状態空間の自律的分割による実ロボットの実時間学習

Reasonable Performance in Less Learning Time by Real Robot Based on Incremental State Space Segmentation

○　高橋 泰岳 (阪大)　正　浅田　稔　(阪大)
准　細田　耕　(阪大)

Yasutake TAKAHASHI, Osaka University, 2-1, Yamadaoka, Suita, Osaka
Koh HOSODA, Osaka University
Minoru ASADA, Osaka University

Reinforcement learning has recently been receiving increased attention as a method for robot learning with little or no a priori knowledge and higher capability of reactive and adaptive behaviors. However, there are two major problems in applying it to real robot tasks: how to construct the state space, and how to reduce the learning time. This paper presents a method by which a robot learns purposive behavior within less learning time by incrementally segmenting the sensor space based on the experiences of the robot. The incremental segmentation is performed by constructing local models in the state space, which is based on the function approximation of the sensor outputs to reduce the learning time and on the reinforcement signal to emerge a purposive behavior. The method is applied to a soccer robot which tries to shoot a ball into a goal. The experiments with computer simulations and a real robot are shown. As a result, our real robot has learned a shooting behavior within less than one hour training by incrementally segmenting the state space.
Key Words : reinforcement learning, soccer robot, state-space construction

1　はじめに

近年，反射的かつ適応的な行動を獲得できるロボットの学習法として，強化学習が注目されている[1)]．強化学習ではロボットの置かれている環境やロボット自身に関する先見的知識をほとんど必要としないという好ましい特徴がある．しかしながら，実際のロボットタスクに強化学習を適用する場合，幾つかの問題点が生じる．

1. ロボットと環境を記述するための情報(状態)の選択の問題が生じる．センサーからの出力をそのまま状態を記述する量として選択していたのではセンサーの数が増えるにつれ状態を記述する次元が増え収拾がつかなくなる．

2. 選択された情報をどのように離散化するかという問題が生じる．人間が適当に離散化した状態空間がロボットにとって最適なものとなっている保証はない．状態の離散化が粗すぎると，一つの状態に対する最適な行動が獲得されない(「知覚的見せかけ問題」[2)]と呼ばれている)．また不必要に細かく状態を分割すると，十分近傍で本来同一と見なして良い状態の学習が進まないので経験の汎化が期待できず，また学習時間が状態数に応じて指数関数的に増加し，結果として膨大な学習時間を要することになる．

これまで連続性が仮定できる状態空間を分割する研究はほとんど外部からの報酬信号のみを状態分割に使用し，センサー出力については状態の認識のみにしか使用していなかった．そこで本論文ではセンサー情報をより積極的に使ってセンサー空間を逐次的に分割し，同時に目的の行動を獲得する手法を提案する．

2　強化学習

我々の手法の説明に入る前に簡単に強化学習について述べる．

ロボットが識別することができる状態の集合を$\mathbf{S}$とし，環境に対してとり得る行動の集合を$\mathbf{A}$とする．環境は現在の状態とロボットの行動によって確率的に遷移するマルコフ過程に従うものとする．状態と行動の組(s,a)に対しては報酬$r(s,a)$が定義される．

一般的な強化学習の問題は，減衰する報酬の積算を最大にする政策を見つけることである．環境のダイナミクスを学習しながら政策を決定するものとして，WatkinsのQ学習アルゴリズム[3)]は有効な手法である．

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．

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　連続な空間の自律的分割

本稿では状態空間の分割と学習時間の問題を取り扱う．逐次的に状態空間の分割と統合を行い，また学習時間を短縮するために局所モデルを導入する．

センサー空間の連続性およびセンサの変化量の局所関数近似性を仮定すると，モーターコマンドを起動した時，センサー変化量を計算し記録することで自分のモーターコマンドとセンサー出力の変化量の対応がつく．この様な局所的に近似したモデルをここでは局所モデルと呼ぶ．この局所モデルを持つことで必要な状態遷移をするモーターコマンドを生成することで無駄な探索を削減できる．

