サッカーロボットの技能学習

◯内部 英治　　浅田 稔　　細田 耕
大阪大学大学院 工学研究科 電子制御機械工学専攻

A Purposive Behavior Acquisitoin for a Soccer Robot Using Reinforcement Learning

◯Eiji Uchibe　　Minoru ASADA　　Koh HOSODA
Dept. of Mech. Eng. for Computer-Controlled Machinery Osaka University

1　はじめに

人工知能とロボティクスの研究の究極の目標は，動的環境に適応するために，センサ情報から自身の内部構造を組織化する自律的な知能ロボットを実現することである．これまでの熟考型のアプローチの限界から，行動規範型のアプローチが注目されているが[3]，その問題点として，現在までに実現されたものが，反射型行動に限定されること，個々の行動モジュールのコーディングが人手に頼っていることなどが挙げられる．真の意味でのロボットの知能を実現させるためには，ロボット自身の感覚や行動による環境との密な相互作用を通して，種々の行動を学習し，それらを状況に応じて統合することが必要である．

近年,「ロボットと環境との相互作用を通して，与えられたタスクを達成する合目的な行動を獲得」する手法として，強化学習が注目されている．しかし，それらの多くはコンピュータシミュレーションのみの研究であり，実ロボットを用いた研究[5]は少ない．これらの研究では，センサとしてバンパーセンサ，ソナーなどの近接センサのみを使用しているので，行動が局所的かつ反射的で，簡単なタスクの実験しか行っていない．

また，複数のタスク達成のための行動を強化学習を用いて獲得した研究[10,14]もある．しかし，これらの研究は問題設定が理想的であり，サブタスク間で状態空間が干渉する時には適用できない．また，ロボットは鳥瞰図的なセンサを想定しており，自律ロボットへの適用に関して考慮されていない．

それに対し Asada et al. は視覚センサを搭載した実ロボットに対し，強化学習を適用し,「ボールをゴールにシュートする」行動を獲得した[2]．また，強化学習の本質的な問題の一つである「状態・行動空間の構成問題」に対して行動ベースの状態空間の構成法を提案し，人間が構成するよりも良い結果が獲得できた[8]．次に，タスクを拡張して,「キーパーロボットとの衝突を回避する」行動との協調問題を取り扱い，タスク間で状態空間が干渉する場合，干渉する状態を追加することで対処した[1,12]．しかし，提案された手法には二つの問題点がある．一つ目の問題として，ロボットは状態空間の直積全体を再学習する必要があることである．もう一つの問題は，隠れ状態[6]の検出を人間が事前に行なっていることである．

そこで本稿では，サッカーロボットが複数の行動を調整しながら，合目的な行動を獲得するために，二つの手法を提案する．一つ目は学習時間と獲得されるパフォーマンス間のトレードオフを考慮した複数行動強調のための「モジュール統合による強化学習法」である．二つ目は，隠れ状態推定のための「情報量規準に基づく推定」法である．1対1の簡単なサッカーゲームに本手法を適用し，その有効性を検証する．なお，状態空間の構成問題に関しては，他の文献[11]を参照されたい．

2　強化学習

2.1　強化学習の枠組

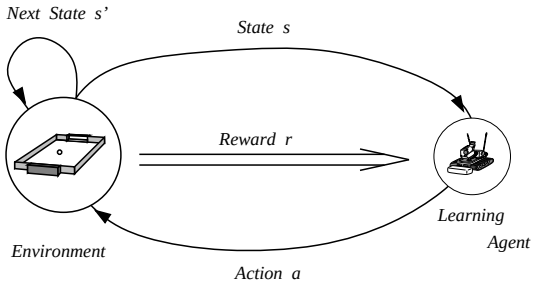


Fig.1 The basic model of robot-environment interaction

強化学習の基本的な枠組では，ロボットと環境はそれぞれ離散化した時間系列過程で同期した有限状態オートマトンとしてモデル化される．ロボットが観測できる環境の状態の集合を $\boldsymbol{S}$，実行できる行動の集合を $\boldsymbol{A}$ とする．ロボットは現在の環境の状態 $s(\in \boldsymbol{S})$ を感知し，それに基づき行動 $a(\in \boldsymbol{A})$ を実行する．結果として環境は次の状態 $s'(\in \boldsymbol{S})$ に遷移し，報酬とよばれる環境からのスカラーの評価値 r をロボットに与える．このような環境とのインタラクションを繰り返すことで，ロボットは与えられたタスクを遂行する目的行動を獲得する (**Fig.1**参照).

