GA により探索空間の動的生成を行う Q 学習

A Study of Q-learning: Dynamic Structuring of Exploration Space Based on Genetic Algorithm

伊藤 一之

Kazuyuki Ito

東京工業大学大学院総合理工学研究科知能システム科学専攻

Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

kazuyuki@cs.dis.titech.ac.jp, <http://cs.dis.titech.ac.jp/~kazuyuki/>

松野 文俊

Fumitoshi Matsuno

(同 上)

matsuno@dis.titech.ac.jp, <http://cs.dis.titech.ac.jp/>

keywords: Q-learning, dynamic structuring of exploration space, reinforcement learning, genetic algorithm

Summary

Reinforcement learning has recently received much attention as a learning method for complicated systems, e.g., robot systems. It does not need prior knowledge and has higher capability of reactive and adaptive behaviors. However increase in dimensionality of the action-state space makes it difficult to accomplish learning. The applicability of the existing reinforcement learning algorithms are effective for simple tasks with relatively small action-state space. In this paper, we propose a new reinforcement learning algorithm: "Q-learning with Dynamic Structuring of Exploration Space Based on Genetic Algorithm". The algorithm is applicable to systems with high dimensional action and interior state spaces, for example a robot with many redundant degrees of freedom. To demonstrate the effectiveness of the proposed algorithm simulations of obstacle avoidance by a 50 links manipulator have been carried out. It is shown that effective behavior can be learned by using the proposed algorithm.

1. は じ め に

近年、ロボットの行動獲得法として強化学習 [Sutton 88] が注目されている [宮崎 99a, 宮崎 99b, 片山 99, 木村 00, Svinin 99, Ushio 01]. 強化学習は学習に際し予備知識を必要とせず、高い環境適応性をもつ学習法である。しかし、学習速度や行為・状態空間の大きさなどの問題により、複雑な実ロボットへの適用は困難であり、多自由度を有するロボットへの適用例は少ない。

強化学習の代表例である Q 学習 [Watkins 92] は、最適性に関する理論的枠組みを持ち、小規模な問題に対しては非常に有効であるが、多自由度ロボットなどの大規模な問題においては、行為・状態空間の増加が学習を困難にしまい、事実上適用不可能である。

分類子システム [Holland 86] は、GA によりルールを生成するため、探索空間が広いタスクに対して学習を試みることは可能であるが、条件部が示す状態の集合と、行動部によって遷移可能な状態の集合との整合をとることが難しく、多自由度ロボットなど、探索空間の大きなシステムに対する有効な政策の獲得は困難であると考えられる。また、"don't care" は対象を部分観測としてしまうという問題を持つこと、さらに、汎用的なルールと特殊なルールの共存が困難であることなどが指摘されて

いる [Wilson 89, Westerdale 97, 林 99]. このような問題に対する解決案として、GLS [林 99] や IBCG [中村 99] などが提案されている。GLS では、don't care symbol によって異なる状態が同一の状態として扱われるという問題を部分観測マルコフモデルとして捉え、この問題をメモリーなし確率的政策 [Jakkola 94] を用いることで解決している。GLS は、部分観測となる小規模な環境においては非常に有効であると考えられるが、行為・状態空間の大きなシステムに対しては、同一の状態とみなされる異なる状態が非常に多くなること、また、GLS では、ピッツバークアプローチを用いていることなどから、必要とする計算時間が膨大となり、事実上学習が困難であると思われる。また、GLS では、対象に依存したルール数を設計者が与える必要があり、大規模なシステムでは、適切なルール数を選択することが困難である。中村らは、自律移動ロボットにおける状態空間の考察から、分類子システムにおけるルール生成機構を見直し、IBCG [中村 99] を提案している。IBCG では、実観測領域を対象に学習を行い、経験した状態・動作のペアをルールとして記憶していくため、分類子システムに比べ効率がよく学習が速い。しかし、IBCG では各行為についてルールが作られ、各行為は、その状態部が全ての状態とマッチする無限低ルールを必ず 1 つ保持する。したがって、ルー

ル総数は少なくとも行為総数以上となり、行為数の大きなシステムへの適用は困難である。

一方、探索空間を動的に生成する手法として、状態空間のみを適応的に生成または分割する手法が提案されている [高橋 99, Yamada 01]。しかし、これらの研究では、行為の動的生成は行われず、行為集合が固定されているため、行為空間の大きなシステムへの適用は困難である。

これらの問題をふまえ、本研究では、強化学習の新しいアルゴリズムとして、Q-learning with Dynamic Structuring of Exploration Space Based on Genetic Algorithm を提案する。本手法では、環境はマルコフ決定過程であるとする。本手法は、多自由度ロボットのように行為・内部状態により構成される空間が広いシステムに対して設計されたものである。本論文で提案する学習器は階層構造を持ち、各内部状態をローカルに制御する低層と低層における内部状態の目標値の軌跡のプランニングを行う上層とからなる。上層における学習機構は、強化学習を行う行動獲得ダイナミクス (Q-learning) と、強化学習における行為空間を生成する行為空間生成ダイナミクス (GA) の 2 つの構造からなる。GA は広範囲な行為空間の中から有用な部分集合を選択するために用いられる。GA によって構成された行為空間をもとに探索空間が動的に生成され、生成された探索空間に対し Q-table を構成し Q 学習を行う。Q 学習により強化された Q-table をもとに各行為に対し適合度が計算され、これをもとに再び GA により行為の集合が再編される。以上を繰り返すことにより有効な行為の集合とその行為による有効な政策が獲得される。