いまロボットが経験した範囲で状態空間を構成していくとなると，状態空間を分割する時のヒューリスティックな指標が必要となる．適切なヒューリスティックな指標が指定されればそれによって分割される状態空間と行動によってタスクが遂行できるようになると考えられる．

このヒューリスティックな指標としてが考えられる．

(**A**): センサー出力の予測が合っていない．

(**B**): 同じ状態から同じ行動を取っても行きつく先の状態がばらつくところ

前者によってロボットはそれまでの経験に基づきできるだけ少ない状態数で環境における状態を知ることができる．またセンサー出力の予測が正しいかぎり無駄な探索を削減でき，速く強化学習が進む．またロボットは逐次的に分割していくので環境の動的な変化にも対応できる．しかしながら (**A**) によってゴール状態を考慮しているわけではない．予測が正しくても同じ状態から同じ行動を取っても違う状態に移るのなら意味が無い．同じ行動が望ましい状態遷移をするように分割されるべきである．(**B**) によってこれが達成される．

(**A**) によってゴール状態から離れている状態空間を粗く分割する．一方，(**B**) によってゴール状態付近の状態空間分割が行われる．

4 アルゴリズム

本稿ではセンサー出力の変化量の関数近似として線形モデルを採用する．つまり

$$\dot{\boldsymbol{s}} = \boldsymbol{C}\boldsymbol{s} + \boldsymbol{d}$$

.

4.1 行動空間とデータ構造

従来の強化学習において，行動は離散化されたモーターコマンドをある一定時間出力することであった．実際には，この定義による行動では「状態と行動のずれ問題」が生じることを Asada et. al. が指摘している [4]．そこで先で定義された行動を基本行動とし，状態が変化するまでの基本行動のシーケンスを行動と定義しなおす．

データの集合 D のなかの一つのデータ $d_i \in D \ \ i = 1, 2, \cdots, n$ は，基本行動の集合 M の中の基本行動, $m_{d_i} \in M$，センサー出力 $\boldsymbol{s}_{d_i}$ とセンサー出力の変化量の予測値 $\dot{\boldsymbol{s}}_{d_i}$ で構成される．

学習が始まってから得られてデータ全てを記録していたのではデータ量が時間に比例して増大していくので現実的でない．またセンサーのノイズや環境の変動，モーターコマンドの不確実性により同じセンサー出力，同じ基本行動でもセンサー出力がばらつき，正しいデータが取れない場合がある．そこで新しく得られたデータを d_j としその時の基本行動，センサー出力，センサー出力の変化量の予測値をそれぞれ $m_{d_j} \in M$, $\boldsymbol{s}_{d_j}$, $\dot{\boldsymbol{s}}_{d_j}$ と置くともし

$$|\boldsymbol{s}_{d_i} - \boldsymbol{s}_{d_j}| < \epsilon \text{ かつ } m_{d_i} = m_{d_j}$$

であれば，そのセンサー出力の変化量の予測値 $\dot{\boldsymbol{s}}_{d_i}$ を次式で更新する．

$$\dot{\boldsymbol{s}}_{d_i} = (1-\beta)\dot{\boldsymbol{s}}_{d_j} + \beta\dot{\boldsymbol{s}}_{d_i} \tag{2}$$

ただし $0 < \beta < 1$．こうすることでデータの記憶容量をある値までに抑えることができる．また上の $|\cdot|$ はノルムを表すが，ここでは重み付ユークリッドノルムを使う．

4.2 局所モデル

基本的なモーターコマンドを固定した時のセンサー出力 $\boldsymbol{s}$ とセンサー出力の変化量 $\dot{\boldsymbol{s}}$ の線形モデルで状態空間 (センサー出力空間) を分割していく．

1. 同じ基本行動を出力した時のデータを集める．
2. そのデータのセンサー出力及びその変化量から最小自乗法で線形モデルパラメータを求める．
3. 得られた線形モデルパラメータに対して不変分散を計算し，ある閾値よりも大きい時，センサー出力とセンサー出力の変化量の予測値をデータとし，重みつきユークリッド距離を類似度として最短距離法と呼ばれるクラスター分析を行い二つに分割して 2. に戻る．そうでなければ分割を止めて終る．

