岡山理科大学紀要 第42号 A pp. 115-124(2006)

マルチエージェント環境における知覚情報の分割を有する強化学習エージェントの有効性

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. まえがき

近年，家庭用の実用ロボットが登場してきた．例えば，Roomba[1]という掃除ロボットは自律的に部屋の中を動き床の掃除を行う．そのロボットは，製品では初の本格的なロボットとして注目されている[13]．そのようなロボットが活躍する場所は家庭内である．家庭内の環境は，部屋の配置，部屋にある物体の位置，人間の動作によって生じる環境の変化など，動的かつ複雑な環境である．このように，実世界において遭遇する現実的な問題の多くは，多くの不確実性を含んでいる．そのような不確実性は人間が考慮できる範疇を超越している場合が殆どである．そのため，現実的な問題に対して，人間が設計した行動群に従うエージェントを予めプログラム化することには限界がある．そのような，困難な問題に対して，近年，強化学習(Reinforcement Learning)[3]による対応が注目されている．

強化学習とは，予め設定された目標を達成するために自律エージェント自身が行った試行錯誤による行動に対して環境から報酬が与えられ，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学習制御の枠組みである．また，最近では，複雑かつ動的な環境を多くの自律的なエージェントが協調動作をすることで解決しようとするマルチエージェント環境下での研究が進められている．しかし，マルチエージェント環境での多くは人間が設計した行動群によって制御されており，対象やタスク，個体間の相互作用が複雑になるにつれ設計が困難になる．そのため，エージェント自身の学習や適応能力が求められている．このような背景からマルチエージェント環境下での強化学習法が注目されている[5][10]．

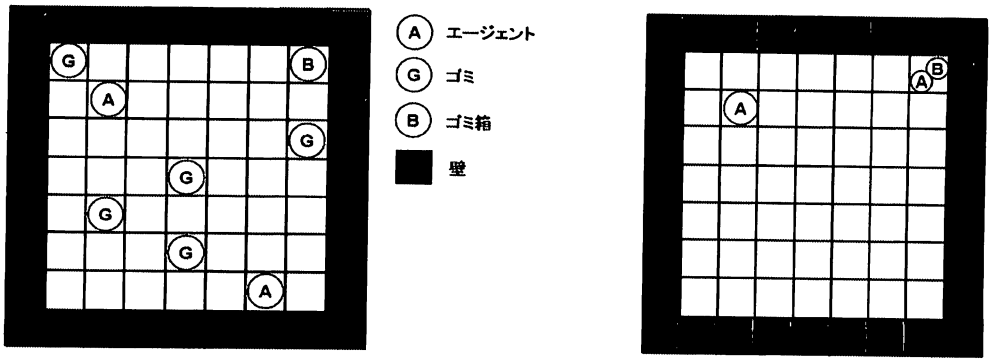
マルチエージェント強化学習において，エージェントは他のエージェントも環境の一部として観測する．そのため，エージェント数の増加に伴い知覚する状態数が指数的に増加し，知覚する状態を保存しておく記憶領域の大きさ，また，その状態数の増加に伴う学習速度の遅さ等が問題になる．そこで，その問題点を改善する手法として，知覚情報を粗視化した学習器と完全知覚の学習器の二つを並行に学習させ，学習初期では粗視化した学習器で行動選択を行い，学習途中に切り替えることで，学習後期の学習性能を劣化させることなく，学習速度を高速化する手法[8]などが提案されている．

しかしながら，上述した手法では，知覚情報を粗視化した学習器と完全知覚の学習器を用いるため，多くの記憶領域を必要とする．我々の目的は，記憶領域を小さくしつつ，学習後期での学習性能を劣化させないことである．本論文では知覚情報の分割を有する強化学習エージェントを提案するとともに，クリーンナップ問題を対象とし提案法の有効性を検討する．

2. クリーンナップ問題

クリーンナップ問題におけるエージェントは，初期状態から目標状態へ到達するために，ゴミを拾い，そのゴミをゴミ箱に捨てるといった，時系列的時間差を有する複数のタスクを解決しなければならない．そのため，クリーンナップ問題は，これまで多く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タスクの追跡問題[7][8][9][10][11]に比べ，目標状態への到達が比較的困難な問題である．

本論文では以下に基づくクリーンナップ問題を対象とする．図1(a)に示すような$n \times n$格子状の環境を設定し，ここに，m体のエージェントとk個のゴミ，j個のゴミ箱を配置する．各エージェントは同時に移



