計測自動制御学会東北支部 第188回研究集会 (2000.6.23)
資料番号 188-16

# 部分観測環境における移動ロボットの強化学習

# Reinforcement Learning for a Mobile Robot in Partially Observable Environments

○釜谷博行*, 李海妍**, 阿部健一**

○Hiroyuki Kamaya*, Hae Yeon Lee**, Kenichi Abe**

*八戸工業高等専門学校, **東北大学大学院工学研究科

*Dept. Electrical Eng., Hachinohe National College of Technology,
**Dept. Electrical and Communication Eng., Tohoku University

**キーワード**： 強化学習 (Reinforcement Learning), スイッチングQ-学習 (Switching Q-learning), 部分観測マルコフ環境 (Partially Observable Markov Decision Processes), CMAC (Cerebellar Model Arithmetic Computer), ロボットナビゲーション (Robot Navigation)

**連絡先**： 〒039-1192 八戸市田面木字上野平16-1 八戸工業高等専門学校 電気工学科
釜谷博行, Tel.: (0178)27-7283, Fax.: (0178)27-9379, E-mail: kamaya-e@hachinohe-ct.ac.jp

## 1. はじめに

Q-学習やTD-学習に代表される強化学習とは, 報酬という特別な入力を手掛かりに, 環境に適応するための行動決定戦略を獲得する機械学習の一種である[1)2)]。設計者はタスクのゴールを報酬という形で与えるのみで, 学習システムはゴールへ向かう方策を自動的に獲得する。Q-学習では, 環境の性質にマルコフ性を仮定しており, ある条件の下で最適性が保証されている[3)]。強化学習の最大の利点は, 不確実性のある環境, 報酬に遅れが存在する場合にも適用可能なところにある。

しかし, ロボットなどへの応用を考えた場合, ロボットに取り付けられたセンサなどの制約によって環境状態を完全に観測することができない。この場合には, 部分観測マルコフ決定過程 (POMDP) 問題となる。POMDP環境に対する代表的な学習アルゴリズムとして, 確率的な政策を用いる方法, 環境モデルを構築する方法, 有限長の過去のシーケンスを記憶する方法などが提案されている。しかし, これらの多くは, 現状では比較的規模の小さな問題にしか適用されていない。

著者らはこれまで,POMDP環境に対する新たな強化学習法としてスイッチングQ-学習 (SQ-学習)[4)5)6)] を提案し, 比較的規模の大きなエージェントの迷路探索問題に適用し, その有効性について検討してきた。

本研究では, このSQ-学習を連続値のセンサデータを扱うため小脳モデル神経回路を用いて拡張し, 目的地へ向かう行動を自動的に獲得する移動ロボットのナビゲーション問題へ適用する。なおこのとき, ロボットの自己位置は認識できないものとする。本発表では, シミュレーション実験により提案する手法の有効性を検討する。

## 2. 強化学習

離散マルコフ決定過程 (MDP) 問題では, 各時点 $t \in \{0,1,2,\ldots\}$ において, 環境状態が $s_t \in S$ のとき, その状態観測に基づき, エージェントが行動 $a_t \in A$ をとったとすると, 報酬 $r_t$ を受け取り, 環境状態は未知遷移確率でつぎの状態 $s_{t+1}$ に遷移する。エージェントの目標は, 報酬の割引期待利得 $E\{\sum_{t=0}^{\infty}\gamma^t r_t\}$ を最大にすることである。ここで,$\gamma(0 \leq \gamma \leq 1)$ は割引率を表す。

部分観測マルコフ決定過程 (POMDP) 問題では, 各時点でそのときの状態を知ることはできず, 観測情報 $o_t \in O$のみが与えられる。エージェントの目標は, MDP と同様に報酬の割引期待利得を最大にすることである。

