岡山理科大学紀要 第40号 A pp. 129-136(2004)

Kheperaロボットを用いた強化学習手法の比較

馬場 安彦・片山 謙吾*・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
(2004年9月30日受付、2004年11月5日受理)

1. まえがき

現実世界には宇宙空間や深海，被災地など人間では作業し難い環境が多々ある．そのような環境でロボットに人間の代わりとして作業させることを目的とした研究が現在盛んに行われている[5)19)]．ロボットのほとんどは設計者によって与えられた制御則に従って行動する．この制御則をロボットに与えることは設計者がその環境を熟知していることが前提である．しかし，設計者がその前提を満たすことは困難である場合が多い．そこで設計者がロボットに制御則を与えるのではなく，ロボットが環境に適応した制御則を自律的に獲得する手法として強化学習 (Reinforcement Learning)[7)8)9)10)13)14)15)17)18)] が注目を集めている．

強化学習は，学習者 (エージェント) が試行錯誤を通して環境に適応する学習制御の枠組みである．従来扱われてきた強化学習の問題はマルコフ決定過程 (Marcov Decision Process:MDPs)[1)] であり離散的な環境が多かった．しかし，現実世界のほとんど問題は非 MDP で連続的な環境である．ロボットを扱う環境は現実世界であるため，強化学習をロボットに適用するためには非 MDPs の問題を対象とした手法が必要不可欠であると考えられる．強化学習の手法として環境同定型と経験強化型が提案されている．代表的な手法として環境同定型の Q-learning[16)] と経験強化型の Profit Sharing[2)12)] が知られている．

本研究は，超小型移動ロボット (KheperaII) に強化学習を適用し，ロボットが環境に適応した制御則を自律的に獲得することを目標としている．実ロボットの学習には多大な時間が必要であるため，実機を用いる前段階としてシミュレータ (Webots: KheperaII を高い水準でシミュレートする) を用いる．本論文では，シミュレータを用いてエージェントに強化学習の代表的な手法である Q-learning と Profit Sharing を適用し，迷路問題を対象として両学習法を比較検討する．

本論文は，第2章を強化学習の概要，第3章を強化学習手法の紹介，第4章をエージェントに強化学習手法を適用した実験，第5章をむすびとする．

2. 強化学習

強化学習とは，エージェントは図1のように環境との相互作用を繰り返し，環境に適応する学習制御の枠組である．教師付き学習とは異なり，状態入力に対する正しい行動出力を明示的に示す教師が存在しない．エージェントは教師のかわりに報酬というスカラーの情報を手がかりに学習するが，報酬にはノイズや遅れがある．そのため，行動を実行した直後の報酬をみるだけでは，エージェントはその行動が正しかったかどうかを判断できないという困難を伴う．

ここでは，強化学習の学習の主体となるエージェントとマルコフ決定過程について説明する．

2.1 学習者 (エージェント)

エージェントは予め環境に関する知識を持たず，状態遷移を繰り返し，やっと目標にたどり着くような段取り的な行動を行う．

エージェントは図2のように3つのモジュールにより構成されている．状態認識器はエージェントが現在存在する状態を認識する．そして状態認識器から学習器に現在の状態情報を渡す．学習器は強化学習を適用するモジュールである．学習器には各状態における行動の重みが蓄えられている．そして学習器から行動

選択器に重みを渡す．行動選択器はエージェントの次の行動の選択をする．

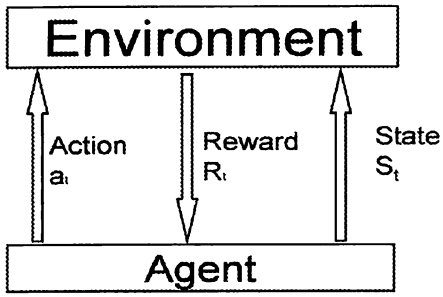

図 1　エージェントと環境の関係

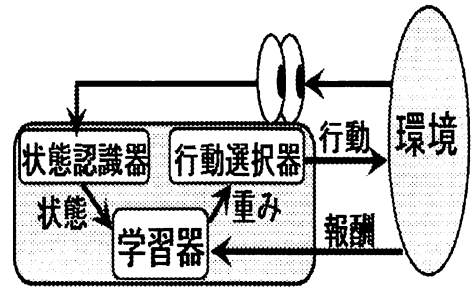

図 2　エージェントの構成

2.2　マルコフ決定過程 (Markov Decision Process:MDPs)

