利益を追求すればするほど（合理的に行動するほど）実際に得られる利益が減少するため、個々は非合理的に行動することが求められる。社会科学の分野では[1]、この場合に何らかの権威者（政府など）が法律や税制などを整備することによりジレンマ状況を解決する研究がなされている。しかし、例えば我々人間の場合は状況に応じて非合理的に行動できるため、法律などによる細かな規制がない場合でも社会生活を成立させることが可能である。

そこから、権威者による上位の制御がない場合でもジレンマ状況を解決できる人間のような非合理的エージェントを構築することが考えられる。しかし、一方では従来の評価を最大化する合理的行動の学習に関する研究の蓄積があり、それらを有効利用することが望ましい。そこで本研究では、自分に与えられた外部評価を操作して内部評価を生成し、それを従来の学習手法に適用することにより非合理性を実現するエージェントを構築することを目的とする。

確かに、例えば自己の内部評価が他者の利益を含むように予め設計者が設計すればエージェントはジレンマ状況を克服できるだろう。しかし、我々の人間社会のような現実のマルチエージェント環境は常にジレンマ状況であるとは限らず、むしろ動的に変化する。そのため、時と場合によっては他者への影響が無視できる状況も存在し、その中では上記のように設計されたエージェントは明らかに非効率なものとなる。また、そのようなエージェントは開かれたマルチエージェント環境に起こり得る状況の変化には対応できない。つ

Construction of a Learning Agent Handling Its Rewards According to Changes of Environmental Situations.

Koichi Moriyama, 東京工業大学 情報理工学研究科 計算工学専攻, Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology. koichi@nm.cs.titech.ac.jp

Masayuki Numao, 同上, ditto. numao@cs.titech.ac.jp

まり、現実の開かれたマルチエージェント環境で適切に行動するエージェントを構築する場合には、予め設計者が設計するのではなく、エージェント自身がその環境中でジレンマ状況を克服することが求められる。

これらの目的の下で筆者らは、外部評価から内部評価を生成する能力に加えて環境の状況を識別する能力をエージェントに付与し、状況に応じて自己の内部評価を生成するエージェントを構築した [4]。そしてそのエージェントを集めた同質なマルチエージェント環境において適切にジレンマ状況が克服できる一方、非ジレンマ状況においても比較的良好な結果が得られることを示した [4]。しかし、そこでは異なる状況それぞれについて実験を行なっており、実験途中で状況が変化する場合を扱っていない。そこで本稿では途中で状況が変化するような実験を行ない、文献 [4] で提案した手法について検証する。

本稿は以下のような構成になっている。2 節で本研究で対象とするマルチエージェント環境を分類し、3 節では文献 [4] で提案したエージェントの構築手法について述べる。4 節では共有地の悲劇 [2] [8] をモデル化したゲーム [3] と筆者オリジナルの狭路問題という 2 種類の問題について、途中で状況が変化する実験を行なった結果を述べる。5 節でその実験結果について考察を行ない、6 節でまとめと今後の課題を示す。

2　マルチエージェント環境

本研究では個々のエージェントとエージェント社会全体の関係に着目して、環境を次の 3 つの状況に分類する [4]。

- 非干渉状況: 全てのエージェントが利益を追求する時に社会全体が最良
- 泥沼状況: 全てのエージェントが利益を追求しない時に社会全体が最良
- 競合状況: 一部のエージェントが利益を追求し、残りが利益を追求しない時に社会全体が最良

それぞれの状況で望ましい行動を 2 人ゲームの表の形で **表 1** に示す。

非干渉状況は個々が単に利益を追求すれば良いため非ジレンマ状況に当たる。泥沼状況は個々が自分の利益をより良くしようと合理的に行動するほど全体が悪化し個々の利益も減少する状況であり、囚人のジレンマ問題はこれに当てはまる。競合状況は多くの個体が同時に合理的に行動してもそれらの要求を満たせないタイプの状況であり、狭い空間などの排他的な資源によって競合が発生する状況に相当する。