Q-学習では, ある状態 s において行動 a をとるときの評価値を Q 値と呼び,Q(s,a) で表わす。Q 値の大きさ応じて, エージェントは状態 $s$ において実行すべき行動 $a$ を決定する。環境と対峙したエージェントは試行錯誤を繰り返しながら, 割引期待利得の最大化を目的として, 各時点で得られる報酬 $r_t$ に基づいて Q 値を更新していく。

## 3. スイッチング Q-学習

スイッチング Q-学習 (SQ-学習) の基本的な考えは, 複数個の Q-モジュールを用意しておき, ある観測値 (サブゴール) が得られたときに Q-モジュールを順次切り換えることで, 隠れマルコフ環境をマルコフ環境に近似しながら学習を行なうところにある。この SQ-学習は, 原理的には HQ-学習[7]と同じクラスに属する。未知環境を想定しているため, 設計者が予めサブゴール系列を決定することは難しい。このため, エージェントの行動学習と同時にサブゴールとなる観測値の学習も行なう。

エージェントの行動学習には, MDP における有効な強化学習法のひとつとして知られている Sarsa($\lambda$)[8]を利用する。また, サブゴール学習には

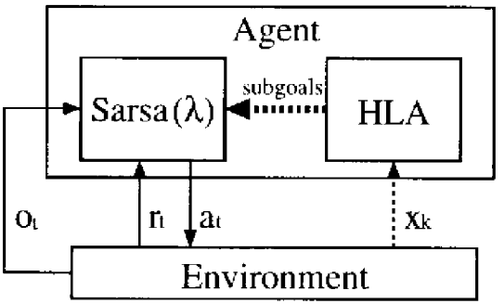



Fig. 1　SQ-$_{HLA}$学習のアーキテクチャ

階層構造学習オートマトン (HLA) を利用する。このため, この SQ-学習を特に $SQ_{HLA}$-学習と呼ぶ。$SQ_{HLA}$-学習の構成を Fig.1に示す。なお, これ以降 $s_t = o_t$として話を進める。

### 3.1 Sarsa($\lambda$)

$SQ_{HLA}$-学習では,Q-学習の変形バージョンである Sarsa($\lambda$) を用いる。Sarsa($\lambda$) では, 遷移先の状態 $s_{t+1}$において選択した行動 $a_{t+1}$に対応する Q 値を評価値として用いる。$< s_t,a_t,r_t,s_{t+1},a_{t+1} >$の 5 項組を用いて,Q 値はつぎのように更新される。

$$\delta_t = r_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha\delta_t e(s,a) \qquad \text{for all } s \text{ and } a$$

ここで,$\alpha$は学習率 $(0 \leq \alpha \leq 1)$ を, $\gamma$は割引率を表わす。また,$e(s,a)$ は *eligibility trace* と呼ばれ,

$$e(s_t,a_t) \leftarrow 1$$

$$e(s,a) \leftarrow \gamma\lambda e(s,a) \qquad \text{for all } s \neq s_t \text{ or } a \neq a_t$$

ここで,$\lambda$ は $0 \leq \lambda \leq 1$ なる実数である。なお, Q 値の更新に先立って *eligibility* の計算が行なわれる。

行動選択には Max-Boltzmann 分布による方法を用いる。これは, 確率 $p_{max}$で greedy action を, $1 - p_{max}$で Boltzmann 分布により行動を選択するものである。Boltzmann 分布では, 状態 $s_i$において行動 $a_i$が選択される確率は,

$$Prob(a_i) = \frac{\exp\frac{Q(s_i,a_i)}{Temp}}{\sum_k \exp\frac{Q(s_i,a_k)}{Temp}}$$

