人　工　知　能　99－1
（１９９５．　３．　６）

マルチエージェントシステムにおけるゲーム環境の利用に関する研究

吉村　潤　，　嘉数　侑昇
yosimura@hupe.hokudai.ac.jp

北海道大学工学部

本論文は，競争的ゲーム環境を用いたマルチエージェントシステムにおける学習の効果とゲーム環境の利用性について議論する．近年、強化学習を用いてより適応的な行動を獲得できるかという研究が多く成され，分散的な問題解決のアプローチであるマルチエージェントシステム的な環境での学習も行われている．しかし，より適応的な行動獲得させる点で，競争の概念によるエージェントの創発的行動の獲得については議論されておらず，競争的関係をもつマルチエージェント環境が学習を促進させると期待できる．本論文では，ゲームの特性である競争的概念を用いたゲーム環境下での学習法を提案し，ゲーム環境をもちいることによって明示的に競争関係を定め，学習にあたえる影響を議論する．

A Study on The Utility of Game Environments in Multi-agent Systems

Hiroshi YOSHIMURA , Yukinori KAKAZU
yosimura@hupe.hokudai.ac.jp

Department of Precision Engineering, Hokkaido University

In this paper, we discribe the learning effect and the utility of Game Environments in Multi-agent Systems under competitive game environments. Recently many reserchers notice Reinforcement Learning to acquire more adaptive behaviors, and apply to the Multi-agent Systems. Because there are little disscusion about emergent behaviors in competition, we make agents to promote its learning by using Multi-agent environments. We propose the learning method in the competitive environments using the competitive comcept that is the characteristic in games, and we finaly discuss the effects of learning influenced by Game Environments.

1.はじめに

本研究は,マルチエージェントシステムにおける学習について,エージェント同士のゲーム環境を用いることによる有効性について議論する.

近年,複数のエージェントによる協調動作の学習や,複雑な問題解決のアプローチとしてマルチエージェントシステム[1)-3)]の概念が提案されている.これらのシステムは,複数のエージェント同士が協調することで一つの動作を学習することや,単一エージェントでは解決できない問題を複数のエージェントにより解決しようとするアプローチである.

また,自律移動ロボットのようなエージェントを未知の環境に適応させるための学習法の一つに,強化学習(Reinforcement Learning)[4)]がある.この強化学習とは環境からの評価（強化信号）に従って,状態と行動の対である戦略を獲得していく学習法である.その中でも特にWatkinsの提案したQ-Learning[5)]は非常に優れた学習法として知られ,様々な研究がなされている.このような強化学習に対してマルチエージェントシステムの概念を拡張させた研究が,Tan[6)]やMataric[7)]らによって成されている.これらはエージェント間の通信や社会的な関係を用いた学習法であり,エージェントの環境に対する適応的な行動を獲得するものである.しかし,このようなマルチエージェント的な環境における学習において,個々のエージェントの創発的な行動能力がいかなる性質のものかについては議論されていない.

一方,エージェントの行動や振る舞いの性質について議論するとき,ゲーム環境を利用して環境からの明示的な評価をする方法がある.これはゲームの特性である"勝敗"を用いることで,エージェントの持つ行動特性をエージェント間の"優劣"として評価することができ,また単なる単一エージェントのみの学習よりもエージェント同士の相互作用による創発的な行動の獲得が期待できる.

そこで本論文ではゲーム環境を利用したマルチエージェントシステムの環境の下で,エージェント同士の相互作用的な学習による創発的な行動の獲得を試み,その性質について議論を行う.具体的にはマルチエージェント的な環境としてTag Game [8)]と呼ばれるゲーム環境を設定し,Q-Learningとクラシファイヤーシステム[9)]による強化学習を用いた場合の有効性について議論する.

以下では,まず強化学習の枠組みを説明し,次に本論文で扱うマルチエージェントシステムの概念を説明する.さらにゲーム環境を設定し,このゲーム環境に対する学習法を示し,簡単な計算機実験によってその有効性について議論する.