マルコフ決定過程 (Markov Decision Process:MDPs) とは，現在の状態 $s(t)$ が一つ前の状態 $s(t-1)$ と行動 $a(t-1)$ にのみ依存し，それ以前の過去の状態と行動に依存しないことである．グリッド環境を例に挙げて詳しく説明する．グリッド環境とは図 3 のように格子状 (マス) で区切られた空間のことある．エージェントは図 3 で上下左右に移動できるとする．MDPs の場合，図 4 のようにエージェントが現在いるマスに移動する以前のマスは必ず上下左右のマスのうちのどれかである．しかし，図 5 のようにエージェントが風の影響を受け次の状態が予測できない状況などは非 MDPs である．要するに MDPs ではエージェントの現在の状態とエージェントの行動則から前の状態が予測できるが，非 MDPs ではエージェントに他の影響が加わり現在の状態とエージェントの行動則から前の状態が予測できないということである．また，エージェント自身において摩擦や認識のずれなどがある場合も非 MDPs になる．すなわち，エージェントが移動するような現実世界の問題の多くは非 MDPs であると考えられる．

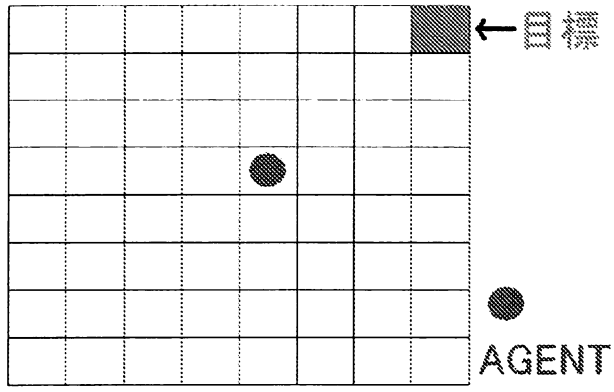

図 3　グリッド環境

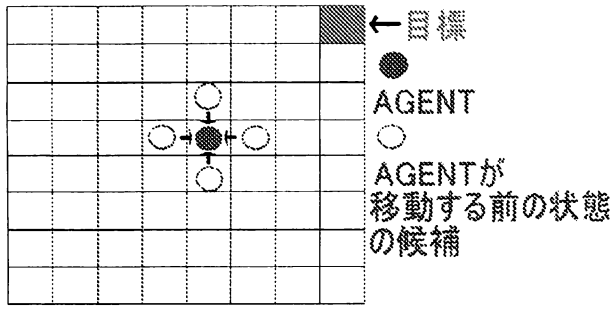

図 4　MDPs

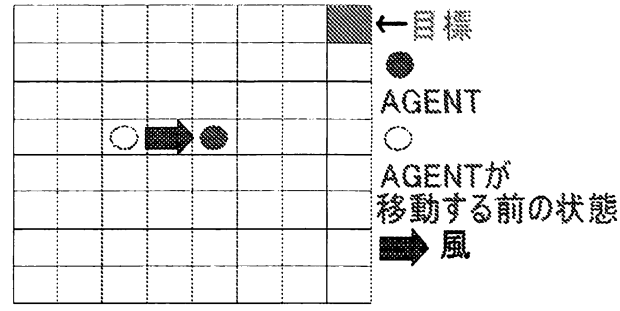

図 5　非 MDPs

3. 強化学習手法

ここでは，強化学習の手法である環境同定型と経験強化型，そして強化学習でよく用いられる行動を選択する手法について説明する．

3.1 環境同定型

環境同定型は環境をすべて探索することで最適解を導き出す．しかしその前提としてMDPsを満たしていなければならない．また最適解を導くには環境すべてを探索する必要があるので学習時間は膨大となる．

環境同定型に属する手法としてTD学習，そしてTD学習を発展させたQ-learningとActor-Criticなどがある．次の節よりそれらの手法について説明する．

TD学習

TD学習 (Temporal Difference Learning) は，経験から直接学習し，目標到達しなくても次の状態の行動価値 $V(s_{t+1})$ により現在の行動価値 $V(s_t)$ を更新する．以下の更新式を用いて行動価値 $V(s_t)$ を更新する．

$$V(s_t) \leftarrow (1-\alpha)\, V(s_t) + \alpha\,(r_{t+1} + \gamma V(s_{t+1}))$$