表 1　各状況における望ましい行動組み合わせ

(a) 非干渉状況

	利益追求	利益不追求
利益追求	○	
利益不追求		

(b) 泥沼状況

	利益追求	利益不追求
利益追求		
利益不追求		○

(c) 競合状況

	利益追求	利益不追求
利益追求		○
利益不追求	○	

我々の人間社会のような現実の開かれたマルチエージェント環境は、これらの状況が複雑に組み合わさっていると考えられる。従って、そのような環境で行動するエージェントには、上記のそれぞれの状況で適切な学習を行なう能力とともに状況を識別する能力が必要となるだろう。

3　学習エージェントの構築

本研究ではエージェントの学習手法として、明確な教師を必要としない強化学習 [6] を用いる。強化学習は報酬というただ 1 つのスカラー強化信号を用いて、それを可能な限り大きくする行動を選択させようとする学習手法である。以下では強化学習のうちの代表的手法である Q-learning [7] を用いる。これは状態 s と行動 a の組み合わせに伴う価値（報酬の見込み値）$Q(s,a)$（Q 値）を学習サイクルごとに以下の式により更新し、行動選択の際にこの値を用いるものである。r_{t+1} は今回の行動提示で得られた報酬を、

α, γ はそれぞれ学習率と割引率を表す†1。

$$Q_t(s,a) = Q_{t-1}(s,a) \quad \text{if } s \neq s_t \text{ or } a \neq a_t,$$
$$Q_t(s_t,a_t) = (1-\alpha)Q_{t-1}(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_b Q_{t-1}(s_{t+1},b)). \quad (1)$$

しかし、これは報酬 r_{t+1} を最大化する合理的行動を選択するように学習する手法である。そのため、マルチエージェント環境、特に泥沼状況や競合状況では各エージェントはどのように学習を行なえば良いだろうか。

その 1 つの解として筆者らは学習の際に報酬値を調整することを提案した [4]。そこで用いた手法は、エージェント A_i の時刻 t の報酬 $r_{i,t+1}$ にパラメータ $\lambda_{i,t+1}$ を加えたものを A_i の学習用報酬 $r'_{i,t+1}$ とし、これを (1) 式の r_{t+1} の代わりに用いるものである。なお、以下では表記の繁雑さを避けるために自分（A_i）を表す添字 i を省略する。

$$r'_{t+1} = r_{t+1} + \lambda_{t+1}. \quad (2)$$

本研究ではまず泥沼状況に有効な λ_{t+1} として

$$\lambda_{t+1} = \sum_{A_k \in N_i \backslash A_i} r_{k,t+1} \quad (3)$$

を用いる。この式で $N_i \backslash A_i$ はエージェント A_i の近隣のエージェントからなる集合 N_i のうち自分自身を除いたものを表している。(2) 式と (3) 式を合わせると、これは自分及び近隣他者の現在の報酬の和を学習用報酬として用いることを意味し、自分が合理的に行動することにより周囲に悪影響が現れる環境において有効である。

また、競合状況に有効な λ_{t+1} として

$$\lambda_{t+1} = r_{t+1} - r_t \quad (4)$$

を用いる。これは報酬の時間変化を考慮したもので、前回の行動による報酬と今回のそれの差異を強調するものである。

これらの λ_{t+1} はそれぞれ対応する環境については有効であるが、異なる環境ではあまり芳しくない。例えば (3) 式を用いた場合、非干渉状況では通常の強化学習のみを用いた場合よりも結果が悪い。また (4) 式を用いた場合には、泥沼状況では文字通り泥沼にはまってしまう。ところが非干渉状況でこの式を用いるとエージェントの合理性をより高める結果となり、良好な結果が得られる†2。これらの点から、現実の開かれたマルチエージェント環境で行動するエージェントは λ_{t+1} を適切に選択しなければならない。すなわち、環境が泥沼状況の際には (3) 式を、非干渉状況・競合状況の際には (4) 式を自動的に選択するような能力をエージェントは保持しなくてはならない。

そこで筆者らはこの両者を状況によって自動的に選択する手法を考案した [4]。それには、個々のエージェントがどのようにして状況を識別するかということが重要になるわけだが、提案手法では以下 2 つの状況識別仮定を置き、少なくとも 1 つが満たされる場合に泥沼状況と認識するようにした。

$$Q_{t-1}(s_t,a) < 0 \quad \text{for all } a. \quad (5)$$
$$r_{t+1} < Q_{t-1}(s_t,a_t) - \gamma \max_b Q_{t-1}(s_{t+1},b). \quad (6)$$