(a) 初期状態　　(b) 目標状態

図1　クリーンナップ問題

動し，上下左右斜めに一コマ進む行動を一つ選択する．格子の外枠を全て壁とし，エージェントは壁を越えることができない．ゴミ，もしくは，ゴミ箱と同座標に行く事でゴミを拾う，もしくは，ゴミを捨てることができる．なお，エージェントが持つことのできるゴミの数は一つである．また，エージェント同士は同一マスに存在することができる．本論文でのクリーンナップ問題の設定において，エージェントが協調するためには，以下の二つを最適化しなければならない．

- 報酬獲得ステップ数均等化
- ゴミ捨て個数均等化尺度

強化学習を現実問題に適用することを考えると，すべての環境を認識することは困難である．そのため，本研究では部分観測環境下のもとで実験を行い，エージェントの視界は$v \times v$とし，自らの周囲v^2マスが見える．また，エージェントは視界内で観測できる「他のエージェント」,「ゴミ」,「ゴミ箱」,「壁」の位置，及び「ゴミを所有している，所有していない」の判別を可能とする．目標状態は図1(b)に示すようにすべてのゴミをゴミ箱に捨てた状態である．

3. 強化学習

強化学習は，最適な行動を人間がエージェントに教えるのではなく，自律エージェント自身が行った試行錯誤による行動に対して，環境から報酬が与えられ，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学習制御の枠組みである．しかし，行動を実行した直後の報酬を見るだけでは，エージェントはその行動が正しいかどうか判断できないという困難を伴う．強化学習が注目を集めている理由は以下の2つである[12]．

不確実性のある環境

多くの実世界の制御問題では，予め不確実性を考慮し制御することは困難である．しかし，強化学習はエージェント自身が環境との試行錯誤を通して学習するので，不確実性を含む環境でも有効に制御することが可能となる．

離散的な状態遷移も含んだ段取り的な制御

設計者が目標状態で報酬を与えるという形で，タスクをエージェントに指示しておくことで，ゴールへの到達方法はエージェントの試行錯誤によって自律的に獲得される．すなわち，設計者がエージェントに「何をすべきか」を報酬という形で指示おくことで，エージェント自身が「どのように実現するか」を学習によって獲得する枠組みである．

強化学習における研究は，主に「環境のクラス」と「接近の指向性」の二つの観点から分類される[16]．「環境のクラス」では，状態遷移がマルコフ的であるか否かで特徴付けられ，現在の状態から未来の状態が予

測可能なマルコフ決定過程 (Markov Decision Processes, MDP) と MDP としてのモデル化が困難である非 MDP 環境が知られている．一方,「接近の指向性」は，最適性を重視する環境同定型と，学習途中においてもなるべく報酬を得続けるという効率性を重視する経験強化型が知られている．環境同定型の代表例として Q-Learning[4]，経験強化型の代表例として Profit Sharing[2] が挙げられる．本論文では，Profit Sharing 強化学習を用いる．

3.1 Profit Sharing 強化学習

Profit Sharing は，遺伝的アルゴリズム (genetic algrithm, GA) を併用するクラシファイアシステム (classifier system) での信用割り当て (credit assignment) の方法として 1980 年代後半に提唱された[9]．現在，Profit Sharing は GA だけではなく強化学習の枠組みにおいても利用可能であり，さらに，現在の状態から未来の状態のモデル化が困難である非 MDP 環境（マルチエージェント環境）においても有用であると期待されている[6]．また，Profit Sharing は他の強化学習手法に比べ，学習の立ち上がりが素早く，不完全知覚状態に対しても有効であることが示されている[18][19]．これらの理由から，本論文では Profit Sharing を学習手法として用いる．

3.2 Profit Sharing の合理性定理

Profit Sharing は，報酬に至るまでに使用された状態 s と実際に行った行動 a のルール系列を記憶しておき，報酬を得たときにそれまでのルール系列の評価値 $w(s,a)$ をエピソード単位で強化する手法である．エピソードとは，初期状態あるいは報酬を得た直後から次の報酬までのルールの選択系列のことである．

