岡山理科大学紀要 第42号 A pp. 75-83(2006)

複数タスクの問題に対するマルチエージェント強化学習の報酬に関して

太田 真由美・金重 徹・片山 謙吾*・南原 英生*
成久 洋之*

岡山理科大学大学院工学研究科情報工学専攻
*岡山理科大学工学部情報工学科

(2006年10月2日受付、2006年11月6日受理)

1 はじめに

複数の自律エージェントが相互に作用しあいながら問題を解決するマルチエージェントシステム (Mutliagent System) の研究が盛んである [14]. エージェントとは，行動を行うことによって，自分がおかれている環境に対して影響を与えることのできる自律的主体を指す [11]. マルチエージェントシステムの応用例として，災害現場での活躍が期待されるレスキューロボットがあげられる. 1995年に起こった阪神淡路大震災は，想像を絶する大災害であったため，被災者の救助活動は困難を極めた. このような災害現場において，人間の代わりに迅速に被災者の救助活動を行うことができる自律ロボットの実現が強く求められている. 自律ロボットの活躍が期待される災害現場のように，マルチエージェントシステムがおかれる環境として想定されるのは，大規模で複雑（動的・未知）な環境である場合が多く，そのような環境でタスク（被災者を救助するという仕事）を達成するためには，複数のエージェント同士が協力してタスクを達成する協調行動の実現が非常に重要な課題となる. しかし，大規模で複雑な環境に適応可能であり，協調行動を行うことができるマルチエージェントを設計することは非常に困難である. 設計者が予め起こりうる全ての状況を予測し，知識をプログラム化して，エージェントに与えておくことは事実上不可能だからである. よって，各エージェントが自身の経験を通じてタスクを達成する方法を学習できる機能を備えていることが望ましいといえる. そのような学習機能として，設計者が目標達成時に与える報酬の設定をするだけで，あとはエージェントが自律的に環境との相互作用を通して，報酬を最大にする適応行動を獲得していく強化学習 (Reinforcement Learning) [4, 12] による機械学習アプローチが注目を集めている [2, 5, 10].

現在多くのマルチエージェント強化学習の研究では，エージェント間の協調の獲得を前提として，単一タスクを有する問題を対象にする場合がほとんどであるが，現実的な問題においては複数タスクが存在する場合が多い. 例えば，マルチエージェントシステムの代表的な問題であるサッカーゲームでは，通常，ゲームに勝利する，またはゴールするというタスク（主目標）を達成するまでには，パスやドリブルなどを成功させるという複数のタスク（副目標）が連鎖的に存在している. このような問題を対象とするマルチエージェント強化学習においては，一般的に主目標達成時に報酬を与えることが適切と考えられている. しかし，問題が大規模で複雑であれば，非常に多くの試行錯誤行動を余儀なくされるため，報酬が得られるまでの時間が膨大になることがある. そのような場合には，学習途中に補助的な報酬を与えるなどの工夫が考えられているが，その補助的な報酬が強化学習に与える影響などについては不明な点が多い.

そこで本研究では，ゴミを拾うタスクと拾ったゴミをゴミ箱に捨てる複数のタスクが連鎖的となるクリーンナップ問題を対象にして，主目標達成時に報酬を与える方法と副目標達成時に報酬を与える方法の報酬の与え方の違いによるマルチエージェント強化学習の特性に関して検討する. その際，後述する協調確認の尺度を用いて協調行動の有無を確認し，結果として，副目標達成時に報酬を与える方法は，主目標達成時に報酬を与える方法よりも良好な学習を行い，かつエージェント間の協調行動も獲得できることを示す.

