強化学習によるロボットの行動獲得のための状態空間の自律的構成

◯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 by Using a Reinforcement Learning

◯Shoichi NODA　　Minoru ASADA　　Koh HOSODA
Osaka University

Abstract – An input generalization problem is one of the most important ones in applying reinforcement learning to real robot tasks. To cope with this problem, we propose an action-based state space construction. We apply this method to a task in which a soccer robot shoots a ball into a goal and give the experimental results and a discussion.

1　はじめに

実世界を動き回る自律ロボットには，動的な環境に対応して敏速に行動し，自律的に作業を遂行することが望まれる．近年，このような適応的，反射的かつ合目的な行動を環境やロボット自身のモデルを前提とせずに獲得する手法として，強化学習が注目されてきている [1].

しかしながら，強化学習の研究のほとんどはコンピュータシミュレーションによる実験を示すのみであり，これらの研究では，格子空間などの大局的な情報や，単純なセンサのオン・オフの情報から，理想的な状態空間を構成しているものが多い [2–6]．このような理想的な状態空間では，ロボットの行動と行動の結果としての状態遷移が1対1に対応するように構成されている．だが，実際の環境内で動く自律ロボットを考えるとき，状態空間はロボット自身が持つセンサによって構成されるべきである．このように構成された状態空間は，ロボットの行動空間と1 対1に対応するとは限らず，強化学習を適用するにはこの “状態と行動のずれ” 問題を解決しなければならない [7].

また，強化学習においては，学習に要する時間は状態の数に対して指数関数的に増大する [8]．ロボットが持つセンサの分解能で，観測し得る全ての状態を区別すれば，状態数が不必要に増えてしまうので，ある程度の粗さをもった状態空間を構成しなければならない．この問題は，入力の一般化問題と呼ばれるものである [9, 10]．このため，複雑な環境下で動作する実ロボットが学習するためには，学習時間の短縮と，状態空間の適切な構成による状態数の軽減が不可欠である．

本研究では，入力の一般化問題を解決するために，ロボットが状態空間を自律的に構成する手法を提案する．この手法では，ロボット自身の経験を通して状態空間を構成するため，人間が機械的に分割して与える状態空間と比べて無駄な状態がなく，効率的な学習をすることができる．この手法の有効性を，ロボットがボールをゴールに入れるタスクを例に，シミュレーションおよび実ロボットによる実験を通して検証する．

2　強化学習

強化学習の代表的手法の一つとしてQ学習 [11] 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．

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

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 状態空間の自律的構成

ロボットが識別する状態空間を，人間が適当に分割しても，それがロボットの行動空間に対応するとは限らず，タスクにとって最適な分割になっている保障はない．このような状態空間では，

- 状態遷移のばらつき
- 不必要に分割された状態

が存在する恐れがある．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，タスクの達成に障害となる．また，不必要な分割がなされている場合は，状態数が多くなり，学習に時間がかかる．そこで，ロボットが自らの経験を通して状態空間を構成することにより，上記の問題の解決を図る．

状態空間は，与えられたタスクに応じて，状態遷移が行動とできる限り 1 対 1 に対応するように構成すべきである．そのために，**Fig.1**に示すような，ゴール状態と行動に基づいた状態空間を考える．ゴール状態に一つの行動で到達できる状態をその行動の種類ごとに，$s_{1,k}, \quad k = 1,2,3,...,n \leq |\mathbf{A}|$ とし，それらの集合を $\mathbf{S_1}$ とする．さらに，$\mathbf{S_1}$ に一つの行動で到達できる状態集合を $\mathbf{S_2}$ とする．同様にして，ゴール状態に最低 m 個の行動で到達できる状態集合は $\mathbf{S_m}$ となり，ゴールに到達可能なものは，いずれかの状態集合に含まれることになる．このような状態空間が構成されていれば，政策行動と状態遷移が 1 対 1 に対応することになり，パフォーマンスの向上が期待される．