次式を用いて状態と行動の組に対する評価値 $w(s_t,a_t)$ を更新する．ここで，$w(s_t,a_t)$ はエピソード系列上の t 番目の評価値，r は報酬値，f は強化関数である．

$$w(s_t,a_t) \leftarrow w(s_t,a_t) + f(r,t)$$

あるエピソードで，同一の感覚入力（状態）に対して異なるルールが選択されているとき，その間のルールを迂回系列という．常に迂回系列上にあるルールを無効ルールと呼び，それ以外を有効ルールと呼ぶ．無効ルールと有効ルールが競合するならば，無効ルールを強化すべきでないと考えられる．また，宮崎らによって，政策の局所的合理性を保証する必要条件である合理性定理が証明されている[17]．ここで，政策とは，ある状態において実行可能な行動の中で何が適切であるかを示すものである．

$$f(r,t) = \frac{1}{S} f(r,t-1),\ t = 1,2,\cdots N-1.$$

N はエピソードの最大長，S は報酬割引率である．なお，報酬割引率は $S \geq L+1$ とする（L は同一感覚入力下に存在する有効ルールの最大個数である）．Profit Sharing の合理性定理は，最適性を保証していないが，MDP の過程を必要としないためマルチエージェント環境のような非 MDP 環境に対しても適用できる点に特徴がある．

3.3 ルーレット選択法

Profit Sharing の学習過程における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．ルーレット選択法は，ある状態 s において，各行動の評価値 $w(s,a_t)$ を全ての行動の評価値の合計 $\sum w(s,a_t)$ で除算し，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．また，ルーレット選択法は確率的に政策を自然に実現する枠組みであり，非 MDP 環境における行動選択法として有効である[6][9]．以上の理由から本論文では，行動選択法としてルーレット選択法を使用する．

$$P(a_t|s) = w(s,a_t) / \sum w(s,a_t)$$

3.4 マルチエージェント環境における強化学習適用の際の諸問題

強化学習はマルチエージェント環境に対応できる手法として注目されている．しかしながら，マルチエージェント環境における強化学習適用の際にはシングルエージェント環境に無い様々な問題が生じる．以下にその代表例を示し簡潔に述べる．

- 同時学習
  同時学習とは，複数のエージェントによる相互の政策の変更，学習途中で定常な政策を持たないことで，自己の行動による状態遷移先を変動させることにより生じる問題である[5][15]．

- 報酬配分問題
  マルチエージェント環境ではエージェント間の連帯行動に対する報酬は定義できるが，各エージェントの個別の行動に対する報酬を定義することは困難である．目標達成のために多大な貢献をしたエージェントとあまり貢献しなかった（できなかった）エージェントに同値の報酬を与えることは多くの場合適切ではないと考えられている．例えば，ランドマークプロジェクトである RoboCup サッカーのような環境において，相手ゴールまでのパスを考える．

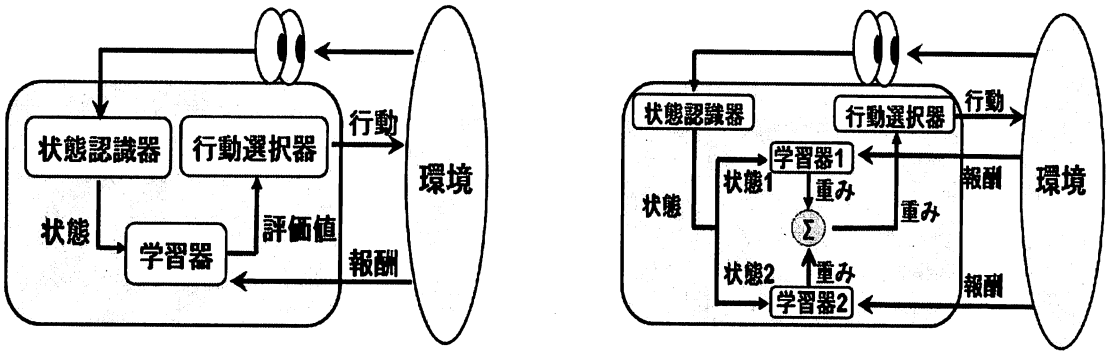
$$\text{Agent1} \rightarrow \text{Agent2} \rightarrow \cdots \rightarrow \text{AgentG} \Rightarrow \text{GOAL}$$

  ここで，直接報酬（ゴールを決めた）に寄与した AgentG のみに報酬を与えると，AgentG 以外は全く学習しないことは明らかである．また，すべてのエージェントに報酬を与えた場合，GOAL までに関係ない行動までも強化してしまう恐れがある．そのため，報酬配分問題について様々な研究が進められている[11)14)20)]．