で表される。$Temp$は温度係数 (temperature) と呼

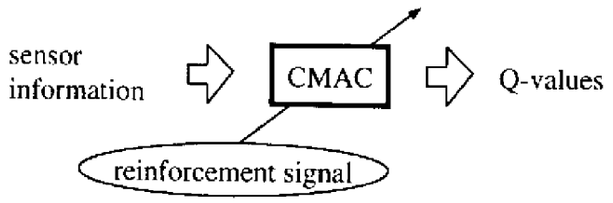



Fig. 2　CMACの入出力データ

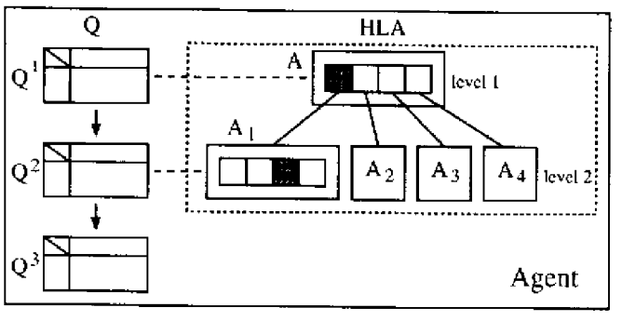



Fig. 3　階層構造学習オートマトン(HLA)

ばれ, 行動選択のランダムさの度合を調整するパラメータである。

### 3.2　CMACによる状態表現

SQ-学習において連続的な状態を扱うため, 小脳モデル神経回路 (CMAC:Cerebellar Model Arithmetic Computer) を用いてQ-モジュールを実装する。CMACは汎化能力を備えているが基本的にはテーブル参照法であり, バックプロパゲーションネットワークに比べて正確なデータを学習することができるという利点をもっている[9]。CMACへの入力はセンサデータであり, 出力は各行動に対するQ値である (Fig.2)。

### 3.3　階層構造学習オートマトン(HLA)

単体の学習オートマトンは行動数の増加につれてその学習回数が急速に増大することが知られている。これに対して, HLAは行動数が多い場合にもより少ない学習回数で最適解を見出すことができるという特徴をもつ。HLAの各レベルでの行動を, 各Q-モジュール$Q^y$のサブゴールに対応づける (Fig.3)。

HLAには, 各試行$k \in \{1,2,3,\ldots\}$においてゴールの到達に成功したかどうかの2値の評価値(応答)$x_k \in \{0,1\}$が与えられ, 行動確率ベクトル$p_j$は$L_{R-I}$強化法を用いてつぎのように改変される。今, 行動としてサブゴール$sg_i$を選択したとする。$x_k = 1$(成功) ならば,

$p_i \leftarrow p_i + \alpha_A(1 - p_i)$

$p_j \leftarrow (1 - \alpha_A)p_j \qquad (j \neq i)$

とする。ここで,$\alpha_A$は$0 < \alpha_A < 1$なる実数である。$x_k = 0$(失敗) ならば,$p_i$を変更しない。HLAの各レベルにおける行動確率ベクトルの効率的な計算方法については, 文献[10][11]を参照されたい。

### 3.4　サブゴール選択リスト

全く観測されない観測値をサブゴール選択リストから除外するため, 各試行毎にサブゴール選択の指標$\rho$をつぎのように計算する。

if(前回の試行中に$s_i$を観測)

　　$\rho_i \leftarrow 1$

else

　　$\rho_i \leftarrow \beta\rho_i$

サブゴールとして選択するかどうかは, つぎの条件を用いて判定する。

if($\rho_i > C_{sg}$)

　　$s_i$をサブゴール選択リストへ挿入

なお,$\beta$(実験では0.95) は減衰定数を, また,$C_{sg}$(実験では0.2) は選択基準となる定数を表す。$\rho_i$の初期値はすべて1とする。観測されないサブゴールの$\rho_i$は試行とともに減衰していく。サブゴールが観測されると$\rho_i$は1にリセットされる。

## 4.　シミュレーション実験

シミュレーション実験では, 著者らが開発した自律型移動ロボット開発支援システムのシミュレータを用いる[12]。移動ロボットとしてNOMADIC社のNomad200を想定した。このロボットは全方向移