2 強化学習

2.1 強化学習エージェント

強化学習を用いて学習する自律エージェントを強化学習エージェントと呼ぶ．強化学習エージェントは，環境の状態を認識し，それに対してエージェントが可能な行動群の中から行動を一つ選択して実行する．この状態認識と行動を繰り返した結果，目標状態に達したとき，環境から報酬が与えられる．エージェントは報酬をもたらす行動を優先するように環境への適応を目指す．

2.2 マルコフ決定過程 [13]

強化学習では，環境のダイナミクスをマルコフ決定過程（Markov decision process:MDP）として定式化し，アルゴリズムの解析を行うのが一般的である [7]．現状態を時刻 t の状態 s_t，次状態を時刻 $t+1$ の状態 s_{t+1}，現状態での行動を時刻 t の行動 a_t とし，現状態から次状態への遷移確率を $P^{a_t}_{s_t s_{t+1}}$ と表すと，式1に示すような条件つき確率で記述できる．これは，過去の状態 s_0 において行動 a_0 を実行し，その遷移先である状態 s_1 において行動 a_1 を実行するというプロセスを繰り返し，最終的に状態 s_t において行動 a_t を実行したときに，状態 s_{t+1} に遷移する確率を表している．

$$P^{a_t}_{s_t s_{t+1}} = P\{s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \cdots, s_0, a_0\} \tag{1}$$

ここで，もし P が現在の状態と行動のみに依存するだけなら，式2のように簡略化できる．つまり，状態 s_t において行動 a_t を実行したときに，状態 s_{t+1} に遷移する確率となる．

$$P^{a_t}_{s_t s_{t+1}} = P\{s_{t+1}|s_t, a_t\} \tag{2}$$

式2のように1ステップ前の現象（ここでは時刻 t の状態と行動）のみから次の事象（ここでは時刻 $t+1$ の次状態）が決定できる性質を（単純）マルコフ性（Markov property）と呼び，そのマルコフ性を持った確率過程を（単純）マルコフ過程（Markov process）という．上記の例のように状態を離散値として扱っているマルコフ過程において，各時刻ごとに状態が定義されているとき，エージェントが何らかの決定をすることによって状態の遷移が起こり，それに依存した利得や費用が発生する場合，適切な決定を下す必要がある．このような離散状態のマルコフ過程を基にした逐次（あるいは多段）決定過程のことをマルコフ決定過程と呼ぶ．

マルチエージェント環境では，他のエージェントの行動や内部状態を正確に知ることができないという設定が一般的であり，その環境では単純マルコフ性は仮定できなくなる．なぜならば，再び同じ状況で同じエージェントに出会ったとしても，そのエージェントが今までに学習した内容によって依然とは異なる行動を実行する可能性があるからである．なお，このような環境はマルコフ性を仮定できないという意味で，非マルコフ決定過程（non Markov decision process）環境と呼ばれている．このように，マルチエージェント環境では単純マルコフ決定過程を基にした学習メカニズムでは完全に対処できないことが多い．非マルコフ決定過程の代表例として，セミマルコフ決定過程（Semi-Markov Decision Process:SMDP）や部分観測マルコフ決定過程（Partially Observable Markov Decision Process:POMDP）があげられる．SMDP は時間が連続な MDP であり，報酬が時間積分される点を除けば，基本的な取り扱いは MDP と大きな相違はない．一方，POMDP とは，実際には異なる環境の状態がエージェントにとっては同一の感覚入力として知覚される，いわゆる不完全知覚状態を有する問題クラスのことをいう．

2.3 Profit Sharing

本研究では，強化学習手法として，マルチエージェント環境において有効とされている [1] Profit Sharing を用いる．Profit Sharing は，報酬に至るまでのエピソードにおける状態 s と実際に行った行動 a の対からなるルール系列を記憶しておき，報酬が得られたときにそれまでの系列上のルールを一括して強化する学習方法である．ルール系列は次式を用いて強化する．

$$w(s_i, a_i) \leftarrow w(s_i, a_i) + f(r, i) \tag{3}$$

$$f(r, i) = \beta^{W-i} r \tag{4}$$

ここで，$w(s_i, a_i)$ はエピソード系列上の i 番目のルールの重み，f は強化関数，r は報酬値，$\beta(0 \leq \beta \leq 1)$ は報酬割引率，W はエピソードの最大長である．