- 状態爆発
  マルチエージェント環境では，エージェントは他のエージェントも環境の一部として観測する．そのため，エージェント数の増加に伴い知覚する状態数が指数的に増加し，知覚する状態を保存しておく記憶領域の大きさ，また，その状態数の増加に伴う学習速度の遅さ等が問題になる．さらに，現実問題のような複雑な環境に対し強化学習を適用することを考えると，エージェントが知覚する状態数は膨大になると考えられる．そのため強化学習をそのような問題に適用することすら困難になる．これらの事から我々は，状態爆発の問題点に着目し，状態爆発を改善する手法を提案する．

4. 提案する強化学習エージェントの構成

まず，一般的な強化学習エージェント（従来の強化学習エージェント）の構成図を図 2（a）に示す．強化学習におけるエージェントは，状態認識器，学習器，行動選択器の三つのモジュールを用いることで，環境との相互作用を通して学習する．強化学習は，環境から観測可能な状態を認識し，その状態における各行動評価値に基づき行動を選択する．行動した結果，環境が目標状態であれば報酬を与え，各行動評価値を更新する．エージェントは報酬を得た後，再び状態認識，行動といったサイクルを繰り返し，同じ状態に至ったとき，その更新された各行動評価値を行動選択に利用する．強化学習エージェントはそのサイクルを繰り返すことで環境に適応していく．



(a) 強化学習エージェントの構成図　　(b) 提案手法の強化学習エージェントの構成図

図 2　各強化学習エージェントの構成図

図 2(b) は我々が提案する知覚情報の分割を有する強化学習エージェントの構成である．エージェントは従来の強化学習と同様に環境から状態を認識する．知覚する状態数を削減するために知覚した状態を分割する．そして，分割したそれぞれの状態を別々の学習器を用いて学習を行う．行動選択は，各学習器からの各行動の評価値を加算し，その重みに基づき行動選択を行う．行動した結果，環境が目標状態になれば各学習器に報酬を与え評価値を更新する．

4.1 状態認識

4. 節で述べたように，強化学習は各状態に対する可能な行動の評価値をもとに行動を選択し，その評価値を更新することで学習を行う．すなわち，強化学習において状態認識は重要な 1 つの役割を担っている．以下では，従来法の状態認識と提案法の状態認識について述べる．

従来法の状態認識

図 3 は従来の強化学習エージェントの状態認識である．エージェントは，格子に振り分けられた格子番号によって物体の位置を認識する．

本論文におけるクリーンナップ問題の設定 (2. 節参照) では，エージェントは視界内で観測できる「他のエージェント」,「ゴミ」,「ゴミ箱」,「壁」の位置を格子状に振り分けられた番号によって認識し,「ゴミを所有している，所有していない」の判別を可能とする．すなわち，エージェントはそれらの番号の組み合わせによって状態を認識する．エージェントは認識した状態を入力とし，この状態に対する各行動（上下左右斜め）の評価値を出力し行動選択に用いる．

提案法の状態認識

図 4 は提案法の状態認識である．状態 1 は，グリッド上の列を認識し，状態 2 は，グリッド上の行を認識する．また，図 4(a) は Alpha の範囲が 1 マスであり，エージェントの周辺を細かく認識することができる（以後 Division 1 と記す）．図 4(b) は Alpha の範囲が図 4(a) の 3 倍（3 マス）であり，Division1 に比べ，Alpha の範囲は粗く認識するが，Alpha 以外の範囲を細かく認識することができる (以後 Division 3 と記す)．このように提案法は，Alpha の範囲を変更することで様々な状態認識を行うことができる．また，従来法の状態認識と同様にエージェントは，格子に振り分けられた格子番号によって物体の位置を認識し，その番号の組み合わせによって状態を認識する．エージェントは認識した状態を入力とし，この状態に対する各行動の評価値を出力し行動選択に用いる．