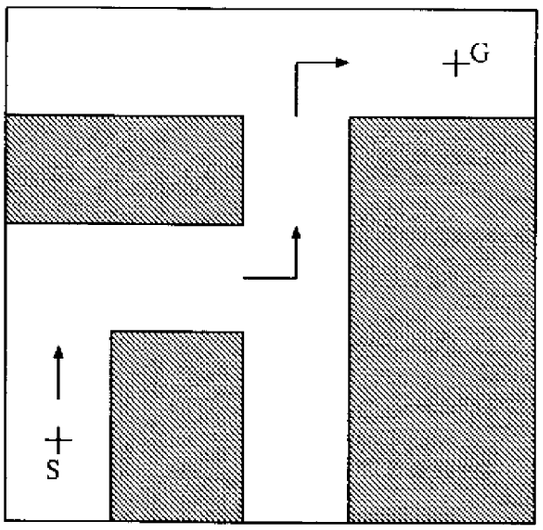



Fig. 4　環境 1

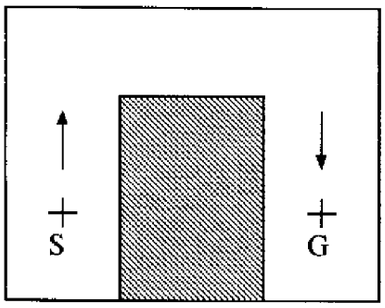



Fig. 5　環境 2

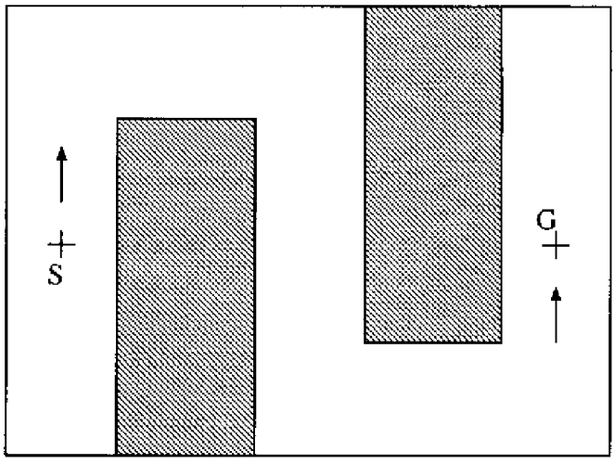



Fig. 6　環境 3

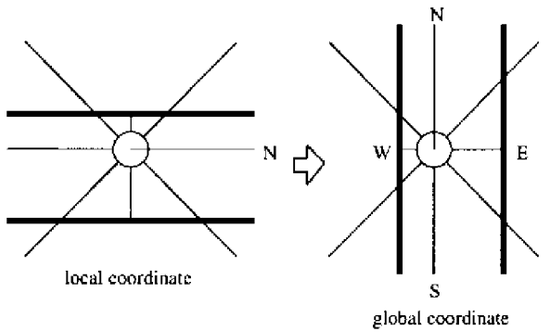



Fig. 7　超音波センサ情報の座標変換

動車で, その大きさは直径 46[$cm$] である。ロボットには 3 種類のセンサが取り付けられている。ロボットのグローバルな方位を計測するセンサ, 壁までの距離を計測する 16 個の超音波センサ (22.5[$deg$] 間隔), 衝突を検出するリミットスイッチである。特に, 超音波センサは, センサ主軸と壁面との角度が 25[$deg$] を越えると壁を検出できないとしてモデリングした。シミュレーションでは 0.016[$sec$] 毎に, ロボットの移動と 1 つの超音波センサの計測が行なわれる。16 個のセンサ情報をすべて取得するには $0.016 \times 16 = 0.256[sec]$ が必要となる。

移動ロボットのナビゲーションで利用する環境は環境 1(Fig.4), 環境 2(Fig.5), 環境 3(Fig.6) の 3 つで, それぞれ大きさは $10[m] \times 9.5[m]$, $6.5[m] \times 5.5[m]$, $6.5[m] \times 5.5[m]$, $11[m] \times 8[m]$ である。ロボットの目的は, スタート地点 S からゴール地点 G へ向かう行動を超音波センサ情報のみを用いて獲得することにある。