最後に、50 リンクからなる超冗長マニピュレータの障害物回避問題を取り上げ、シミュレーションにより学習を行い提案手法の有効性を検証する。

2. 提案手法

2.1 階層構造と探索空間の動的構成

§1 内部状態と外部状態

本論文では、内部状態と外部状態を次のように定義する。内部状態とは、ローカルに直接制御可能な状態の集合であるとし、全状態集合から、内部状態の集合を除いたものを外部状態とする。例えば、マニピュレータの障害物回避を考えると、障害物の位置などが外部状態であり、マニピュレータの各関節角度などが内部状態である。

§2 階層構造

提案手法では学習器は 2 階層からなる階層構造を持つ。低層のエージェントは、制御対象の各内部状態をローカルに直接制御する。上層のエージェントは、全内部状態の目標値のプランニングを行い、低層のエージェントへ指令する。実装例として、マニピュレータの制御に適用した場合を図 1 に示す。低層のエージェントは関節数と同

図 1 マニピュレータへの適用

数存在し、各関節の角度を上層のエージェントから与えられた目標値に追従するようローカルな制御を行う。これにより、上層のエージェントは各関節にその目標値を与えるだけでマニピュレータの姿勢を制御でき、タスクを達成するための軌道計画のみを行えばよいことになる。

階層構造が導入可能な条件は、状態を外部状態と内部状態に分離可能であることであり、これは、システムの中にローカルに制御可能な状態が含まれるという条件に相当する。一例として、多自由度ロボットの場合について考えると、各関節をローカルに制御可能であれば、これが内部状態となり、ロボット全体の重心の位置、姿勢や障害物の位置などの環境に依存する状態が外部状態となり、階層構造が導入可能である。一般に、十分な能力を備えたアクチュエータを用いることで、各関節をローカルに制御可能である。また、一部の関節に多大な力がかかる場合や、一部のアクチュエータの能力が十分でない場合には、十分な能力を備えたアクチュエータをもつ関節を内部状態と設定し、アクチュエータの能力が十分でない関節を外部状態と設定することで、階層構造を導入可能である。従って、多自由度ロボットのみについて考えても、提案手法の適用可能な対象のクラスは十分に広いと考えられる。

ロボットの行動獲得に階層構造を用いた従来研究として森本らの提案する手法をあげることができる [森本 00]。森本らの手法と我々の手法を比較すると、下位層が関節角度を制御し、上位層が全体のプランニングを行うという点は、類似している。しかし、その目的、適用可能クラスにおいて、両者は異なると考えられる。森本らの手法は、学習の高速化を目的としたものであるのに対し、我々の手法は、多自由度ロボットをはじめとする内部状態の大きなシステムへの適用を図ったものである。森本らの手法では、下位層における学習器は一つであり、一つの学習器で、全体の関節を制御する学習を行う。従って、各関節間の相互作用を調整することが可能な反面、自由度が増加した場合には適用不可能である。一方、我々の提案する手法では、各関節を独立したエージェントが制御することを考えており、十分な能力を持ったアクチュエータを搭載し、相互作用を無視できるようなシステムであれば、多自由度なシステムであっても提案手法を適用することが可能である。

図 2 フローチャート

本論文では、低層のエージェントについては、制御工学をはじめ、ニューラルネットワーク、強化学習など既存の方法で十分実装可能なため、主に上層のエージェントにおける学習法について考え、低層のエージェントについては十分な性能を持ち、各状態は各ステップ間に目標値に追従するものとする。

§ 3 探索空間の動的構成

学習器の階層構造から、低層のエージェントの働きにより、システム全体の内部状態は、上層のエージェントの指令した目標値と等しくなる。ここで、上層のエージェントが目標値を出力することを行為と考え、システムが遷移可能な内部状態の集合は、その行為が示す内部状態の集合と等しくなる。従って、行為空間を限定することで内部状態空間も限定され、結果として探索空間が限定される。ここで我々は、行為に着目し、行為を動的に生成することを考える。この場合、状態空間を構成する際に、全ての内部状態を必要とせず、動的に生成された行為に対応する内部状態のみで構成すればよいこととなる。従って、行為の動的生成とともに、状態空間も動的に生成される。行為の動的生成法および学習アルゴリズムについては次節で述べる。本論文では以降、全ての行為・状態によって構成される空間を全探索空間と呼び、動的に構成された行為をもとに動的に構成された行為・状態空間を単に探索空間と呼ぶこととする。

2.2 学習アルゴリズムの概要

本節では、主に上層での学習アルゴリズムについて説明する。図 2 にフローチャートを示す。本手法は、GA による “行為生成ダイナミクス” と Q 学習による “行動獲得ダイナミクス” からなる。各行為は、遺伝子としてコーディングされ GA により生成される。まず初期個体がランダムに生成され、生成された個体の表現形である行為とその行為によって遷移可能な状態を要素とする Q-table が生成される。この Q-table を用いて Q 学習が行われ、Q-table が強化される。Q 学習が十分行われた段階で Q 学習を終了し、強化された Q 値をもとに各行為（個体）の適合度が計算される。計算された適合度をもとに選択

