5Q-7

移動ロボットにおける走行特性を組込んだ方策の強化学習

鈴木亮次† 五十嵐治一‡ 石原聖司* 田中一基**
芝浦工業大学† 芝浦工業大学‡ 近畿大学* 近畿大学**

1. はじめに

人間が移動型ロボットを操縦する際には，予めロボットの走行特性を頭に入れた上で操縦を行うことが熟練者の技とされる．ところが，自律移動型ロボットの場合は，毎時刻，自らが自分自身の走行特性を考慮し，移動方向や回転速度を決定する必要がある．

一般に，適切な行動決定（方策）を強化学習により獲得する際には，こうした走行特性は環境ダイナミクスの中に含まれており，環境の状態遷移確率の中で表現されている．強化学習の中でよく知られている TD 学習や Q 学習[1]においては，こうした状態遷移確率を直接学習するわけではなく，状態価値関数 V(s)や行動価値関数 Q(s,a)の値として間接的に学習することになる．すなわち，ロボットの走行特性の知識も問題解決のための知識も一体となって，価値関数の値の中で表現されている．

本研究では，この走行特性と，ロボットの走行特性に依存しない問題解決のための一般的な知識（以下，行動知識）とを区別する．その上で，両者の知識を方策の中で分離して表現する．さらに，走行特性を予め測定しておくことや，行動知識としてヒューリスティクスを用いることにより，学習の効率化を図るとともに，走行特性に応じたより適切な行動知識の獲得が可能であることをシミュレーションにより示す．

2. 方策勾配法による行動学習

2.1 目的関数と方策

前章で述べた走行特性と行動知識の分離については，すでに方策勾配法における分離方式が提案されている[2]．本研究でもその方式を採用する．以下，概略を説明する．方策勾配法では TD 法や Q 学習などとは異なり，状態価値関数や行動価値関数を求めることなく，方策中のパラメータを直接学習することができる．また，環境や方策のマルコフ性を仮定することもなく，方策の表現に柔軟性がある[3]．

まず，各時刻における行動決定問題をある目的関数 $E(a)$の最小化問題として定式化する．ただし，学習中の行動決定時には，次のボルツマン型の確率的方策により行動を決定する．すなわち，状態 $s\in S$ におけるエージェントの行動 $a\in A$ を決定する方策$\pi(a;s,\omega,\theta)$を，

$$\pi(a;s,\omega,\theta)\equiv\frac{e^{-E(a;s,\omega,\theta)/T}}{\sum_{b\in A}e^{-E(b;s,\omega,\theta)/T}} \quad (1)$$

と定義する．目的関数 $E(a;s,\omega,\theta)$としては，

$$E(a;s,\omega,\theta)\equiv-\sum_{s'}\omega(s,s';a)\theta(s') \quad (2)$$

を採用する．ここで，$\omega(s,s';a)$ は状態 s で行動 a を選択したときに状態 s'へ遷移する度合いを表す走行特性パラメータを，$\theta(s)$ は状態 s の価値を表す行動知識パラメータである[2]．(2)では，行動 a の選択により得られる遷移先の状態 s'の状態価値$\theta(s')$に基づいて，状態 s における行動 a の妥当性を目的関数により評価している．

2.2 状態価値パラメータの学習則

本研究ではエピソードごとのパラメータ更新を考える．各エピソードは実際に選択した行動列 $\{a(t)\}$ と実現された状態列 $\{s(t)\}$ で表される．エピソードごとに与えられる報酬 r の期待値 $E[r]$ を極大化するパラメータを求めるために，(2)の目的関数 $E(a;s,\omega,\theta)$中の状態価値パラメータ$\theta(s)$による $E[r]$ の勾配を計算する．この勾配を用いて，$\theta(s)$の学習則は以下のように与えられる[2]．

$$\Delta\theta(s)=\varepsilon\cdot r\cdot\sum_{t=0}^{L-1}e_{\theta(s)}(t) \quad (3)$$

ただし，$\varepsilon(>0)$ は学習係数であり，$e_{\theta(s)}$は特徴的適正度

$$e_{\theta(s)}(t)\equiv\partial/\partial\theta(s)\ln\pi(a(t);s(t),\omega,\theta) \quad (4)$$

$$=\frac{1}{T}\left[\omega(s(t),s;a(t))-\langle\omega(s(t),s;a)\rangle_{\pi}\right] \quad (5)$$

である．$\langle\cdots\rangle_{\pi}$は(1)の分布$\pi$による期待値を表す．また，本研究では走行特性に関する知識$\omega(s,s';a)$は予め測定し，固定しておくものとする．

3. 学習実験

3.1 問題設定

計算機シミュレータ上で Fig.1 に示すような走行特性を持った全方位移動型ロボットを考え

Reinforcement Learning of Mobile Robots with Policies Including Environmental Dynamics
† Ryoji Suzuki, Shibaura Inst. of Tech.
‡ Harukazu Igarashi, Shibaura Inst. of Tech.
* Seiji Ishihara, Kinki Univ.
** Kazumoto Tanaka, Kinki Univ.