それぞれの基本行動における分割されたセンサー空間の論理積をとり，それぞれを一つの局所モデルと見る．

ゴール状態付近で，つまり報酬が与えられる付近で状態を細かくみることで正確にタスクを遂行できるようになると期待される．そこで報酬を与えられたときから数ステップ前までのセンサー出力を記録し，タスクを遂行できたかどうかで状態空間を分割する．

これらの方法で分割された局所モデル一つ一つが表す領域を状態をとみる．分割されたセンサー空間はそれぞれ複数の代表点で表現される．この代表点はそれまでに取ったデータ $d_i \in D$ のセンサー出力 $\boldsymbol{s}_{d_i}$ である．あるセンサー出力 $\boldsymbol{s}_q$ が入って来た時，この $\boldsymbol{s}_q$ に一番近い点を含む状態を $\boldsymbol{s}_q$ の状態と分類する．この方法は NN 識別法 (nearest neighbour classification) と呼ばれる．

4.3 行動の生成

4.1節で述べたように強化学習を適用する際の行動を “状態が変化するまでの複数の基本行動のシーケンス” と定義する．複数の基本行動のシーケンスは以下のように生成される．

現在のセンサー出力 $\boldsymbol{s}_n$ と目的の状態のセンサー出力 $\boldsymbol{s}_d$ から目標のセンサー変化量 $\dot{\boldsymbol{s}}_d$ を計算する．

$$\dot{\boldsymbol{s}}_d = \boldsymbol{s}_d - \boldsymbol{s}_n$$

その状態における局所モデルからそれぞれの基本行動を生成した時のセンサー変化量 $\dot{\boldsymbol{s}}_{m_i}$ が得られる．目標のセンサー変化量と一番近いセンサー変化量を起こす基本行動を生成する．

$$m_d = \arg\min_{m_i} (\dot{\boldsymbol{s}}_d - \dot{\boldsymbol{s}}_{m_i})^2 \tag{3}$$

センサー出力の連続性を仮定しているが，センサーの検出範囲に対象物が存在しない時センサーによって検出できない．現在のセンサー出力と目的の状態のセンサー出力の間のどちらかにおいて対象物がセンサーの検出範囲を越えていた時，式 (3) は適用できない．その様な場合でも状態遷移を起こす行動が必要なので，状態が変わるまで同じ基本行動を生成することも状態遷移をする行動とする．

4.4 経験によって得られた知識の再利用

状態空間を逐次的に分割するごとにそれぞれの状態が変化していくが，それを理由にそれまでに獲得した行動価値関数をリセットするのは行動の学習が進まず，好ましくない．そこでそれまでに獲得した行動価値関数をそのまま残し，その初期値として利用することで，行動の学習を速められると考えられる．

つぎに古い状態空間とそのQ値から新しく分割された状態空間の行動価値関数の計算方法を考える．4.1節に述べたように状態は複数の代表点で表現される．S^{old}とS^{new}はそれぞれ古い状態空間と新しい状態空間を示す．$\boldsymbol{s}_{d_i}(i=1,2,\cdots,n)$, ${s_j}^{old}(j=1,2,\cdots,n^{old})$と${s_i}^{new}(i=1,2,\cdots,n^{new})$はそれぞれ貯えられたデータ$d_i$のセンサー出力，古い状態空間における状態，新しい状態空間における状態を示す．それぞれの代表点の古い状態から新しい状態への変換の表Tは以下の様に計算できる．

1. 2次元配列Tを0に初期化
2. $i=1,2,\cdots,n$
3. $\boldsymbol{s}_{d_i}$からs^{old}，s^{new}を判別
4. $T(s^{old},s^{new})$に1をたす．

そして新しい状態空間のそれぞれの状態の$Q^{new}(s_i^{new},a)$値は次式で計算する．ただし，aは行動．

$$Q^{new}(s_i^{new},a)=\frac{\sum_{j=1}^{n^{old}} Q^{old}(s_j^{old},a)\cdot T(s_j^{old},s_i^{new})}{\sum_{j=1}^{n^{old}} T(s_j^{old},s_i^{new})} \quad (4)$$