本稿では，ロボットの行動に対して，

- 固定長の行動要素系列からなる行動，
- 可変長の行動要素系列からなる行動

の 2 つの定義を考える．行動要素とは，ロボットのアクチュエータに対するコマンドであるとする．

この 2 つの行動の定義のいずれにも適用可能な，状態空間を構成するためのアルゴリズムを以下に示す．

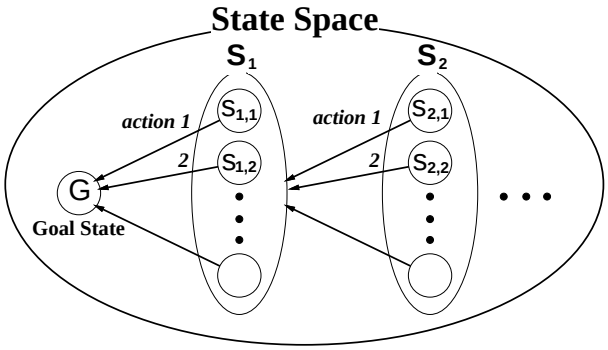


Fig.1 The goal-directed and action-based state space construction

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ベクトル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，状態の分布は m 次元の楕円体内で一様な分布とする．状態変数ベクトル $\boldsymbol{x}$ の平均ベクトルを $\boldsymbol{\mu}$，分散共分散行列を $\boldsymbol{\Sigma}$ とすると，楕円体の境界は，

   $$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = m+2$$

   で表される [12].
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 (マハラノビス距離)

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

   の近い方をとる．
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．

3.1 固定長行動系列

一定時間続けた行動要素の系列を一つの行動と定義し，状態空間の構成アルゴリズムを行う．

3.2 可変長行動系列

状態が変化するまでとり続けた行動要素系列を一つの行動と定義し，状態空間の構成アルゴリズムを行う．初期状態は，目標状態を除いてロボットの認識できる状態空間全てを覆う一つの状態であるとする．

固定長行動系列の場合には，同じ行動をとってゴールできる状態でも，行動をとる時間の長さによって細かく分かれる．そのような状態が，可変長の行動では一つの状態とみなすことができる．

この可変長行動系列による状態空間の構成法の有効性を示すために，簡単なシミュレーションを行なった．シミュレーションは，100×100 の空間の中心に直径 10 の円があり，前後左右に毎ステップ 1.0 だけ進む行動要素をとることができるロボットの中心が円内に入ればゴールとする．初期配置はゴールの円内に入らぬようにランダム，ロボットがフィールドから出ればリセットされる．状態パラメータはロボット中心の絶対座標 (x, y) の 2 次元である．

状態空間の分割結果を **Fig.2**に示す．分かれた状態数は全部で 12 個であり，分かれた順が早いほど明度が低くなっている．また，図中の矢印は，その状態がどの行動でラベリングされて分割されたかを示している．

もし，固定長行動系列による行動の場合でこのシミュレーションをすれば，一つの行動にかける時間や，その行動の速度，フィールドの大きさなどによって，得られる状態数などのシミュレーション結果が大きく異なってしまう．しかし，可変長行動系列を行動と定義すれば，そのようなパラメータにあまり依存せずに状態空間が構成される．このことは，この行動の定義とアルゴリズムの組合せが，与えられたタスクに対して普遍的な状態空間を構成する上で有効であることを示している．

4 タスクと仮定

実ロボットの例として，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．サッカーロボットとその環境を表したものを **Fig.3**(a) に示す．環境内にはロボットの他に，ボールとゴールしか存在しないものとする．ロボットが得られる情報は，ロボットに搭載されたカメラからのボールとゴールについての画像情報のみである．ボールやゴールの大きさや距離などの三次元情報，カメラパラメータ，ロボット自身の動特性などの先験的な知識は一切与えられていない．

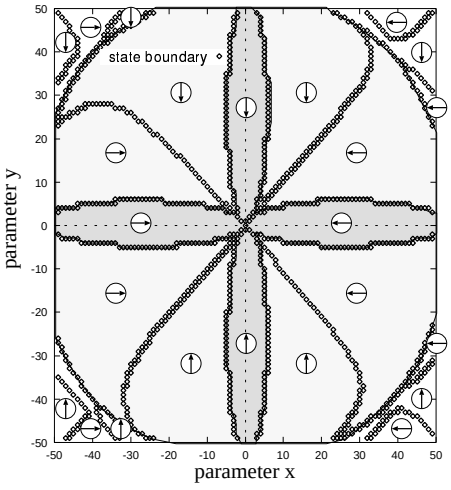
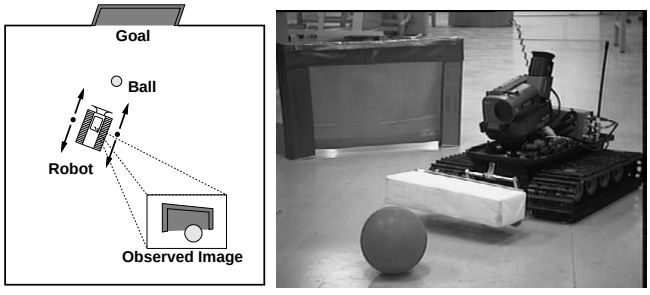