Profit Sharing の学習過程におけるエージェントの行動選択法としては，ルーレット選択法が良い性能を示すことが知られている [1]．このことから本研究では，エージェントの行動選択法としてルーレット選択法を使用する．ルーレット選択法は，ある状態 s において，各行動の重み $w(s, a_t)$ を全ての行動の重みの合計 $\sum w(s, a_t)$ で割り，確率 $P(a_t|s)$ を求め，その確率により行動を決定する方法である．

$$P(a_t|s) = w(s, a_t) / \sum w(s, a_t) \tag{5}$$

3 マルチエージェント強化学習における問題

マルチエージェント強化学習には，マルチエージェントシステム特有の困難な問題が発生する．本章では，マルチエージェント強化学習において特に考慮すべき問題である，(1) 不完全知覚問題，(2) 同時学習問題および (3) 報酬分配問題の 3 つについて述べる [2, 13]．

3.1 不完全知覚問題

不完全知覚問題（perceptual aliasing problem）とは，エージェントの知覚情報が限られている，あるいは不完全であるために，異なる状態を同じ状態として知覚することで，学習に悪影響を及ぼす問題である．不完全知覚問題を有する環境のクラスは，上述したように POMDP と呼ばれている．この POMDP 環境下において，各エージェントに Profit Sharing を適用し，陽に定義された通信による情報の交換を行うことなしに，マルチエージェント間で不完全知覚問題を解消し協調行動を創発できたことが報告されている [2]．

3.2 同時学習問題

同時学習問題（concurrent learning problem）とは，自分の行動した結果が自分の行動のみによるものか，他のエージェントとの協調行動によるものかを判断することが困難なために，適切な学習が行えない問題である．マルチエージェント環境では，他のエージェントと協調して，効率的にタスクを達成することが求められており，状態の遷移先は自己の行動のみによるものではなく，他のエージェントとの連係行動による場合が多い．しかし，複数のエージェントが同時に学習するため，自己の行動による環境の状態遷移先を特定するのは難しい．よって，他のエージェントの学習の影響により，適切な学習が行えない場合が生じる．

3.3 報酬分配問題

報酬分配問題（reward sharing problem）とは，複数のエージェントが協力して目標を達成した場合，どのエージェントにどれだけの報酬を与えるべきか適切な判断が難しい問題である．例えば，サッカーゲームにおいて，ゴールを決めたエージェント，ゴールを決めたエージェントに絶妙なパスをしたエージェント，何もしていなかったエージェントがいた場合，ゴールを決めたエージェントのみに報酬を与えると，他のエージェントは全く学習をしない．逆に，全てのエージェントに同じ報酬を与えると，目標達成に貢献しなかったエージェントが冗長な行動を学習する問題が発生する．適切に報酬が分配されなければ，各エージェントの学習だけでなく，システム全体に悪影響を及ぼすことになる．

4 クリーンナップ問題

クリーンナップ問題とは，エージェントがある環境中に存在するゴミをすべてゴミ箱に捨てることを目標とする問題である．よって，クリーンナップ問題での主目標はすべてのゴミをゴミ箱に捨てることであり，副目標はゴミを拾う，またはゴミを捨てることである．クリーンナップ問題は，ゴミを拾い，そのゴミをゴミ箱に捨てるという連鎖的な複数のタスクが存在するため，これまでのマルチエージェント強化学習の研究で対象とされてきた単一タスクの追跡問題 [5] に比べ，目標状態への到達が困難な問題である．以下，本研究で利用するクリーンナップ問題の設定に関して記述する．

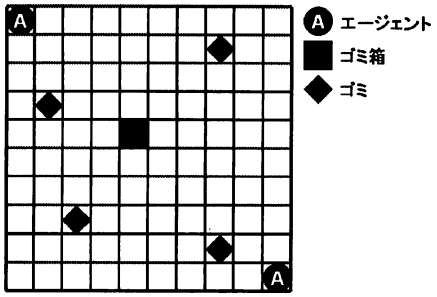


図 1: クリーンナップ問題の環境の例

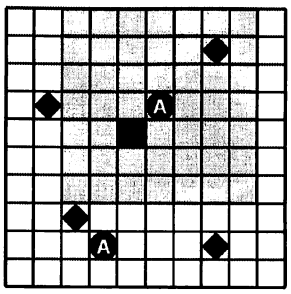


図 2: エージェントの視界

$n \times n$ の 2 次元格子状の環境を設定し，格子の外枠を壁とする．この環境にゴミ箱を N_{gc} 個，ゴミを N_g 個，エージェントを N_A 個配置する．図 1 に $n = 10$，$N_{gc} = 1$（中央付近に固定），$N_g = 4$，$N_A = 2$（左上と右下の端が初期位置）の例を示す．クリーンナップ問題におけるエージェントは，自分が置かれている環境と以下に示す相互作用を行う．

