移動ロボットのためのナビゲーションシステム ～その1：分割予測報酬を取り入れた強化学習～*

6J－1

溝口文雄†　佐藤淳一†　長谷部正信†　平石広典†

東京理科大学 理工学部‡

1 はじめに

移動ロボットにおいては，動的に変化する周囲の状況へのすばやい対応が重要である．一般に，多くの場合において強化学習が使われるが，報酬など，定義が難しい部分が多く存在する．

そこで，本稿では，高速にその場に適した動作を行なえるようにした新たな強化学習システムを示す．

ここで示す新たな考え方は，その場に適した報酬の再定義などに基づいており [1]，これまでの強化学習には無い，報酬遅れの問題の解決や，速い収束を可能にする報酬分割，報酬予測などの特徴を持った学習システムとなっている．さらに，これまでに開発した強化学習システム [2] と比較しても，より柔軟な動作決定を行なうことが可能である．

2 本システムのアルゴリズム

2.1 概要

本システムは，新たな報酬の与え方を取り入れたものであり，図1に示すようなものである．

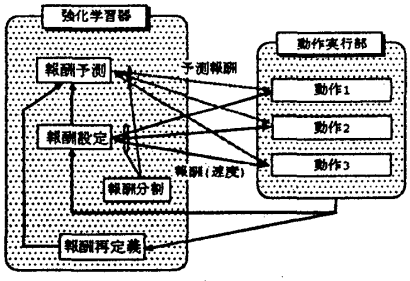

図 1: 新たな強化学習アプローチ

なお，ここで用いるロボットは，Nomadic Technologies,inc. の Nomad200 と呼ばれる移動ロボットである．また，本強化学習システムが用いるデータは，ロボットの周囲に配置されている 16 個のソナーセンサーから得られる障害物までの距離データである．

2.2 報酬分割

報酬を各移動方向ごとの速度に分割すると，その方向の成分について強化を行なうことが可能となる．そこで，移動ロボットの形状およびソナーセンサーに合わせ，これから移動する方向を仮定した場合のそれぞれの予測報酬 (後述) を求める．

ここでは，左回転，直進，右回転のそれぞれの方向ごとに分割する．$reward\_c$=直進方向報酬，$reward\_r$=右方向報酬，$reward\_l$=左方向報酬，$Sonar\_no$=障害物までの距離が最小のソナー番号とると，予測報酬を求める関数 F(x) は，

$$reward\_c = F(0)$$
$$reward\_r(reward\_l) = F(Sonar\_no)$$
$$reward\_l(reward\_r) = F(16 - (Sonar\_no - 1))$$

と定義される．

$reward_r$と $reward_l$を求めるために使われる $Sonar\_no$ は，図 2に示した位置関係になる．これにより，その場に適した移動方向成分ごとの報酬が設定可能となり，細かな速度の再定義が可能となる．

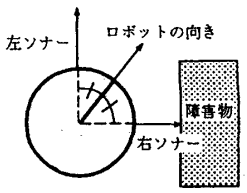

図 2: 使用するソナーの位置関係

2.3 報酬予測

1サイクル先の報酬予測をすることによって，これまでの強化学習の問題点であった報酬遅れを解決する．

ここでは，予測報酬をそれぞれの分割報酬へ適用する．予測報酬を求める関数を $F(x)$ とすると，

$$F(x) = \sum_{n=x-1}^{x+1} Stc[n] \quad (rewc \text{ の場合})$$

*Navigation system for mobile robot. -Reinforcement learning system using splited and estimated rewards-

†Fumio MIZOGUCHI, Junichi SATOH, Masanobu HASEBE, Hironori HIRAISHI

‡Faculty of Sci. and Tech. Science University of Tokyo

$$Stc(n) = \begin{cases} St(n) & (St(n) \leq AvgSt) \\ AvgSt & (St(n) > AvgSt) \end{cases}$$

$$F(x) = \sum_{n=x-1}^{x+1} St[n] + rdeg \quad (rewl, rewrの場合)$$

と定義する．ただし，$AvgSt$=センサーからのデータの平均，$rdeg$=ロボットの向きとゴール方向角度差である．また，x=次のステップの進行方向を示す値 (0から，15 の間の整数)，$St[n]$=ソナー番号に対応した距離データ (実際のデータを定数倍したデータ)，n=ソナーセンサー番号である．

この処理を前で述べた報酬分割により，3パターンにおいて実行することにより，次のステップでの報酬の予測が可能となり，前もって適した動作が可能となる．

2.4 速度再定義

これまでに示した報酬を用いることにより，各方向ごとに分割した速度を細かく調節することが可能になる．

動作速度決定関数を $G(x)$ とすると，

$$G(REW) = \begin{cases} f(REW) & (\max(REW) = reward\_c) \\ g(REW) & (\max(REW) \neq reward\_c) \end{cases}$$

なお，$REW = reward_l, reward_c, reward_r$である．

$f(REW)$ と $g(REW)$ は，共に前ステップの報酬と現在の報酬の差から速度に変換する関数であるが，それぞれは，$f(REW)$:直進方向速度を決定する関数，$g(REW)$:回転方向速度を決定する関数である．ここで，それぞれの関数は以下のように示される．

$$f(REW) = \alpha \cdot (reward\_c - REWARD\_c)$$

$$g(REW) = \beta \cdot (reward\_l \ \ (reward\_r) - REWARD\_t)$$

ここで，$REWARD\_c$,$REWARD_t$は，それぞれ前回の動作を行なった結果得られた直進方向，回転方向のスピードに対する報酬である．また，α, βは，定数である．

これにより，従来の強化学習の動作選択とは違い，速度の大きさとして学習結果を反映させることが可能となる．

2.5 報酬の再定義

細かな動作を行なわせることと同時に，その場で適した報酬を与えることが可能となるように，予測報酬自体の再定義を行なうことを考える．

予測報酬を Rとすると，新たな予測報酬の定義は，以下のように行なう．

$$R = \alpha \cdot F(x)$$

このように，直接変更するのは報酬を求める $F(x)$ ではなく，その重みの係数αである．ここでは，直進の報酬設定にのみ行ない，最小回避距離以下の場合，$\alpha = 0.5$ であり，その他の場合，$\alpha = 1.0$ である．これにより，状況にあった報酬の定義が可能となる．

3 実験

3.1 環境

これまでに示した，強化学習システムの有効性を示すために，実際の移動ロボットに適用した実験を行なった．ここでの実験は，局所的な始点と終点のみ設定し，2 地点間の移動を行なったものである．

3.2 結果

実際に本強化学習システムを用いた局所的な最適動作決定の実験環境とその結果を図3に示す．左が実験環境，右がロボットの経路とロボットからのソナーデータをプロットした実験結果である．左上の点がスタート，右下の点がゴールである．

図 3: 実験環境と結果

始点と終点のみ与えているが，図 3の右図に示したように，本強化学習システムのみで，障害物を回避して最適な行動を行なっていることがわかる．また，その場に適した速度決定が行なえた．

4 まとめ

本稿では，新たな考え方を取り入れた，新たな移動ロボットのための強化学習システムについて述べた．実験結果から，本システムは，局所的な部分での動作決定をより柔軟に，高速かつより細かく行なえることが可能であるということが分かった．

参考文献

[1] Lonnie Chrisman, "Reinforcement Learning with Perceptual Aliasing: The Perceptual Distinctins Approach", *CS-CMU PA 15213*,1992

[2] 溝口，佐藤，長谷部，平石，" 移動ロボットにおけるエージェントのための強化学習", 人工知能学会全国大会 (第9回) 論文集,1995