2.2　Q 学習

Q 学習はもっとも良く知られた強化学習の一手法であり，減衰した積算報酬

$$J^{\pi}(s_0) = \lim_{n\to\infty} E(\sum_{t=0}^{n-1} \gamma^t R^{\pi}_{s_t,a}), \tag{1}$$

を最大化する政策 π を獲得する学習則である[13].

Q 学習では状態と行動の組 (s,a) に対して行動価値関数 $Q(s,a)$ が定義され

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'\in \boldsymbol{A}} Q(s',a')), \tag{2}$$

と更新される．ここで，α は学習率 $(0<\alpha<1)$，γ は減衰係数 $(0<\gamma<1)$ である．

2.3　R 学習

R 学習は最近注目されている強化学習の手法であり，Q 学習と異なり，減衰しない平均化した積算報酬

$$J^{\pi}(s_0) = \lim_{n\to\infty} \frac{1}{n} E(\sum_{t=0}^{n-1} R^{\pi}_{s_t,a}), \tag{3}$$

を最大化する政策 π を獲得する学習則である [9].

R 学習も Q 学習と同様に，状態と行動の組 (s,a) に対して行動価値関数 $R(s,a)$ が定義され，

$$R(s,a) \leftarrow (1-\alpha)R(s,a) + \alpha(r - \rho + \max_{a' \in \boldsymbol{A}} R(s',a')), \tag{4}$$

と更新される．ここで ρ は平均報酬と呼ばれ，行動価値関数と同時に更新される．更新式は

$$\rho \leftarrow \rho + \beta(r + \max_{a' \in \boldsymbol{A}} R(s',a') - \max_{a' \in \boldsymbol{A}} R(s,a') - \rho), \tag{5}$$

で与えられ，β は ρ の学習率である．R 値の更新と異なり，ρ の更新は最適行動を選択した場合にのみ行なわれることに注意されたい．

2.4 強化学習を複数タスクに適用する場合の問題点

Q, R 学習などの環境同定型の強化学習には，学習時間が状態数の指数関数に比例するという問題点がある．そのため，強化学習をそのまま複数タスクに適用するのは，原理的には可能だが，事実上困難である．

そのため，Asada et al. は個々のタスクをそれぞれ独立に学習し，獲得された行動を初期行動として，最終的な行動を再学習によって獲得する手法を提案した [1,12]．しかし，この手法には二つの問題点がある．一つ目はロボットは最終的には全ての状態を学習しなければならないことである．二つ目は隠れ状態は設計者によって事前に追加されていることである．

これらの問題点を克服するために，二つの方法を提案する．前者の問題に対しては，学習時間とパフォーマンスの間のトレードオフを考慮した複数行動協調のための「モジュール統合による強化学習」を 3 節に示す．後者の問題に対しては，情報量規準に基づいた隠れ状態検出法を 4 節に示す．

3 モジュール統合による強化学習

複数行動協調のための「モジュール統合による強化学習」について述べる．この方法の特徴は，それぞれ独立に学習された結果を統合する時の先験的知識として利用し，学習結果をそのまま利用したり，必要であれば再学習をすることができる点にある．

3.1 モジュールの生成

Fig.2 にモジュール生成の概要を示す．はじめに，n 個のモジュールから，対応する新しい状態空間 ${}^c\boldsymbol{S}$ を構成する (Fig.2は $n=2$ の場合である)．最適性のためには全状態 ${}^c\boldsymbol{S}$ を学習しなければならないが，現実的でなく，大規模な問題に適用できない．また，学習時間の割には行動はそれほど改善されないという問題もある．

そこで，独立に学習することで既に獲得された行動価値を利用することを考える．すなわち，学習時間を削減するために，全状態をそれぞれ独立に獲得された行動価値の最大値に基づき，次の二つのカテゴリに分類する:

学習の必要がない領域： 独立に獲得された行動がそのまま利用できる状態の集合.

