知 能 と 複 雑 系 128−24
（２００２．５．２４）

# 社会的ジレンマを解決する内部評価生成法の分析

森山 甲一　沼尾 正行

東京工業大学 情報理工学研究科 計算工学専攻

koichi@nm.cs.titech.ac.jp, numao@cs.titech.ac.jp

**概 要**

筆者らは社会的ジレンマを解決するために、内部評価を生成する能力と社会の状況を識別する能力を与えたエージェントを提案した。そしてそのエージェントが同質なエージェントからなるマルチエージェント社会の状況に応じて適切な行動を行なうことを確認した。しかし、その際の評価基準がエージェント社会全体の報酬和の平均のみであったため、結果の有意性や個々のエージェントの挙動などが不明瞭になっている。そこで、本稿では提案した内部評価生成法について、その挙動の詳細な分析を行なう。

# Analysis of an Internal Reward Generation Method for Solving Social Dilemmas

Koichi Moriyama and Masayuki Numao

Department of Computer Science,
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology.

koichi@nm.cs.titech.ac.jp, numao@cs.titech.ac.jp

**Abstract**

The authors previously proposed an internal reward generation method for solving social dilemmas. We observed that an agent having the method was able to cope with dilemma and non-dilemma situations of a Multi-Agent environment composed of homogeneous agents. However, the result we observed was only summed rewards in the environment and we paid no attention to individual effects of the method. Thus, in this paper, we analyze the result in detail.

## 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、エージェントが複数存在するマルチエージェント社会では社会的ジレンマの問題がある。

ゴミの分別の問題を考えてみよう。ある街ではゴミの焼却の際に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ことが求められており、ゴミが分別されずに焼却場へ運ばれた場合には税金を使って人を雇い分別させるとする。この場合、住民がきちんと自分のゴミを分別すれば税支出は最小になる。しかしゴミの分別は面倒で、きちんと分別しても自分が払う税金は減らないため、実際には分別をしない住民も現れる。これが社会的ジレンマの一例である。

このような状況で行動する自律エージェントを構築することを考える。自分のゴミの分別に正のコストを見積もると、見返りがないために分別しないことが合理的な行動となる。しかしこれは社会的ジレンマを引き起こすため、個々のエージェントは非合理的にゴミの分別を行なわなくてはならない。これを学習により解決するためには、エージェントに与えられる外部評価から内部評価を生成し利用することで非合理性を実現することが考えられる。では、常にゴミの分別を行なうように内部評価生成法を設計すれば良いのだろうか。

ある時点でこの街に新しく高温でゴミを燃やせる焼却炉が出来たとする。ダイオキシンは低温（約400°C）でゴミを燃やすことで発生するため、新しい炉では発生しない。すなわち、もはや可燃ゴミと不燃ゴミの分別は必要ない。しかし、予め設計されたエージェントでは変化に対応できずゴミの分別を続けることとなり非効率である。つまり、エージェントが現実社会で適切に行動するためにはジレンマ克服法を予め固定するのではなく、状況に応じてそれを変化させなくてはならない。

これらの条件の下で筆者らは、外部評価から内部評価を生成する能力に加えて社会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生成法を切替えるエージェントを構築した [1, 2]。そしてそれが同質なエージェントからなるマルチエージェント社会において適切にジレンマ状況を克服する一方、非ジレンマ状況においても比較的良好な結果を得ることを示した [1]。しかし、それらの実験では評価基準としてエージェント社会の報酬和のみを利用しているため、行動選択のランダム性に起因する有意性の問題や手法の欠点、各エージェントの挙動などが不明瞭になっている。そこで、本稿では文献 [1] に示した実験の結果についてより詳細な分析を試みる。

## 2 提案手法

本研究の目的は、各エージェントが外部評価を元に内部評価を生成して学習することにより、個々の利益追求行動がマルチエージェント社会において不適切な場合に各エージェントが自らその行動を制限することである。以下本節では、筆者らがこれまでに提案した手法 [1, 2] について簡単に紹介する。詳細はこれらの文献を参照されたい。

**表 1**　各状況における望ましい行動組合せ（○）

(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 | ○ | |

本研究ではまず、個々のエージェントの行動とそのエージェント社会における好ましさの関係に着目し、社会を「非干渉状況」「泥沼状況」「競合状況」の 3 種の状況に分類する（**表 1**）。