が行われ、選ばれた親に対し、交叉・突然変異が行われ、新たな個体集合が生成される。以上を繰り返すことにより、有効な行為の集合とその行為を用いた政策が獲得される。

ここで、用語を次のように定義しておく。図 2 における start から end までを、1 学習とよび、学習 1 回につき、1 つの目標を達成する行為が獲得される。GA における繁殖回数を世代とよび、初期集団を第 1 世代とする。Q 学習において、初期状態からゴールまたはペナルティを得て再び初期状態に戻るまでを 1 試行と呼ぶ。1 試行において、Q 値が更新された回数をステップと呼ぶ。また、初期状態からゴールに至るまでに遷移した状態とその状態で使われた行為の一連を系列と呼ぶ。以下に提案手法の詳細を述べる。

2.3 遺伝子のコーディングと個体の生成

本手法で扱う GA では、各個体は行動獲得ダイナミクス（Q 学習）における実行可能な行為を表し、染色体は各行為と対応するようにコーディングされる。

本手法では、行為の全集合のなかから、GA により部分集合を構成し、その部分集合のなかで Q 学習を行う。Q 学習によって部分集合の各行為が評価され、GA によって部分集合が再構築される。GA における個体数は、Q 学習における行為の最大数に一致する。

2.4 個体から行為への変換と Q-table の構成

各個体の表現形が行為である。しかし、各個体すべての表現形をそのまま行為としてしまうと、同一の表現形をもつ個体が複数存在した場合、Q-table が冗長となってしまう効率が悪い。本手法では、冗長性を排除し、同一の表現形をもつ個体は 1 つとみなして行為空間を構成する。図 3 に一例を示す。ここで、個体集合の中に action1 を表現形とする個体が 3 個体含まれるが、Q-table の構成においてこれらは同一とみなされ、Q-table には、 a_1 が一つ含まれるのみである。GA における繁殖を繰り返すことにより同一の表現形を持った個体が増加し、Q-table の構成においてこれらは同一とみなされるので、その結果、行為数が減少し、世代を重ねるごとに学習効率が改善される。従って、本手法では、設計者は、タスクの複雑さに対し十分に大きな数の行為数を個体数というかたちで与えるだけでよく、実際に必要な行為数は学習によって自動的に得られる。

本手法では、状態空間は次のように構成される。外部状態については、環境がマルコフ決定過程として観測されるよう適切に構成されるものとする。内部状態は、低層のエージェントの働きにより、行為が示す目標状態に遷移する。従って、内部状態は、GA により作られた行為に対応する状態のみで構成すればよく、行為の動的構成に伴って動的に構成されるものとする。従って、探索空間のうち内部状態と行為からなる空間は、GA により動

Population Phenotype		Q-table							
111011	action 1	⇒	a ₁	S ₁	S ₂	S ₃	S ₄	S ₅	...
101110	action 2	⇒	a ₂						
101010	action 3	⇒	a ₃						
101111	action 4	⇒	a ₄						
111011	action 1	⇒	}						
111011	action 1	⇒							

図 3 Q-table の構成

的に生成されることとなり、世代を重ねるにつれて、タスク遂行に必要な大きさまで減少していく。

2.5 Q 学習

行為獲得ダイナミクスでは、現在提案されている Q 学習に関する研究成果の多くが適用可能である。本論文では、簡単のため、従来の典型的な Q 学習を用いる。以下に更新式を示す。

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha\{r(s, a) + \gamma \max_{a'} Q(s', a')\} \quad (1)$$

s : 状態, a : 行為, s' : a によって s から遷移した状態, $r(s, a)$: 状態 s で行為 a を実行したときに得られる報酬, α : 学習率, γ : 割引率

なお, Q 値の更新は、低層のエージェントの行動が全て終了した段階で行うものとする。

2.6 適合度の算出

強化された Q 値をもとに、各個体の適合度を算出する。適合度の算出は、Q 値の規格化、Q 値による適合度の算出、使用頻度による適合度の算出の 3 手順からなる。以下に詳細を記す。

§ 1 Q 値の規格化

規格化は以下の手順で行う。まず、各状態の価値の最大値 V_{max} 、最小値 V_{min} を次のように定義する。

$$V_{max}(s) = \max_{a'} Q(s, a')$$

$$V_{min}(s) = \min_{a'} Q(s, a')$$

次に以下の式により、規格化された Q 値である Q' を求める。

$$\text{if } Q(s, a) \geq 0 \quad \text{then} \quad Q'(s, a) = \frac{1 - p}{V_{max}(s)} Q(s, a) + p \quad (2)$$

$$\text{else } (Q(s, a) < 0) \quad Q'(s, a) = -\frac{p}{V_{min}(s)} Q(s, a) + p \quad (3)$$

ここで、 p は報酬とペナルティーの比重を決める定数である。

この規格化により、次の 2 つのことが行われている。第一に、状態空間についての規格化である。Q 学習では、報酬に近い状態ほど状態の価値が高く評価される。Q 値をそのまま行為の価値とした場合、報酬に近い状態で実行された行為が高く評価されてしまい、その状態に至るまでの行為は淘汰されてしまう可能性があり、タスクの獲得が困難となる。このため、タスク達成に必要な行為の集合は、その行為が実行された状態の価値にかかわらず評価される必要がある。このため、状態についての規格化が必要となる。

第二に、ペナルティーについての扱いである。Q 学習では、負の値を持つペナルティーを設定することが可能であり、Q 値が負の値をもつ場合がある。一方、GA では適合度は正の値である必要がある。本手法では、報酬とペナルティーの比重を決める定数 p ($0 < p < 1$) を導入し、正の Q 値は規格化された Q 値の 1 から p の間に (式 (2))、負の Q 値は p から 0 の間に変換している (式 (3))。これにより、報酬だけでなく、ペナルティーを回避する行為についても評価される。また、 p の値を変えることにより、報酬を得るための行為の価値とペナルティーを回避するための行為の価値の比重を変えることが可能である。

