報酬変動型ゲームにおける安定な状態遷移をもたらすエージェントの性質に関する研究

○（株）ジャパンテクニカルソフトウェア　稲垣　裕伸
室蘭工業大学　魚住　超　　小野　功一

この研究では，強化学習を行うエージェントが，動的な環境で活動するシミュレーションを行い，それぞれが協調するための条件について議論した．本研究では，各エージェントの行動によって環境が変化するような繰り返しゲームの考え方を導入し，各エージェントが協調できるのはどのような状況かをまとめた．

１．はじめに

自然界では，多くの生物が各々の生命を維持するために競合したり強調したりする様子が見られる．そのため生物種の個体数や環境などは常に変動しているが，生態系全体の恒常性が維持されている例が多く見られる．このような複雑システムでは，その各構成員は生命を維持するなどの，各々の利己的な目標をかなえるための行動をとっていて，その行動の結果が巡りめぐって自身や他の構成員に影響を及ぼしていると考えられる．

本研究では，自己の目標を達成しようと行動する構成員をエージェントとして扱うが，このようなエージェントの関わりを解析する手法にゲーム理論がある．しかしこれまでのゲームは，各エージェントの利得値が変化しないことを前提にしているが，本研究で注目する環境は，一定で変動しないことを想定することはできない．むしろ，エージェントを取り巻く環境が変化し，それに依存してエージェント同士の関係も変化していくことを想定する必要がある．

そこで本研究では，利得の変化する繰り返しゲームで動的環境を表現し，その環境において強化学習を行う複数のエージェントが，各々独自の目標を達成しようとするシミュレーションを行った．そして各エージェントが，動的な環境の中で強調していくための条件を，各エージェントの性質と目標の観点から考察した．特に，環境や他者に容認される寛容な性質を持ったエージェントが存在する場合，全体を安定に保つことができるかどうかを確認した．

２．実現方法

以下にシミュレーションの枠組みを示す．各エージェントは，環境にある共通の対象に興味を持っており，その対象の状態は，各エージェントの行動によって変動させることができるものとする．時間 t における興味対象の状態を x(t)とし，x は一次元の数値とする．各エージェントは，x に対する各々の目標値を持っており，x(t)の値を各目標値に近づけようと行動している．これらの複数のエージェントをエージェント A，B，…として，各エージェントの行動が x(t)に及ぼす結果を，それぞれ e(A)，e(B)，…，とすると，次の時間の状態 x(t+1)は，

$$x(t+1) = x(t) + e(A) + e(B)$$

となるように，状態の遷移を定める．当然，各エージェントとも，x の値を各々の目標値に近づけられることができたときに正の利得値が得られるものとする．勿論各エージェントの選択できる行動は，複数あるので，相手の行動選択に対して自分がどのような行動を取ればよいかという選択は，そのときの x(t)の値に依存して変化する．

次にエージェントの実装について説明する．上で示したシミュレーションの枠組みや，全章で述べた背景から，ゲーム理論のように，現在の x の値から各エージェントの行動との利得表を作成して解析していくのは，現実的ではないと考える．むしろエージェントに適応能力を持たせ，試行錯誤で適切な行動パターンを獲得させる方が自然である．よって，本研究ではゲーム理論による解析ではなく，各エージェントの性質と環境の状態 x の遷移に注目していくことにする．

そのため，エージェントに強化学習の能力を持たせる．強化学習は，環境からの状態入力によって行動を決定し，環境からその行動によって罰や報酬が得られるというものである．今回は，事例に基づく強化学習を導入した[1]．これは，各時間の x の値とそのとき選択した行動のペアを事例として記憶し，その行動によってもたらされた報酬（罰）も記憶する．エージェントは，各時間の x の値に最も近い値を持つ事例を検索し，報酬をもたらしてくれたものであるなら，その行動を再び選択し，そのときの事例を記憶していく，というものである．もし報酬をもたらしてくれた事例が無ければ行動はランダムになる．

次に，エージェントの性質であるが，これはエージェントの利得関数として表現できる．利得関数は，各エージェントが，そのときの x の値と自己の目標から，報酬や罰を決定するものである．つまり，どの程度 x を目標値に近づけることができると成功とみなすか，ということを決定するものである．