Step1 エージェントは時刻 t において，状態 s_t を認識する．ただし，図 2 の網掛けで示すように，エージェントの視界は $m \times m$ で与え，自分の位置とその周囲 m^2 コマに存在するゴミ箱，ゴミ，他のエージェント，壁の状態を認識することができる．

Step2 エージェントは入力として認識した状態に対し，上下左右の方向に 1 コマ進む，または停止の 5 つの行動の中から 1 つの行動を選択し，出力として実行する．

Step3 エージェントの行動 a_t により，環境の状態 s_t が次状態 s_{t+1} に遷移する．

Step4 次状態 s_{t+1} が目標状態となれば，エージェントは環境から報酬 r が与えられる．目標状態でなければ，時刻 t を $t+1$ に進め，Step1 へ戻る．

ただし，各エージェント $A_j (j = 1, \cdots, N_A)$ は同時に行動するものとする．エージェントは壁への移動を選択することはできない．エージェントがゴミを拾うときは，ゴミが存在するマスの上で停止の行動を選択しなければゴミを拾うことができない．また同様に，エージェントが拾ったゴミをゴミ箱に捨てるときは，ゴミ箱が存在するマスの上で停止の行動を選択しなければゴミを捨てることができない．エージェント同士が同一の場所に存在する，またはすれ違うことはできない．いずれの場合もエージェント間の衝突となり，衝突前の位置に戻ることとする．ただし，あるマスでゴミを拾う，またはゴミを捨てること（停止の行動）を選択したエージェントと，そのマスに進む行動を選択したエージェントが衝突した場合，そのマスに進んだエージェントのみがマスに進む前の位置に戻ることとする．エージェントが行動した単位時間を 1 ステップ，目標状態に達するまでを 1 エピソードとする．

5 クリーンナップ問題における協調確認の尺度

マルチエージェントによる協調行動を適切に確認することは一般的に難しく，学習後の各エージェントの行動を人間が視覚的に観察し，協調めいた行動が達成されているか否かを主観的に判断する場合が多い．それに対しクリーンナップ問題は，協調行動の有無を数値的に表しやすい問題である．本研究では，クリーンナップ問題におけるマルチエージェントの協調獲得を確認する尺度として，以下に示す 3 つを考えた．協調をしているならば，学習の進行に伴って以下に示す尺度を最適化することとなる．

報酬獲得ステップ数均等化尺度

N_g/N_A の剰余が 0 の時，A_j が報酬を得るまでのステップ数 N_{A_j} が（ほぼ）均等になる．

衝突回数最小化尺度

1 エピソードでエージェント同士が衝突する回数 N_c が 0 に近づく．

ゴミ捨て個数均等化尺度

N_g/N_A の剰余が 0 の時，A_j が捨てるゴミの数が均等（N_g/N_A 個）になる．

6 実験結果と考察

報酬の与え方の違いによるマルチエージェント強化学習の特性を検討するために，4章で記述したクリーンナップ問題を対象に，次の3つの報酬の与え方による学習の比較を行う．

Method1 すべてのゴミがゴミ箱に捨てられたときに報酬を与える．ただし，各エージェントに対する報酬は，各エージェントが最後にゴミを捨てた時点から与えることとする．

Method2 エージェントがゴミを捨てたときに報酬を与える．

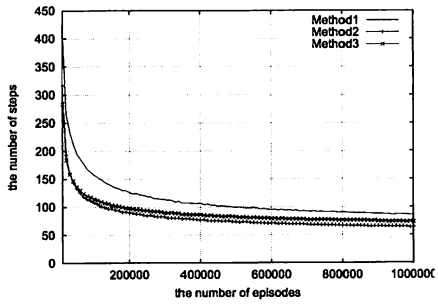
Method3 エージェントがゴミを拾ったときとゴミをゴミ箱に捨てたときに報酬を与える．