本研究ではエージェントの学習手法として、強化学習のうちの代表的手法である Q-learning [3] を用いる。これは状態 $s$ と行動 $a$ の組合せに伴う報酬見込み $Q(s,a)$ を逐次更新し、行動選択に利用するものである。Q-learning はその性質から、たとえ個体の利益追求行動がマルチエージェント社会では不適切な場合でも自分の利益を最大化するような学習を行なう。そのため個々が常に利益を追求することが社会的に好ましくない泥沼状況や競合状況では、各エージェントはどのように行動を学習すべきだろうか。

その 1 つの解として筆者らは、外部評価から内部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 $A_i$ の内部評

価 $r'_{i,t+1}$ とし、これを強化信号として Q-learning に適用する。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。

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

泥沼状況に有効な $\lambda_{t+1}$ として「近隣報酬」

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

を、競合状況に有効な $\lambda_{t+1}$ として「報酬差分」

$$\lambda_{t+1} = r_{t+1} - r_t \tag{3}$$

を用いる。そして本研究ではこれらの $\lambda_{t+1}$ のうち適切なものを使用して学習するために、状況識別仮定として以下の 2 つの式[1]

$$Q(s_t, a) < 0 \quad \text{for all } a \in \mathcal{A}_t. \tag{4}$$

$$r_{t+1} < Q(s_t, a_t) - \gamma \max_{a \in \mathcal{A}_{t+1}} Q(s_{t+1}, a). \tag{5}$$

を導入し、少なくともどちらか 1 つが成立する場合に泥沼状況と見なして (2) 式を、どちらも不成立の場合に競合状況と見なして (3) 式を用いて学習を行なう「自動選択」を提案する。

# 3 実験

提案手法を用いたエージェントによるマルチエージェント社会とそれを用いないエージェントによる社会を比較するために、非干渉状況と泥沼状況を再現した「共有地の悲劇ゲーム」と競合状況を再現した「狭路ゲーム」を用いて実験を行なう。学習は学習率 0.5 割引率 0.5 の Q-learning により行なわれ、行動選択にはボルツマン選択を用いる。また、内部評価を用いないものを「通常」と表す。

## 3.1 実験概要

以下では共有地の悲劇ゲームと狭路ゲームの説明を行なう。これらのゲームは文献 [1] に述べられているため、詳細は文献を参照されたい。

[1] $s_t, s_{t+1}$ はそれぞれ遷移前後の状態、$a_t$ は状態遷移を起こした行動、$\mathcal{A}_t$ は $s_t$ で実行可能な行動の集合、$\gamma$ は割引率を表す。

**共有地の悲劇ゲーム** 10 台のエージェントからなる社会を考える。各エージェントは利己的・協力的・利他的の 3 種の行動を選択する。エージェント $A_i$ の行動 $a_i$ には基本報酬 $r(a_i)$ と社会に対するコスト $c(a_i)$ が付随するものとし、基本報酬は利己的・協力的・利他的な行動に対して $3, 1, -3$ の各値を、コストはそれぞれ $+c, \pm 0, -c$ の値を取るものとする。全てのエージェントが行動を選択すると、各エージェントは報酬 $r_i \triangleq r(a_i) - \sum_j c(a_j)$ を獲得する。10 台のエージェント社会では $c > 1/5$ で各々が利己的な行動を採る場合より全体が協力的な行動を採る場合に各々の報酬 $r_i$ が大きくなり、$c > 2/5$ では全体が利他的な行動を採る場合に報酬 $r_i$ が最大となる。しかし、この場合でも個々の立場では利己的行動が常に最大の報酬をもたらすため、泥沼状況となる。以下では $c = 1$ としたものを「共有地（泥沼）ゲーム」、$c = 0$ としたものを「共有地（非干渉）ゲーム」と称する。これらはそれぞれ前述のゴミ分別問題における新焼却炉建設前後に相当する。1 回の行動提示から報酬獲得までを 1 サイクルと称し、各エージェントの状態は 4 台のエージェントの行動の組合せとする。