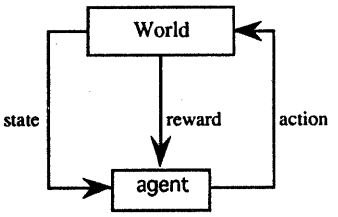

Figure.1:Overview of Reinforcement Learning.

2.強化学習

強化学習とはある状態(state)に対して最適な行動(action)を選択するように,環境からの評価(報酬,reward)によってその選択確率の増減を行ない環境に適応する行動を獲得していく手法である.強化学習の枠組みをFigure.1に示す.

以下では代表的な強化学習法であるQ-Learningとルールベース型のクラシファイヤーシステムについて説明する.

2.1Q-Learning

Q-Learning(QL)は各状態行動対にQ-値と呼ばれる評価値を与え,このQ-値に基づいてある状態に対して最適な行動を選択し,次の状態の有用度によってQ-値を更新していくアルゴリズムでありWatkinによって提案された[5)].これは次のように記述される.

ある状態xに対し適切な行動を選択する政策を$\pi(t)$とする.状態xと行動aに対する有用度$Q(x,a)$は次のようになり,これをQ-値と呼ぶ.

$$Q(x,a)=E\left\{\sum_{n=0}^{\infty}\gamma^n r(n+1)\,|\,x_0=x,a_0=a\right\}, \tag{1}$$

ここで,γ,$0\le\gamma\le1$は割引率,$r(t)$は時刻tにおける環境からの報酬である.このQ-値に基いて状態xの有用性$U(x,a)$,および状態xに対しての最適な行動をとるための$\pi(x)$は次のようになる.

$$U(x,a)=\max_{b\in A}Q(x,b), \tag{2}$$

$$\pi(x)=a:Q(x,a)=\max_{b\in A}Q(x_t,b), \tag{3}$$

ここでAは行動集合である.これより政策$\pi(t)$は最大のQ-値を持つ行動を選択することである.

前述の行動選択を実現する際には,同時に最適行動の提案をも考察にいれる必要があり,そのため一般的には競争的選択を取り入れる.ここでは次のようなボルツマン分布に従った確率によって行動の選択を行なう.

$$\Pr(x,a)=\frac{\exp(Q(x,a)/T)}{\sum_{b\in A}\exp(Q(x,b)/T)}, \tag{4}$$

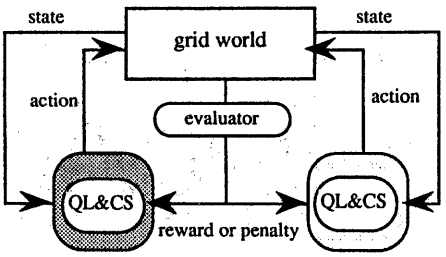

Figure. 2 : Overview of Reinforcement Learning for Multi-agent System.

ただし,Tは温度定数である.

このような確率に基づきQ-値は次のような更新式によって更新される.

$$Q_t(x,a)\leftarrow(1-\alpha)Q_{t-1}(x,a)+\alpha[r(t)+\gamma U(y)], \quad (5)$$

$$U_{t-1}(y)\equiv\max_{b\in A} Q_{t-1}(y,b), \quad (6)$$

ここで,α,$0\leq\alpha\leq1$は学習率である.ただし,報酬は状態xが目標状態であれば,1,それ以外では0であるように設定されるのが一般的である.

2.2クラシファイヤーシステム

クラシファイヤーシステムとはクラシファイヤーと呼ばれるビット列の文字列で表わされた条件部と行動部からなるルールの集まりを用いたルールベース型の強化学習であり,Holland[9][10]らによって提案されている.この学習法はエキスパートシステムのようなif-then型のルールを用い,ルールの有用度に基づいて状態に対する行動を選択する学習法である.さらに,特徴として条件部のビット列を遺伝的アルゴリズム(GA)の遺伝子とみなし,遺伝子操作によって各ルールの改良,及び新ルールを発生させる点がある.