ここでは，クリーンナップ問題のパラメータを $n=10, m=3, N_{gc}=1, N_g=4,6, N_A=2,3$ と設定して実験を行った．ただし，ゴミ箱は中央付近に固定して配置し，エージェントの初期位置は $N_A=2$ のとき格子の右上と左下の端に，$N_A=3$ のときランダムに配置している．また，Profit Sharing の初期のルール重みを 0.1，報酬割引率 $\beta=0.9$，報酬値 $r=1.0$ とする．

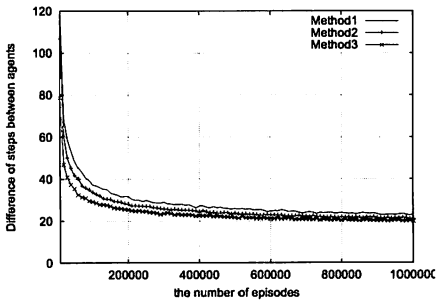
$N_A=2, N_g=4$ に設定した実験結果を図3に示す．さらに，ゴミの数を増やして $N_A=2, N_g=6$ と設定した実験結果を図4に，エージェントの数を増やして $N_A=3, N_g=6$ と設定した実験結果を図5に示す．図3から図5の（a）は，各エピソードで目標状態に達するまでに費やしたステップ数をプロットした図である．図3と図4の（b），図5の（b）から（d）は，各エージェントが最後に報酬を獲得したステップ数の差をプロットした図である．図3と図4の（c），図5の（e）は，各エピソードでエージェントが衝突した回数をプロットした図である．図3（d）は50000エピソードごとに，図4（d）は20000エピソードごとに，図5（f）は5000エピソードごとに，エージェントがゴミを（N_g/N_A）個捨てた割合をプロットした図である．クリーンナップ問題ではマルチエージェント環境である上に，ゴミ箱，ゴミ，壁が存在する．このためゴミの数やエージェントの数が増えるほど，エージェントが認識する状態の数が指数関数的に増加し，計算機のメモリ不足を引きおこして，実行不能になることもあった．

図3（a）より，副目標を達成した時点で報酬を与えるMethod2と3は，主目標を達成した時点で報酬を与えるMethod1よりも，目標状態に到達するまでに費やしたステップ数が少ない結果となっており，学習が速く進んでいることがわかる．これは図4（a），図5（a）についても同様の結果が得られている．学習が進むにつれてMethod1，2，3の差は小さくなっているが，図3（a）から観測できるように，1000000回学習を行ってもなおMethod2と3の方がステップ数が少ない結果となっている．

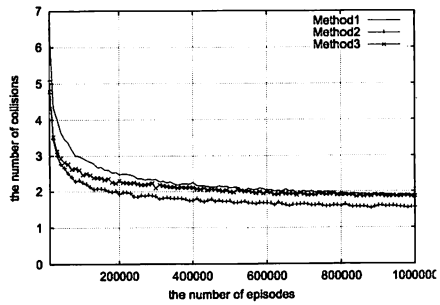
学習が早く進行しているMethod2と3がマルチエージェントシステムにおいて重要な協調行動を獲得できているか確認するために，上述した尺度に基づいて考察する．図3（b），(c）より，エージェント間のステップ数の差，エージェント間の衝突回数は学習の進行とともに徐々に減少している．図4（b），(c），図5（b），(c），(d），(e）も同様の傾向を示している．これらのことから，どのパラメータの実験結果も報酬獲得ステップ数均等化尺度，衝突回数最小化尺度を最適化する学習が行われていることがわかる．また，ほぼ全ての結果で，副目標達成時に報酬を与えるMethod2と3が，主目標達成時に報酬を与えるMethod1よりも良好な結果を示している．図3（d）より，学習の進行とともに A_j がゴミを (N_g/N_A) 個捨てた割合が上昇している．図4（d）も同様の傾向があらわれている．しかし，図5（d）から（h）は，学習が足りないために収束に向かっていない．これは，前述した図5（b），(c），(d），(e）が報酬獲得ステップ数均等化尺度，衝突回数最小化尺度を最適化する学習の傾向が見られることから，学習が進行すれば (N_g/N_A) 個捨てた割合が上昇するのではないかと考えられる．これらのことから，どのパラメータの実験結果も，学習の進行に伴いゴミ捨て個数均等化尺度を最適化する学習になっていることが確認できる．また，学習が足りないと考えられる図5（d）から（h）の結果を除いては，Method2と3がMethod1よりも高い割合を示す結果となっている．以上の結果より，副目標達成時に報酬を与えるMethod2と3は，協調行動を獲得でき，さらに主目標達成時に報酬を与えるMethod1よりも良好な学習を行っていることが確認できる．



