3F4-06

各種のゲームで適切な行動を導く自己評価の生成法

Generating Self-Evaluations for Learning in Various Games

森山 甲一[*1] 沼尾 正行[*2]
Koichi Moriyama Masayuki Numao

[*1]東京工業大学大学院 情報理工学研究科 計算工学専攻
Department of Computer Science, Tokyo Institute of Technology

[*2]大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

Most of multi-agent reinforcement learning algorithms aim to converge to a Nash equilibrium, but a Nash equilibrium is not a good result in the prisoner's dilemma (PD). On the other hand, there are several methods aiming to depart from bad Nash equilibria, but they are effective only in PD-like games. In this paper, we construct an agent learning appropriate actions in both PD-like and non-PD-like games through self-evaluations. The agent has two conditions for judging whether the game is like PD or not and two methods which generate self-evaluations according to the judgement. We empirically show that the agents properly learn good actions in both PD-like and non-PD-like iterated games.

1 はじめに

本稿では複数の行動主体（エージェント）が同一の環境中で相互作用を行なうマルチエージェント環境での機械学習、特に強化学習を扱う。マルチエージェント環境における強化学習の研究は既にいくつも行なわれている [4, 1] が、そのほとんどはゲーム理論のナッシュ均衡をもたらすことを目的としている。しかし、囚人のジレンマ [8] に代表されるようなゲームでは、必ずしもナッシュ均衡が最適な解を意味しない。一方で、各エージェントが学習に用いる報酬を操作して、好ましくないナッシュ均衡以外の解をもたらす試みが行なわれている [5, 13]。しかし、これらを囚人のジレンマタイプでないゲームで用いると通常の強化学習よりも結果が悪くなる。

本研究では自己評価に基づく強化学習により、囚人のジレンマタイプのゲームではナッシュ均衡に反してより良い結果を求める非合理的な行動を、それ以外のゲームではナッシュ均衡に相当する合理的な行動を学習するエージェントを構築し、実験により有効性を確認する。

2 自己評価の生成法

2.1 ゲームの分類

本研究では、全てのエージェントの採り得る行動の集合と利得関数が等しい対称なゲームを扱う。本節では対称なゲームを分類する。分類にはゲーム理論のナッシュ均衡とパレート最適の概念を利用する。全てのエージェントの戦略が等しい純粋戦略からなるナッシュ均衡を対称純粋戦略ナッシュ均衡 (symmetric pure strategy Nash equilibrium) と定義し、この対称純粋戦略ナッシュ均衡の集合 S^*_{sp} とパレート最適な戦略の集合 P を用いてゲームを以下の 4 種に分類する。

非干渉状況: $S^*_{sp} \neq \emptyset,\ S^*_{sp} \subseteq P$
全ての対称純粋戦略ナッシュ均衡はパレート最適

泥沼状況: $S^*_{sp} \neq \emptyset,\ S^*_{sp} \cap P = \emptyset$
全ての対称純粋戦略ナッシュ均衡は非パレート最適

選択状況: $S^*_{sp} \neq \emptyset,\ S^*_{sp} \cap P \neq \emptyset,\ S^*_{sp} \not\subset P$
パレート最適と非パレート最適な対称純粋戦略ナッシュ均衡が存在

競合状況: $S^*_{sp} = \emptyset$
対称純粋戦略ナッシュ均衡が存在しない

非干渉状況とは、個々の合理的な行動によりもたらされる均衡が全てのエージェントにとって平等かつ全体にとっても望ましいという意味で、各エージェントの勝手な行動が許される、特別な制約を考慮しなくても良い状況である。泥沼状況は、個々の合理的な行動によりもたらされる、全てのエージェントにとって平等となるどのような均衡も全体として望ましくないという意味で、個々が合理的に振る舞えば振る舞うほど社会全体が苦しく、もがけばもがくほど泥沼にはまり込んでしまうという状況である。選択状況とは、複数の対称純粋戦略ナッシュ均衡を持つがパレート最適なものとそうでないものがあるために、望ましい均衡を選択するにはどうすれば良いかという問題が生じる状況である。競合状況とは、まず勝ち負けの利得得失が等しいジャンケンなどのゼロ和ゲームのように、純粋戦略ナッシュ均衡が存在しない場合がある。そもそも、ゼロ和（定和）ゲームでは自分の利益は他者の損失であるためパレート最適の概念が意味を持たない。更に道を譲る場合などのように両者が同じ行動をしても困るが、先に通るほうが有利となる状況が挙げられる。この場合には一方が待って他方が進むことがナッシュ均衡かつパレート最適であるが、有利な方をどちらが選択するかで競合が生じる。