16 個の超音波センサ情報を用いて状態を表現する。前処理として, ローカルなセンサ座標系から方位 N を基準としたグローバルな座標系へと距離情報を変換する (Fig.7)。その後,CMAC の 4 つのタイリングにおいて, 4 方向の方位情報と 4 段階の距離情報に離散化される (Fig.8)。なお, ロボットはその座標 $(x, y)$ やゴール方向を認識できないものとする。このため, 環境 1 は MDP 環境であるが, 環境 2,3 は POMDP 環境となる。

方位 N を基準としたロボットの移動方向を行動とする (Fig.9)。行動選択は 1.28[$sec$] 毎に行なわれ, 各時点でロボットは 8 方位中,1 方位のみ移動方向を選択できる。選択された移動方向からロボットへの指令 (移動速度 $tv$ とステアリング速度 $sv$) はつぎのように計算される。

選択した移動方向 $\theta_T$ とロボット姿勢 $\theta_R$ との角度

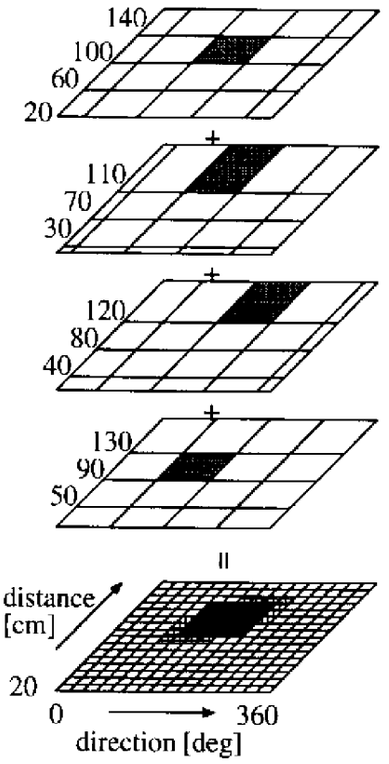



Fig. 8　CMAC による状態表現

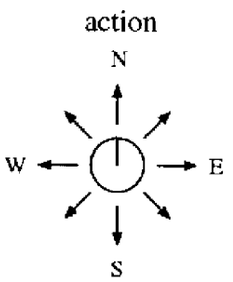



Fig. 9　移動ロボットの行動 (移動方向)

差を$\Delta_\theta = \theta_T - \theta_R$とする。移動速度$tv$は,

$$tv = tv_{max} - (tv_{max} - tv_{min})\frac{|\Delta_\theta|}{\theta_{min}}$$

if$(tv < tv_{min})$

$$tv = tv_{min}$$

ここで,$tv_{max}$は最高速度,$tv_{min}$は最低速度を表わし, 実験ではそれぞれ $20[cm/sec]$,$5[cm/sec]$ とした。$|\Delta_\theta|$が大きくなるにつれて, 移動速度は$\theta_{min}$(実験では $45[deg]$) まで直線的に減少し, その後は最低速度 $tv_{min}$で一定となる。