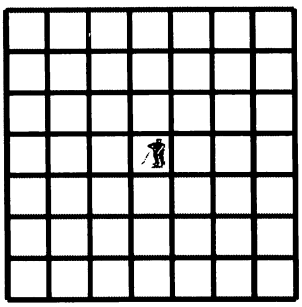
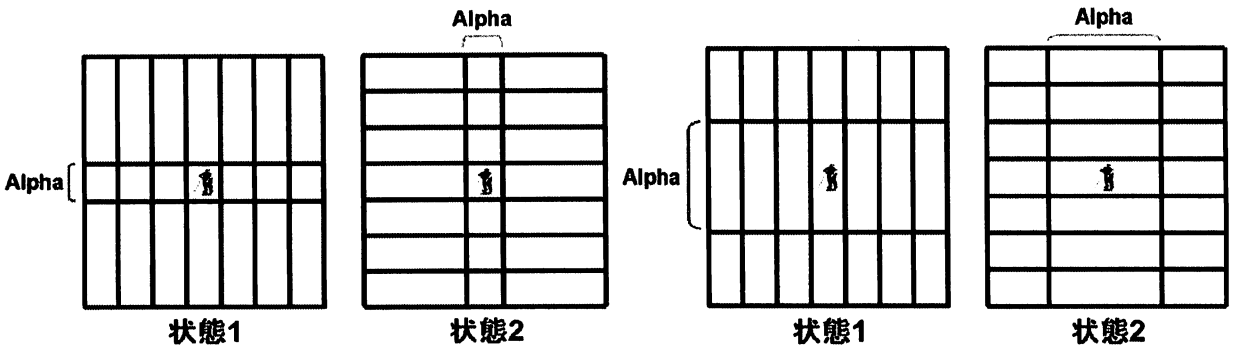


図 3　知覚情報の分割を行わない状態認識 (従来法)



(a) 知覚情報の分割を行う状態認識 (Division 1)　(b) 知覚情報の分割を行う状態認識 (Division 3)

図 4　知覚情報の分割を有する状態認識

5. 実験

本節では，提案法の有効性を検討するために，2. 節に示した基本設定に基づいたクリーンナップ問題を対象とし，様々な問題設定のもとで実験を行い考察する．

5.1 実験設定

実験設定は次のようにする．エージェントの視覚サイズを 7×7(エージェントの周辺 7^2 マスを知覚する)とし，Profit Sharing 強化学習のパラメータ設定は，報酬値を 1.0，初期ルールの評価値を 0.1，公比 0.9 の等比減少関数とする．また，報酬はすべてのゴミがゴミ箱に捨てられた際に発生するが，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ものとする．

クリーンナップ問題の設定は，エージェント数を二体，ゴミ箱を一つとする．知覚する状態数による学習性能を考察するためにゴミの数を一つから五つまでの環境で実験を行う．なお，エージェントが知覚する状態数は，ゴミの数が一つの環境が最も少なく，ゴミの数が五つの環境が最も多い．すなわち，状態数の観点から言えばゴミの数が一つの環境の場合，知覚する状態数が少なく学習を行いやすい．各エピソード毎にエージェント，ゴミ箱，ゴミをそれぞれランダムに配置する．これを初期状態とする．ここで，エピソードとは初期状態から目標状態へ到達するまでの期間を指す．

5.2 実験結果と考察

各実験環境 (ゴミの数が一つから五つまでの環境) に対する学習の試行回数を 5 試行とし，1 試行の学習回数を 100 万エピソードとする．

まず，各実験環境において，従来法 (Not Division) と提案法 (Division 1 及び Division 3) におけるエージェントが知覚した状態の数（状態数）を表 1 に，従来法の知覚した状態数を基準とした，提案法との知覚した状態数の割合を表 2 に示す．

表 1　各実験環境における知覚した状態数

Environment	Not Division	Division 1	Division3
ゴミの数一つ	270148.2	140894.8	163991.8
ゴミの数二つ	712983.6	384112.8	472906.6
ゴミの数三つ	1398637.4	761106.8	965235.6
ゴミの数四つ	2495271.0	1320420.8	1665400.6
ゴミの数五つ	3828403.6	2058096.6	2544919.2

表 2　各実験環境における知覚した状態数の割合

Environment	Not Division	Division 1	Division3
ゴミの数一つ	1.000	0.522	0.607
ゴミの数二つ	1.000	0.539	0.664
ゴミの数三つ	1.000	0.544	0.690
ゴミの数四つ	1.000	0.529	0.667
ゴミの数五つ	1.000	0.538	0.665

表 1 の結果から，ゴミの数が増加するに従いエージェントの知覚する状態数が増加することが分かる．また，表 2 の結果から，すべての環境において，Division1，Division3 ともに従来法に比べ，状態数が約半分まで減少していることが観測できる．この事から，提案法は記憶領域の面において有効である．

次に，各実験環境に対する学習性能を図 5 から図 9 に示す．なお，各図ともに，縦軸は目標達成に到達するまでの行動数（ステップ数），横軸はエピソード数である．また，表 3 に学習後期における 500 エピソード分の平均ステップ数を示す．