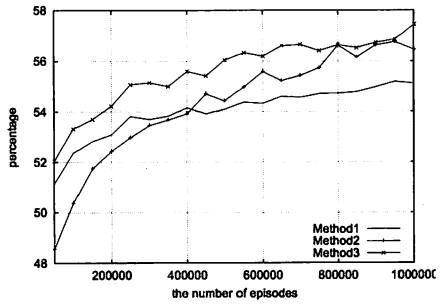
(a) 1エピソードのステップ数



(b) A_0 と A_1 の最後の N_{A_j} の差

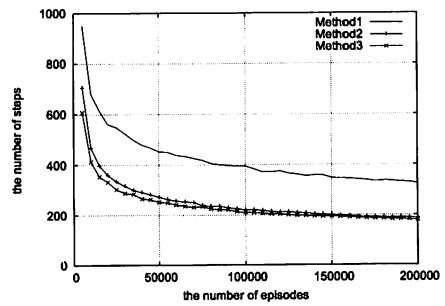


(c) エージェント間の衝突回数

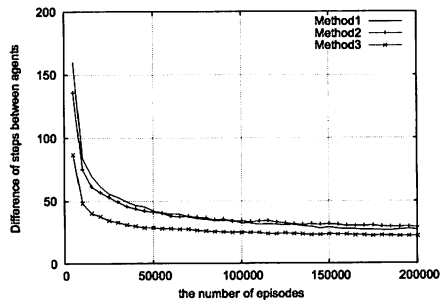


(d) A_0 が (N_g/N_A) 個のゴミを捨てた割合

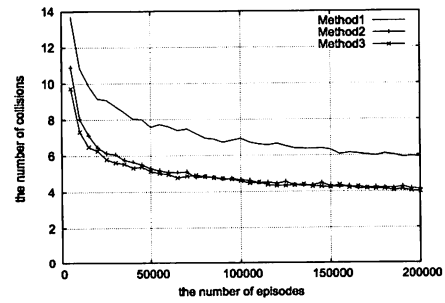
図 3: $N_g = 4$, $N_A = 2$ の実験結果（10000 エピソードごとの平均）



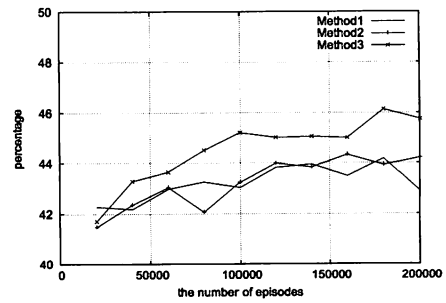
(a) 1エピソードのステップ数



(b) A_0 と A_1 の最後の N_{A_j} の差

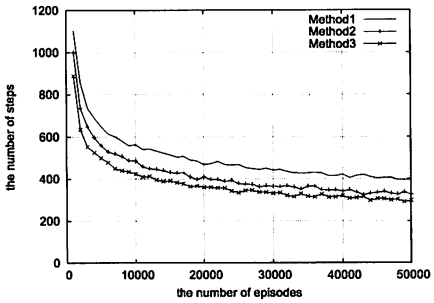


(c) エージェント間の衝突回数

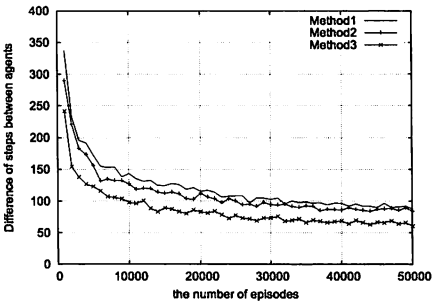


(d) A_0 が (N_g/N_A) 個のゴミを捨てた割合

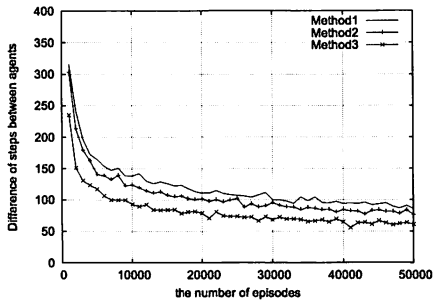
図 4: $N_g = 6$, $N_A = 2$ の実験結果（5000 エピソードごとの平均）