再学習が必要な領域： 複数の行動間でビヘービアの競合が発生している状態の集合.

まず，変数を次のように定義する.

- ${}^iT$: i 番目のモジュールに対応するサブタスク, $(i=1,\cdots,n_T)$
- ${}^i\boldsymbol{S}$: i 番目のサブタスクの状態空間,
- ${}^is_k$: i 番目の状態空間 $\boldsymbol{S}^i$ の中の k 番目の状態,
- ${}^iQ$: i 番目のサブタスクの行動価値関数,
- ${}^c\boldsymbol{S}$: n_T 個の状態空間の直積によって構成される状態空間,
- ${}^cs$: 直積の状態空間 ${}^c\boldsymbol{S}$ における状態.

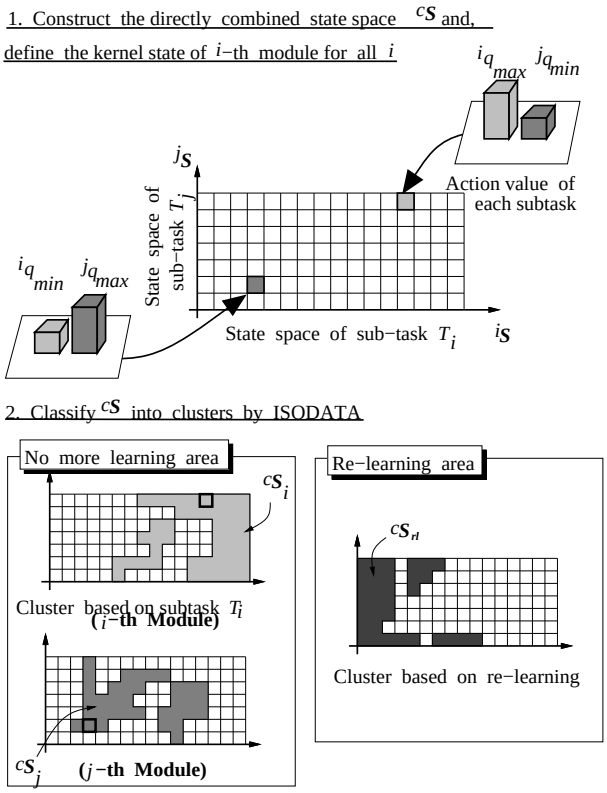


Fig.2 An overview of module construction procedure

全状態 ${}^c\boldsymbol{S}$ をクラスタリングするために，それぞれのモジュールの核となる状態 ${}^cs_{kernel}$ を定義する．いま，サブタスク T_i の状態 ${}^is_k \in {}^i\boldsymbol{S}$ に対して

$${}^iq^k_{max}({}^is_k) = \max_{b \in \boldsymbol{A}} {}^iQ({}^is_k, b),$$

を計算する．もし状態 ${}^cs$ が

$${}^cs = \arg\max_{{}^is_k \in {}^i\boldsymbol{S}} {}^iq_{max}({}^is_k), \text{ and} \tag{6}$$

$${}^cs = \arg\min_{{}^js_k \in {}^j\boldsymbol{S}} {}^jq_{max}({}^js_k) \text{ for all } j \neq i, \tag{7}$$

であれば，${}^cs$ を i 番目のモジュールの核となる状態 ${}^cs_{kernel}$ とする.