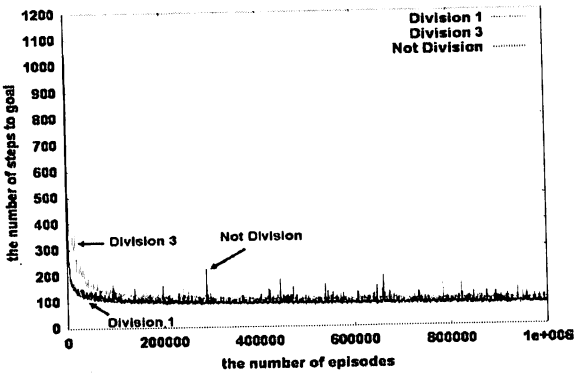


図5　ゴミの数一つの環境における学習結果

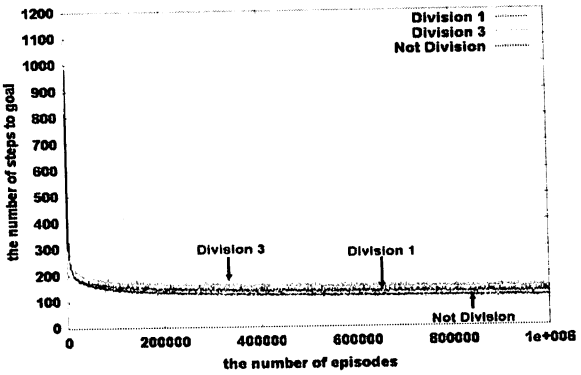


図6　ゴミの数二つの環境における学習結果

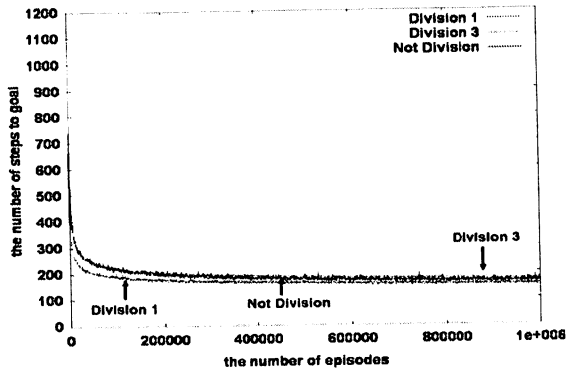


図7　ゴミの数三つの環境における学習結果

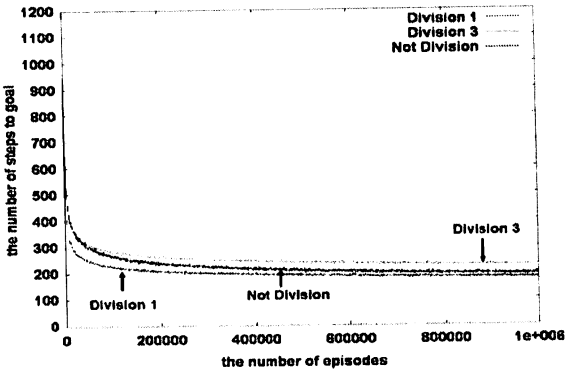


図8　ゴミの数四つの環境における学習結果

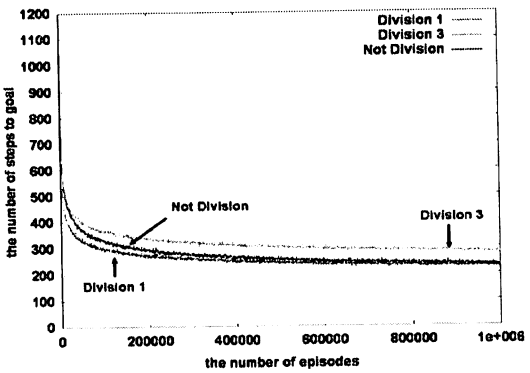


図9　ゴミの数五つの環境における学習結果

表3　各実験環境における学習後期の平均ステップ数

Environment	NotDivision	Division 1	Division 3
ゴミの数一つ	89.03960	**79.9080**	99.4916
ゴミの数二つ	**111.20240**	133.05640	147.0972
ゴミの数三つ	156.65360	**152.08440**	170.5556
ゴミの数四つ	189.49960	**175.03000**	222.7168
ゴミの数五つ	228.20160	**225.16000**	283.3848