る．このロボットを一定の距離（l=75cm）だけ離れた円形のゴール領域（半径ρ=15cm）へ最短時間で到達させたい．ただし，ロボットはある時間間隔(=0.5s)ごとに移動方向ϕだけを 1°刻みに決定するが，実際の移動後の位置と姿勢はϕごとに定められた走行特性に従うものとする．

3.2 走行特性

ロボットの形状は前方が平らな円形状（半径15cm）である．初期位置は Fig.1 の座標軸の原点上で，y 軸の正の方向を向いている．進行方向 ϕ はロボット正面方向から左回りにはかる．Fig.1 には，ϕ=0,30,60,…,330 の場合の 0.5s ごとの位置と姿勢が示されている．前方向よりも後方向への走行速度の方がかなり速い．

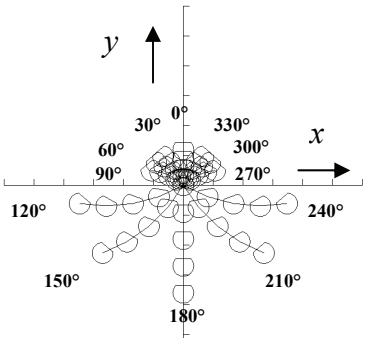


Fig.1 ロボットの走行特性

3.3 学習条件

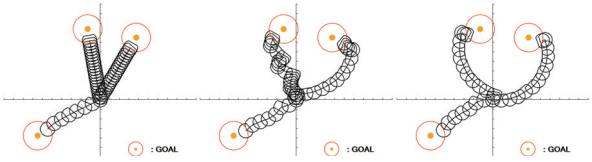
学習時には，初期状態におけるゴールへの進行方向ϕ_g を 0°～360°の間で 1°単位でランダムに設定した．環境の状態 s は，ロボットを中心とする相対的な極座標系上でのゴール位置を表す．ただし，角度を 10°ごとに，距離を 2.5cm ごとに分割したセル状の離散的状態表現を採用する．移動方向 ϕの選択は 1°単位で 0.5s ごとに行う．1 エピソード内での選択回数をエピソード長 L とする．したがって，到達時間 t_L は t_L=0.5L(s)である．報酬 r として $L\leq 60$ の場合は $1/t_L^2$ を与え，L>60 の場合は，r=0 とした．また，温度 T = 0.2，学習係数ε = 0.0005 とし，最大 15000 エピソードの学習を行った． 学習対象のパラメータ$\theta(s)$の初期値としては，ロボットの原点を中心とする円錐状のポテンシャルを与えた．値は中心で 1，150cm 離れた地点で-1 であるが，セル内で同じ値を取るように状態と同様に離散化されている．

3.4 実験結果

Fig.2 と Fig.3 に， ϕ_g = 0°,10°,120°,330°の場合の greedy 法(毎時刻ϕをゴール方向にとる)，学習前と学習後の軌跡を示す．Fig.4 には 10°ごとのϕ_g に対する到達時間(s)を示す．ただし，このときは T= 0.001 としたが，状態の離散化の影響から確率的要素が入るので， 10 回の平均値を示した．

(a) greedy 法　(b) 未学習　(c) 学習後

Fig.2 ϕ_g = 0°の軌跡の例(T= 0.001)



(a) greedy 法　(b) 未学習　(c) 学習後

Fig.3 ϕ_g = 10°, 120°,330°の軌跡の例(T= 0.001)

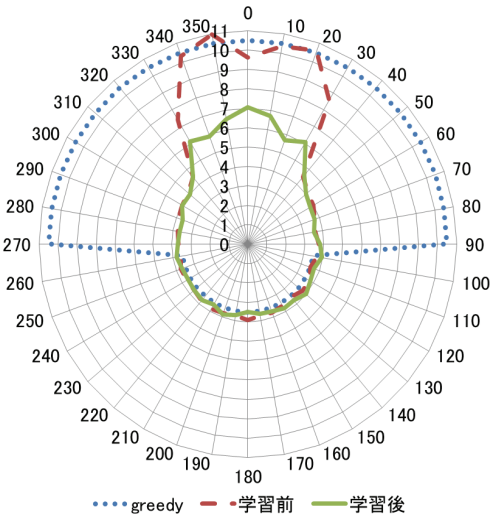


Fig.4 ϕ_g が 10°ごとの到達時間(s)の比較（10 回の平均）

4. 考察

Fig.2～Fig.4 より， ゴールが前方の場合，greedy 法では直線的に進み到達時間が長くかかる．走行特性と行動知識とを与えると途中からそれらを利用して移動速度が速い横方向や後方への移動行動を行っている．さらに強化学習により行動知識を微調整すると，かなり早い時期からこれらの方向への移動行動を取ることにより到達時間を 1/3～1/2 程度までに短縮している．

文献

[1] R.S.Sutton, A.G.Barto: Reinforcement Learning, MIT Press, 1998.

[2] 石原聖司，五十嵐治一，“方策こう配法を用いた行動学習―環境のダイナミクスと行動知識の分離－”，電気学会論文誌 C, 129 巻 9 号，pp.1737-1746(2009).

[3] 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こう配法の適用−追跡問題−”，電子情報通信学会論文誌 D-I, Vol.J87-D1, No.3, pp.390-397(2004).