ここで t は現在の時間，s_t は現在の状態，s_{t+1} は次の状態，r_t は環境から得られる報酬，$\alpha\,(0 < \alpha \leq 1)$ は学習率，$\gamma\,(0 \leq \gamma < 1)$ は減衰率である．

Q-learning

Q-learningは，現在の行動価値 $Q(s,a)$ を現在の状態から遷移可能な状態の最大行動価値を減衰した値を反映させ強化する手法である．環境との試行錯誤による相互作用の繰り返しを通して行動価値 $Q(s,a)$ を推定する．次式を用いて行動価値 Q を更新する．

$$Q(s,a) \leftarrow (1-\alpha)\, Q(s,a) + \alpha \left(r + \gamma \max_{a' \in A} Q(s',a') \right)$$

ここで s は現在の状態，a は現在の状態における行動，s' は次の状態，a' は次の状態における行動の候補，r は環境から得られる報酬，$\alpha\,(0 < \alpha \leq 1)$ は学習率，$\gamma\,(0 \leq \gamma < 1)$ は減衰率である．

Actor-Critic

Actor-Critic[6)11)] は，行動を司るActor部と評価を司るCritic部に分かれている．Actor部で行動を選択し，Critic部で行動の評価を行う．以下の更新式を用いてActor部で行動優先度 $P(s_t,a_t)$，Critic部で状態評価値 $V(s_t)$ を更新する．行動優先度及び状態評価値の更新はエージェントが行動する度に行われる．Actor-Criticの学習モデルは図6である．

行動優先度

$$TD - Error = r_t + \gamma\ V(s_{t+1}) - V(s_t)$$

$$P(s_t,a_t) \leftarrow P(s_t,a_t) + \alpha\ TD - Error$$

状態評価値

$$V(s_t) \leftarrow (1-\alpha)\, V(s_t) + \alpha\,(r_{t+1} + \gamma V(s_{t+1}))$$

ここで s_t は状態，r_t は環境から得られる報酬，a_t は選択された行動，$gamma\,(0 \leq \gamma < 1)$ は減衰率，$\alpha\,(0 < \alpha \leq 1)$ は学習率，$P(s_t,a_t)$ は行動優先度，$V(s_t)$ は行動前の状態評価値，$V(s_{t+1})$ は行動後の状態評価値，$TD - Error$ はTD誤差である．行動優先度は，状態 s_t で行動 a_t のそれぞれを選択する (優先させる) 傾向を与える値である．TD誤差は，選択された最新の行動 a_t を評価するのに使われる．ある行動に対しTD誤差が正の場合ならその行動を選択する傾向を強め，負の場合ならその行動を選択する傾向を弱める．

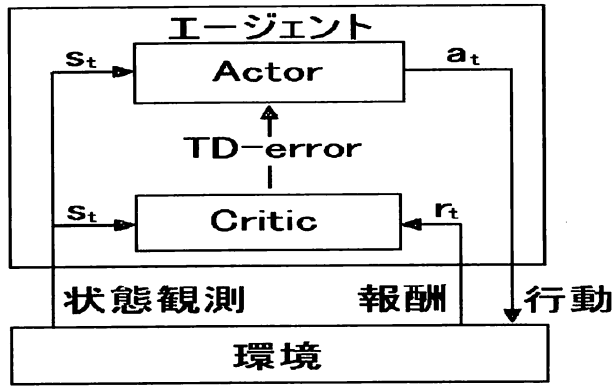

図6 Actor-Critic のモデル

3.2 経験強化型

報酬を獲得できる行動を優先して選択するため最適性は保障されない．ただし，環境同定型に比べて非MDPs の場合でも学習しやすく，学習速度が速い．

Profit Sharing

Profit Sharing は，報酬を得たときにそれまでに使用した状態行動対 s_t,a_t を一括して強化する手法である．次式を用いて行動価値 W を更新する．

$$W(s_t,a_t) \leftarrow W(s_t,a_t) + f(t,r_T,T)$$

$$f(t,r_T,T) = \beta^{T-t-1} r_T$$

ここで f は強化関数と呼ばれる関数であり，r は報酬，$\beta\,(0 \leq \beta \leq 1)$ は減衰率，T は報酬が発生した時刻である．

3.3 行動選択法

行動選択法とは，エージェントの行動選択器を司る部分である．上述した Q-learning と Profit Sharing にはよく用いられる行動選択法がある．ここではその行動選択法を説明する．