2.2 自己評価の生成

本研究では、学習の際に外部報酬から自己評価を生成して学習に利用することを提案する。具体的には、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 の自己評価 $r'_{i,t+1}$ とし、これを Q 学習 [12] に適用する。なお、以下では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。

$$r'_{t+1} \triangleq r_{t+1} + \lambda_{t+1}. \tag{1}$$

以下では「近隣報酬 (neighbors' rewards)」と「報酬差分 (difference of rewards)」と呼ぶ 2 種の λ を提案する。前者は泥沼

連絡先: 〒 152–8552 東京都目黒区大岡山 2–12–1, 東京工業大学大学院 情報理工学研究科 計算工学専攻 森研究室, 森山 甲一. 03(5734)2510. moriyama@mori.cs.titech.ac.jp

状況で、後者は非干渉状況で有効に働く。

$$\lambda_{t+1}^{NR} \triangleq \sum_{A_k \in N_i \setminus A_i} r_{k,t+1} \tag{2}$$

$$\lambda_{t+1}^{DR} \triangleq r_{t+1} - r_t \tag{3}$$

(2) 式の $N_i \setminus A_i$ 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。「近隣報酬」は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状況において有効である。「報酬差分」は前回の行動による報酬と今回のそれの差異を強調することにより、通常の強化学習のみの場合よりも報酬の変化に敏感になるので、エージェントの合理性が増すと思われる。

「近隣報酬」λ^{NR}・「報酬差分」λ^{DR} はそれぞれ対応する状況では有効であるが、泥沼状況と非干渉状況は対立するものであるため、対応するものとは逆の状況に用いるとかえって悪影響を与える。従って、エージェントはゲームに応じて λ を適切に選択しなければならない。すなわち、ゲームが泥沼状況の際には「近隣報酬」λ^{NR} を、非干渉状況の際には「報酬差分」λ^{DR} を λ として自動的に選択する能力をエージェントは保持しなくてはならない。

そこで、λ^{NR}, λ^{DR} のうち適切なものを使用して学習するために各エージェントが現在の状況をどうやって判断するかを考える。例えば、各エージェントの持つ環境状況モデルについて

条件 1 現在どのような行動を選択しても将来の見込みが無い

条件 2 現状が既に獲得しているモデルより悪い

が満たされる場合、そのような現状を招いた利己的な行動を抑えるべきであると考えられる。各エージェントの持つ環境状況モデルは Q 関数で表されるため、条件を数式で表現すると

$$Q_{t-1}(s_t, a) < 0 \quad \text{for all } a. \tag{4}$$

$$r_{t+1} < Q_{t-1}(s_t, a_t) - \gamma \max_a Q_{t-1}(s_{t+1}, a) \tag{5}$$

となり、(4) 式が条件 1、(5) 式が条件 2 に対応する。(4) 式は現在どのような行動を採っても正の報酬が得られる見込みがないことを意味し、(5) 式は実際の報酬が、時間差分誤差 (TD error) [11] が 0、すなわち学習が収束した場合に計算される報酬の予測値よりも小さいことを意味する。そして、少なくとも 1 つが成立する場合に環境が泥沼状況であると見なして「近隣報酬」λ^{NR} を、どちらも不成立の場合に「報酬差分」λ^{DR} を用いて学習を行なうことにする。以下では (4) 式を条件 1、(5) 式を条件 2 と呼ぶ。