また,行動選択に関してQLとCSでは次のような違いが考えられる.すなわち,QLでは行動選択とその獲得という点では非常に優れているといえるが,複雑な環境状態では多数の行動が与えられるとQ-値テーブルの状態爆発が起こり,その結果学習能力が落ちる場合がある.一方,CSでは必要な数のルール群を維持し,さらにGAによって創発的なルールの出現が期待される.よってこの2つの学習法の融合が望ましいと考えられる.

3.マルチエージェントシステム

マルチエージェントシステム[1]とは,各エージェントが全体のシステム目標に対し,独立的に個別目的や環境,および,他のエージェントの行動を認識することにより全体として分散的に行動を決定しつつ,全体のシステムとしての機能を維持させようとするものである.本論文では,このようなマルチエージェントシステムに適用する環境として,明示的に勝敗が与えられるゲーム環境を用いて,そこでの適応行動獲得について議論を行なう.このようなゲーム環境におけるマルチエージェント的な強化学習の概念をFigure.2に示す.ここでの報酬の与えられ方はゲームの勝敗に依存し,一方のエージェントが早く目的を達成したとき,すなわち勝利をおさめたとき,リワードを受け,その他のエージェントへはペナルティが与えられるものとする.

4.ゲーム環境

本論文で扱うゲーム環境として,Tag Game [6]と呼ばれるゲームを用いる.Tag Gameとはいわゆる鬼ごっこである.ゲームはふたり以上のプレーヤーから構成され,あるプレーヤーは他のプレーヤーを追いかけ,追いかけられる側は逃げなければならない.すなわち,各々のプレーヤーの目的は追いかけることと,逃げることである.ここで,その目的が各々独立していることから,全体の目的としてマルチエージェントシステムとみなすことができ,本環境下における学習は非常に興味深いと思われる.すなわち,ゲームによって与えられる競争関係が,より共進化的な学習の促進をさせるものと期待される.

本論文では学習法として各プレーヤーを,状態行動対としてCS,その強化手法としてQLの学習機構をもったエージェントとみなしてゲームを行う.すなわち,CSでの行動選択のために用いる強度としてQ-値を用い,QLのアルゴリズムによるルールの選択を行う.具体的には,エージェントの目的は目標物に接近してこれを捕獲するような行動ルール群となるCSを獲得することである.そして,ゲームの勝敗の結果からエージェントの評価を行ない,QLに基づく適応的な行動の獲得を試みる.以上の評価の概念をFigure.3に示す.また,単一エージェントによる学習との比較によって,このようなゲーム環境を用いたときの行動特性について議論を行う.

5.計算機実験

5.1環境設定

ゲーム環境としてFigure.4のような10*10の格子空間(grid world)を設定する.マルチエージェントシステム的なゲーム環境として二人ゲーム(two agents game),三人ゲーム(three agents game)を考える.前者の目的はエージェントがふたりで静止している目標物の獲得を競いあうこと,後者は互いが目標となり追いかけあうことを目的とする.各エージェントはQLとCSによって行動の獲得を行わせる.

これらエージェント間の競争関係に基づいて得られた行動と,競争関係のない単一エージェントのみの学習によって得られた行動を比較することによってゲーム環境が学習に与える有効性を示す.ここで,単一エージェントのみの学習とは,Figure.4の

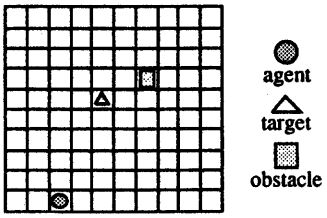

Figure. 3: Grid world.

Table 1:Parameters of experiments.

学習率 α	0.15
割引率 γ	0.95
温度定数 T	0.01
総classifier数	3000

ような格子空間において単一のエージェントだけで目標物の捕獲を学習することである.これをシングルゲーム(single agent game)と呼ぶことにする.

各環境において学習によって得られた行動の比較を次のように行なう.目標物と障害物が一つずつ存在し,これらがランダムに動き回る環境の中で,前述の三つのゲームで各々学習したエージェントがどれくらい目標物を捕獲できるかを比較する.これを比較ゲームと呼ぶ.