表１に，エージェントの性質と，行動の種類，その効果についてまとめた．

表１．エージェントの基本パラメータ

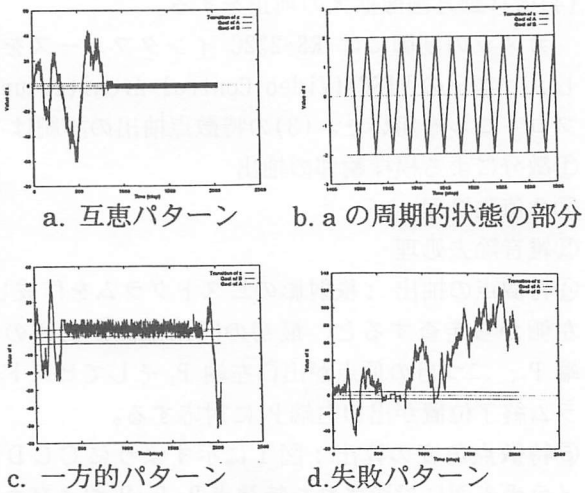
選択できる行動	a1, a2, a3, a4, a5,
xへの行動の効果	それぞれ -3, -1, 0, 1, 3
目標値 g	{0,2,4} （この値を変えてシミュレーション）
性質の種類	寛容，普通，心の狭い
寛容な性質 Tolerant	1 （|x(t) -g|≦１） 0 （|x(t) -g|＞１）
普通の性質 Normal	1 （|x(t) -g|≦１） 0 （1<|x(t) -g|≦5） -1 （5<|x(t) -g|）
心の狭い性質 Selfish	1 （|x(t) -g|≦１） -1 （|x(t) -g|＞１）

心の狭いな性質というのは，自分の性質以外全て失敗とみなすものである．寛容というのはその逆で，成功以外のすべての状態を全く気にしないというものである．

３．シミュレーション

各々独自の目標値を持ったエージェントについて，それぞれの性質を当てはめてシミュレーションを行った．シミュレーションの初期状態でx=0であり，2000ターンの推移を見た．

いずれの結果も，図のような３つのパターンに分類できる．



a．互恵パターン　b. a の周期的状態の部分

c．一方的パターン　d.失敗パターン

図１．各パターンの分類

１つ目は，xの値が各エージェントの目標値の間を周期的に推移するもので，周期的にそれぞれ報酬を分け合っている互恵的状況である．このパターンになるとどのエージェントも，目標を達成した状態が続くので，xの遷移も安定して推移する．２つ目は，xが一方の目標値の近くでのみ遷移するパターンである．これは，どちらか一方のエージェントだけが，目標を達成するための有効な行動パターンを獲得できた状況で，そのエージェントだけが利益をあげている状況である．この状況は，不安定であり，優位性が反転したり，次の失敗パターンに移行したりすることもある．３つめのパターンは，xの値がどの目標値よりもかけ離れて全くランダムに遷移している状況である．どのエージェントもxを目標値に近づけることができずに，失敗している状態である．

２つのエージェントに，独自の目標値と３つの性質をそれぞれ当てはめてシミュレーションを行った結果を表２に示す．

表２．エージェントが２つの時の結果

組合せ (A×B)	(A の目標値，B の目標値) (0, 0)	(2, -2)	(4, -4)
T×T	26/0/4	18/3/9	12/4/14
T×N	26/0/4	18/2/10	0/5/25
T×S	20/0/10	3/1/26	0/2/28
N×N	21/0/9	12/1/17	0/9/21
N×S	14/0/16	2/0/28	0/5/25
S×S	4/0/26	1/0/29	0/3/27

全体的に，目標が離れるにつれ互恵パターンが少なくなり失敗パターンが増えているが，目標値が近くても心の狭いエージェントがある場合は，失敗パターンの数が多くなり，目標値が離れていても寛容なエージェントだと互恵パターンが増えている．

同様のシミュレーションをエージェントが３つの場合も行ってみた．組合せの数が複雑にはなるが，大まかな傾向は、エージェント２つの場合とほぼ同様の結果が得られた．

４．まとめ

本研究における寛容な性質とは，目標値が離れている場合で，他者が目標を達成しているときにそれを気にしないということとみなすことができる．本研究では，xの遷移が安定しており各構成員の目標も同時に達成できていることから，３パターンのうち互恵パターンが，理想的であると考えることができる．今回シミュレーションによって，適応能力を持ったエージェントが独自の目標を達成しようと行動するとき，寛容な性質が安定な互恵パターンをもたらすという結果を確かめることができた．

５．参考文献

[1]畝見達夫：事例に基づく強化学習，人口知能学会誌，Vol.9, No.6, pp.697-707(1992).