ところが、本研究では (1) 式により調整した自己評価 r' を学習に用いるため、時間差分誤差の式から求められる条件 2 の左辺は獲得報酬 r ではなく自己評価 r' でなくてはならない。しかし、この条件による状況判定の結果から r' が計算されるため、ここで r' を用いることは出来ない。そこで、以下では 3 種の代替手法を考案する。まず、条件 2 の左辺に来るべき r' を r で代用する手法が考えられる [6]。この場合に条件判定に利用される Q 関数は r' によって学習されるエージェントの行動を決定するものなので、以下では Q^{act} と称し、この方法で状況を判定する手法を「自動選択 AA (auto-select AA)」と呼ぶ。一方で、通常の r による Q 関数を用いて状況を判定し、それから r' によって Q^{act} を学習するようにすれば、この問題は解消する。以下ではこの判定用 Q 関数を Q^{recog} と称し、それによって状況を判定する手法を「自動選択 RR (auto-select RR)」と呼ぶ。更に、条件 1 はこの問題から自由であるため、条件 1 については行動選択用の Q^{act} を、条件 2 については状況判定用の Q^{recog} を利用することも可能である。これを「自動選択 AR (auto-select AR)」と呼ぶ。これらをまとめると表 1 となり、次節で実験的に比較する。

表 1: 自動選択 AA・AR・RR：判定に用いる Q 関数

自動選択	条件 1	条件 2
AA	Q^{act}	Q^{act}
AR	Q^{act}	Q^{recog}
RR	Q^{recog}	Q^{recog}

3 実験

本節では 2.2 節で提案した学習手法を用いて、繰り返しゲームを用いて同種の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実験を行なう。以下では 4 人 2 行動ゲームを扱う。実験に用いたエージェントは、行動選択をランダムに行なう「ランダム (RD)」、通常の強化学習を行なう「通常 (NM)」、2.2 節で導入した「近隣報酬 (NR)」「報酬差分 (DR)」「自動選択 AA (AS-AA)」「自動選択 AR (AS-AR)」「自動選択 RR (AS-RR)」の合計 7 種である。実験では 25 回の試行を行ない、各試行において Q 関数は 0 で初期化する。Q 学習のパラメータは学習率 $\alpha = 0.5$・割引率 $\gamma = 0.5$ であり、温度 $T = 1$ の softmax 法 [11] により行動を選択するとする。

2.1 節の 4 種の状況それぞれに相当するゲームを用いて、4 台のエージェントによる実験を行なう。各エージェントは 2 種の行動 C, D を選択するものとする。各エージェントは利得関数を知らず、繰り返される行動と報酬の関係から採るべき行動を Q 関数の形で学習する。行動選択と報酬獲得をまとめて 1 サイクルと称し、サイクルごとに学習を行なう。「近隣報酬」で用いられる近隣エージェント集合 N_i はゲームに参加する全てのエージェントとする。従って、各エージェントはゲームに参加する全てのエージェントが得た報酬を知ることが出来る。各エージェントはゲームの参加エージェントの行動組合せを状態変数として Q 学習を行なう。つまり、各エージェントは $2^4 = 16$ 種類の状態に対して学習する。

非干渉状況（表 2a）は、他者の行動に関係なく D を選択すると常に 2 の利得が得られ、C を選択すると常に –2 の利得が得られるゲームである。泥沼状況（表 2b）は囚人のジレンマと同じく D が常に C よりも大きな利得をもたらすが、全員が D を選択すると最悪の結果になるゲームである。両者とも唯一のナッシュ均衡は全員 D であるが、前者はそれがパレート最適である一方、後者は異なる。選択状況（表 2c）では全員の行動を一致させることを目的とするが、一致した行動により利得が異なる。対称純粋戦略ナッシュ均衡は全員 C と全員 D であるが、前者のみパレート最適である。競合状況（表 2d）は志願者のジレンマとして知られるゲームで、少なくとも 1 人が利得の少ない C を選択すれば良いが、全員が D を選択すれば最悪の結果となるゲームである。このゲームは対称純粋戦略ナッシュ均衡を持たない。