§ 2 Q 値による適合度の算出

規格化された Q 値から適合度を算出する。まず、規格化された Q 値 ($Q'(s, a)$) を行為を固定し状態に対してソートし、これを Q'_s とする。例えば、行為 a_i について表すと、 $Q'_s(1, a_i)$ は、 a_i における $Q'(s, a_i)$ の最大値であり、 $Q'_s(N_s, a_i)$ は最小値である。ここで、 N_s とは各世代の Q 学習における状態の総数である。行為 a_i を表現形とする遺伝子の適合度は以下の式により与えるものとする。ここで、 a_i は Q-table を構成する行為集合の任意の要素である。

$$\begin{aligned} fitQ(a_i) &= w_1 \frac{Q'_s(1, a_i)}{1} + w_2 \frac{Q'_s(1, a_i) + Q'_s(2, a_i)}{2} + \\ &\dots + w_{N_s} \frac{Q'_s(1, a_i) + Q'_s(2, a_i) + \dots + Q'_s(N_s, a_i)}{N_s} \\ &= \sum_{j=1}^{N_s} \left(w_j \frac{\sum_{k=1}^j Q'_s(k, a_i)}{j} \right) \end{aligned} \quad (4)$$

w_j ($j = 1, \dots, N_s$) は重み係数であり、 w_j の値により、一般的な行為と特殊な行為の価値の重みを決めることができる。例えば、第一項は規格化された Q 値の最大値を示しており、これは、行為 a_i が最も高く評価される状態での a_i の価値を示している。したがって、 w_1 を大きくすることにより、特殊な状態で有効な特殊な行為が高く評価される。一方、第 N_s 項は平均値を示しており、これは、様々な状態においての a_i の価値の平均を意味し、 w_{N_s} を大きくすることにより、様々な状態で有効な一般的な行

為が高く評価される。

§3 使用頻度による適合度の算出

使用頻度による適合度を導入することにより、有効な系列の維持を考える。§2 で述べた Q-table による適合度の算出では、各状態における行為の価値が評価され、その和が適合度となるため、系列における行為の重要性は評価されない。従って、有効な系列を構成する行為であっても、他の状態で実行するとペナルティーを受ける場合には、低く評価され淘汰されてしまう可能性がある。また、これは系列の破壊を意味する。ここで、使用頻度の適応度を考えてみると、系列の核となる行為、つまり、複数の系列で使われる行為は他の行為に比べ使用頻度が高く、適応度が高く評価され、系列の破壊を防ぐことができる。また、試行回数を十分に確保した場合には Q 学習が収束するにつれて有効な系列が繰り返し使われ、結果として有効な系列を構成する行為の集合が次世代に残りやすくなる。以上の理由から使用頻度による価値は重要であると考えられる。

本論文では、使用頻度による適合度を以下のように定義する。

$$fit_u(a_i) = \frac{N_u(a_i)}{\sum_{j=1}^{N_a} N_u(a_j)} \quad (5)$$

ここで、 N_a はある世代での行為の総数であり、 $N_u(a_i)$ は、その世代での行動獲得ダイナミクス (Q 学習) において行為 a_i が使われた試行回数である。(5) 式は、行動獲得ダイナミクスにおいて実行された全試行回数に対してその行為が少なくとも 1 回は使用された試行の回数の割合、つまり、使用頻度を表している。

ここで、(5) 式は行動獲得ダイナミクスが終了した後一括して計算され、行為生成ダイナミクスにおいてのみ使われるため、行動獲得ダイナミクスに影響を与えないという点が重要である。Q 学習の性質が保たれ、前回の行動獲得ダイナミクスにおける使用頻度とは無関係に、現在の行為集合のなかで最適な行動に向かって Q 学習が行われ、新たな系列が作られる。

§4 適合度の算出

§2, 3 で導入した適合度の和をとって各遺伝子の適合度を以下の式で与える。

$$fit(a_i) = fit_Q(a_i) + k_f \cdot fit_u(a_i) \quad (6)$$

ここで、 $k_f (k_f \geq 0)$ は、 fit_Q と fit_u の影響力の割合を決めるための定数である。

2.7 選択・繁殖

計算された適合度をもとに、選択・繁殖を行う。これらは、現在提案されている様々な手法が適応可能であり、タスクや、コーディングされた遺伝子をもとに、適切に選択する必要がある。本論文では、従来のシンプル GA

を基本とする。詳しい実装法についてはシミュレーションによる学習の章で述べる。

3. コスト関数の導入

本章では、コスト関数を導入することにより、学習における遷移と、それに要する時間、エネルギーなどの物理量を対応させ、その物理量を評価の基準とするための報酬の設定について述べる。