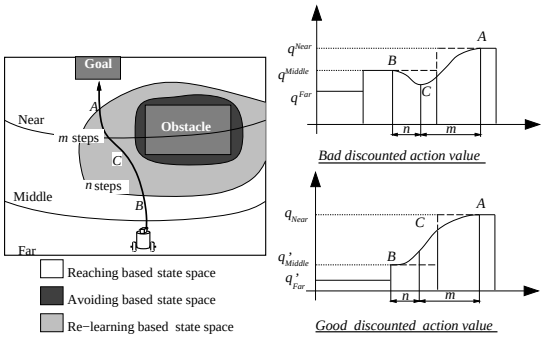
まず，全状態 ${}^cs \in {}^c\boldsymbol{S}$ を核となる状態 ${}^cs_{kernel}$ とのマハラノビス距離に応じて分類する．分類のための手法には，非階層型のクラスタリング手法である ISODATA アルゴリズムを用いた．結局，合成した状態空間 ${}^c\boldsymbol{S}$ は学習の必要がない領域 ${}^c\boldsymbol{S}_i, i=1\cdots n$ と再学習領域 ${}^c\boldsymbol{S}_{rl}$ に分類される．

3.2 学習スキーマ

もし現在の状態 s と遷移した状態 s' がともに再学習領域であれば，単純に学習則を適用できる．また，s がモジュールベースドな領域であれば，行動価値関数を更新する必要はない．問題となるのは，再学習領域からモジュールベースドな領域に遷移した場合である (**Fig.3** 参照)．2 節で示した二つの強化学習のどちらを使用するかによって，このモジュール間での「行動価値関数のずれ問題」への対処方法は異なる．

Q 学習の場合は，減衰係数 γ を学習前に決定しなければならないため，実際の物理ステップを用いて

$$Q_{rl}(s,a) = (1-\alpha)Q_{rl}(s,a) + \alpha(r + \gamma V_i(s')), \tag{8}$$



(a) Simple navigation task

(b) Action value for the re-learning area

Fig.3 Unbalance problem between action value function of modules

と計算する．ここで $V_i(s')$ は i 番目のモジュールの学習の必要がない領域における行動価値関数であり，

$$V_i(s') = \gamma^{steps(s')} \max_{b \in \boldsymbol{A}}, Q_i(s', b) \quad s' \in \boldsymbol{S}_i, \qquad (9)$$

である．ここで，$steps(s)$ は状態 s から目標状態までの物理ステップ数の見積りである．Q 学習では，このような行動価値関数を適切に調整する枠組が必要である．一方 R 学習の場合は，平均報酬 ρ を行動価値関数 $R(s,a)$ の学習と同時に行なえるため，特別な学習アルゴリズムを用いる必要はない．

4 情報量規準に基づく隠れ状態の検出

3 節の方法では，隠れ状態について考慮されていない．最適な行動を獲得するためには，自律的に隠れ状態を検出する枠組が必要となる．本節では，その方法について述べる．

4.1 自己回帰モデルへの当てはめ

隠れ状態は，異なる状態をロボットが識別できないことが主な原因として挙げられる．よって隠れ状態での行動価値は，複数の状態での行動価値の平均となる．いま，与えられた n_f 個の時系列の行動価値を p 次の自己回帰モデル (AR(p)) に当てはめる．ここで問題となるのが，自己回帰モデルの次数 p である．ここでは，時系列解析の分野で用いられている赤池の情報量規準 (AIC) を用いて p を決定する [7]．AIC は

$$AIC = -2 \times MLL + 2 \times p, \qquad (10)$$

と計算できる．ここで MLL は最大対数尤度である．AIC を計算すると同時に t 値と呼ばれる安定性の指標

$$t(c) = \frac{\sqrt{n_f(n_f-2)}c}{\sqrt{(1+M^2/V)r_p}}, \qquad (11)$$

も計算する．ここで M，V，c はそれぞれ時系列データの平均値，分散，回帰モデルの係数の推定値を示す．自己回帰モデルの次数 p は AIC を最小化するように決定される．

4.2 隠れ状態の判定条件と対処

状態 s が隠れ状態のとき，行動価値の推移は不安定になり，

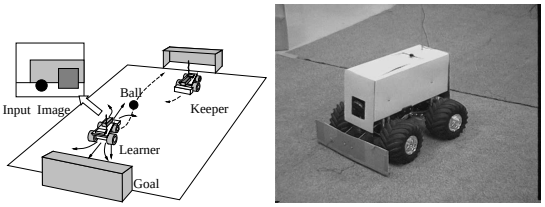
- 自己回帰モデルの次数 p が 2 以上，かつ
- t 値の絶対値が 2 以上，

