4L－6

# 移動ロボットの迷路学習

## — 確率的傾斜法によるアプローチ —

名平 光宏*　高橋 貞夫

芝浦工業大学大学院 電気工学専攻

## 1 はじめに

強化学習の一手法である確率的傾斜法 [2] は，Q-learning 法 [1] に代表される従来の学習法に比べ，様々な利点がある．代表的な応用例は RoboCup エージェントの学習に用いた Andhill[3] である．しかし，迷路学習など他の手法で頻繁に取り上げられている問題については，有効性の検証は十分ではない．

そこで，迷路学習問題について，自律移動ロボットの行動決定を階層型ニューラルネットで構成し，その学習における確率的傾斜法の有効性を示した．

## 2 問題設定

迷路内をロボットが走行するタスク（**迷路学習問題**）を設定する．ロボットは，スタートから移動を開始し，自分のセンサ情報をもとに行動を決定する．そして，ゴールに到達すると報酬を獲得する．

これを繰り返しながら学習を行い，スタートからゴールまで最短経路で移動する行動パターンを得ることが目標である．

### 2.1 エージェント（ロボット）の設定

本論文では，A.K.Peters 社製の自律移動ロボット，Rug Warrior[4] を想定する．Rug Warrior は，実装されたセンサ群により，任意距離走行及び任意角度の回転が可能である．従って，迷路内で 4 方向への 1 マス単位での移動が容易に実現できる．そこで，エージェントを次のように設定する．

**センサ情報** 現在地点の座標値を認識することはできないが，迷路をセル単位で識別可能である．

**行動** 4 方向へ 1 セル単位で移動を行う．また，触覚センサーを持ち，壁に衝突すると，行動する直前の位置に戻り再度行動選択をやり直す．

**学習** 行動を実行するたびに逐次，学習を行う．

### 2.2 迷路学習の研究課題

迷路学習はマルコフ決定過程下における．強化学習の典型的な問題として，頻繁に取り上げられており．以下のような課題がある．

**報酬遅れ** 報酬を受け取れるのが，ゴール地点のみであるので，報酬が直接得られない地点での効果的な行動評価・学習が必要になる．

**未知の環境下での学習** 学習する迷路についての予備知識（迷路の大きさ・ゴール位置・ゴールまでの距離等）は持たない．また，これらの知識を学習中に直接得ることもない．

**不必要な行動の抑制** 壁に向かう行動や，ゴールから離れていく行動など，報酬獲得に関係ない行動を，自律的に抑制させる必要がある．

## 3 強化学習システム

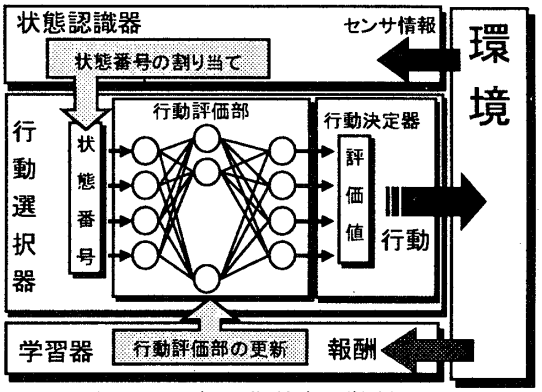


図 1: 逐次認識型強化学習システム

本論文で提案する強化学習システムを図 1 に示す．このシステムの特徴は，入力層ユニットと状態の対応関係を設計者が作成するのではなく，状態を知覚するたびに逐次的に状態番号を割り当てていることである．

まず外部環境からのセンサ情報は，状態認識器に送られる．状態認識器では，エージェントが今までに体験した状態（**既知状態**）か，初めて経験する状態（**未知状態**）かを判断し，状態番号を出力する．

行動評価部は 3 層の階層型ニューラルネットによって構成されている．入力は状態番号である．例えば，図 1 において，状態番号が 3 ならば，ニューラルネットの入力値は 0011 となる．出力は，入力された状態番号に対する各行動の評価値である．

行動決定器では，評価値に基づいて，確率的に行動を決定する．つまり，行動 $a_i$ の行動選択確率 $\pi(a_t, W, X_t)$ は，以下のようになる．

$$\pi(a_t, W, X_t) = \frac{O_i}{\sum_k O_k} \quad (1)$$

*Reinforcement Learning of Maze environment using Stochastic Gradient Ascent
Mitsuhiro Nahira,Sadao Takahashi
Postgraduate Course of Electrical Engineering,
Shibaura Institute of Technology,
3-9-14 Shibaura,Minato-ward,Tokyo,108-8548,Japan