(5) 式は現在どのような行動を実行しても今後正の報酬が見込めないことを意味している。(6) 式は (1) 式から考案したものであり、左辺は実際に獲得した報酬を表している。一方で通常の Q-learning の場合、Q 値が収束した際にはこの両辺が一致することから、右辺が現在までに学習した報酬の予測値を意味していると考えられる。すなわち、(6) 式は実際に得られた報酬が予測値よりも小さいことを表しており、これは報酬の予測値を学習した時点よりも現在の環境が悪化していることを意味している。

ところが、本研究では (2) 式により調整した学習用報酬 r'_{t+1} を Q 値の学習に用いているため、(6) 式の左辺は学習用報酬 r'_{t+1} でなくてはならない。しかし、これは r'_{t+1} を計算するためにどちらの λ_{t+1} を選択するかを決めるための式であるので、r'_{t+1} をここで用いることは出来ない。そのため実際の獲得報酬 r_{t+1} で代用していることに注意を要する。

筆者らは、以上の仮定を満たす場合に (3) 式を、満たさない場合に (4) 式を用いて報酬を操作することを提案している [4]。

†1　以下、時刻の表記は Sutton らの流儀に依る [6]。そこでは時刻 t の状態 s_t で行動 a_t を実行すると、状態 s_{t+1} に遷移し報酬 r_{t+1} が与えられるとする。

†2　非干渉状況では合理的行動が望ましいことに注意。

4 実験

本研究では実験環境として、共有地の悲劇（n 人囚人のジレンマ）[2] [8] をモデル化したゲーム [3] と筆者オリジナルの狭路問題を用いる。いずれの実験も同質なエージェント 10 台で行なう。実験に用いるエージェントは、行動選択をランダムに行なうランダムエージェント、通常の強化学習を行なう通常エージェント、(3) 式を用いて報酬を操作する近隣報酬合計エージェント、(4) 式を用いる報酬差分エージェント、更に (3) 式と (4) 式を (5)(6) 式を用いて自動的に選択する自動選択エージェントの 5 種である。学習率 α・割引率 γ はともに 0.5 とする。行動選択には Q 値から求めた温度 $T = 1$ のボルツマン分布

$$p_t(a) = \frac{\exp\bigl(Q_{t-1}(s_t, a)/T\bigr)}{\sum_b \exp\bigl(Q_{t-1}(s_t, b)/T\bigr)} \qquad (7)$$

による確率 $p_t(a)$ を用いる [6]。

4.1 共有地の悲劇ゲーム

これは共有地の悲劇（n 人囚人のジレンマ）[2] [8] をモデル化したゲーム [3] である。以下ではこのゲーム環境において泥沼状況と非干渉状況を作り出し、エージェントの集合がどのように振る舞うかを確認する。

各エージェントは利己的・協力的・利他的の 3 種の行動のいずれかを一斉に提示し、エージェント社会全体における行動の組み合わせにより決められる報酬を獲得する。この一連の流れを 1 サイクルと称する。エージェント A_i の行動に対して社会全体に対するコスト c_i が付随する。ここでは利己的・協力的・利他的な行動についてそれぞれ $+\Delta r_c, 0, -\Delta r_c$ の各値をとるものとする。泥沼状況においては $\Delta r_c = 1$、非干渉状況においては $\Delta r_c = 0$ とした。一方、エージェント A_i の報酬はそれぞれ $3 - r_c, 1 - r_c, -3 - r_c$ とする。ここで r_c は共有コストと称し、$r_c \triangleq \sum_i c_i$ と定義する。この場合に、泥沼状況では全てのエージェントが利他的に行動した場合に全体の報酬合計が最大 70 となり、全てが利己的に行動した場合に -70 と最小になる。一方で非干渉状況では $r_c \equiv 0$ となり、各エージェントの獲得する報酬は他のエージェントに依存しないため、全てのエージェントが利己的に行動した場合に報酬合計は最大 30 となる。この環境で重要な点は、個々のエージェントの立場から見ると、泥沼状況においても他者の行動に関係なく利己的行動が常に最大の報酬をもたらすことである。