ステアリング速度 $sv$は,

$$sv = c_{sv}\Delta_\theta$$

if$(|sv| > sv_{max})$

$$|sv| = sv_{max}$$

実験では$c_{sv} = 1.0$, $sv_{max} = 10[deg/sec]$ とした。

報酬は, つぎのように定めた。

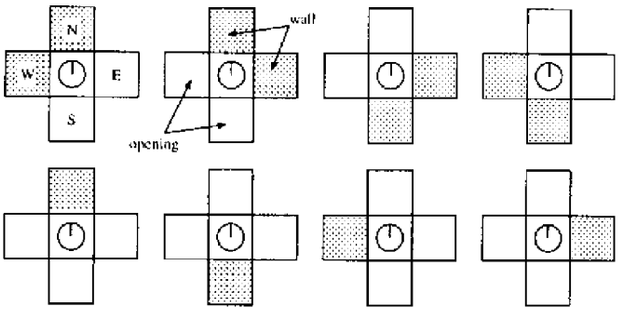



Fig. 10　サブゴール状態

| 報酬 | 利得 |
|---|---|
| ゴール到達時 | 100 |
| 壁との衝突時 | −1 |
| それ以外 | +0.1 |

サブゴール数の増加にともない学習が遅くなるという問題がある。このため, 建物内の幾何学的な性質を利用し, N,S,E,W の 4 方位の通路の有無からサブゴール状態を決定した。具体的には, 超音波センサからの距離情報がある距離 (実験では $200[cm]$) 以上の場合には, その方位に通路があるものとみなした。通路の有無の 2 ビットで表現すると, サブゴール数は $2^4 = 16$ パターンとなる。今回の実験では,Fig.10に示すようにサブゴール状態を 8 パターンに限定した。

## 5.　実験結果

学習システムのパラメータは, $\gamma = 0.99$, $\alpha = 0.1$, $\lambda = 0.9$, $\alpha_A = 0.05$ とする。行動選択において$p_{max}$の値は, 初期値を 0.9 とし, 最後の試行で 1.0 になるように直線的に増加させる。最大ステップ数 $T_{max} = 500$ で, Boltzmann 分布の温度係数は $T = 0.2$( 一定), サブゴール系列数は $n_s = 2$ とした。1 つの試行は, ロボットがゴールへ到達したとき, ロボットが壁と接触したとき, 予め決められたステップ数の上限を越えたときのいづれかで終了

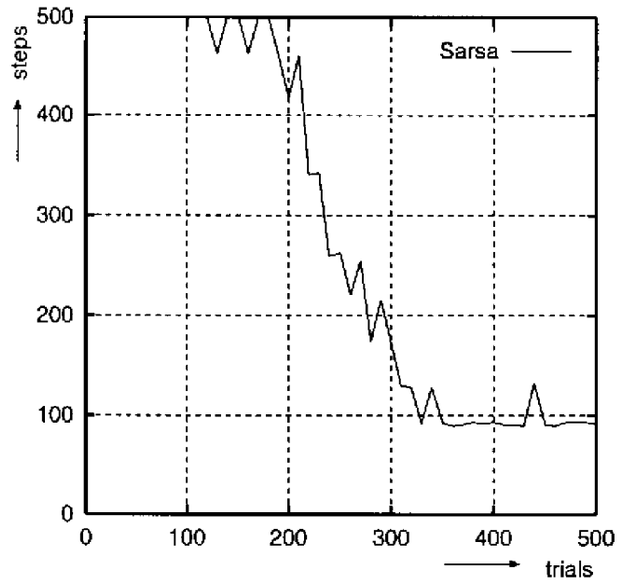


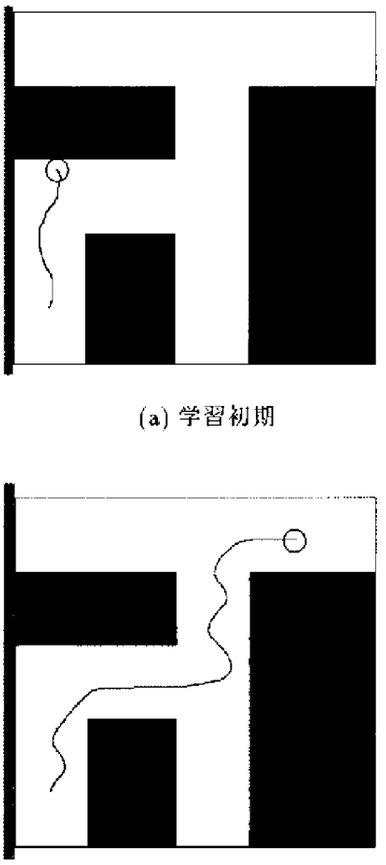

Fig. 11　学習性能曲線 (環境 1)