である時，その状態を隠れ状態であると判定する．

もし，状態 s が隠れ状態であると判定されれば，過去の履歴 $\{s_{prev}, a_{prev}, s\}$ を一つの状態とみなして，合成した状態空間に追加する．ここで s_{prev} は一つ前の状態，a_{prev} は s_{prev} でとった行動である．

5 タスクと想定

タスクとして，学習ロボットがキーパーロボットとの衝突をできるだけ回避しながら，ボールをゴールにシュートするタスクについて考える．ロボットに搭載されているのはカメラだけであり，自身の幾何学的パラメータや動的特性などに関する知識はもっていない．



(a) Environment (b) Soccer robot

Fig.4 The task of shooting a ball into the goal with a keeper robot

また，状態・行動空間の構成は一般に困難で，強化学習を適用する際に本質的な問題である [4] が，ここでは環境内にはボール，キーパーロボット，ゴール，ラインだけが存在すると仮定し，それぞれ，画像上で位置や大きさなどについて適当に量子化し，その組合せで構成する．詳細は他の文献 [1,12] を参照されたい．

6 シミュレーションおよび実ロボットにおける実験

まず，提案した手法をコンピュータシミュレーションによって検証した結果を示す．**Table 1** は従来法と本手法を比較した結果である．ここで，

- 従来法 1:行動価値関数の単純和 [1,14]．
- 従来法 2:再学習法 [1]．

であり，用いた強化学習は Q 学習である． 表からわかる通り，従来法 1 よりも本手法の方が良い結果が得られている．また，モジュール統合による強化学習では従来法 1 と 2 のトレードオフを取っていることになり，実際パフォーマンスをそれほど低下させることなく，学習時間を約 1/3 にすることができた．また，再学習する状態数も削減できた．今回の実験では，Q 学習と R 学習の間に有意な差は見られなかった．今回の実験では，学習中の行動戦略と学習法の関係が明らかになっていないため，今後はこの点を含めて，更に考察する必要がある．

次に，検出された隠れ状態の一例を **Fig.5** に示す．これは状態が粗すぎるために発生した状態であり，具体的にはゴール近傍でボールが画面の端にある場合である．矢印のところで隠れ状態と判定され，履歴を用いて判別された．これによって，不安定な行動価値関数が安定したことがわかる．

Table 1 Average performance over 10 runs measured

	previous work 1	previous work 2
success of shooting (%)	36.1	60.8
mean steps to shooting	172.3	128.3
mean steps to collision	231.2	5048.5
learning time	***	480.4
# of relearning-states	***	11132
	modular (Q)	modular (R)
success of shooting (%)	57.3	58.4
mean steps to shooting	138.6	139.4
mean steps to collision	3624.8	3982.0
learning time	128.9	143.5
# of relearning-states	395	403

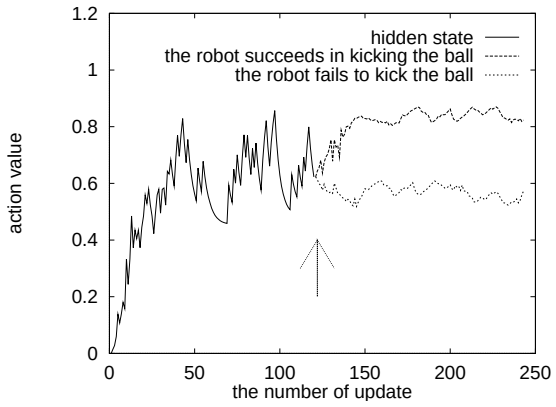


Fig.5 Transition of action value of hidden states detected by AR model enhanced by *AIC*

最後に実ロボットに適用した結果を **Fig.6** に示す．現時点ではキーパーロボットは静止しているが，技術的な問題は既に解決されており，現在2台同時に行動する場合の実験を計画中である．

7 おわりに

複数行動の協調のために，学習時間とパフォーマンスのトレードオフを考慮した「モジュール統合による強化学習」を提案した．また，行動価値関数の履歴を自己回帰モデルに当てはめ，モデルの次数を情報量規準を用いて決定し，そのときの次数から隠れ状態を判定する手法を提案した．今後は，サッカーゲームを実現するための，ロボット間の協調・競合問題への対処が考えられる．

参考文献

[1] M. Asada, E. Uchibe, S. Noda, S. Tawaratsumida, and K. Hosoda. Coordination Of Multiple Behaviors Acquired By A Vision-Based Reinforcement Learning. In *Proc. of the 1994 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 2, pp. 917–924, 1994.

