マルチエージェント強化学習による社会規範の発現

中尾圭佑† 菅原俊治†

† 早稲田大学基幹理工学研究科情報理工学専攻

1 序論

本研究では，マルチエージェント強化学習によるエージェントの社会規範の発現について報告する．一般にエージェントは、それぞれの効用に基づいて理性的かつ自律的判断により行動を決定する．しかし，行動結果が相互に影響し合うマルチエージェント環境では，自分の効用に基づく行動が必ずしも最良の結果をもたらすとは限らない．また，具体的にどの行動をとることで，自分及び全体が最適化されるのかを予め指定することは難しい．そこで本研究では，全体のエージェントが効率的に動作するルールを規範と考え，それを強化学習によって抽出することを試みる．すべてのエージェントに同じ利得行列を持たせ，それに基づき行動を決定させることで社会規範の習得の成否を試みる．本研究を通して，未知の環境において独立した判断をしていても，全体として徐々に効率的になるような仕組みを検討する．

2 提案モデル

2.1 環境

実験環境として，狭路ゲーム [1] 及びセルオートマトンモデル 184[2] を組み合わせた，図 1 のようなモデルを提案する．狭路部における自動車の通行をモデル化したものであり，グリッドは道路に，エージェントはその道路を通行する車に見立てている．各マスに存在できるエージェントは 1 体のみであり，方向 0 の場合は水平の正方向に，方向 1 の場合は水平の負方向にしか進むことができない．また，水平方向においてトーラス状となっている．狭路部では 1 体のエージェントしか通過できず，狭路部を挟んでエージェントが向き合った場合は，「進む」を選択しても，相手が「待つ」を選択した時のみ通過できる．

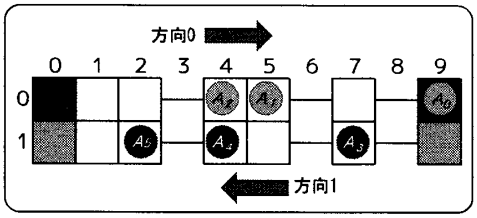

図 1 提案モデル概略図

An Experiment of Generation of Social Norms by Multi-Agent Reinforcement Learning

Keisuke Nakao†, Toshiharu Sugawara†

†Department of Computer Science and Engineering, Waseda Univercity

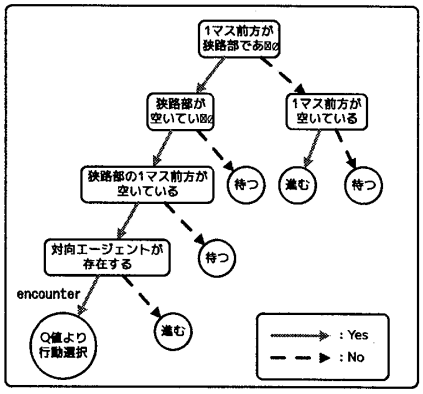

図 2 エージェントの行動アルゴリズム

表 1 利得行列 1，2 の報酬

行動		報酬	
自分	相手	利得行列 1	利得行列 2
進む	待つ	+3	+3
	進む	-5	-1
待つ	待つ	-0.5	-0.5
	進む	0	-0.5

2.2 エージェントの行動

エージェントは 1 ステップごとに「進む」か「待つ」の行動を選択・実行する．「進む」を選択したエージェントは 1 マス進むことができ，「待つ」を選択したエージェントは何もせずその場に留まる．図 2 にエージェントの行動アルゴリズムを示す．この中で encounter の状態で，各エージェントは自身の学習結果に基づき行動を決定する．

2.3 学習

本研究では，エージェントの学習アルゴリズムとして以下の Q 学習を用いる．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)]$$

なお，本研究では行動選択手法に $\epsilon-$ グリーディ手法 ($\epsilon = 0.05$) を採用し，その他のパラメータについては，学習率 $\alpha = 0.1$，割引率 $\gamma = 0.95$ とした．

状態 s は，狭路部での連続待ちステップ数とする．連続待ちステップ数には最大値を定め，最大値以降は 1 つの状態として扱う．また，表 1 に示す利得行列をエージェントに持たせ，これを基に報酬を受け取る．

3 シミュレーション

3.1 シミュレーション環境

シミュレーションの設定を表 2 に示す．本研究では，2 種類の利得行列を用意し，それぞれを保持するエー

表2　シミュレーション環境

道路長	50
狭路部数	5
狭路部位置	9, 16, 26, 35, 40
エージェント数	20
エージェントの初期配置	エピソードごとにランダム
待ちステップ数の上限	4
試行回数	1000