する。つぎの試行は, ロボットをスタート地点へ戻して行なわれる。なお, 実験はすべて超音波センサにノイズがない理想的な状況で行なった。

Fig.11は環境 1 において通常の Sarsa($\lambda$) を用いた学習性能曲線である。試行回数 $K_{max} = 500$ とした。グラフは乱数の初期値を変えた 10 シミュレーションの平均値を横軸の 20 データ毎にプロットしたものである。横軸が試行回数, 縦軸がゴールまでのステップ数を表わす。ステップ数が小さいほど良好である。なお, 衝突による失敗の場合にはステップ数を $T_{max}$ として示した。実験結果をみると, 試行回数の増加とともにゴールまでのステップ数が減少しており学習に成功していることがわかる。

Fig.12は学習初期と学習後のロボットの行動軌跡である。学習初期では壁に衝突しているが, 学習後はうまくゴールへ到達していることがわかる。

Fig.13は環境 2 において Sarsa($\lambda$) と SQ-学習の学習性能曲線を比較したものである。試行回数は環境 1 と同様に設定した。環境 2 は極めて単純な構造であるが, 通常の Sarsa($\lambda$) ではうまく学習できないことがわかる。これに対して,SQ-学習の学習結果は良好である。



(a) 学習初期

(b) 学習後

Fig. 12　ロボットの軌跡

Fig.14は環境 3 における SQ-学習の学習性能曲線である。試行回数 $K_{max} = 1{,}000$ として実験を行なった。時折, 試行回数が 1,000 回に達しても, ゴールへ向かう行動の獲得に失敗する学習結果が得られた。このため, グラフ上ではゴールまでのステップ数の平均値が良好な場合に比べて大きくなっている。失敗の原因は, 衝突しなければ+0.1 の報酬が得られるためロボットがある一定地点にとどまる行動を学習してしまうところにある。つまり, ゴールへの到達回数が少ないと衝突回避学習のみが優先されることになる。このことから, ゴールまでのステップ数が大きくなるにつれてゴールへ到達するための学習が難しくなる傾向にある。

## 6.　おわりに

本研究では, 連続値のセンサデータを扱うため SQ-学習を CMAC を用いて拡張し, 目的地へ向かう

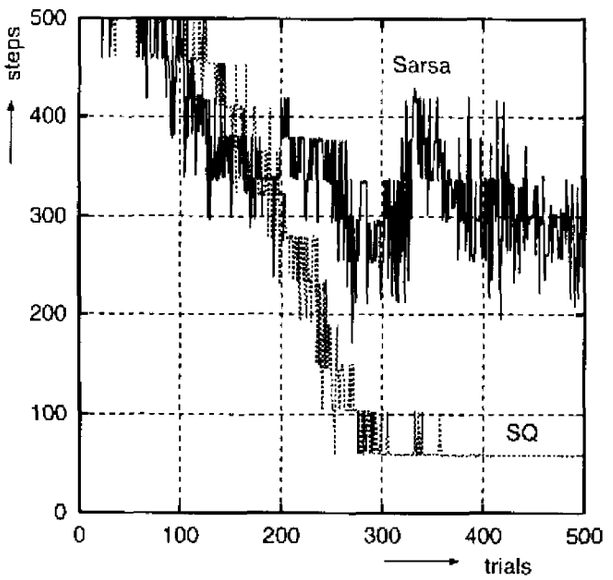



Fig. 13　学習性能曲線 (環境 2)