エージェント 4 台の 3000 サイクル時点における報酬和の 25 試行平均を図 1 に、25 試行の 2996–3000 サイクルにおける合計 125 回の行動のうち、「ランダム」を除く手法のナッシュ均

表 2: 4 人 2 行動ゲーム：各行の $\#D$ は他者のうち D を選択した人数を表し、C, D の各行は利得を表す

(a) 非干渉状況

$\#D$	0	1	2	3
C	−2	−2	−2	−2
D	2	2	2	2

(b) 泥沼状況

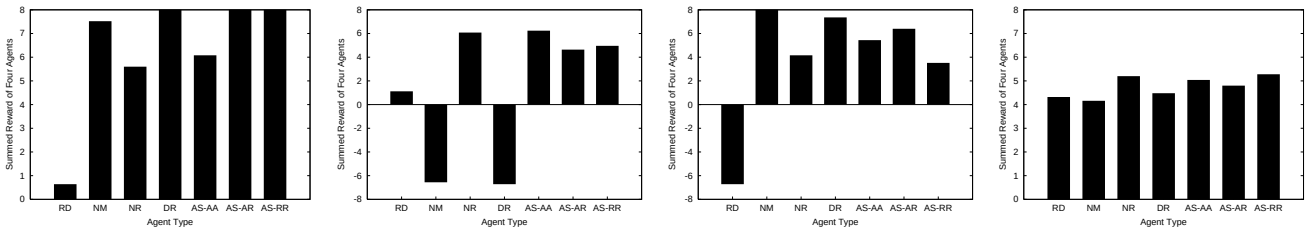
$\#D$	0	1	2	3
C	2	0	−2	−4
D	4	2	0	−2

(c) 選択状況

$\#D$	0	1	2	3
C	2	−2	−2	−2
D	−2	−2	−2	0

(d) 競合状況

$\#D$	0	1	2	3
C	0	0	0	0
D	2	2	2	−2



(a) 非干渉状況　(b) 泥沼状況　(c) 選択状況　(d) 競合状況

図 1: 4 人 2 行動ゲームの結果：エージェント 4 台の 3000 サイクル時点における報酬和の 25 試行平均

衡とパレート最適な行動組合せの割合を表 3 に示す。

図と表、統計的検定から以下のことが判明した。なお、以下の記述では「ランダム」を除いている。

- 非干渉状況では「通常」「報酬差分」「自動選択 AR」「自動選択 RR」の報酬和がほぼ最大値の 8 である。「近隣報酬」「自動選択 AA」はこれらより明らかに劣る。行動割合からも「近隣報酬」「自動選択 AA」以外は 9 割以上の割合でナッシュ均衡かつパレート最適な全員 D を選択していることが分かる。

- 逆に、泥沼状況では「通常」「報酬差分」は悪い結果となるが、「近隣報酬」「自動選択 AA」は良好な結果を得ている。「自動選択 AR」「自動選択 RR」はこちらでも同様に良好である。行動割合からも「通常」「報酬差分」がナッシュ均衡である全員 D が多く、残りの手法はパレート最適な組合せが多いことが分かる。

- 選択状況では全ての手法がナッシュ均衡にほぼ収束している。「通常」「報酬差分」の報酬和はほぼ最大値 8 となっている。「自動選択 AR」は「近隣報酬」より良いが、残りの「自動選択」は「近隣報酬」と同様である。

- 競合状況では全ての手法がほぼ同様の結果となっている。

4 考察

実験結果は、提案手法である「自動選択」、特に「自動選択 AR」「自動選択 RR」が正反対である非干渉状況と泥沼状況の両方でパレート最適な解を導く行動を学習できることを表しており、提案手法が本研究の目的を満たすことが示された。しかし、選択状況で結果の良くない「近隣報酬」と差があまり見られないことは問題であり、その修正は今後の課題である。