**狭路ゲーム** 路上駐車があって狭くなっている道路を想定する。その両端に同時に車が現れた場合に、各々の車は「進む」「待つ」のどちらかを選択する。両者が同じ行動を示した場合には状況は変化せず、異なる行動を示した場合には「進む」を提示した車が通過に成功する。この場合に両者が出来るだけ早く狭い部分を通過することが目的である。10 台のエージェントをランダムに 2 台ずつ 5 組に分けて実験を行なう。各エージェントの状態は相手の ID と自分の組の残っている車の台数により計算される。各エージェントは行動提示の際、通過に成功したら報酬 $r_i = 1$ を、それ以外では $r_i = -0.5$ を獲得する。両者が通過に成功するか、両者の行動の組合せと同数である $2^2 = 4$ 回の行動提示をしたらその組は終了とする。組分けから全ての組が終了するまでを 1 サイクルとする[2]。

[2] 各エージェントにとって、1 サイクルに行動提示・学習が 1 回だけとは限らないことに注意。

**表 2**　社会全体の報酬和 $R(3000)$ の基本統計量：10 回の実験による平均値・標準偏差・最大値・最小値

(a) 共有地（泥沼）ゲーム

| 手法 | 平均 | 偏差 | 最大 | 最小 |
|---|---|---|---|---|
| 通常 | −35.2 | 16.66 | −12 | −62 |
| 近隣報酬 | 12.6 | 11.04 | 32 | −2 |
| 報酬差分 | −41.2 | 13.54 | −14 | −62 |
| 自動選択 | 14.8 | 5.98 | 24 | 6 |

(b) 共有地（非干渉）ゲーム

| 手法 | 平均 | 偏差 | 最大 | 最小 |
|---|---|---|---|---|
| 通常 | 28.2 | 1.14 | 30 | 26 |
| 近隣報酬 | 20 | 2.49 | 26 | 18 |
| 報酬差分 | 30 | 0 | 30 | 30 |
| 自動選択 | 26.6 | 3.27 | 30 | 22 |

(c) 狭路ゲーム

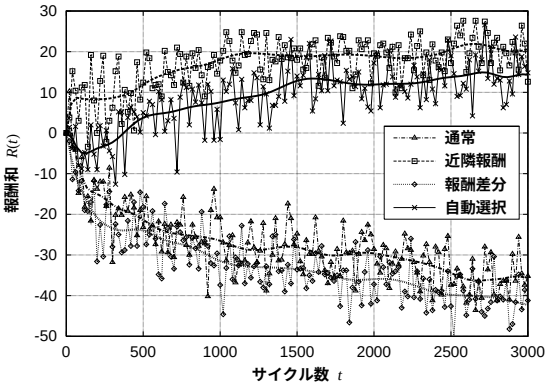
| 手法 | 平均 | 偏差 | 最大 | 最小 |
|---|---|---|---|---|
| 通常 | 1.3 | 2.37 | 4.5 | −1.5 |
| 近隣報酬 | 2.75 | 3.78 | 6.5 | −6.5 |
| 報酬差分 | 2.65 | 2.65 | 5.5 | −3 |
| 自動選択 | 4.55 | 2.06 | 7 | 0 |

## 3.2　実験結果の分析

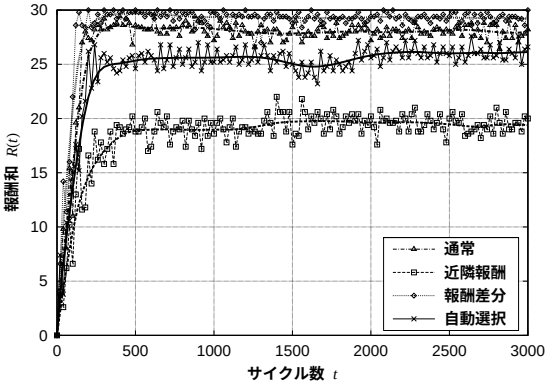
**社会報酬和の基本統計量と検定**　文献 [1] ではマルチエージェント社会全体の報酬和 $R(t) \triangleq \sum_i r_{i,t+1}$ を評価の対象とした。10 回の試行によって得られた平均値を 20 サイクルごとにプロットし **図 1** に示す。以下では $R(t)$ の基本的な統計量を示し、それらの情報を用いて統計的検定を行なう。