これらの結果から，Division 1 は従来法に比べ学習初期の段階から立ち上がりが早く，さらに，表 3 の結果から学習後期においてもゴミの数が二つの環境を除き，良好な性能を示している．この事から Division 1 は有効であると考えられる．しかしながら，Division 3 は従来法及び Division 1 に比べ，状態数が増加する (ゴミの数が増加する) に従い，学習性能の劣化が顕著に現れている．これは，4.1 節の図 4 に示すように，Division 1 は Alpha の値が 1 マスであり，エージェントの周辺を細かく知覚できるのに対し，Division 3 は Alpha の値が 3 マスであり，エージェントの周辺を粗く知覚したため学習性能に悪影響を与えたと考えられる．

また，図 5 の結果から，従来法はゴミの数一つの環境において，ステップ数にばらつきが多いことが分かる．この理由を考察するために，従来法と提案法における，ゴミを捨てた回数が多い Agent(MTA) と，ゴミを捨てた回数が少ない Agent(LTA) 別の，各環境でのゴミ捨て均等化尺度を表 4 から表 6 に示す．この結果から，ゴミの数が一つの環境において，従来法の MTA，LTA のゴミを捨てた割合の差が提案法に比べ

少ないことが分かる．すなわち，従来法は，提案法に比べ両エージェントが均等に目標状態に貢献した（報酬を得た）こととなる．しかし，ゴミの数が一つの環境では，他の環境に比べ，各エージェントが報酬を得る機会が少ない．さらに，従来法は状態数が多いため，多くの学習回数を必要とする．そのため，ステップ数にばらつきが生じたと考えられる．

表 4　従来法におけるゴミ捨て均等化尺度

環境	環境においてゴミを捨てる回数	MTA のゴミを捨てた回数	LTA のゴミを捨てた回数	ゴミを捨てた回数の差	MTA のゴミを捨てた回数の割合	LTA のゴミを捨てた回数の割合	割合の差
ゴミの数一つ	1000000	565728.0	434272.0	131456.0	0.565	0.434	0.131
ゴミの数二つ	2000000	1096034.8	903965.2	192069.6	0.548	0.451	0.097
ゴミの数三つ	3000000	1553358.0	1446642.0	106716.0	0.517	0.482	0.035
ゴミの数四つ	4000000	2114311.2	1885688.8	228622.4	0.528	0.471	0.057
ゴミの数五つ	5000000	2598720.4	2401279.6	197440.8	0.519	0.480	0.039

表 5　Division 1 におけるゴミ捨て均等化尺度

環境	環境においてゴミを捨てる回数	MTA のゴミを捨てた回数	LTA のゴミを捨てた回数	ゴミを捨てた回数の差	MTA のゴミを捨てた回数の割合	LTA のゴミを捨てた回数の割合	割合の差
ゴミの数一つ	1000000	599741.4	400258.6	199482.8	0.599	0.400	0.199
ゴミの数二つ	2000000	1091935.0	908065.0	183870.0	0.545	0.454	0.091
ゴミの数三つ	3000000	1583270.4	1416729.6	166540.8	0.527	0.472	0.055
ゴミの数四つ	4000000	2117912.8	1882087.2	235825.6	0.529	0.470	0.059
ゴミの数五つ	5000000	2577173.8	2422826.2	154347.6	0.515	0.484	0.031

表 6　Division 3 におけるゴミ捨て均等化尺度

環境	環境においてゴミを捨てる回数	MTA のゴミを捨てた回数	LTA のゴミを捨てた回数	ゴミを捨てた回数の差	MTA のゴミを捨てた回数の割合	LTA のゴミを捨てた回数の割合	割合の差
ゴミの数一つ	1000000	576259.2	423740.8	152518.4	0.576	0.423	0.153
ゴミの数二つ	2000000	1120378.4	879621.6	240756.8	0.560	0.439	0.121
ゴミの数三つ	3000000	1593334.4	1406665.6	186668.8	0.531	0.468	0.063
ゴミの数四つ	4000000	2101341.4	1898658.6	202682.8	0.525	0.474	0.051
ゴミの数五つ	5000000	2593682.4	2406317.6	187364.8	0.518	0.481	0.037

6. むすび

実世界において遭遇する現実的な問題の多くは，複雑かつ動的な変化を伴うような非常に複雑な環境である．強化学習はそのような問題に対して対応できる手法として注目されている．しかし，そのような問題に対して強化学習を適用する際，エージェントの知覚情報が膨大になり，非常に多くの学習回数，もしくは，学習が進行しないという問題点がある．そこで本論文では，その問題点を改善するために，知覚情報の分割を有する強化学習エージェントを提案し，クリーンナップ問題を対象とし提案法の有効性を検討した．