2.1 節において、行為を遷移先の目標状態として定義することにより、状態空間を動的に構成することが可能であることを述べた。この場合、各行為において遷移に要する時間、エネルギー等が異なる場合が多い。提案手法のなかで用いている Q 学習では、割引報酬の期待値が最大となるように学習が進む。ゴール状態でのみ報酬を与えた場合、できる限り少ない遷移回数でゴールに到達する行動が獲得される。しかし、各行為によって遷移に要する時間、エネルギー等が異なる場合には、遷移回数は、それに要する時間やエネルギーなどの物理量を直接意味するものとならない。このため、得られた行動が必ずしも設計者の意図するものとならない可能性がある。以上を踏まえ、コスト関数 ($cost(s, a)$) を導入し、遷移と実際の物理量との対応を図る。コスト関数は、状態 s において行為 a を実行した場合に消費する物理量と定義し、状態と行為の関数として表すものとする。例えば、できる限り短い時間でタスクを達成させたい場合は、コスト関数は、状態 s において行為 a を実行するのに要する時間となる。コスト関数は、 s と a にのみ依存する値をとることから、Q 学習における報酬と同様に扱うことができる。ここで、これまでのゴールにおける報酬や、好ましくない行動に与えられるペナルティーを $r'(s, a)$ とし、実際の報酬 $r(s, a)$ を次のように再定義する。

$$r(s, a) = r'(s, a) - k_c \cdot cost(s, a) \quad (7)$$

これにより、より少ないコストでタスクを達成する行動が獲得される。ここで、 k_c は $cost(s, a)$ の影響を示す正の定数である。

4. シミュレーションによる学習

本手法を多自由度マニピュレータの制御問題に適用し、シミュレーションによりその有用性を検証する。

4.1 マニピュレータによる障害物回避

§1 タスク

平面内を運動する多自由度 (10 ~ 50 リンク) のマニピュレータによる障害物回避問題を扱う。図 4 に概要を示す。マニピュレータの初期姿勢を直線とし、固定端を原点、初期姿勢の直線を x 軸とする慣性座標系を定義する。目標タスクは、移動する障害物を回避しつつ、マニピュレー

図 4 タスク

タの先端を中心座標 $(0.5, 0.5)$ 半径 $0.1[m]$ の目標円内に到達させることである。障害物は半径 $0.1[m]$ の円であり、 $(0.5, 0.5)$ を中心とする半径 $0.3[m]$ の円上を反時計周りに等速円運動する。障害物の初期位置は、各試行ごとに円軌道ランダムに配置される。マニピュレータの全長は $1.5[m]$ とし、リンク長は全長が $1.5[m]$ となるようにリンク数によって変化させる。リンク数が 30 のときリンク長は $0.05[m]$ である。これは、リンク数が増加した場合でも、目標や障害物とマニピュレータの相似関係が保たれるよう配慮したためである。

マニピュレータの関節にはステップモータが取り付けられているとする。低層の各エージェントは、対応する関節を一定の角速度で目標角度まで運動させるものとする。従って、各関節は、目標角度に至ったものから順次停止し、全ての関節が目標角度に至った段階でマニピュレータは停止する。

§2 GA の構成

提案手法における GA 部を次のように構成する。最初にコーディングについて述べる。

1 つの染色体は関節数と同数の遺伝子をもつ。遺伝子は、関節の角度を表し、遺伝子と関節は、1 対 1 に対応している。遺伝子は 0 から 8 までの 9 種類をとるものとする。各関節の可動範囲を $\pm 40[\text{deg}]$ とし、遺伝子の 0 は $-40[\text{deg}]$ を表し、以降、 $10[\text{deg}]$ ずつ増加し、8 は $+40[\text{deg}]$ を表すものとする。マニピュレータ自身が交差している形状を表現形とする遺伝子は、致死遺伝子とする。

初期集団はランダムに生成され、個体数を 300 とする。選択はルーレット選択を用い、初期集団と同数の個体を選択すると設定した。

交叉は 1 点交叉を用いる。交叉確率 P_c の確率で、個体集合から親を選び、その中からランダムにペアをつくり交叉を行う。交叉により作られた子個体の全ビットに対し、突然変異確率 P_m で変異させる。交叉・突然変異によって作られた子個体は親個体との交換は行わず、元の個体集合に加えられる。従って個体数は交叉により増加し、こ

の集合に対して Q 学習による評価が行われる。従って、交叉・突然変異により行為が消滅することがなく、集合全体としてのパフォーマンスは、交叉・突然変異により減少することはない。なお、増加した個体は、選択により一定数まで減少するため、繁殖を繰り返すことにより、増加し続けることはない。

§3 行動獲得ダイナミクス (Q 学習) の構成

外部状態は、障害物の位置とし、障害物が移動する円軌道における x 軸からの角度とする。実装では、これを $18[\text{deg}]$ ごとに 20 分割し、外部状態とする。内部状態は、マニピュレータの姿勢とし、動的に生成された行為によって遷移可能な姿勢と初期姿勢からなり、動的に構成される。報酬は、マニピュレータの先端が目標円内に達した場合 100 を与える。マニピュレータが障害物に接触した場合、または、移動中にリンクが交差した場合、 -50 を与える。各ステップ間における各関節の回転角度の絶対値の和をコスト関数とし、実際の報酬を (7) 式で定義する。選択にはボルツマン分布による確率的選択を用い、温度定数を 1 とする。学習パラメータは、学習係数を 0.5、割引率を 0.9 とし、10000 回の試行を行う。各試行は、報酬またはペナルティーが与えられるまで行われ、与えられた時点で初期状態に戻され再び次の試行が行われる。ただし、1 試行の最大ステップ数を 10 とし、10 ステップとなった時点で、初期状態に戻し、次の試行を行う。上層が目標値を出力してから、全関節の移動が終了するまでを 1 ステップとする。

§4 行動獲得

50 リンクのマニピュレータを対象とし、シミュレーションによりタスクの学習を行う。世代数を 30 とし、その他の設定パラメータを $p=0.01$, $w_1=0.5$, $w_i=0 (i=2, \dots, N-1)$, $w_N=0.5$, $k_f=1.0$, $k_c=0.5$ と設定した。(以後、 $w_i=0 (i=2, \dots, N-1)$ とする。)