## 4 学習アルゴリズム

確率的傾斜法で階層型ニューラルネットが学習できるように，拡張したアルゴリズムを以下に示す．

1. 時刻 $t$ における，状態を観測
2. $\pi(a_t, W, X_t)$ の確率で行動を実行
3. 環境から報酬 $r_t$ を受け取る
4. ニューラルネットの結合係数 $w_i$ について $e_i(t)$ と $\overline{D}_i(t)$ を求める

$$e_i(t) = \frac{\partial}{\partial w_i} \ln\{\pi(a_t, W, X_t)\} \quad (2)$$

$$\overline{D}_i(t) = e_i(t) + \gamma \overline{D}_i(t-1) \quad (3)$$

5. $\Delta w_i$ を求め，$w_i$ を更新

$$\Delta w_i(t) = (r_t - b)\overline{D}_i(t) \quad (4)$$

$$w_i(t+1) = w_i(t) + \alpha \Delta w_i(t) \quad (5)$$

$e(t)$ については，BP 法と同様に出力層から入力層へと修正量の伝播を行い，さらにユニットの出力関数にシグモイド関数を用いることで，以下のように算出する．

中間–出力層間の結合係数 $v_{kj}$ の $e(t)$

$$e_{v_{kj}}(t) = \frac{1}{O_n}\frac{\partial}{\partial v_{kj}} O_n - \frac{1}{\sum_k O_k} O_k(1-O_k)H_j \quad (6)$$

入力–中間層間の結合係数 $w_{ji}$ の $e(t)$

$$e_{w_{ji}}(t) = (1-H_j)I_i \sum_k v_{kj} e_{v_{kj}}(t) \quad (7)$$

## 5 実験

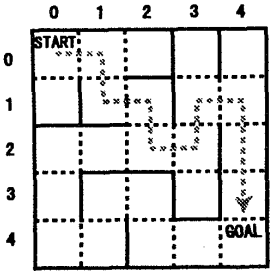


図 2: 学習に用いた迷路

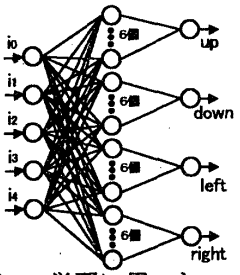


図 3: 学習に用いたニューラルネット

図 2 の迷路を Q-learning 法・確率的傾斜法を用いて学習を行い，両者を比較した．スタートからゴールまでの最短ステップ数は 10 ステップである．確率的傾斜法の場合は図 1 の強化学習システムを用いて学習を行う．なお，行動評価部は図 3 のニューラルネットを用いて，学習率 $\alpha = 0.10$, 減衰率 $\gamma = 0.80$, 報酬 $r = 10.0$, 定数 $b = 0.01$ とした．結合係数の初期値は $\pm 0.05$ 以内とした．

Q-learning 法の場合は，$\alpha = 0.10$,$\gamma = 0.80$,$r = 1000$ とした．そして，行動評価部はニューラルネットの代わりに Q 値を（状態数）×（行動数）だけ用意し，その初期値を 50 とした．

両者とも，100 個のサンプルについて，それぞれゴール到達を 10000 回行うまで学習を行い，毎回ごとのゴール到達までに要した行動回数の平均値を測定した．なお，スタートから $10^5$ ステップ経過してもゴールに到達しないサンプルについては，局所解に陥ったものとみなし，学習を打ち切った．

## 6 実験結果

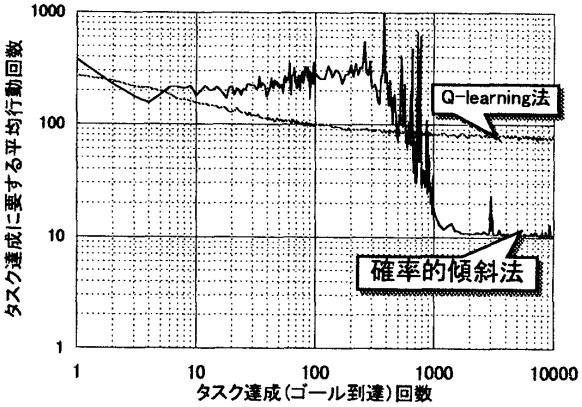


図 4: 学習曲線

図 4 に実験結果を示す．図 2 の最短ステップ数は 10 ステップであるから，平均行動回数が 10 に収束すると，最適な政策に収束したことになる．

学習の初期段階において，確率的傾斜法の平均行動回数が多いのは，直接 Q 値を更新する Q-learning 法に対して，ニューラルネットの結合係数を 1 回の行動につき 1 度だけ更新しているためである．しかし，図 4 から明らかなように，確率的傾斜法は，Q-learning 法よりも少ないタスク数で最適政策に収束していることがわかる．

また，ニューラルネットの学習方法の共通の課題でもある，局所解への収束が，確率的傾斜法でも生じてしまい，サンプルによっては，ゴール到達にかなりの行動回数を要したり，ゴールに到達できない場合があった．この点の改善が今後の課題である．

## 7 おわりに

本論文は，確率的傾斜法のニューラルネットへ適応した場合の有効性をコンピュータシミュレーションで示した．

## 参考文献

[1] Watkins,C.J.C.H.,and Dayan, Technical Note:Q-learning, Machine Learning 8,pp.55–68(1992)

[2] 木村元・山村雅幸・小林重信, "部分観測マルコフ決定過程下での強化学習：確率的傾斜法による接近", 人工知能学会誌,Vol.11,No.5,pp.761–768(1996)

[3] T.Andou, "Refinement of Soccer Agents' Positions Using Reinforcement Learning", RoboCup-97:Robot Soccer World Cup1, pp.373–388,Springer(1998)

[4] J.L. ジョンズ, A.M. フリン（熊切康雄 訳）, "移動ロボット 基礎理論と応用", トッパン (1996)