ここで一例としてゴミの分別を考えてみよう。焼却の際のダイオキシンの発生を防ぐために、可燃ゴミと不燃ゴミを分別するように求められているとする。ダイオキシンは塩素化合物を含むゴミを約 400°C で燃やすことにより生じ、発ガン性が疑われている [5]。この場合に上記の利己的・協力的・利他的の 3 種の行動はそれぞれ「分別しない」「自分のゴミは分別する」「集積場にある他者のゴミも分別する」となる。もし分別しなかった場合には、自治体はダイオキシンの発生を防ぐために税金を用いて人を雇い、ゴミの分別をさせることになるだろう。この場合に使われる税金が上記の共有コストに相当する。一方では、たとえ利他的に集積場の他者のゴミまで分別するとしても、その人が自治体に支払う税金が減るわけではない。そうするとゴミの分別は面倒なので、分別をしないという利己的な行動が最も好ましいと思う人が生じても不思議ではない。しかし、もし住民全員が利己的な行動を採ったらどうなるだろうか。これは増税をもたらすかもしれないため、泥沼状況となるだろう。

しかし、ダイオキシンは約 400°C で燃やすことにより生じるため、これよりも高い温度でゴミを焼却するようにすれば発生を防ぐことが可能である。また、ダイオキシンを除去する装置も世の中に存在する [5]。従って、高温でゴミを焼却するダイオキシン除去装置付きの焼却炉を自治体が建設した場合には、それ以後ダイオキシンの観点からは住民は可燃ゴミと不燃ゴミを分別する必要がなくなる。そうすると分別をしないことによる共有コストの増加は 0 となり、非干渉状況に移行したと考えられるだろう。

(3) 式に現れる近隣エージェント集合 N_i は

$$N_i \triangleq \bigl\{ A_k \,|\, k = (i+j) \bmod 10, j = 0, 1, 2, 3 \bigr\} \qquad (8)$$

で定義する [3]。更に、各エージェントの保持する状態は近隣エージェントの行動の組み合わせとする。

最初は泥沼状況だが 30000 サイクル後に突然非干渉状況へ変化した場合の実験結果を **図 1** に示す。こ

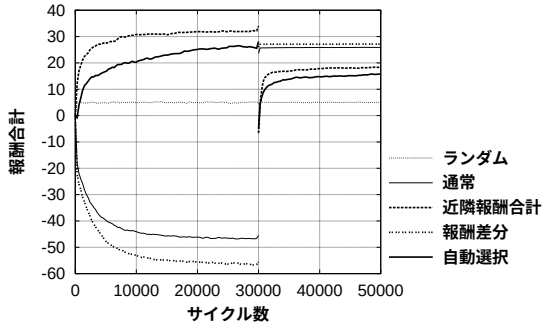

図 1　共有地の悲劇ゲームの実験結果

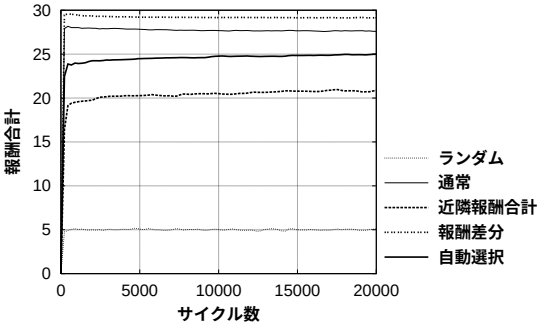

図 2　共有地の悲劇ゲーム（非干渉状況）の実験結果

のグラフは 100 回の実験の平均をベジエ曲線で近似したものであり、状況が変化する 30000 サイクルでグラフを分割している。縦軸が 10 台のエージェントの報酬合計、横軸がサイクル数である。30000 サイクルまでを見てみると「通常」は時間とともに報酬和が減少している。これは個々が合理的行動をするほど社会が悪化して個々の利益も減少する泥沼状況が生じていることを意味する。また、「近隣報酬合計」が最良の結果をもたらす一方で、「報酬差分」は「通常」よりも悪い結果となっている。「自動選択」は「近隣報酬合計」よりは遅れるが上昇している。