本タスクでは、状態・行為空間の大きさは、行為数 $= 9^{50}$ 、状態数 $= 9^{50} \times 20$ となり、Q 学習をはじめとする従来の強化学習では、実装不可能である。また、本論文では、学習器が事前知識を持たない場合を考えているため、従来のニューラルネットワークや、シンプル GA によって解くことは困難である。

図 5 に獲得された行動の一例を示す。ここで各図は、初期状態からゴールに至るまでの運動を等時間間隔で抜き出したものであり "No." は各図の遷移順序を示す。図 5 から障害物を回避し、目標へマニピュレータの先端を移動させる行動が獲得されている事がわかる。この他にも、様々な行動パターンが創発することが確認された。

図 6 に世代に対する行為数、ゲインの推移を示す。ここで、ゲインとは各世代において、行動獲得ダイナミクス (Q 学習) の最終の 100 試行 (9900 試行から 10000 試行) の間に獲得した全報酬と全ペナルティーの総和を総ステップ数で割ったものであり、各世代において獲得された行動の 1 ステップあたりの平均報酬に相当する。図

図 5 獲得された行動

6 から、世代を重ねるにつれてゲインが増加し、それに伴って行為数が減少していることがわかる。第 1 世代において行為数が少ないのは、ランダムに生成された初期個体のなかに致死遺伝子が多く含まれていたためである。第 2 世代では、交配により行為数が増加し、多様な遺伝子による幅広い探索が行われている。また、第 2 世代から第 10 世代にかけてゲインの増加とともに行為数も減少している。これは、学習の主体が、有効な行為が存在しない場合の GA による幅広い探索から、有効な行為が生成された後の Q 学習による行動の学習に移っていくことを示している。

次に、ゲインに注目すると、第 5 世代まではゲインが負であることがわかる。これは、第 5 世代以前では、タスクを達成するために必要な行為がそろっておらず、タスクを達成していないことを示している。しかし、第 1 世代から第 5 世代にかけてもゲインは増加している。これは、ペナルティーを回避する行動が獲得されていることを示している。第 6 世代において、タスクを達成するのに必要な行為が生成され、ゲインが正となりそれとともにゲインが著しく増加し、第 10 世代においてほぼ収束している。次に、ゲインが激しく変化している第 5 世代から第 9 世代におけるタスクの成功数を考察する。図 7 に第 5 世代から第 9 世代までの各世代について、Q 学習における成功回数の推移を示す。横軸が Q 学習の試行回数であり、縦軸が 100 試行ごとの成功回数である。図 7 から、第 6 世代から第 9 世代にかけて、Q 学習の収束

図 6 行為数とゲインの推移

図 7 成功回数の推移

が世代を重ねるごとに速くなっていくことがわかる。これは、行為数の減少と、Q-table を構成する行為に占める有効な行為の割合の増加のためであると考えられる。

複数回の学習を行った結果、30 世代において有効な行動の獲得に至っていない (ゲインが負である) 場合も 2 割程度確認された。この原因については、次のようなことが考えられる。提案手法では、タスクを達成するために必要な行為が存在しない場合には、ペナルティーを回避する行為が高く評価される。しかし、今回のタスクでは、ペナルティーを回避する行為どうしの交叉によって生まれる子個体は、必ずしもタスク達成に有効とはいえない。したがって、タスク達成のために不可欠な行為の最初の個体の発生については、交叉が有効に働かず、確率に頼らざるを得ないためであると考えられる (ただし、致死遺伝子の出現する可能性などから考えると、ランダムに選ぶよりは有効であると考えられる)。この問題は、GA に依存する問題であり、提案手法によって新たに発生したものではない。多くの場合には、コーディング方法や交叉方法などを工夫することにより改善可能であると考えられる。本タスクのコーディング法等の改善については今後の課題である。また、このような問題が回避できないタスクについては、初期個体数を大きくとることで、ある程度対応が可能であると考えられる。本手法では、世代を重ねるごとに自動的に行為数が必要な数まで減少するため、初期世代における個体数を増やし、必要な行為が生まれる確率を増加させても、その後の学習時間を著しく増加させることはない。

表 1 リンク数に対する，計算時間・行為数の変化

Number of links	10	20	30	40	50
Total time [s]	502.0	1263.8	3344.8	5571.4	6868.0
Time of QL [s]	51.4	33.4	35.2	50.0	40.4
Time of GA [s]	5.6	4.6	3.4	5.2	3.0
Number of action	43.0	53.2	50.2	68.8	57.0

§5 リンク数に対する計算時間，行為数の変化

リンク数を変化させた場合の計算時間の変化，行為数の変化についてシミュレーションにより検討する．リンク数を 10 から 10 づつ 50 まで変化させた場合についてそれぞれ，同一のタスクの学習を行う．突然変異によって変化する個体の割合の期待値をリンク数によらず一定にするため，突然変異確率を 10 リンクの場合から 50 リンクの場合まで，それぞれ，0.02, 0.01, 0.0067, 0.005, 0.004 とした．これにより，交叉で誕生した個体のうち約 2 割が突然変異する．表 1 にリンク数に対する時間と行為数の変化を示す．行為数は，各リンク数において最終世代における行為の数を表したものである．表の各値は 5 回のシミュレーション結果の平均値である．ここで，Total time とは，シミュレーションを行うプログラムをスタートさせてからそれが終了するまでの時間であり，マニピュレータの運動を計算する時間を含む．Q 学習に要した時間 (Time of QL) は図 2 の Q 学習の部分の計算に要した時間から，その間にマニピュレータの運動を計算するのにかった時間を引いた時間であり，各世代にかかった時間の 30 世代分の合計である．GA に要した時間 (Time of GA) は，Total time から，Q 学習に要した時間，マニピュレータの運動の計算に要した時間を引いたものである．シミュレーションを行った計算機環境は，CPU : Pentium 800[MHz]，メモリー : 256[MB]，OS : Windows2000 である．