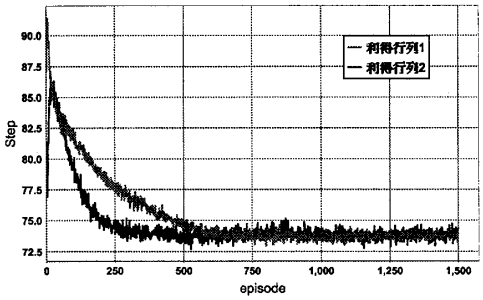

図3　所要ステップ数の推移

ジェント毎にシミュレーションを行った．利得行列1は，自分が「進む」を選択した場合，相手が「待つ」を選択すれば狭路部の通過に成功するので報酬が最も大きいが，相手も「進む」を選択するとお互い引き返さなければならないのでマイナスで最も小さい．また，自分が「待つ」を選択した場合，相手も「待つ」を選択すると状況は変わらないので，相手が「進む」を選択する場合よりも報酬を小さくした．利得行列2は，利得行列1よりも「進む」という行動に比重を置いており，やや利己的なエージェントを想定した．

3.2　結果

図3は，全エージェントが1周するのに要したステップ数の推移である．この図より，どちらの利得行列の場合でも，全エージェントが1周するのに要するステップ数はエピソードの経過とともに収束していることが分かる．

次に社会規範の発現について分析する．方向0を進むエージェントの「進む」のQ値の推移を調べた．図4は利得行列1の場合，図5は利得行列2の場合である．利得行列1の場合は，試行2，4のQ値が高くなっていることから，試行2，4では方向0のエージェントが狭路部では優先して通行するという規範が発現しており，反対に試行1，3，5では方向1のエージェントが狭路部では優先して通行するという規範が発現したと考えられる．試行4で方向0を進むエージェントの行動選択割合に注目すると (表3参照)，エピソードの進行に伴い「お互いに進む」，「自分が待ち，相手が進む」という行動の選択割合が減少し，「自分が進み，相手が待つ」という行動の選択割合が増加しており，1500エピソード目では，相手は必ず「待つ」を選択している．

一方，図5より，利得行列2については，3000エピソード目でもQ値は収束していない．また，実際の行動選択割合についても収束は見られない (表4参照)．これは，利己的なエージェントのみでは社会規範が発現しなかったことを示している．

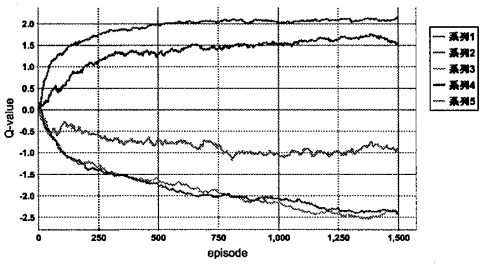

図4　利得行列1におけるQ値の推移

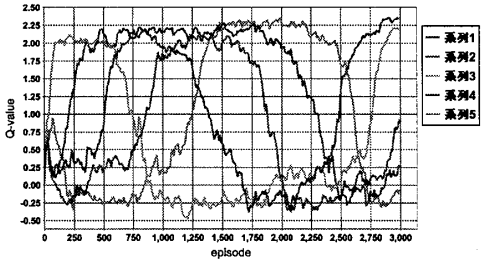

図5　利得行列2におけるQ値の推移

表3　利得行列1，試行4，方向0における行動選択割合

自分の行動	進む		待つ	
相手の行動	待つ	進む	待つ	進む
1エピソード目	36.2%	14.5%	30.5%	18.8%
500エピソード目	40.0%	8.8%	48.8%	2.4%
1500エピソード目	67.4%	0%	32.6%	0%

表4　利得行列2，試行1，方向0における行動選択の推移

自分の行動	進む		待つ	
相手の行動	待つ	進む	待つ	進む
1エピソード目	22.1%	20.8%	33.8%	23.4%
500エピソード目	15.0%	5.0%	5.0%	75.0%
1000エピソード目	46.3%	27.8%	9.3%	16.7%
1500エピソード目	57.1%	11.9%	11.9%	19.0%
2000エピソード目	24.1%	5.2%	20.7%	50.0%

4　結論

本研究では，強化学習による社会規範の発現を目的に，狭路部での自動車の通行をモデルとしたシミュレーションを行った．その結果，Q値の収束状態では社会規範の発現が確認できた．しかし，利己的エージェントの世界では社会規範の発現は確認できなかった．今後の課題としては，異なる情報利得を保持しているエージェントが混在している環境におけるシミュレーションを行い，社会規範の発現に与える影響を分析する．

参考文献

[1] Koichi Moriyama, Masayuki Numao. Self-Evaluated Learning Agent in Multiple State Games. ECML-2003 (LNAI 2837), pp.289-300 (2003)
[2] 宮入洋介．交通流解析における流体モデルとセルオートマトンモデルの比較，新潟工科大学情報電子工学科卒業論文 (2003)