**表 2** に、10 回の試行から得られた $t = 3000$ における社会全体の報酬和 $R(3000)$ の平均・標準偏差・最大値・最小値を示す。これによると、共有地（泥沼）ゲームと狭路ゲームの標準偏差が大きいことが判る。

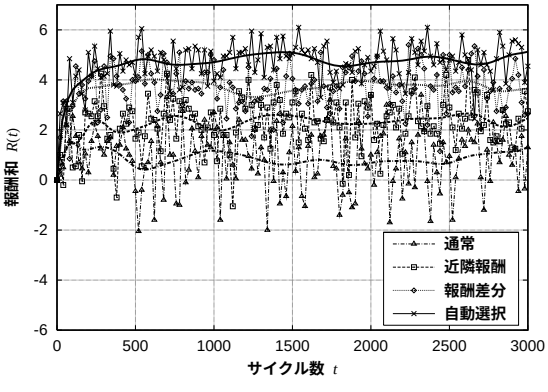
続いて、自動選択による結果が他の結果とどれだけ異なるのかを調べるために、$R(3000)$ の統計情報を基に $t$ 分布による両側検定を行なった。こ




(a) 共有地（泥沼）ゲーム




(b) 共有地（非干渉）ゲーム




(c) 狭路ゲーム

**図 1**　実験結果：10 回の試行による社会全体の報酬和 $R(t)$ の平均値を 20 サイクルごとにプロット

**表 3**　$R(3000)$ について自動選択を対立仮説とした場合の検定結果（○：有意差あり ×：有意差なし）

(a) 共有地（泥沼）ゲーム

| 帰無仮説 | 有意水準 5% | 有意水準 1% |
|---|---|---|
| 通常 | ○ | ○ |
| 近隣報酬 | × | × |
| 報酬差分 | ○ | ○ |

(b) 共有地（非干渉）ゲーム

| 帰無仮説 | 有意水準 5% | 有意水準 1% |
|---|---|---|
| 通常 | × | × |
| 近隣報酬 | ○ | ○ |
| 報酬差分 | ○ | ○ |

(c) 狭路ゲーム

| 帰無仮説 | 有意水準 5% | 有意水準 1% |
|---|---|---|
| 通常 | ○ | ○ |
| 近隣報酬 | ○ | × |
| 報酬差分 | ○ | × |

れはそれぞれの帰無仮説について得られた平均値を母平均と見なして自由度 9 の $t$ 分布で調査したものである。その結果を **表 3** に示す。共有地（泥沼）ゲームでは、**図 1** で全体として自動選択より結果が良い「近隣報酬」とは有意水準 5%でも差が見られず、悪い「通常」とは有意水準 1%でも差が見られる。共有地（非干渉）ゲームでも同様に、結果が良い「通常」とは有意水準 5%でも差が見られず、悪い「近隣報酬」とは有意水準 1%でも差が見られる。しかし最良の「報酬差分」とは有意水準 1%でも差が見られる。また、狭路ゲームでは自動選択が最良の結果となっていたが、それに続く報酬差分とは有意水準 1%では差が認められないものの 5%では差が認められる。近隣報酬とも有意水準 1%では差が認められない。

**自動選択の状況認識**　自動選択エージェントが状況認識に用いる (4)(5) 式の妥当性を確認するために、実験中に各ゲームで実際に行なわれた状況認識のうち泥沼状況と認識された回数と割合を **表 4** に示す。狭路ゲームで通過に成功した場合とその他の場合の状況認識についても内数で示す。

**表 4**　自動選択エージェントが泥沼状況と認識した回数と割合

| ゲーム | 回数 | 割合 |
|---|---|---|
| 共有地（泥沼） | 2061 / 3000 | 68.7% |
| 共有地（非干渉） | 2579 / 3000 | 86.0% |
| 狭路 | 4333 / 5922 | 73.2% |
| （通過成功） | 1768 / 2881 | 61.4% |
| （その他） | 2565 / 3041 | 84.3% |

表から読み取れるのは、全体的に泥沼状況と認識する割合が高いことである。特に競合状況と認識されるべき共有地（非干渉）ゲームにおいて割合が最大となっている。これは設計上好ましくない点であるため今後の改良が必要である。また、狭路ゲームについて通過成功とその他の 2 つのケースに分けて見ると、成功時に泥沼状況と認識する割合が小さく、それ以外では泥沼状況と認識する割合が大きいことが注目に値する。

