修士論文要旨

木構造を用いることによるQ学習の環境変化への適応法

（知能情報システム学）　山本　健弘

1. 序論

あるシステムを環境の変化に適応させようとするとき、起こりうる環境の変化をあらかじめ全て想定し、エージェント知識を設計する事は困難であるので、ロボットのような人工システムにとっても、環境の変化に適応するために学習することが重要である。

現在、様々な環境下で複雑な行動政策を獲得させる研究が盛んに行われているが、個々の環境下で行動政策を学習させ、環境ごとにそれらを切り替えるという従来手法の場合、以下のような問題点がある。

- ある環境で行動政策を獲得したとき、その環境でのタスク遂行には優れているが、その環境に特化した行動政策であり、汎用性がない。そのため、環境ごとに行動政策を持たなければならず、扱う環境が増加すれば膨大な記憶領域が必要となる。
- 強化学習では、扱う環境が複雑になり状態空間が大きくなれば学習時間が増加するので、環境ごとに独立して行動政策を学習させる場合、扱う環境が増加すればさらに学習時間も膨大になる。

これらの問題に対し港[1)]は、ある環境でロボットが獲得した行動政策を保持し、タスク成功率の低下により環境変化を認識し、行動政策を一部修正する手法を提案している。具体的には、失敗状態から1 step前までの状態と、状態空間上でそれらと隣接する状態で再学習をし、目標とする成功率が得られれば再学習を終え、得られなければさらに再学習を行う状態空間を広げていくといった手法である。しかし、この手法では、環境変化の認識に数百エピソードの時間がかかるため実機への搭載は不向きである。高井[2)]は負の報酬が与えられたとき、つまりタスク遂行を失敗したときに再学習することで、環境変化を認識するまでの時間を削減し、実機への搭載を実現している。しかし、大きく環境が変化する場合1)の手法では環境認識までの時間だけでなく、再学習に要する時間も膨大になる。また、2)の手法では、再学習する状態空間を広げていくことを行わないので、タスクの遂行ができなくなる環境が多い。そこで、木構造を用い効果的に再学習することにより、大きな環境の変化にも対応できる学習手法を提案する。

2. 方法

2-1 Q学習について

Q学習とは強化学習の一種で、ロボットが識別できる状態の集合をS、とり得る行動の集合をAとしたとき、時刻tでの状態$s_t \in S$において行動$a_t \in A$をとり、その後は状態集合Sから行動集合Aへの写像である行動政策$\lambda : S \rightarrow A$に従って行動したときの期待報酬である行動価値関数$Q^{\lambda}(s_t, a_t)$を最大にするために、どの行動を選択すればよいかを学習するもので、Q値は行動ごとに以下の式により更新する。

$Q^{\lambda}(s_t, a_t) \leftarrow Q^{\lambda}(s_t, a_t) + \alpha\,(r_{t+1} + \gamma \max_{a' \in A} Q(s_{t+1}, a') - Q^{\lambda}(s_t, a_t))$

ここで $\max_{a' \in A} Q(s_{t+1}, a')$ とは、状態 s_t で行動 a_t をとることにより遷移した次状態 s_{t+1} で、最も行動価値の高い行動 a における行動価値関数である。r_{t+1} は、状態 s_t で行動 a_t をとることにより得られた報酬値であり、α ($0<\alpha \leqq 1$) は学習率と呼ばれるパラメータで行動価値関数の更新の度合いを表し、γ ($0 \leqq \gamma \leqq 1$) は減衰係数または割引率と呼ばれるパラメータで将来の報酬が現在どれだけの価値があるのか表す。このように Q 値を更新する強化学習手法を Q(0)学習と言い、一般に Q 学習と言うときは Q(0)学習を指す。

2-2 行動政策

Q 学習では、学習段階においては以下の max-ボルツマン選択により状態 s_t のとき確率 $P(s_t, a_t)$ で行動 a_t を選択するのが一般的である。

$P(s_t, a_t) = [\exp\{\beta Q(s_t, a_t)\}] / [\sum_{a' \in A} \exp\{\beta Q(s_t, a')\}]$