5 タスクと仮定

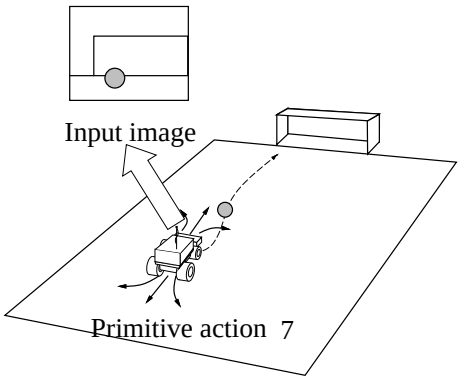

Fig.1 A task is to shoot a ball into the goal

ここではセンサー出力の連続性のみを仮定する．この条件下で，ロボットは観測されたセンサーデータから適切なモーターコマンドを生成しなくてはならない．そのために，ロボットが自分で行動をとりながら状態空間を自律的に分割していく必要がある．

本研究ではタスクの例としてサッカーロボットを扱うが，以下を設定する．サッカーロボットのタスクはボールをゴールにシュートすることである(Fig.1参照)．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．ロボットは画像からボールやゴールの基本的な特徴量が得られる．

また状態を表わすセンサー出力は５次元にした．ボールは画面上の位置と大きさを使用した．ゴールについては位置，大きさに加えその傾きを使用した．これらの特徴量は画像データを主成分分析をすることで得られる．

6 シミュレーション

シミュレーションにおける環境はFig.1に示すような，長方形のフィールドで上辺の中央にゴールがあり，カメラを乗せたロボットがボールを蹴る． また，このシミュレーションでは画像処理にかかる遅れ時間は考慮していない．

ロボットがとることのできる基本行動はFig.1に示すように左前進，前進，右前進，左後退，後退，右後退，停止のあわせて7通りである．

ロボットがシュートに成功したり，ロボットやボールがフィールドの外に出たり，ロボットがゴールポストに衝突するまでを一つの試行とし，試行が終わると，ロボット，ボールをリセットする．

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-0.1とした．学習率αの値は0.25，減衰係数γの値は0.9 とした．

Fig.2に状態空間の分割とシュート行動の獲得を同時に獲得させた時の成功率と分割された状態の数を示す．ただし成功率は最近の20試行の間に成功した回数である．

Fig.3は試行回数1,110で最終的に得られた状態空間のボールとゴールが同時に見える時の様子を示す．ただし，5次元の状態空間をボールの位置は画面中央，ゴールは傾きが無く真ん中に見えている状態での断面をとり，ボールの大きさとゴールの大きさの2 次元で表現したものである．また最終的に得られた状態数は28，獲得されたデータ数141であった．

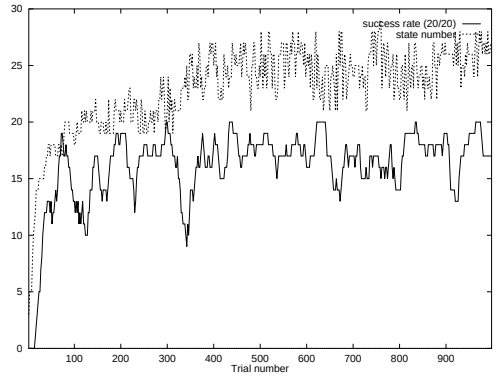

Fig.2 The success rate and the state number

また環境が動的に変化しても対応できることを示すために試行回数500でボールの大きさを倍に変化させた時の成功率と状態数を示す図をFig.4に示す．

7 実機での実験

実際に構築したシステムは文献[5]を参照のこと．

Fig.5は試行回数72で最終的に得られた状態空間のボールとゴールが同時に見える時の様子を示す．(Fig.3同様)