どちらのゲームでも、非干渉状況において「自動選択 AA」は過剰に「近隣報酬」を選択する傾向が見られる。これは、「自動選択 AA」が「近隣報酬」を選ぶと自己評価 r' が大きくなるために Q^{act} が大きくなることが理由として考えられる。もし Q^{act} が $r/(1-\gamma) = 2r$ より大きいならば常に「自動選択 AA」の条件 2 が満たされるため、「近隣報酬」を選択する頻度がより高くなる。紙面の都合により詳細は省くが、3 種の「自動選択」について条件の満足回数を調べると「自動選択 AA」の条件 2 の満足回数が常に多く、この考察が正しいことが分かる。

「近隣報酬」は非干渉状況において「通常」よりも結果が悪い。「近隣報酬」は自分の報酬も計算に入れているため、自分の報酬が大きくなればなるほど自己評価も大きくなり、結果として非干渉状況にも適するはずである。しかしそうなってはいない。理由としては、もし自分が “悪い” 行動をしたとしても、近隣の他者の報酬が大きければ自己評価が大きくなることが挙げられる。これにより、学習に際して “良い” 行動と “悪い” 行動の差があいまいになるため、結果として行動にランダム性が生じてしまうのである。

本研究では、各種のゲームにおける適切な行動を自己評価を用いた Q 学習によって学習するエージェントを構築することを目的とした。筆者らの知る限り、強化学習の分野で同様の研究は存在しない。確かに、利得関数を操作する研究はいくつかあるが [5, 13]、これらはいずれも単一の状況に対応するものである。従って、もしこれらの手法をエージェントの構築の際に用いるならば、我々人間があらかじめ状況を判別しなくてはならない。一方で、もし本研究の手法を用いるならば、状況を判別するという負担を減らすことが出来るだろう。

本研究は適切な学習のための状況判定条件すなわちメタ規則を導入した。それは泥沼状況について判定するものであり、実験により有効性は確認された。しかしこれが最適か否かは不明である。そのため、異なる条件が要求されることもあると思われるが、それは簡単に導かれるものではない。従って、そのような条件を自動的に設定する機構が要求されるだろう。メタ学習や遺伝的アルゴリズムによる探索などを利用することが考えられる。

5 関連研究

強化学習の分野では報酬を操作する手法は数少ないが、遺伝的アルゴリズム (GA) の分野ではいくつか行なわれている [2, 7]。また、石田ら [3] は「価値観」と称するメタ規則を GA で獲得することを試みている。阪口ら [10] は Q 関数を状況判定に用いる研究を行なっているが、単一エージェント環境を対象とする。

表 3: 4 人 2 行動ゲーム：2996–3000 サイクルの行動組合せ割合。各表題のカッコ内左側は純粋戦略ナッシュ均衡、右側はパレート最適な行動組合せ

(a) 非干渉状況（全員 D　全員 D）

学習手法	ナッシュ均衡	パレート最適
通常	92.8%	92.8%
近隣報酬	38.4%	38.4%
報酬差分	100 %	100 %
自動選択 AA	54.4%	54.4%
自動選択 AR	97.6%	97.6%
自動選択 RR	97.6%	97.6%

(b) 泥沼状況（全員 D　全員 C または 1 台のみ D）

学習手法	ナッシュ均衡	パレート最適
通常	61.6%	1.6%
近隣報酬	0 %	100 %
報酬差分	83.2%	0 %
自動選択 AA	0 %	96 %
自動選択 AR	0 %	86.4%
自動選択 RR	0.8%	83.2%

(c) 選択状況（全員 C または全員 D　全員 C）

学習手法	ナッシュ均衡	パレート最適
通常	96.8%	96 %
近隣報酬	99.2%	55.2%
報酬差分	100 %	93.6%
自動選択 AA	100 %	69.6%
自動選択 AR	98.4%	79.2%
自動選択 RR	97.6%	53.6%

(d) 競合状況（1 台のみ C　1 台のみ C）

学習手法	ナッシュ均衡	パレート最適
通常	40.8%	40.8%
近隣報酬	64.8%	64.8%
報酬差分	70.4%	70.4%
自動選択 AA	61.6%	61.6%
自動選択 AR	46.4%	46.4%
自動選択 RR	64.8%	64.8%