Q 学習に要した時間を見るとリンク数によらずほぼ横ばいとなっていることがわかる．一般にリンク数が増加すれば全探索空間が指数的に増加し，学習時間もそれに伴って増加するが，提案手法においては，Q 学習を行う探索空間は GA により作り出されるため，実際の全探索空間の大きさによらず一定である（これは，表 1 において行為数が増加していないことから確かめられる）．従って提案手法ではリンク数の増加により学習時間が増加していないと考えられる．

次に GA に要する時間では，リンク数の増加により遺伝子長が長くなるため，線形オーダーで増加していく計算量が含まれるが，今回の結果からは，その傾向は見られなかった．これは，線形に増加する計算量に対し，適合度の計算などの行為数に依存する計算量が大きいためであると考えられる．従って，リンク数が増加しても総学習時間（Q 学習に要する時間と GA に要する時間の和）はほぼ横ばいとなると考えられ，シミュレーション結果からもそう読み取ることができる．これは，リンク数が増加すれば全探索空間が指数的に増加することを考えると極めて優れた性質である．

図 8 環境を変化させた場合

最後に，行為数について考える．表 1 の行為数の変化を見ると，リンク数によらずほぼ横ばいとなっている．これは，全行為空間の大きさによらず，そのタスクに必要な大きさの空間に自動的に収束していくことを示していると考えられ，必要な行為数は学習により自動的に獲得される．また，このことは，リンク数の増加に対し Q 学習に要する計算時間が増加しないことにも貢献している．つまり，提案手法では，タスク遂行に必要な行為数は，そのタスクの複雑さによって決まり，その行為に対応する探索空間において学習が行われるため，計算時間も学習対象の自由度の大きさではなく，タスクの複雑さに依存すると考えられる．

4.2 環境変化に対する頑強性

環境変化に対する頑強性を検討するため，学習が収束した段階で，環境を変化させた場合についてシミュレーションを行う．

障害物の移動する軌道の中心を (0.8,0.5) に移動させた場合と，目標となる円の中心を (0.5,0.8) に移動させた場合の 2 つの環境変化に対し，学習を行わせる．各環境とも，図 4 と同じ環境でタスクを 30 世代まで学習させ，30 世代が終了した時点で，それぞれの環境に変化させ，引き続き学習を行わせた．リンク数を 10 とし，学習パラメータは，図 4 と同じとした．図 8 にゲインの推移の一例を示す．それぞれ，5 回の学習を行った結果，同様の傾向が見られた．図 8 から，障害物の軌道を変化させた場合には，変化後最初の世代である 31 世代においてもゲインが正であり，タスクを達成していることがわかる．これは，環境の変化に対し，行為を使うタイミングを変化させることで適応可能であり，現行で維持している探索空間内における強化学習のみで環境変化に適応可能であるためであると考えられる．一方，目標となる円の中心を移動させた場合には，31 世代においてゲインが負となり，世代を重ねることで再び環境に適応していくことがわかる．これは，タスクを達成するためには，移動した目標円内にマニピュレータの先端を移動させることのできる新たな行為が必要となり，探索空間を再構成することで環境に適応しているためであると考えられる．一般に，環境の変化が，現行で維持している探索空間内の変

図 9 k_f を変化させた場合

化であれば、強化学習のみで迅速に適応可能であり、環境の変化が、探索空間外への変化である場合には、世代を重ね、探索空間を再構成することで、自動的に適応可能であると考えられる。

4.3 パラメータの学習結果への影響

提案手法において導入したパラメータの学習に与える影響について考察する。

最初に、コストについて考える。 k_c を 1.0 および 5.0 とした場合について、それぞれ 5 回の学習を行わせた。その他のパラメータは、 $p = 0.01, w_1 = 0.5, w_{N_s} = 0.5, k_f = 1.0$ とし、リンク数は 10 とした。この結果、 $k_c = 1.0$ の場合には、5 回の学習全てに対して、タスクが達成されたが、 $k_c = 5.0$ の場合には、5 回の学習全てに対してタスクが達成されなかった。これは、行為を使うことにより生じたコストが、与えられた割引報酬より大きい場合には、その行為がタスクの達成に有用であっても、負の評価を受けるため、コストが大きい場合には、タスクを達成する政策よりコストを抑える政策が高く評価されるためであると考えられる。従って、設定した報酬に対し、コストが十分小さくなるように k_c を設定する必要がある。また、 $k_c = 1.0$ の場合と、コストを用いない評価関数 ($k_c = 0$ に相当) の下で学習させた結果と比較すると、 $k_c = 1.0$ の場合に得られた政策の方が、マニピュレータの運動が小さい傾向にあることが確認された。

次に、使用頻度の影響について考える。 k_f を 0 および 5.0 とした場合について学習を行わせた。その他のパラメータは、 $p = 0.1, w_1 = 0, w_{N_s} = 1.0, k_c = 0.5$ とした。図 9 に世代とゲインの変化の一例を示す。5 回の各学習の結果全てにおいて同様の傾向が見られた。