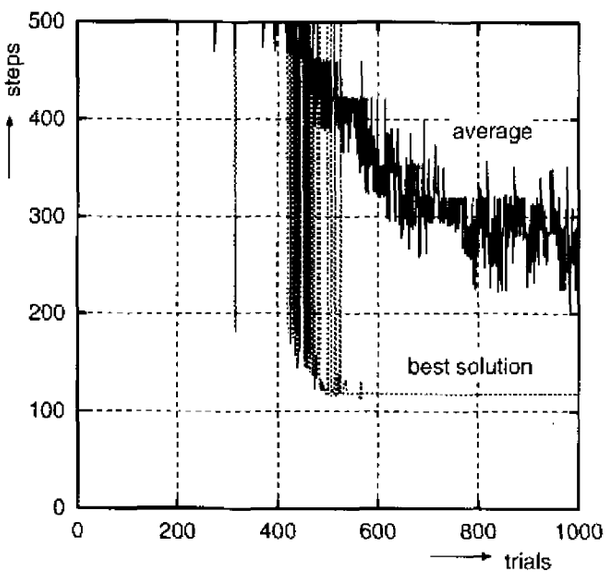



Fig. 14　学習性能曲線 (環境 3)

行動を自動的に獲得する移動ロボットの POMDP ナビゲーション問題へ適用し, シミュレーション実験によりその有効性を確認した。

今回の実験では, 衝突時にロボットをスタート地点へ戻してつぎの試行を開始している。学習効率を向上させるために衝突した場所から学習を継続させるようにする。また, 学習システムにロボットの制御アルゴリズムを全く組み入れずに学習実験を行なった。このため, 単純な環境にもかかわらず学習にかなりの時間を要している。学習量を軽減させるために壁のある方向への移動を禁止したり, 壁に沿うなどの要素行動を予め用意する方法も考えられる。さらに, サブゴール状態として生のセンサデータを利用しているため, 通路を誤認識するという問題がある。サブゴールの認識精度を向上させるため, センサデータからロボット周囲のローカルマップを作成し, このローカルマップから通路を識別させる方法も考えられる。

その他, センサ情報にノイズが含まれた場合や実機を利用した実験などが今後の検討課題として挙げられる。

## 参考文献


[1] Sutton,R.S. and Barto,A.G.,"Reinforcement Learning: An Introduction", MIT Press (1998)

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

[3] Watkins,C.J.C.H. and Dayan P.,"Q-learning", *Machine Learning*, **8**, 279/292 (1992)

[4] 釜谷博行, 李海妍, 阿部健一, "隠れマルコフ環境におけるスイッチング Q-学習", 電関学東北支部連合大会講演論文集, pp.187 (1999)

[5] 釜谷博行, 李海妍, 阿部健一, "隠れマルコフ環境におけるスイッチング Q-学習", 計測自動制御学会東北支部 35 周年記念学術講演会予稿集, 7/8 (1999)

[6] 釜谷博行, 李海妍, 阿部健一, "部分観測マルコフ環境におけるスイッチング Q-学習", 第 27 回知能システムシンポジウム講演論文集, 101/106 (2000)

[7] Marco,W. and Juergen,S., "HQ-Learning," Adaptive Behavior, **6**-2, pp. 219-246, 1997.

[8] Loch,J. and Singh,S.P., "Using Eligibility Traces to Find the Best Memoryless Policy in Partially Observable Markov Decision Processes", ICML98 (1998)

[9] 齊藤史倫, 福田敏男, "強化学習による実ロボットの運動制御", 日本ロボット学会誌, Vol.13, No.1, 82/88 (1995)

[10] Thathachar,M.A.L. and Ramakrishnan,K.R., "A Hierarchical System of Learning Automata", *IEEE Trans.*, SMC-**11**-3, 236/241 (1981)

[11] 阿部健一, 竹田宏, "学習オートマトンとその応用", 電気学会雑誌, 105 巻, 4 号, 333/336 (1985)

[12] 釜谷博行, 本間弘一, 阿部健一, "オブジェクト指向設計に基づいた自律型移動ロボットの開発支援システム", 電気学会論文誌 C, Vol.115-C, No.6, 819/828 (1995)