ゲーム理論では囚人のジレンマにおける合理的プレイヤーは互いに裏切ることになっているが、実際の人間を対象とした実験では結果が異なることが知られている。Rilling ら [9] は、囚人のジレンマゲームを行なっている 36 人の女性の脳の働きを fMRI で観察した結果、脳内の報酬処理関係部位が活動することを報告し、脳内で報酬が生成されることにより協調行動が生じると結論づけている。これは、実際の人間の脳内処理と本研究の提案手法の類似性を示すものである。

6　まとめ

本研究は、複数のエージェントが同一の環境で行動するマルチエージェント環境において、個々の合理性に基づくナッシュ均衡とパレート最適解が共有部分を持たない囚人のジレンマのような状況ではナッシュ均衡ではなくパレート最適な行動組合せを学習し、それ以外の状況では個々の合理性を生かす学習を行なう強化学習エージェントの構築を目的とした。

そのため、まず 2 節でゲーム理論におけるナッシュ均衡とパレート最適の概念からマルチエージェント環境を 4 種の状況に分類した。そして、Q 関数と現在の報酬の関係から環境が泥沼状況であると判定する条件を 2 つ導入し、その条件によって自己評価の生成法を切替える「自動選択 AA」「自動選択 AR」「自動選択 RR」の 3 種のエージェントを構築した。構築したエージェントを用いて、3 節では非干渉状況・泥沼状況・選択状況・競合状況の 4 人 2 行動ゲーム 4 種について実験を行なった。提案手法、特に「自動選択 AR」「自動選択 RR」は非干渉状況と泥沼状況の両方で良好な結果を得る一方、「通常」「近隣報酬」「報酬差分」はいずれもどちらかの状況でパレート最適な行動組合せが少なくなることを示した。

今後の課題としては 4 節で述べた状況判定条件の自動獲得などに加え、理論的考察、エージェントが非対称な場合への拡張、異種エージェントを導入した場合の有効性の確認、近隣エージェントからの情報の不要化、異なる強化学習法の適用と現実の問題への応用が考えられる。

参考文献

[1] J. Hu and M. P. Wellman. Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm. In *Proc. 15th International Conference on Machine Learning, ICML'98*, pp. 242–250, Madison, Wisconsin, U.S.A., 1998.

[2] 石渕, 中理, 中島. 空間型繰返し囚人のジレンマゲームにおける隣接プレーヤ間での信頼関係のモデル化. 電子情報通信学会論文誌, J83-D-I(10):1097–1108, 2000.

[3] 石田, 横井, 嘉数. 競争社会系における価値観群の創発. 人工知能学会誌, 15(5):896–903, 2000.

[4] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. 11th International Conference on Machine Learning, ML'94*, pp. 157–163, New Brunswick, New Jersey, U.S.A., 1994.

[5] 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.

[6] 森山, 沼尾. 環境状況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. 人工知能学会論文誌, 17(6):676–683, 2002.

[7] M. Mundhe and S. Sen. Evolving agent societies that avoid social dilemmas. In *Proc. Genetic and Evolutionary Computation Conference, GECCO-2000*, pp. 809–816, Las Vegas, Nevada, U.S.A., 2000.

[8] W. Poundstone. *Prisoner's Dilemma*. Doubleday, New York, 1992.

[9] J. K. Rilling, D. A. Gutman, T. R. Zeh, G. Pagnoni, G. S. Berns, and C. D. Kilts. A Neural Basis for Social Cooperation. *Neuron*, 35:395–405, 2002.

[10] 阪口, 高野. 環境変化への適応と文脈切替え. 第 16 回生体・生理工学シンポジウム論文集, pp. 157–160, 神奈川県相模原市, 2001.

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

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

[13] D. H. Wolpert and K. Tumer. Collective Intelligence, Data Routing and Braess' Paradox. *Journal of Artificial Intelligence Research*, 16:359–387, 2002.