それぞれのゲームの共通ルールを説明する.各ゲームとも200ステップを1ゲームとし,エージェントは1ステップに行動を一つ選択する.ただし,エージェントは1ステップに1マスのみ移動するものとする.ゲームの終了条件は,1ゲーム以内に目標物を獲得することとし,200ステップを超えるとタイムオーバーとし,ゲームを中断し次のゲームに移る.ただし,格子空間を飛び出したときは,この格子空間をトーラス面とみなし飛び出した側とは反対側に出現させゲームを続行する.また,エージェント,目標物,および障害物のスタート位置は各ゲーム毎にランダムに選ぶ.

エージェントが観測する情報は,自分の位置(x座標,y座標),目標物の方向,障害物あるいは他のエージェントの方向とし,これをクラシファイヤーの条件部へコード化する.エージェントの行動は前進,方向転換(左右45度)の3通りである.エージェントの実験パラメータをTable.1に示す.エージェントへの報酬は,目標物を獲得したとき,1.0,他のエージェントに捕えられたとき,あるいは障害物にぶつかったとき,-0.9,格子空間を飛び出したとき,-1.0,それ以外を-0.01とした.

5.2考察

実験環境中のエージェントを,agent-1,agent-2,agent-3と呼ぶことにする.各環境において2000ゲームを行なった.各環境におけるエージェントの学習評価指標として1ゲーム当たりの目標物のの捕獲率と獲得報酬を計測した.二人ゲーム,三人ゲームの結果をそれぞれFigure.4,Figure.5に示す.以下,各々のグラフから得られた考察を行なう.

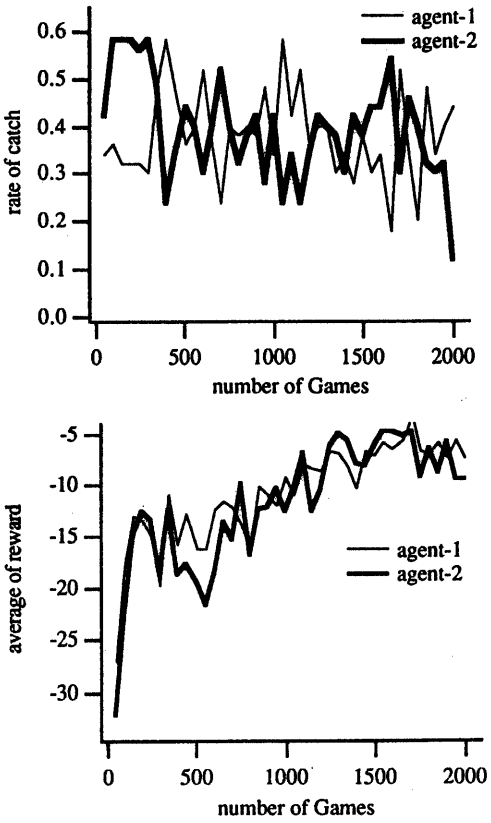

Figure.4: the result of the two agent games.

まず,二人ゲームにおいては学習の初期段階ではagent-2の学習が進み,agent-1は目標物を獲得できていない.しかし,500ゲームを過ぎるころからエージェントの捕獲率が交互に入れ替わるようになった.これはゲームの勝敗によって,報酬の割り当てが変化していると考えられ,エージェントが交互に学習しているといえる.この間,各エージェントの行動は決して悪い評価を受けていないのは1ゲーム当たりの獲得報酬から明らかである.また,エージェントの学習は捕獲率が0.4に収束する傾向が見られた.

次に三人ゲームであるが,二人ゲームの場合と異なり,捕獲率はそれ程高い値にはならなかった.これは三つのエージェントが同時に学習を行なうことから,結果的に互いの足を引っぱりあう状態が起こ