[2] 浅田, 野田, 倹積田, 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, 13(1):68–74, 1995.

[3] R. A. Brooks. A robust layered control system for a mobile robot. *IEEE J. Robotics and Automation*, RA-2:14–23, 1986.

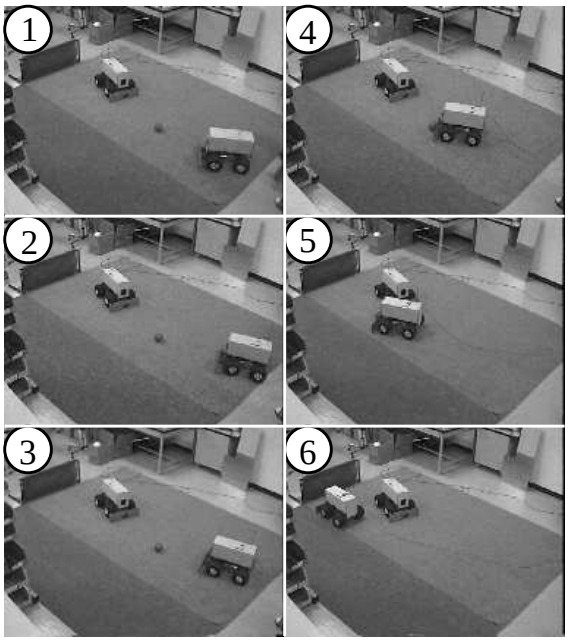


Fig.6 The robot succeeded in shooting a ball into a goal avoiding a stationary keeper robot.

[4] D. Chapman and L. P. Kaelbling. Input Generalization in Delayed Reinforcement Learning : An Algorithm and Performance Comparisons. In *11th International Joint Conference on Artificial Intelligence*, pp. 726–731, Sydney, Australia, 1991.

[5] J. H. Connel and S. Mahadevan. Rapid Task Learning for Real Robot. In *Robot Learning*, chapter 5, pp. 105–140. Kluwer Academic Publishers, 1993.

[6] R. A. McCallum. Instance-Based Utile Distinctions for Reinforcement Learning with Hidden State. In *Proc. of the 12th International Conference on Machine Learning*, pp. 387–395, 1995.

[7] 中溝. 信号解析とシステム同定. コロナ社, 1993.

[8] 野田, 浅田, 細田. 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. 第5回ロボットシンポジウム予稿集, pp. 145–150, 1995.

[9] A. Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In *Proc. of the 10th International Conference on Machine Learning*, pp. 298–305, 1993.

[10] S. P. Singh. Transfer of Learning by Composing Solution of Elemental Sequential Tasks. In *Machine Learning*, Vol. 8, pp. 99–115, 1992.

[11] 高橋, 浅田, 細田. 状態空間の自律的分割による実ロボットの実時間学習. ロボティクス・メカトロニクス講演会論文(掲載予定). 日本機械学会, 1996.

[12] 内部, 浅田, 野田, 細田. 視覚を有する移動ロボットの強化学習による複数タスクの 達成. ロボティクス・メカトロニクス講演会論文, pp. 700–703. 日本機械学会, 1995.

[13] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, University of Cambridge, May 1989.

[14] S. D. Whitehead, J. Karlsson, and J. Tenenberg. Learning Multiple Goal Behavior Via Task Decomposition And Dynamic Policy Merging. In J. H. Connel and S. Mahadevan eds., *Robot Learning*, chapter 3. Kluwer Academic Publishers, 1993.