Fig.2 Result of the variable action-based construction of 2D state space

ロボットは，左右の車輪が二つのモータにより独立に駆動される PWS (Power Wheeled Steering) システムを持っている．左右輪がそれぞれ，前進，停止，後進の 3 段階の速度を出すので，合計 9 通りの行動要素が選択できる．ただし，この内，停止行動は状態の変化をもたらさないので選択せず，残りの 8 通りの中から行動要素を選択する．Fig. 3(b) に，実際に用いた移動ロボット，ボール，ゴールを示す．



(a) The task is to shoot a ball into the goal

(b) A picture of the radio-controlled vehicle with a ball and a goal

Fig.3 A task and our real robot

5　シミュレーション

シミュレーションにおける環境はFig.3(a)に示すような，3.0×3.0[m]の正方形のフィールドで，上辺の中央に幅0.9[m]，高さ0.23[m]のゴールがあり，全長0.45[m]，幅0.31[m]のロボットが直径0.09[m]のボールを蹴る．カメラはロボットの中央部についており，画角は36度である．ロボットの最大速度は1.1[m/s]，最大回転角速度は4.8[rad/s]である．ロボットの質量はボールと比べて十分大きいものとし，はねかえり係数は0.5とした．また，ボールの転がり速度は床との摩擦を考えて，各ステップ毎に0.8を掛けて減衰させている．また，画像処理による33[ms]の遅れおよび，モータの立上りの遅れ時間100[ms]を考慮している．

サッカーロボットの状態空間は，ボールの位置，大きさ，ゴールの位置，大きさ，向きの5次元のパラメータから構成される．画像のサイズは512×480である．ボール，ゴールのいずれかが観測されない状態は，これらのパラメータがわからないので，状態空間の自律的構成を行なう対象はボール，ゴールの両方ともが観測されている場合だけとする．Q学習を適用する際には，状態空間の自律的構成による状態の他に，ボールやゴールが観測されない場合を考慮した状態を付加して学習を行なった [7].

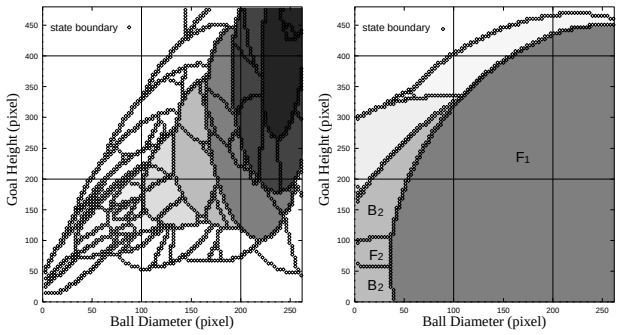
学習率αの値は0.25，減衰係数γの値は0.9で固定とした．報酬としては，ロボットがボールをゴールに入れた状態と行動の行動価値関数値に対して1の値を与え，それ以外は0とした．

また，固定長行動系列により状態空間を構成する場合，ロボットが状態を観測するサンプリング時間である時間ステップ(33[ms])の間とり続けた行動要素系列を行動とした．

Fig.4(a),(b)に固定長行動系列，可変長行動系列で状態空間の構成を行なったときの状態空間を，ボール位置，ゴールの位置，向きがいずれも0(真正面)での断面をとり，ボールとゴールの大きさの2次元で表現したものを示す．図では状態が分割されたのが早い順(ゴール状態に近い順)に明度が低くなっている．また，我々は状態空間を人間が分割した場合についても実験を行なった [7]．その際，5つのパラメータをそれぞれ3つに分割し，ボールとゴールが観測されている場合に対しては$3^5 = 243$個の状態に分割した．図の格子状の線は，人間が分割したときの状態の境界を表している．

固定長行動系列の場合は，ゴールとボールが大きく見えるときに最初の状態が分割されているのに対し，可変長行動系列の場合は，ボールとゴールが真正面にあるので，最初にできた状態の前進で表される大きな状態でほとんどが覆われている．可変長行動系列の場合の図中で，F-と書かれている状態は，その楕円体が前進する行動のサンプルから得られたものであることを示し，B-は後進の行動である．F-,B-の後の数字は，その楕円体が何番目にできたかを示している(1なら状態集合$\mathbf{S_1}$に含まれ，2なら$\mathbf{S_2}$に含まれる)．