また、最初から非干渉状況である場合の実験結果を **図 2** に示す。グラフによると「報酬差分」が最良の結果をもたらしており、それに続くのが「通常」となっている。一方では、「近隣報酬合計」が途中で上昇が止まってしまっており、「通常」より悪くなっている。「自動選択」は「近隣報酬合計」と「報酬差分」のほぼ中間の結果となっている。

ここで、**図 1** の 30000 サイクル以後と **図 2** を比較すると、状況が途中で変化する場合の結果が初めから非干渉状況であるものよりも悪くなっていることが分かる。これは「ランダム」を除く全てのエージェント集合に言えることであるが、(2) 式により学習用の報酬を調整する 3 種のエージェントのうち「自動選択」の悪化が大きいことが見て取れる。

4.2　狭路問題

以下では筆者オリジナルの狭路問題を用いて競合状況と非干渉状況を作り出し、エージェント集合がどのように振る舞うかを確かめる。

狭路問題では、路上駐車のため道路に車がすれ違うことのできない狭い部分があり、狭い部分の両端に同時に 2 台の車が現れたと仮定する（**図 3** (a)）。問題はこの狭い部分を 2 台の車がいかにして早く通過するかというものである。車（エージェント）の取り得る選択肢として「進む」と「待つ」の 2 つがあり、各エージェントにとって最良の結果は相手を待たせて自分が先に通過することである。しかし両者が「進む」を選択すると、狭い部分ですれ違うことが不可能なので結局両者とも通過できない。一方で両者が共に「待つ」を選択すると何も状況は変化しない。そのため、この場合にはエージェントの一部が利益を追求（「進む」を選択）し、残りが利益を追求しない（「待つ」を選択する）ことが求められ、競合状況となる。