ここで β はボルツマン温度 τ の逆数である。

また、学習収束後の行動政策 $a = f(s)$ は以下のように定義される。

$f(s) \leftarrow a \text{ such that } Q(s, a) = \max_{a' \in A} Q(s, a')$

2-3 ロボット

ロボットは K-TEAM 社の KheperaⅡを想定した。直径は 70㎜、車輪間距離は 54㎜ である。センサは CCD カメラと標準で装備されている赤外線センサを想定した。カメラ画像では色情報が扱え、赤外線センサは約 3cm 以内の物体を感知できるので、これらにより状態空間を構成することを想定した。

2-4 タスク

ロボットのタスクは、図 1 に示すような正方形フィールド内で静止障害物を回避しながら静止目標物に到達することである。また環境の変化については、ロボットと目標物と障害物の相対的配置の変化、障害物の個数および形状の変化を想定した。

2-5 学習手法

視野 60° を 20° ずつ左前方、前方、右前方に分け、各領域内で目標物、障害物それぞれが近くにあるか、遠くにあるか、または無いかを認識することにより、状態を 729 に離散化した。また目標物の高さは、障害物の高さより低いものとする。図 2 の状態は、左前領域では目標物は遠く障害物は近い、前領域では目標物は遠く障害物はない、右前領域では目標物は無く障害物は遠いと認識する。行動は、前進(両車輪 9㎜ 前回転)、右前進(左車輪のみ 9㎜ 前回転)、左前進(右車輪のみ 9㎜ 前回転)、右後進(左車輪のみ 9㎜ 後回転)、左後進(右車輪のみ 9㎜ 後回転)、後進(両車輪 9㎜ 後回転)の 6 種類を用意した。また図 3 に示すような木構造を用い、状態 S を、それぞ

れが独立した Q-table を持つ 6 つの状態集合 $S_r \subset S$ に分割した。Q 学習の設定は以下の通りである。

- 学習率 α=0.3、減衰係数 γ=0.9、ボルツマン温度 τ=1/β=1/6.2 とする。
- 報酬は、目標物に到達すれば 1、障害物に衝突する、もしくはフィールドから外れたら-1、その他は-0.08 を与える。
- センサ情報の取得と行動のセットを 1 step とする。
- step を繰り返し、目標物に到達、障害物に衝突、フィールドから外れる、または決められた時間を消化すると 1 episode を終了する。

2-6 再学習手法

1) 初期環境で通常の Q 学習を行い、行動政策 f : S→A を獲得する。
2) 行動政策 f により行動する。
3) -1 の報酬が与えられたら、その状態が属する状態集合 $S_r \subset S_p$ の Q-table を新たに作成し 4)へ。-1 の報酬が与えられなければ 2)へ。ここで $S_p \subset S$ は、行動政策 f による行動で-1 の報酬が得られた状態と、同一の状態集合 S_r に属する状態の集合である。
4) 状態 s_t での行動は

   $s_t \in S_r$ なら Q-table により $P(s_t, a_t)$

   それ以外は $f(s_t)$

   に従う。
5) 行動政策 $f(s_t \in S_r')$ による行動で-1 の報酬が与えられたら、$S_r' \subset S_p$ の Q-table を新たに作成する。
6) 学習が収束するまで 4)、5)を繰り返し、収束後は得られた行動政策を新たに f とし 2)へ。