**学習結果の個体差**　学習の結果として自動選択エージェント間に個体差が生じるのか否かを調査するために、$t = 1$ から 3000 までに各エージェントが獲得した報酬の和 $R_i \triangleq \sum_{t=1}^{3000} r_{i,t+1}$ を計算し、その差が最大になった試行の結果を **表 5** に示す。表からいずれのゲームにおいても少なからぬ個体差が生じていることが判る。全てのゲームで 3000 サイクル終了時には個体の状態遷移が循環しており、それぞれの状態に対応するエージェントの学習結果をまとめると以下のようになる。但し $\max_i R_i$ を獲得したエージェントを「最大エージェント」、$\min_i R_i$ を獲得したエージェントを「最小エージェント」と称する。

- 共有地（泥沼）：最大エージェントは利他的な行動を抑制しており、最小エージェントは利他的な行動のみを選択している。

**表 5**　自動選択エージェントに最大の個体差が生じた試行の結果：個体の報酬和 $R_i$ の最大値・最小値・差

| ゲーム | 最大 | 最小 | 差 |
|---|---|---|---|
| 共有地（泥沼） | 5765 | 1325 | 4440 |
| 共有地（非干渉） | 8686 | −8384 | 17070 |
| 狭路 | 1708.5 | 815.5 | 893 |

- 共有地（非干渉）: 最小エージェントのみが唯一利他的な行動を選択し、その他のエージェントは全て利己的な行動を選択している。
- 狭路: 最大・最小エージェント共に相手とは異なる行動を選択するように学習がされているが、最大エージェントの方が最小エージェントよりも「進む」を選択する場合が多い。

# 4　考察

本節では前節の分析結果について考察する。

まず、**表 2** から判ることとして共有地（泥沼）ゲームと狭路ゲームの標準偏差が大きいことが挙げられる。前者はゲームの性質として、ある 1 台のエージェントが行動を利己的から協力的に変更すると $R(t)$ は 8 増加、協力的から利他的に変更すると 6 増加するというように 1 台のエージェントの行動変化による $R(t)$ の増減幅が大きい点が要因と考えられる。また後者についてはある時刻 $t$ における対戦相手が試行ごとにランダムに決まることが一因と考えられる。

続いて **表 3** にまとめられた検定結果についてである。本研究で提案している自動選択手法は、共有地（泥沼）ゲームでは比較手法中最良の結果をもたらす近隣報酬手法と有意差が見られず、通常の強化学習とは有意差が認められ、共有地（非干渉）ゲームではより良い結果が得られている通常の強化学習と有意差が無く、最悪の結果となっている近隣報酬手法とは有意差が認められている。これらは提案の趣旨からすると好ましい。しかし、前述した通り共有地（非干渉）ゲームにおける状況の誤認識の問題があり、また最良の報酬差分手法とは有意差が見られる。そのため特にこの状況に対して今後の改良が必要である。一方狭路ゲームでは、自動選択手法の次に良い結果を得ている報酬差分手法と若干ながら有意差が認められている。これは **表 4** に見られる通り、不通過時の学習に近隣報酬手法によって対戦相手の報酬を加味することにより、道を譲る行動も学習していることを意味するのではないかと思われる。

最後に **表 4** の状況識別割合である。全体として状況が泥沼であると判定されやすい要因は、2 つの判定条件 (4)(5) 式の少なくとも 1 つが満たされれば泥沼状況と判定するために学習初期ではそう認識されやすく、すると学習に対する他者の報酬の影響が大きいために (5) 式の右辺が大きくなる傾向があり、結果として以後も泥沼状況と認識される悪循環が生じているためと考えられる。

# 5　まとめ

本稿では筆者らが文献 [1] で発表した実験結果をより詳細に分析することを試みた。特に統計的検定を行なうことによって提案手法の効果が示されたのではないかと思われる。しかし、共有地（非干渉）ゲームにおいて状況認識が正しく働いていないなどの欠点も新たに判明した。そこでこれらの知見を元にさらなる手法の改良が必要となろう。また、文献 [2] で筆者らが発表した動的に状況が変化する場合の実験結果の分析も併せて行なう必要があるだろう。

# References


[1] 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.

[2] 森山, 沼尾. 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. 日本ソフトウェア科学会 第 10 回マルチ・エージェントと協調計算ワークショップ (MACC2001), 石川県金沢市, 2001.

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