もし、ある時点でこの路上の車が両方とも移動してしまったとすると（**図 3** (b)）、もはや 2 台の車は相手の車を意識することなくこの場所を通過すること

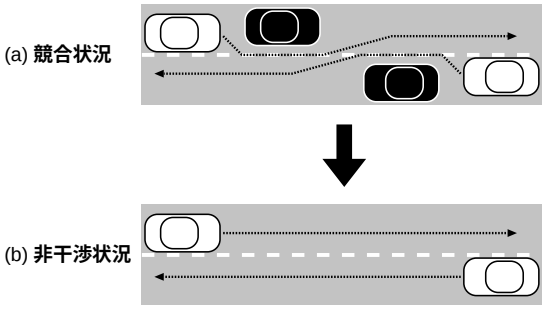

図 3　狭路問題

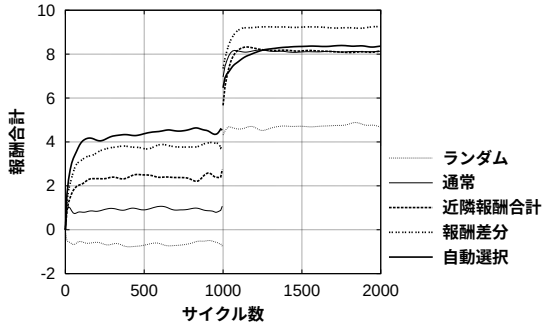

図 4　狭路問題の実験結果

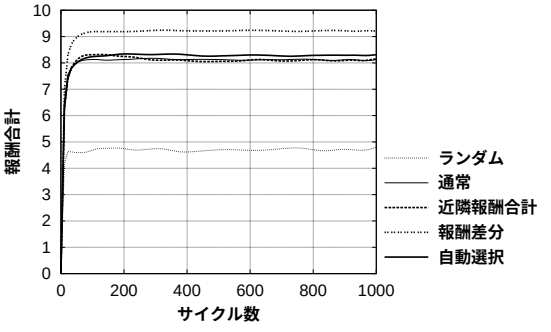

図 5　狭路問題（非干渉状況）の実験結果

が出来るようになる。従って、この場合には非干渉状況となる。

実験では 10 台のエージェントをランダムに 2 台ずつに分け、各組ごとに両者が行動を提示する。競合状況では各組で 1 台のみが「進む」となった場合にその「進む」を提示したエージェントは通過成功となる。一方で「待つ」を提示したエージェントは（相手なしで）再び行動を提示する。また、両者の行動が一致した場合には両者とも再び行動提示を行なう。毎回の行動提示の際に両者には負の報酬が与えられ、「進む」を提示して通過に成功したエージェントには正の報酬が与えられる。非干渉状況では「進む」を提示したエージェントは無条件で通過に成功し正の報酬が与えられ、「待つ」を提示したエージェントには負の報酬が与えられて再び行動提示を行なうことになる。どちらの状況においても $2^2 = 4$ 回の行動提示が行なわれた後にはエージェントが残っていてもその組は終了とする。その場合には正の報酬は与えられない。このようにして全ての組が終了したならばそれを 1 サイクルと称し、再び最初から繰り返す。

エージェントに与えられる報酬は、「進む」を提示して通過に成功したエージェントに $+1$、それ以外には行動提示ごとに -0.5 とした。つまり報酬合計の理論上の最大値は競合状況で $(1 + 0.5) \times 5 = 7.5$、非干渉状況で $(1 + 1) \times 5 = 10$ となる。各エージェントは行動提示ごとに学習を行なう。

近隣エージェント集合 N_i は組分け後の自分と相手とし、もし相手が先に通過に成功した場合には自分のみとする。各エージェントの状態は、自分の組の残りエージェント数と相手 ID の組み合わせとした。

最初は競合状況であったが 1000 サイクル後に非干渉状況となった場合の実験結果を **図 4** に示す。描画方法は **図 1** と同様であり、状況が変化する 1000 サイクルでグラフを分割している。1000 サイクルまでのグラフによると「自動選択」が最良の結果をもたらしており、それに続くのが「報酬差分」となっている。このように、「自動選択」がその元となっている「報酬差分」よりも良い結果となった原因の究明は今後の課題である。「近隣報酬合計」も「通常」よりは良いのだが「報酬差分」より悪いことが分かる。

また共有地の悲劇ゲームと同様に、初めから路上駐車のない非干渉状況における実験結果を **図 5** に示し、これを **図 4** の 1000 サイクル以後と比較する。この場合にはどちらも結果に差はなく、「報酬差分」が最良で「ランダム」を除く他のエージェント集合はほぼ同様の結果となっている。

5　考察

本稿では文献 [4] で筆者らが提案した手法について、途中で状況が変化する環境で実験を行なった。この実験結果について以下で考察を行なう。

まず 4.1 節の共有地の悲劇ゲームでは、「ランダム」を除く全てのエージェントについて、環境が泥沼状況から非干渉状況に移行した後（**図 1** の 30000 サイクル以後）の結果が初めから非干渉状況であった実験の結果（**図 2**）より明らかに悪くなっている。これは状況変化前に学習した結果の影響が出ているためと考えられる。

また、報酬を調整した 3 種のうち「自動選択」の結果の悪化が最も大きく、「近隣報酬合計」よりも劣るものとなっている。この原因としては、状況識別仮定のうちの (6) 式の左辺に r'_{t+1} の代わりに r_{t+1} を用いたことが影響していると考えられる。このゲームにおける泥沼状況では獲得報酬 r_{t+1} が最小となる利他的行動が望まれるため、必然的に r'_{t+1} に対する λ_{t+1} の影響が大きくなり、r'_{t+1} と r_{t+1} が全く異なる値になると思われる。すると学習の結果である Q 値も r_{t+1} から得られる値とは異なるものとなり、結果として (6) 式の判定に誤りが生じるのではないだろうか。またより本質的な問題として、(6) 式では学習が収束した状態を想定しているが、実験開始直後ではそれは望めないという点と、そもそも環境の非マルコフ性により Q-learning では収束性すら保証されていないという点が問題となっているのかもしれない。

一方、4.2 節の狭路問題の場合には、競合状況から非干渉状況に移行した後（**図 4** の 1000 サイクル以後）の結果は初めから非干渉状況であった実験の結果（**図 5**）とほぼ同じである。すなわち、この場合には状況変化前の学習の悪影響が見られない。これは (4) 式を用いれば状況変化前後どちらにも対応できるというように競合状況と非干渉状況は性質が似ているためと考えられる。しかし、非干渉状況において「自動選択」が「近隣報酬合計」とほぼ同じ結果になったということは、提案手法による状況の識別がうまく機能していないことを意味している。