ε-greedy 選択法

Q-learning では，行動選択法として ε-greedy 選択法がよく用いられる．ε-greedy 選択法とは，$\epsilon\,(0 \leq \epsilon \leq 1)$ の確率でランダムに行動を選択し，それ以外の $(1-\epsilon)$ の確率では，現在の状態のおいて最大の評価値を持つ行動を選択する方法である．

ルーレット選択法

Profit Sharing では，行動選択法としてルーレット選択法がよく用いられる．ルーレット選択法は，ある状態 s における各行動価値 $W(s,a_i)$ を全行動価値の合計 $\Sigma_a W(s,a)$ で割り，確率を求め，その確率により行動を選択する方法である．

$$P(a_i|s) = W(s,a_i) / \Sigma_a W(s,a)$$

4. 実 験

本実験の目的は，エージェントによく研究で用いられる Q-learning と Profit Sharing を適用し，シミュレータを用いて迷路問題を対象とし両学習法を比較検討することである．以下では KheperaII の説明，Webots の説明，実験環境 (実験問題)，実験設定，実験結果と考察の順に説明する．

4.1 KheperaII

KheperaII は，強化学習の研究においてよく用いられるロボットである．[3)4)]

エージェントとして扱うロボットKheperaII について述べる．KheperaII を図7に示す．KheperaII の仕様は，直径70[mm]，高さ30[mm]，重さ80[g]，CPU モトローラ68331 プロセッサ24[MHz]，RAM512[Kbyte]，Flash メモリ 512[Kbyte] を搭載している．また、DC モータ (速度2〜60[cm/sec]) を2つ，赤外近接センサと光センサが一体化したものを図8の8箇所に装備している．赤外線センサの有効範囲は70[mm] である．

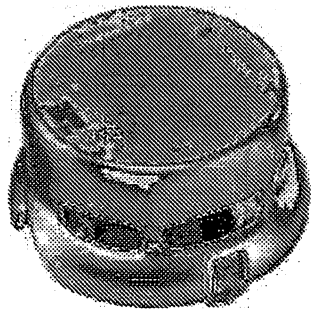

図7　KheperaII

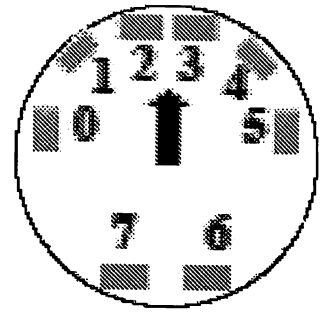

図8　KhepraII のセンサ位置

4.2 Webots

Webots は，知能ロボット研究者や教育者，技術者のための高機能シミュレータであり，KheperaII のシミュレートでよく用いられる．ロボットの自律動作技術，進化ロボット技術などの知的ロボット技術一般の実験や，コンピュータ視覚系，人工知能技術などの研究に適した研究開発ツールである．

4.3 実験環境 (実験問題)

問題として迷路問題を扱う．迷路問題はスタートおよびゴールが与えられ，ゴールまでの道には複数の壁が存在する．実験で用いた迷路は，図9のような強化学習で頻繁に使われる迷路[10)] を用いる．しかし，迷路はグリッド環境ではなく連続的な環境となっている．迷路のサイズは縦60[cm] 横90[cm] とする．スタートとゴールは図9に示す通りである．

4.4 実験設定

エージェントの設定は，KheperaII の外形及びセンサの設定と同一である．1状態における移動方向の候補を図10に示す．1回の行動選択につき，移動可能ならば25[mm] 移動する．状態認識と行動選択と移動を1ステップとする．スタートからゴールまで到達することを1学習とする．Q-learning の設定は，各初期状態行動評価値を0.1，学習率を0.1，減衰率を0.95，報酬を10，ϵ を0.1 とする．Profit Sharing の設定は，各初期状態行動評価値を0.1，減衰率を0.95，報酬を10 とする．両学習法の学習回数は10000 回とする．また，1学習は50000 ステップを超えると終了し，次の学習に移る．