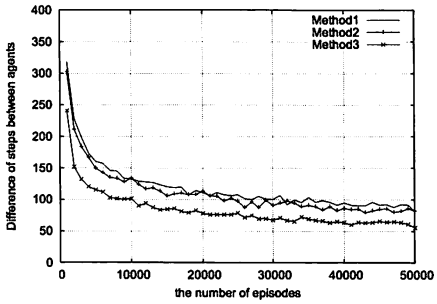
(a) 1エピソードのステップ数



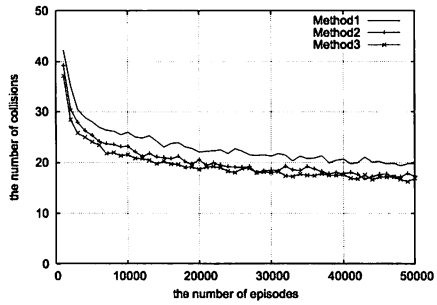
(b) A_0と A_1 の最後の N_{A_j} の差



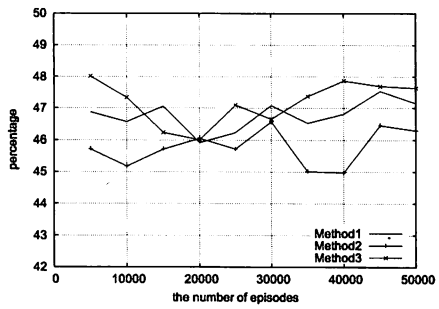
(c) A_1と A_2 の最後の N_{A_j} の差



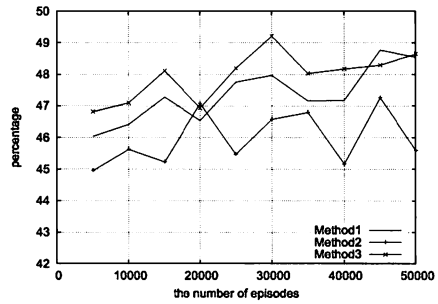
(d) A_2と A_0 の最後の N_{A_j} の差



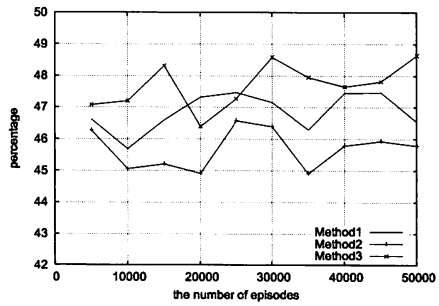
(e) エージェント間の衝突回数



(f) A_0 が (N_g/N_A) 個のゴミを捨てた割合



(g) A_1 が (N_g/N_A) 個のゴミを捨てた割合



(h) A_2 が (N_g/N_A) 個のゴミを捨てた割合

図 5: $N_g = 6, N_A = 3$ の実験結果（1000エピソードごとの平均）

7 むすび

強化学習は，大規模で複雑な環境に適応可能であり，協調行動を実現できるマルチエージェントの設計手法に対するアプローチとして注目を集めている．本研究では，これまでほとんど対象とされていなかった複数タスクを有する問題を対象にして，報酬の与え方の違いによるマルチエージェント強化学習の特性に関して検討を行った．その際，本研究で示した協調確認の尺度を用いて，一般的に確認が難しいとされている協調行動の有無を確認した．結果として，副目標を達成した時点に報酬を与える方法は，主目標を達成した時点に報酬を与える方法よりも良好な学習を行い，さらに協調も獲得できることを示した．

参考文献

[1] 荒井幸代, 宮崎和光, 小林重信, "マルチエージェント強化学習の方法論–Q-learning と Profit Sharing による接近–," 人工知能学会誌, Vol.13, No.5, pp.690–618, 1998.