視覚情報は，近くのものの変化が大きく，遠くのものの変化が小さいという特性があるので，固定長行動系列の定義では遠くのものの状態が細かく分かれる欠点がある．しかし，可変長行動系列の定義はこの特性に対しても対処できていることがわかる．



(a) Fixed-length actions　(b) Variable-length actions

Fig.4 Result of state space construction

Table 1は，人間が格子状に与えた状態空間，固定長行動系列による状態空間，可変長行動系列による状態空間の3つについて，学習時間，状態数，シュート率を比較したものである．学習時間は，ボールとゴールがロボットの前方にランダム置かれるような初期配置で行なった時のものである．状態空間の自律的構成を行なうものについては，括弧内にそれにかかる時間を示した．単位は時間ステップである．また，状態数はボールとゴールの両方が観測されている時の数であり，シュート率は5000回の試行における結果である．

Table 1 Simulation results

State Space	Learning Time (Search Time)	Number of States	Shooting Rate(%)
By programer [7]	500M	243	77.4
Fixed-length	5M (222M)	107	71.5
Variable-length	2.5M (41M)	33	83.3

学習時間は状態数が少ない程短くなっており，可変長行動系列の場合が最も短く済んでいる．可変長行動系列の場合の状態空間の構成にかかる時間が，固定長行動系列の場合と比べて短くなっているのは，可変長行動系列の場合は経験した状態変数ベクトルの履歴を蓄えておき，目標状態に到達した時にそれらをまとめて保存する手法を用いているからである．固定長行動系列の場合では，33[ms] ごとにランダムな行動を発生させるので，この手法を使用することができない．可変長の手法では，一回の目標状態に到達した経験によって，多くの状態変数ベクトルのサンプルをとることができる．しかしその反面，サンプルの数が少ないと偏りが生じる危険がある．固定長行動系列の場合，蓄える状態変数ベクトルのサンプル数は1000個としていたが，可変長行動系列場合は偏りをなくすためにその10倍にしている．また，いずれの場合も，サンプルが全体の3%以下しかない行動に対しては信頼性が低いとして，その行動で区分される状態は作らなかった．

人間が与えた状態空間と比べると，固定長行動系列の場合は状態数は少なくなっているが，これは人間が与えた状態空間に存在していた無駄な状態がなくなっているためであり，量子化の細かさとしては，Fig.4(a) を見てもわかる通り，固定長行動系列の方が細かい．この量子化の細かさにより，時間遅れによる悪影響を受け易くなり，シュート率が悪くなっていると考えられる．可変長行動系列の場合のシュート率が最も高くなっているのは，連続した行動が多いので，時間遅れにあまり左右されないためであると考えられる．実ロボットで学習することを考えれば，学習する環境には必ず遅れがあるので，可変長行動系列による状態空間が有効であると考えられる．

6　実ロボットによる実験

Fig.5に示す実システムを用いて，可変長行動系列による状態空間の自律的構成の実ロボットによる実験を行なった．詳細は文献 [7] を参照されたい．

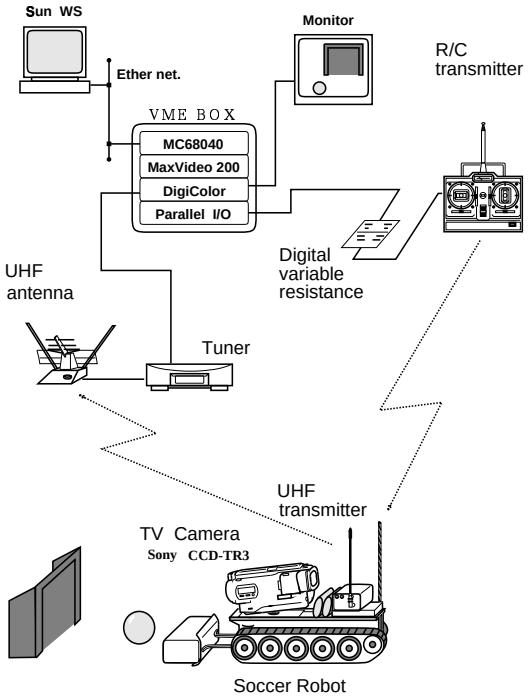


Fig.5 A configuration of the real system

実験は3段階である．実ロボットを動かしてその経験をサンプリングする．つぎに，ワークステーション上でサンプルしたデータを用いて，可変長行動系列に基づく状態空間の自律的構成とQ学習を行なう．そして最後に，学習で得られた政策を用いて実ロボットを動かす．

実ロボットによって得られた経験のサンプルデータは，ノイズが多く，そのままのデータを使うと良好な結果が得られない．これは特にQ学習を行なう時に悪影響があると考えられる．Q学習を行なうと，ノイズがある部分で状態遷移が起こってしまうため，ゴールにたどり着くまでにかなり多くの状態を遷移することになる．このため，本来行動価値関数値Qが高くなるはずの状態であるのに，低い値で留まってしまい，学習がうまくいかない．そこで，サンプルデータ中の明らかなノイズを除去して状態空間の構成および学習を行なった．

ワークステーション上で計算された政策を用いて，実ロボットがシュートする様子を**Fig.6**に示

す.