もし、エージェントの知覚能力がもっと豊かであったらどうであろうか。共有地の悲劇ゲームの例では、自治体が新焼却炉を建設した場合に通常は住民に対して何らかのアナウンスがあるだろう。また狭路問題の例では、運転者は路上駐車がなくなったことを視覚から認識できるはずである。これらのようにエージェントが豊かな知覚能力により状況変化を認識できるならば、エージェントは状況が変化する 1 つの環境を状況が不変な複数の環境に分割することが可能となる。この場合には、それぞれの環境中では文献 [4] における実験と同等となり、過去の学習の悪影響を受けずに学習することが可能である。あるいは (5)(6) 式を用いて状況を推測しなくても、初めから (3) 式か (4) 式のどちらを利用すれば良いかが自明となるかもしれない。しかし、この議論は対象とする問題をエージェントの知覚の問題に転嫁しただけであり、状況変化が知覚できない場合については何も述べていないことに注意が必要である[†3]。本研究はこの点を踏まえて、与えられる外部評価を基に限られた知覚によって学習を行ない、環境に適応するエージェントを構築することを目的としている。

6　まとめ

本稿では筆者らが提案した手法 [4] について新たに行なった実験の結果を示した。文献 [4] ではマルチエージェント環境の 3 種それぞれの状況が不変な場合の実験のみを行なっていたが、本稿では途中で状況が変化する場合の実験を行なった。その結果、共有地の悲劇ゲームにおける泥沼状況から非干渉状況への移行の際には、提案手法は状況の変化に対応する能力が乏しいことが判明した。従って提案手法は、いろいろな状況が組み合わさっている現実の開かれたマルチエージェント環境に対応するためのエージェントを構築するという観点からは不合格であるといえる。しかし、考察で言及した通り、(6) 式の左辺を r_{t+1} で代用した点を改良することによって、少なくとも報酬を調整する個々の手法（(3)(4) 式）と同等な適応力を得られるのではないかと思われる。これは行動選択用の Q 値と状況識別用の Q 値を分けて考えることで可能になると思われるが、単純に考えても空間使用量が倍になる点がデメリットである。一方で狭路問題の場合には、非干渉状況において提案手法による状況の識別がうまく機能しないことが判明した。この原因を究明してより良い状況識別法を構築することも必要となるだろう。

それ以外の今後の課題としては、状況識別の自動化の追求、異質なエージェントからなるマルチエージェント環境における提案手法の有効性の確認、(3) 式における近隣エージェントからの情報の不要化と異なる強化学習法の適用が考えられる [4]。

†3　これは、資源の枯渇の問題などの状況変化が徐々に起こる問題について顕著であると思われる。

References

[1] 出口. 複雑系としての経済学. 社会科学のフロンティア 6. 日科技連, 東京, 2000.

[2] G. Hardin. The Tragedy of the Commons. *Science*, 162:1243–1248, 1968.

[3] S. Mikami and Y. Kakazu. Co-operation of Multiple Agents Through Filtering Payoff. In *Proc. 1st European Workshop on Reinforcement Learning, EWRL-1*, pp. 97–107, Brussels, Belgium, 1994.

[4] 森山, 沼尾. 自己の報酬を操作する学習エージェントの構築. 人工知能学会 第 45 回人工知能基礎論研究会資料 (SIG-FAI-A101), pp. 15–20, 東京都, 2001.

[5] T. G. Spiro and W. M. Stigliani. *Chemistry of the Environment*. Prentice-Hall, Englewood Cliffs, NJ, 1996. （岩田 元彦, 竹下 英一 訳. 地球環境の化学. 学会出版センター, 東京, 2000）.

[6] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[7] C. J. C. H. Watkins and P. Dayan. Technical Note: Q-learning. *Machine Learning*, 8:279–292, 1992.

[8] X. Yao and P. J. Darwen. An Experimental Study of N-Person Iterated Prisoner's Dilemma Games. *Informatica*, 18:435–450, 1994.