ロボットの行動獲得のための自律的な状態空間の構成

◯野田 彰一　　浅田　稔　　細田　耕
大阪大学工学部

Action-Based State Space Construction For Robot Behavior Acquisition

◯Shoichi NODA　　Minoru ASADA　　Koh HOSODA
Osaka University

1　はじめに

近年，ロボット自身や環境のモデルを前提とせずに行動を獲得する手法として強化学習が注目されてきている[1)]．強化学習では，学習時間が状態数に対して指数関数的に増大するため[2)]，周囲の環境が複雑な実ロボットでは，いかに状態空間を構成するかが問題となる．

これまでのコンピュータシミュレーションを中心とした研究では，恣意的にずれのない状態・行動空間を構成しているが，実際のロボットではずれのない理想的な空間を構成することは難しい．文献[3]では画像情報から状態空間を構成し，状態の変化が起こるまでの連続した行動を一つの行動とみなすことで，ずれ問題に対処したが，完全にずれが解消できるとは限らない．

本稿では，逆にロボットの行動コマンドを基に状態空間を自律的に構成し，ロボットのパフォーマンスを向上させる手法を提案し，その有効性を検証する．

2　Q学習

強化学習の代表的手法の一つとしてQ学習[4)]がある．Q学習では，状態$s \in \mathbf{S}$において行動$a \in \mathbf{A}$をとり，次状態s'に遷移した時，行動価値関数値$Q(s,a)$を以下のように更新する．

$$Q(s,a) \Leftarrow (1-\alpha)Q(s,a)+\alpha(r(s,a)+\gamma \max_{a' \in \mathbf{A}} Q(s',a')) \tag{1}$$

ここで，αは学習率，γは減衰係数である．また，$r(s,a)$は報酬であるが，局所解に陥らないように，ゴール状態に対して1，それ以外は0とする場合が多い．

Q値が与えられると，各状態sに対して$Q(s,a)$が最大となる行動aを選ぶことによって政策が定義される．

3　状態空間の構成

ロボットの識別できる状態空間を，例えば人間が適当に格子状に分割したとしても，その状態空間がロボットの行動空間に対応するとは限らない．同じ状態で同じ行動をとっても状態遷移にばらつきが生じる状態空間では，ゴール状態への状態遷移が不確実であるので，ロボットのパフォーマンスが低下することになる．パフォーマンス向上のためには，ゴールに向かう行動の状態遷移確率が高い状態空間を構成する必要がある．そこで，ゴール状態と行動に基づいた状態空間の分割アルゴリズムを以下に示す．

1. ゴール状態を目標状態とする．
2. ランダムに行動し，各行動をとった時，目標状態に到達可能な状態変数$\boldsymbol{x}$を蓄える．ただし，すでに区分された領域内にあるものは蓄えない．
3. 蓄えられた状態変数を各行動ごとに状態として領域に区分する．状態空間がm次元の時，状態の分布が多変量正規分布に従うと仮定し，m次元の楕円体で状態変数をおおう．多変量正規分布の確率密度関数は次式で表される．

$$f(x_1,\cdots,x_m) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

$$\boldsymbol{x} = [x_1,\cdots,x_m]^T,$$

$$\boldsymbol{\mu}: \text{平均}, \quad \Sigma: \text{共分散行列}$$

4. 各行動で区分された領域の論理和をとった領域を次の目標状態とする．重なる領域に対しては，分散で正規化した距離

$$\Delta = (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

の値の小さな方をとる．

5. 目標状態に到達可能な状態変数が無くなれば終了，さもなくば2に戻る．

4 サッカーロボット

実ロボットとして，**Fig.1**に示すボールをゴールにシュートするサッカーロボットを考える．

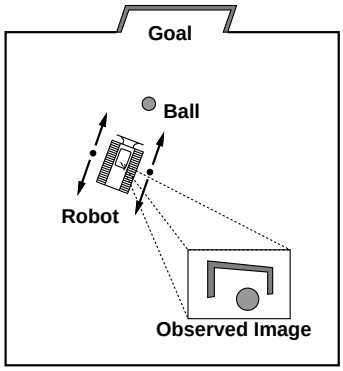

Fig.1 サッカーロボット

ロボットが得られる環境からの情報は，ロボットに固定して搭載されたカメラからの画像情報のみとする．状態空間は画像 (512×480 画素) 上のボールの位置 (重心の x 座標) と直径，ゴールの位置，大きさ，ゴールの傾きの5次元から構成されている．