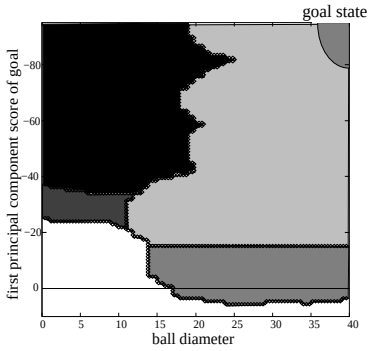

Fig.3 Result of state space construction

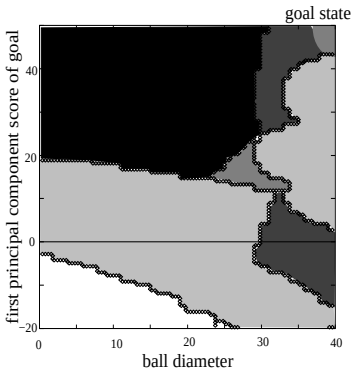

Fig.5 Result of state space construction of real robot experiment

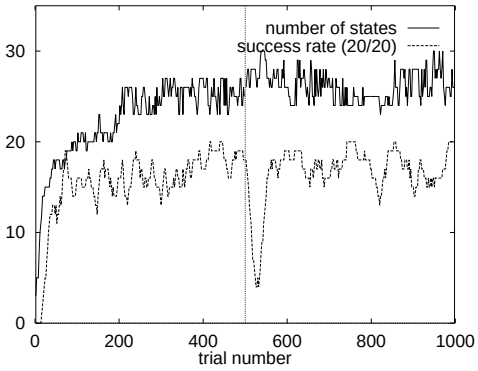

Fig.4 The success rate and the number of state in the case that environment change one the way

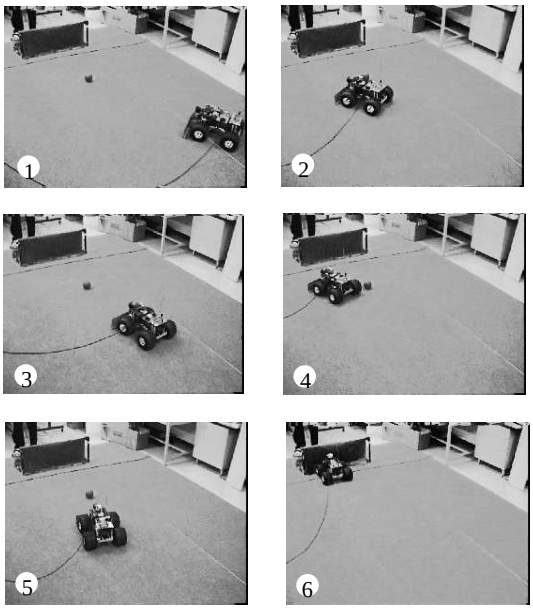

Fig.6 The robot succeeded in shooting a ball into the goal

また最終的に得られた状態数は18，獲得されたデータ数151であった．

Fig.6はロボットがシューティングに成功した例を，6つの時刻に分けて示している．何度かボールを蹴ることに失敗してもバックして切り替えして最後にはシューティングに成功している様子がわかる．

8 終わりに

本研究ではサッカーロボットがボールをゴールにシュートするタスクを例に，ロボット自身の経験に基づいてセンサー空間を分割し，数少ない経験でタスク遂行に必要な状態空間を自律的に構成し，かつ目的行動も同時に獲得するアルゴリズムを提案した．そしてサッカーロボットのシミュレーションおよび実ロボットによる実験を通じて，提案した手法の有効性を検証した．

今後の課題としては3節で述べたヒューリスティックの妥当性の検証，4.3節の状態遷移をするための基本行動のよりよい生成方法の開発などが考えられる．

参考文献

[1] Jonalthan H. Connell and Sridhar Mahadevan. *ROBOT LEARNING*. Kluwer Academic Publishers, 1993.

[2] Steven D. Whitehead and Dana H. Ballard. Active perception and reinforcement learning. In *MLWS*, pp. 179–188, 1990.

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

[4] Minoru Asada, Shoichi Noda, Sukoya Tawaratsumida, and Koh Hosoda. Vision-based reinforcement learning for purposive behavior acquisition. In *Proceedings of 1995 IEEE International Conference on Robotics and Automation*, Vol. 1, pp. 146–153, 1995.

[5] 浅田稔, 野田彰一, 俵積田健, 細田耕. 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, Jan 1995.