[2] 荒井幸代, "マルチエージェント強化学習–実用化に向けての課題・理論・諸技術との融合–," 人工知能学会誌, Vol.16, No.4, pp.476–481, 2001.

[3] 伊藤昭, 金渕満, "知覚情報の粗視化によるマルチエージェント強化学習の高速化–ハンターゲームを例に–," 電子情報通信学会論文誌, (D-I), Vol.J84-D-I, No.3, pp.285–293, 2001.

[4] Kaelbling, L. P., Littman, M. L., and Moore, A. W., "Reinforcement Learning : A Survey," Journal of Artificial Intelligence Research, Vol.4, pp.237–285, 1996.

[5] 片山謙吾, 輿石尚宏, 成久洋之, "強化学習エージェントへの階層化意思決定方の導入–追跡問題を例に–," 人工知能学会論文誌, Vol.19, No.4, pp.279–291, 2004.

[6] 加藤新吾, 松尾啓志, "動的環境下における Profit Sharing," 電子情報通信学会論文誌, (D-I), Vol.J84-D-I, No.7, pp.1067–1075, 2001.

[7] 木村元, 宮崎和光, 小林重信, "強化学習システムの設計指針," 計測自動制御学会, 計測と制御, Vol.38, No.10, pp.618–623, 1999.

[8] 宮崎和光, 木村元, 小林重信, "Profit Sharing に基づく強化学習の理論と応用," 人工知能学会論文誌, Vol.14, No.5, pp.800–807, 1999.

[9] 宮崎和光, 荒井幸代, 小林重信, "Profit Sharing を用いたマルチエージェント強化学習における報酬分配の理論的考察," 人工知能学会誌, Vol.14, No.6, pp.1156–1164, 1999.

[10] 西智樹, 高橋泰岳, 浅田稔, "モジュール型学習機構に置ける例示の理解に基づいた自律的なタスク分解," ロボティクス・メカトロニクス講演会 '05 予稿集, Vol.CD-ROM, 2P1-S-024, 2005.

[11] 大内東, 山本雅人, 川村秀憲, "マルチエージェントシステムの基礎と応用," コロナ社, 2002.

[12] Sutton, R. S. and Barto, A. G., "Reinforcement Learning : An Introduction," The MIT Press, Cambridge, MA, 1998. (邦訳 : 強化学習, 三上貞芳, 皆川雅章 共訳, 森北出版, 2000)

[13] 高玉圭樹, "マルチエージェント学習–相互作用の謎に迫る–," コロナ社, 2003.

[14] 内部英治, 浅田稔, 細田耕, "複数の学習するロボットの存在する環境における協調行動獲得のための状態空間の構成," 日本ロボット学会誌, Vol.20, No.3, pp.281–289, 2002.

[15] 畝見達夫, "強化学習," 人工知能学会誌, Vol.9, No.6, pp.830–836, 1994.

[16] 山村雅幸, 宮崎和光, 小林重信, "エージェントの学習," 人工知能学会論文誌, Vol.10, No.5, pp.683–689, 1995.

[17] Weiss, G., "Multiagent Systems–Modern Approach to Distributed Artificial Intelligence–," The MIT Press, 1999.

On Rewards in Multi-Agent Reinforcement Learning for Multi-Task Problem

Mayumi OHTA, Toru KANESHIGE, Kengo KATAYAMA*,
Hideo MINAMIHARA* and Hiroyuki NARIHISA*

Graduate School of Engineering,
**Department of Information and Computer Engineering,*
Faculty of Engineering,
Okayama University of Science,
1-1 Ridai-cho, Okayama, 700-0005, Japan

(Received October 2, 2006; accepted November 6, 2006)

Research of most multi-agent reinforcement learning has designated on the problem of single task as the subject, but most realistic problems have contained multiple tasks. In this paper, we investigate the characteristic of difference of giving the rewards in multi-agent reinforcement learning for the multi-task problem, called the cleanup problem.

Keywords: multiagent; reinforcement learning; cooperation; reward.