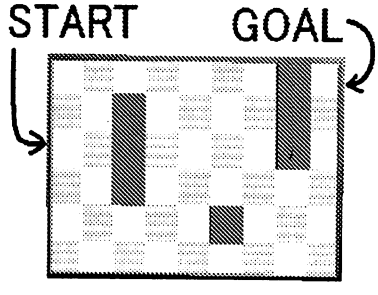

図9　実験問題で使用する迷路

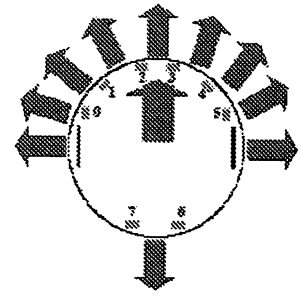

図10　1状態における移動方向の候補

4.5 実験結果と考察

図 11 に Q-learning，図 12 に Profit Sharing の実験の結果を示す．縦軸は 1 学習におけるステップ数，横軸は学習回数を示す．図 11 から，Q-learning は学習を進めてもステップ数は収束は見られない．それに対し，図 12 から，Profit Sharing は学習回数を重ねる毎にステップ数の収束が見られる．実験で用いた環境では，Profit Sharing は Q-learning よりも適していると言える．このような結果になったのは，Q-learning では環境すべてを探索するため報酬の値が全体の行動価値に分散し，Profit Sharing では報酬の値が有効な行動価値に振り分けられ集中したためと考えられる．最終的に Profit Sharing は図 13 のような壁伝いに移動するという行動をエージェントが獲得し，ゴールにたどり着くようになった．

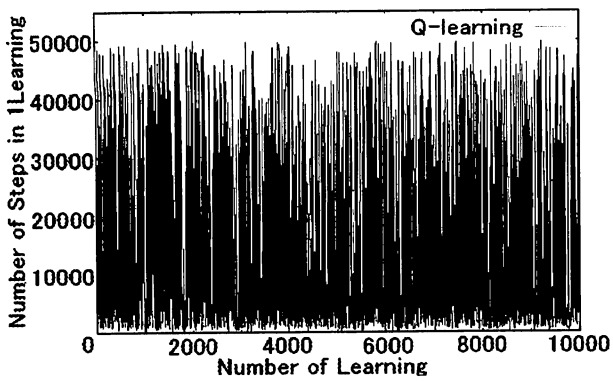

図 11 Q-learning の実験結果

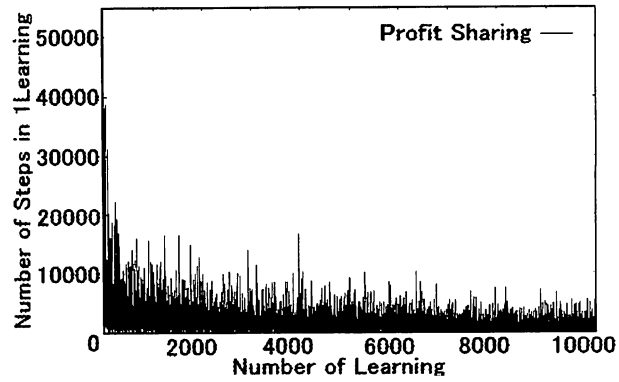

図 12 Profit Sharing の実験結果

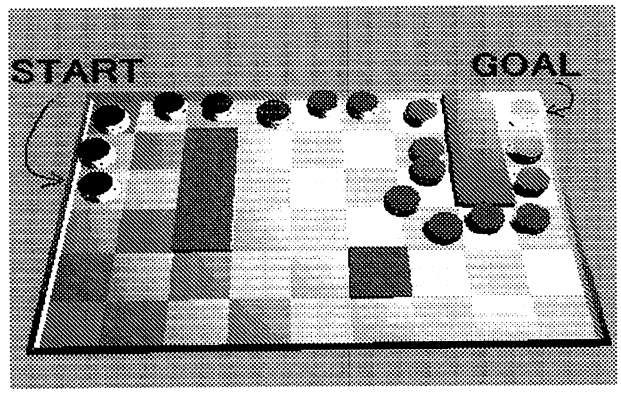

図 13 Profit Sharing によって最終的に得た行動

5. むすび

本論文では，KheperaII に強化学習を導入する前段階としてシミュレータを用い，強化学習手法の代表例である Q-learning と Profit Sharing をエージェントに適用し比較実験をした．迷路問題を対象とした結果，Profit Sharing の方が Q-learning より適していることを示した．

本論文で用いた連続的な環境の迷路問題において KheperaII に環境同定型の Q-learning と経験強化型の Profit Sharing をエージェントに適用したが，10000 回の学習では Q-learning のステップ数の収束は見られなかった．今後の課題として，環境同定型で連続的な環境に適応する可能性のある第 3 章で述べた Actor-Critic をエージェントに適用し，Q-learning や Profit Sharing の結果と比較検討する．

参考文献

1) Bellman, R. E., "A Markov decision process," Journal of Mathematical Mechanics, Vol. 6, 679–684, 1957.