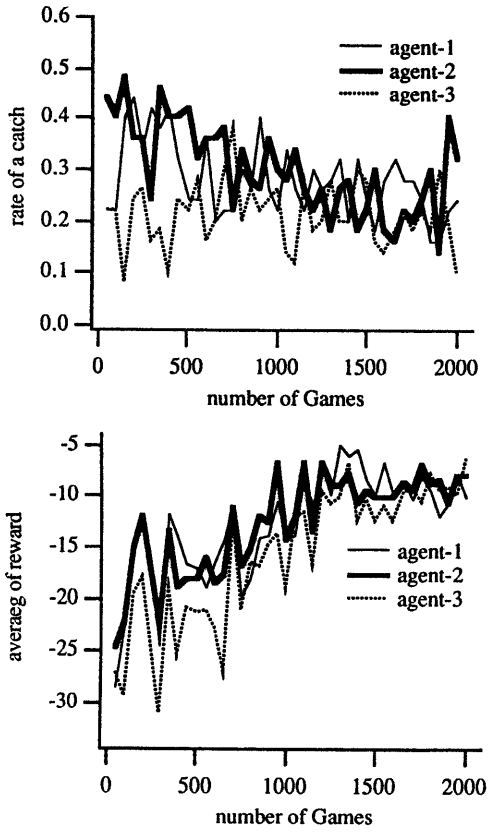

Figure.4: the result of the three agent games.

Table. 2 : Comparison result.

two agent game	683/2000
three agent game	446/2000
single agent game	358/2000

り学習効率が上がりにくくなったと考えられる.しかし,このときも二人ゲームと同様,1ゲーム当たりの獲得報酬から,悪い評価を受けているわけではないことがわかる.

比較実験においては,2000ゲーム中にどれだけ目標物を捕獲できるかを,二人ゲーム,三人ゲーム,そして,シングルゲームの実験より獲得されたクラシファイヤーを用いて比較した.二人ゲーム,シングルゲームよりそれぞれ,agent-1,agent-2のエージェントがもつクラシファイヤーを選んだ.結果をTable.2に示す.これよりシングルゲームに比べ,二人ゲームと三人ゲームの方がより適応的な行動を獲得しているといえる.特に二人ゲームの捕獲数はシングルゲームのそれよりも2倍近い値を示した.これよりシングルゲームのような単なる強化学習では得ることのできない行動を,二人ゲームや三人ゲームのようなマルチエージェント的な環境におけるゲームを利用することで,より様々な環境に適応する行動,すなわち創発的行動を強化学習によって獲得することができることが示せた.

6.おわりに

マルチエージェントシステムにおける学習として,単一のエージェントによる強化学習の枠組みでは得られない行動を,Tag Gameのような競争概念のあるゲーム環境を利用することによってより環境に適応した行動を獲得することができることを示した.

参考文献

1)Martin, M.:The Society ofMind, Simon & Schuster, (1986).

2)Ronald C.A., J. David Hobbs : Dimmensions of Communication and Social Organization in Multi-agent Robotic Systems , The Second International Conference On Simulation of Adaptive Behavier,pp.486-493,(1992).

3) Gerhard, W. : Action Selection and Learning in Multi -Agent Environments,The Second International Conference On Simulation of Adaptive Behavier , pp. 502-510, (1992).

4)Sutton, R. S.(Eds.) : Reinforcement Learning,Kluwer Academic Publishers, (1992).

5)Watkins, J.C.H.and Dayan,O.:Technical Note Q-Learning , Machine Learning 8, pp.279-292(1992).

6)Tan, M. : Multi-Agent Reinforcemant Learning : Independent vs . Cooperative Agents, Machine Learning Proceedings of the 10th International Workshop, pp.330-337,(1993).

7)Mataric, M. J.:Learning to Behave Socially , Proceedings of The Third International Conference On Simulation of Adaptive Behavier,pp.453-462,(1994).

8)Reynolds, C.W. : Competition,Coevolution and the G-ame of Tag,Proceedings of the Force International Workshop on the Synthesis and Simulation of Living Systems,Artifical Life IV,pp.59-69,(1994).

9Goldberg,D.E.:Genetic Algorithms in Search, Optimization,and Machine Learning,Addison-Wesley(1989).

10)J. H. Holland, K. J. Holyyoak, R. E. Nisbett , & P. R.Thagard:Induction:Processof nference , Learning, And Dicovery, MIT Press,(1986).

11)Mataric, M. J. : Reward Function for Accelerated Learning,Machine Learning Proceedings of the 11th International Workshop,(1994).