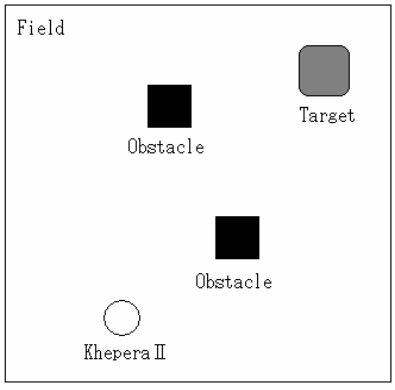

図 1 シミュレート環境

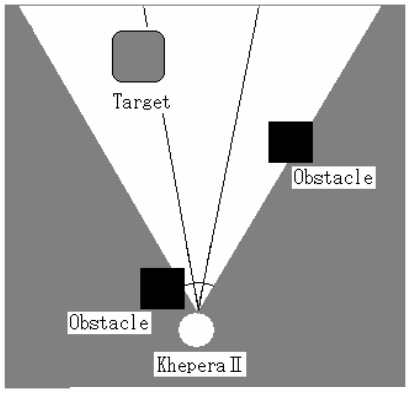

図 2 視野の離散化

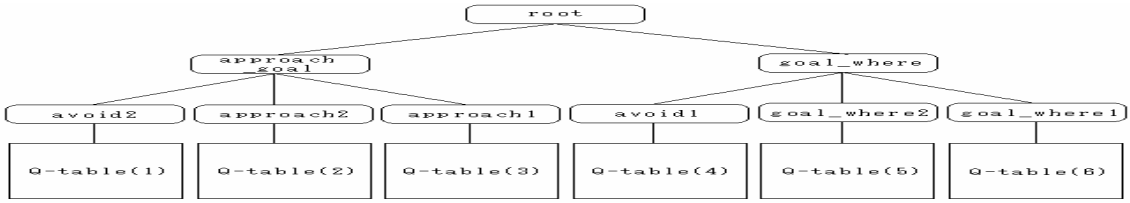

図 3 木構造

3. 結果

図 4 に示す初期環境にて最初の行動政策を獲得させた。その際、スタート地点は三つに分けそれぞれから均等に学習させた。行動政策獲得後のロボットの軌道についても図 4 に示した。次に図 5 に示す環境で、提案手法、Policy Transfer method のそれぞれで行動政策を修正した。ここで Policy Transfer method とは、-1 の報酬が与えられたらその状態に関して以後 Q-table で行動する従来手法である。

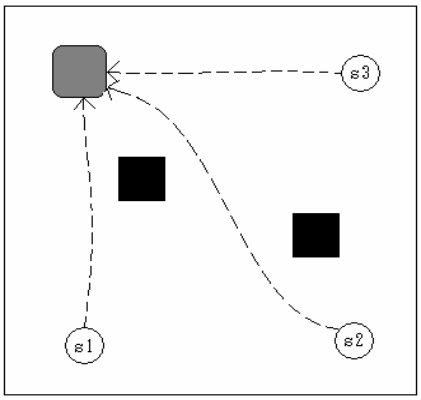

図 4 初期環境

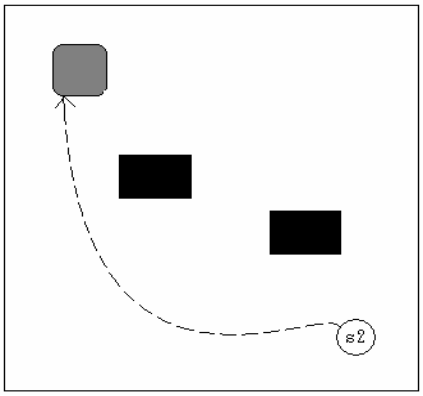

図 5 再学習後の行動

スタート位置が s1、s3 の場合は、どちらの手法でも初期環境とほぼ同一の行動が見られた。スタート位置 s2 の場合は、提案手法、Policy Transfer method ともに図 5 に示すような軌道が得られた。しかし、提案手法は明らかにスムーズにゴールに向かうのに対して、Policy Transfer method は、途中で止まってしまったり、スタート方向に戻ったりといった行動が見られた。これは、Policy Transfer method は-1 の報酬が与えられた状態しか Q-table で行動しないので、二つの状態を行き来することが多いためと考えられる。

4. 結論

高井の手法は、環境変化に応じてサブゴールを設計する必要があった。しかし、環境ごとにサブゴールを設計するには、あらかじめ設計者が環境の変化をすべて想定しなければならない。港の手法では、失敗した状態から順にさかのぼって再学習を行ったが、本研究では木構造を用いることのより順にさかのぼらずとも効果的に再学習すべき状態集合を抽出することに成功した。これにより、環境の変化が大きいほど従来手法に比べ効果が大きいと考えられる。

参考文献

1) 港 隆史,浅田 稔:“環境の変化に適応する移動ロボットの行動獲得”,日本ロボット学会誌、Vol.18,No.5,pp.706-712,2000

2) 高井 悠宇,須鎗 弘樹, “実環境内におけるロボットの Q-learning とその行動政策の逐一修正の検討”,信学技報,NC2004-184,pp.89-93,2005