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

3) 片上大輔, 山田誠二, "対話型分類子システムによる実環境ロボット学習～記述困難なプログラムを人間の教示から自動抽出する～," 第 1 回 MYCOM 資料, pp. 50–53, 2000.

4) 片上大輔, 山田誠二, "対話的進化ロボティクスの観測に基づく教示の設計," システム制御情報学会論文誌, Vol. 16, No. 6, pp. 279–286, 2003.

5) 北村新三, 片山修, "ニューラルネットとロボットの学習," 日本ロボット学会, Vol. 13, No. 1, pp. 63–67, 1995.

6) 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.

7) Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore, "Reinforcement Learning: A Survey," Journal of Artificial Intelligence Research, Vol. 4, 1996.

8) 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," 人工知能誌, Vol. 9, No. 4, pp. 580–587, 1994.

9) 野田彰一, 浅田稔, 細田耕, "強化学習によるロボットの行動獲得のための状態空間の自律的構成," 日本ロボット学会誌, Vol. 15, No. 6, pp. 886–892, 1997.

10) Richard S. Sutton, Andrew G. Barto [著] 三上貞芳, 皆川雅章共訳, "強化学習," 森北出版, 2000.

11) 柴田克成, 西野哲生, 岡部洋一, "Actor-Q アーキテクチャに基づく能動認識学習システム," 信学論, Vol. J84-D-II, No. 9, pp. 2121–2130, 2001.

12) 植村渉, 辰巳昭治, "Profit Sharing 法における強化学習に関する一考察," 人工知能論文誌, Vol. 19, No. 4A, pp. 197–203, 2004.

13) 内部英治, 浅田稔, 野田彰一, 細田耕, "視覚に基づく強化学習による移動ロボットの多重タスクの遂行のための協調行動の獲得," , 日本ロボット学会, Vol. 13, No. 1, pp. 68–74, 1995.

14) 畝見達夫, "実例に基づく強化学習法," 人工知能学会誌, Vol. 7, No. 4, pp. 697–707, 1992.

15) 畝見達夫, "強化学習法とロボットへの応用," 日本ロボット学会, Vol. 13, No. 1, pp. 51–56, 1995.

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

17) 山口智浩, 増渕元臣, 藤原一継, 谷内田正彦, "抽象化副報酬の自動生成による実ロボット強化学習の高速化," 人工知能学会誌, Vol. 12, No. 5, 60–71, 1997.

18) 山田和明, 黒山和宏, 中村陽一郎, Mikhail Svinin, 上田完次, "実例に基づく強化学習の一手法 (Instance-Based Classifier Generator(IBCG) の連続空間への拡張)," 日本機械学, ロボティクス・メカトロニクス講演会 '98, 講演論文集, No. 98, 1998.

19) 山田誠二, 斎藤淳也, "マルチロボットによる箱押しのための明示的通信を用いない適応的行為選択," 日本ロボット学会誌 Vol. 17, No. 6, pp. 818–827, 1999.

Comparison of Reinforcement Learning Methods using Khepera Robot

Yasuhiko BABA, Kengo KATAYAMA* and Hiroyuki NARIHISA*

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

(Received September 30, 2004; accepted November 5, 2004)

Reinforcement learning is known to be a framework of the learning control by which an agent adapts himself to environment through trial and error. The typical reinforcement learning methods are Q-learning and Profit Sharing. In this paper, we compare the reinforcement learning methods of Q-learning and Profit Sharing using a micro move robot KheperaII. The experiment for reinforcement learing using a real robot is impractical. Therefore, the precedent simulation is very important, and we use a simulator called Webots so that the real robot KheperaII efficiently learns an environment. As a learing problem, we adopt a maze problem. The comparison results show that Profit Sharing outperforms Q-learing in term of the learning speed.