ロボットは左右の車輪を独立に動かすことができ，行動は左右輪が前進，停止，後進の合計9通り (ただし左右輪停止を含まない) である．

また，報酬はロボットがボールをゴールに入れた状態と行動に対して1，それ以外は0を与える．

サッカーロボットでは，画像情報を用いているため，ボールやゴールが近い場合には状態変数の変化量が大きいのに対し，遠いものに関しては変化量が小さいという特徴がある．

5 実験結果

多次元の楕円体として状態を記述する時，楕円体が小さいほど政策に沿った行動に対する状態遷移確率は高くなるが，本来含まれるべき状態に入らない状態変数が増え学習に悪影響を及ぼす上に，状態数が増加し学習時間も長くなる．楕円体が大きいとその逆の状況になるため，楕円体の大きさを決める閾値はシュート率，状態数を考慮して定めた．この結果，状態空間分割アルゴリズムを使うと，ボール，ゴール共に見えている状態に対しての状態数は56個になった．先の研究[3)]では，状態空間を構成する5次元のパラメータを人間がそれぞれ3つに分割しており，状態数は $3^5 = 243$ 個であった．ボール位置，ゴールの位置，傾きがいずれも0(真正面) での状態空間の断面を **Fig.2**に示す．図では状態が分割された順 (ゴール状態に近い順) に濃淡が濃くなっており，格子状の線は，人間が分割したときの状態の境界である．

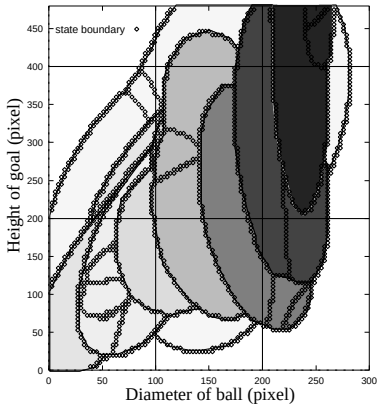

Fig.2 分割された状態空間

また，この状態空間を用いてQ学習を行なった結果のシュート率を **Table 1**に示す．表には，比較のために従来の手動で分けた状態空間の場合も示している．Q学習を行なう時は，ボールやゴールが見えない状態に対して，それぞれに「左に消えた」,「右に消えた」状態を付加して行なった．

Table 1 サッカーロボットのシュート率と状態遷移確率

	this method	previous work[3)]
# of states	56	243
shooting rate(%)	91.2	86.2

6 おわりに

本稿では，強化学習における適切な状態空間をロボットが自律的に構成するアルゴリズムを提案し，その有効性を検証した．ここで行なったことは状態変数ベクトルの量子化であるが，状態変数自体の選択については，今後の課題としたい．

参考文献

[1] J. H. Connel and S. Mahadevan, editors. *Robot Learning.* Kluwer Academic Publishers, 1993.

[2] S. D. Whitehead. “A Complexity Analysis of Cooperative Mechanisms in Reinforcement Learning”. In *Proc. AAAI-91*, pp. 607–613, 1991.

[3] 野田, 浅田, 俵積田, 細田. “強化学習によるロボットの行動獲得の効率化に関する考察—簡単なタスクからの学習LEM—”. 第4回ロボットシンポジウム予稿集, pp. 67–72, 1994.

[4] C. J. C. H. Watkins. *Learning from delayed rewards”*. PhD thesis, King’s College, University of Cambridge, May 1989.