従来法との比較実験の結果，提案法は記憶領域の面では従来法の約半分であった．さらに，エージェントの周辺を詳細に知覚した Division 1 は学習性能においても従来法に比べ，高速に学習を行うことができ，学習後期においても良好な結果を示した．これらの事から，Division 1 は有効であると考えられる．また，エージェントの周辺を粗く知覚した Division 3 は，学習性能の劣化を招く結果であった．この事から，エージェントの周辺を粗く知覚した場合，学習性能に悪影響をもたらすことを明らかにした．

参考文献

1) http://www.irobot.com

2) Grefenstette, J. J., "Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, " Machine Learning, Vol.3, pp.225–245, 1988.

3) Richard S. Sutton and Andrew G. Barto. "Reinforcement Learning: An Introduction," 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).

4) Watkins, C. J. C. H., and Dayan, P., "Q-learning," Machine Learning, Vol.8, pp.279–292, 1992.

5) 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.

6) 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.609–618, 1998.

7) 石原聖司, 五十嵐治一, "マルチエージェント系における行動学習への方策勾配法の適用 —追跡問題—," 電子情報通信学会論文誌, D-1, Vol.J87-D-1, No.3 pp. 390–397, 2004.

8) 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-1, Vol.J84-D1, No.3, pp.285–293, 2001.

9) 片山謙吾, 奥石尚宏, 成久洋之, "強化学習エージェントへの階層化意志決定法の導入 — 追跡問題を例に —," 人工知能学会論文誌, Vol.19, No.4, pp.279–291, 2004.

10) 加藤新吾, 松尾啓志, "動的環境下における Profit Sharing," 電子情報通信学会誌, D-1, Vol.J84-D-1, No.7, pp.1067–1075, 2001.

11) 金重 徹, 奥石尚宏, 片山謙吾, 成久洋之, "マルチエージェント強化学習の報酬分配に関する実験的考察," 平成 16 年度電気・情報関連学会中国支部第 55 回連合大会講演論文集, pp.380, Oct. 16, 2004.

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

13) 小林一樹, 山田誠二, "行為に埋め込まれたコマンドによる人間とロボットの協調," 人工知能学会論文誌, Vol21, No.1, pp.63–72, 2006.

14) 森山甲一, 沼尾正行, "環境状況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," 人工知能学会論文誌, Vol.17, No.6, pp.676–683, 2002.

15) 山口智浩, "人工知能分野における強化学習研究の広がり," 人工知能学会全国大会論文集, Vol.JSAI02, pp.89–90, 2002.

16) 山村雅幸, 宮崎和光, 小林重信, "エージェントの学習," 人工知能学会誌, Vol.10, No.5, pp.23–29, 1995.

17) 宮崎和光, 山村雅幸, 小林重信, "強化学習における報酬割当ての理論的考察," 人工知能学会誌, Vol9, No4, pp.580–587, 1993.

18) 宮崎和光, 小林重信, "離散マルコフ決定過程下での強化学習," 人工知能学会誌, Vol.12, No6, pp.3–13, 1997.

19) 宮崎和光, 木村 元, 小林重信, "ProfitSharing に基づく強化学習の理論と応用," 人工知能学会誌, Vol.14, No.5, pp.800–807, 1999.

20) 宮崎和光, 荒井幸代, 小林重信, "ProfitSharing を用いたマルチエージェント強化学習における報酬配分の理論的考察," 人工知能学会誌, Vol.14, No.6, pp.1156–1164, 1999.

Effectiveness of Reinforcement Learning Agent with the Division of Perception Information for Multi-Agent Environment

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)

Reinforcement learning (RL) is known to be a promising technique for creating agents that can be applied to multi-agent environment in real world problems. When multi-agent RL algorithms are applied to such complex problems, the amount of perception information in the algorithms increases enormously, and the large learning times are required. In this paper, we present a technique that divides the perception information in order to reduce the learning times. We evaluate the profit sharing based reinforcement learning algorithm with the technique for the clean-up problem in the multi-agent environment. The results show that our proposed method is effective and can adapt in the multi-agent environment in faster learning time than traditional method.

Keywords: reinforcement learning; clean-up problem; multi-agent.