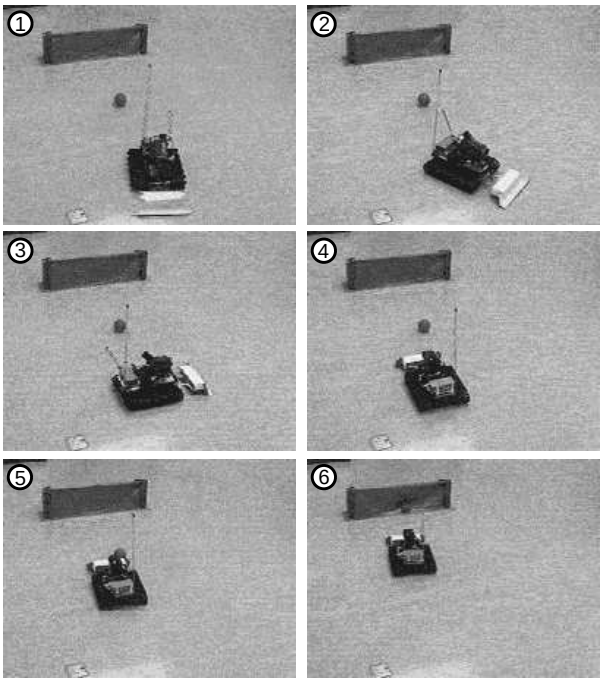


Fig.6 The robot succeeded in finding a ball and shooting a ball into the goal

7 おわりに

本研究では，強化学習の問題点である入力の一般化問題に対する一解決法として，状態空間を自律的に構成するアルゴリズムを提案し，可変長行動系列による状態空間の構成が，実ロボットに対して有効であることを示した.

状態空間の構成時に問題となるのが，楕円体モデルの誤差である．モデル中に含まれるべき空間が含まれなかったり，含まれるべきでない空間が含まれることは理想的には避けたいが，実際のロボットには様々な不確定要因が存在するので完全に誤差がなくなることはない．たとえ誤差の少ないモデルであっても，複雑な計算や多くのメモリを必要とするものは望ましくない．このようなトレードオフを考慮して最適なモデルを選ぶ必要がある.

参考文献

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

2) S. D. Whitehead and D. H. Ballard. “Active Perception and Reinforcement Learning”. In *Proc. of Workshop on Machine Learning-1990*, pp. 179–188, 1990.

3) L. -J. Lin. “Self-Improving Reactive Agents Based On Reinforcement Learning, Planning and Teaching”. *Machine Learning*, Vol. 8, pp. 293–321, 1992.

4) 畝見. 「実例に基づく強化学習法」. 人工知能学会誌, Vol. 7, No. 4, pp. 697–707, 1992.

5) 田野, 三上, 嘉数. 「動的環境における多足歩行機械の適応的歩容計画」. 第11回 日本ロボット学会学術講演会 予稿集, pp. 1103–1106, 1993.

6) 徳本, 三上, 嘉数. 「強化学習を用いた周期的歩行運動の獲得」. 第11回 日本ロボット学会学術講演会 予稿集, pp. 1107–1110, 1993.

7) 浅田, 野田, 俵積田, 細田. 「視覚に基づく強化学習によるロボットの行動獲得」. 日本ロボット学会誌, Vol. 13, No. 1, pp. 68–74, 1995.

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

9) D. Chapman and L. P. Kaelbling. “Input Generalization in Delayed Reinforcement Learning: An Algorithm And Performance Comparisons”. In *Proc. of AAAI-91*, pp. 726–731, 1991.

10) 開, 松原. 「機械学習からみたロボット学習—能動的学習機構に向けて—」. 日本ロボット学会誌, Vol. 13, No. 1, pp. 5–10, 1995.

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

12) H. Cramer. *Mathematical Methods of Statics*. Princeton University Press, Princeton, NJ, 1951.