図 9 から、 $k_f = 0$ の場合には、一度ゲインが正になった後、再びゲインが負となっていることが分かる。これは、タスクを達成する系列が獲得された後、その系列を構成する行為の少なくとも 1 つが淘汰されることにより、系列が破壊された為であると考えられる。一方、 $k_f = 5.0$ の場合には、ゲインが増加傾向にあることが分かる。これは、学習が進むにつれ、タスクを達成する行為全体が

繰り返し使われ、その使用頻度を適合度として評価することにより、有効な系列を構成する行為の集合が評価され、その結果、有効な集合が次世代に受け継がれたためであると考えられる。提案手法では、Q 学習における試行回数を十分多く確保し、GA における淘汰において、Q 学習で得られた最も効果的な系列を構成する行為を全て次世代に残すと仮定すると、以下のような性質が期待される。まず、GA により動的に構成された探索空間の中で最適な政策に向かって Q 学習が進み、最適な系列を構成する行為が高く評価され、次世代に受け継がれる。次世代では、前世代における最適な系列を構成する行為を含む新しい探索空間の中で再び最適な政策に向かって学習が行われる。従って、世代を重ねるごとに、少なくとも前世代に比べ獲得される政策の性能が向上していくことが期待される。これは、分類子システムにはない、提案手法のきわめて優れた性質であると考えられる。

k_f の違いにより、上記のような変化が起こる原因は、提案手法における GA において共進化が行われているためであると考えられる。行為の系列は、GA においては、互いに協調関係にある個体の集合を意味し、 $k_f = 0$ の場合には、この協調関係の維持が困難であることを意味している。一方、 $k_f = 5.0$ の場合における系列の維持は、協調関係にある集合全体の維持を意味している。なお、共進化においては、局所解に陥りやすいという性質が指摘されており、提案手法においても、 k_f を大きくしていくと、同様の性質が生じると考えられる。局所解からの脱出については、Q 値による適合度がそれに関与することが期待される。使用頻度による適合度は、ある、決まった一つの系列を残すように働くが、Q 値による適合度は、全体の中でその行為がどの程度有効であるかを評価するため、最もよく使われる系列を構成する行為以外にも、その集合の中で比較的有効な行為は全て評価される。従って、局所的な状態においては、次世代に残る集合の多様性が増す方向に働くことが期待される。しかし、現状では、有意な違いを持った学習結果を得るには至っていない。大域的最適性の解析については今後の課題である。

最後に、 p, w_1, w_{N_s} について考える。パラメータの影響が顕著に現れるよう、タスクを次のように変更する。図 4 において、障害物であったものを、目標の円と考え、移動する目標へマニピュレータの先端を到達させることを目標とし、障害物は存在しないものとする。この変更により、タスクを達成するための系列の多様性が増加し、多数の行為で多くのゲインを得る政策や、少ない行為で、タイミングを計りタスクを達成する政策など、さまざまな政策が期待できる。パラメータを変化させた場合の、30 世代目におけるゲインと行為数を表 2 に示す。各値は、5 回の学習結果の平均値である。その他のパラメータは、 $k_f = 1.0, k_c = 0.5$ とした。

表 2 の $p = 0.001$ の列に注目すると、 $w_1 = 0, w_{N_s} = 1$ のときに比べ、 $w_1 = 1, w_{N_s} = 0$ のときの方が、ゲイン、

表 2 パラメータの変化に対する, ゲイン・行為数の変化

	p=0.001		p=0.1	
	ゲイン	行為数	ゲイン	行為数
$w_1 = 1, w_{N_s} = 0$	86.5	189.8	87.8	212.8
$w_1 = 0, w_{N_s} = 1$	66.8	42.2	85.8	275.0

行為数ともに大きいことがわかる。これは、§ 2 で述べたように、 w_1 は、特殊な状況で有効な行為の価値を示しており、 $w_1 = 1, w_{N_s} = 0$ の場合には、それぞれの状態で最も有効な行為を集めてタスクを達成する行動が獲得され、その結果として、行為数が増え、大きなゲインが得られたと考えられる。また、獲得された行動を見ると、目標の位置によって異なる行為が使われ、少ない遷移回数で、タスクを達成しており、遷移回数の少なさがゲインの大きさに貢献していると考えられる。一方、 $w_1 = 0, w_{N_s} = 1$ の場合には、様々な状態で有効な行為が高く評価されるため、少ない行為数でタスクを達成する政策を獲得できるものの、行為の少なさを、タイミングを合わせることで補うため、遷移回数が多くなり、結果としてゲインが小さくなったと考えられる。獲得された行動を見ると、目標がある特定の位置まで移動するまでのタイミングを計り、その位置に向かってマニピュレータを移動させることでタスクを達成していた。

次に、 p の変化に注目する。 p が 0.001 から 0.1 に変化すると、 $w_1 = 1, w_{N_s} = 0$ の場合に比べ、 $w_1 = 0, w_{N_s} = 1$ の場合には、行為数、ゲインともに大きく増加していることがわかる。 $w_1 = 1, w_{N_s} = 0$ の場合には、 w_1 は、規格化された Q 値の最大値に対する重みであり、タスクの達成に関与した行為の規格化された Q 値の最大値は、 p より大きくなり、 p に依存しない値となる。行為数が若干増加しているのは、タスクの達成には関与しないが、ペナルティーが少ない行為の価値が p の増加により増加したためであると考えられる。一方、 $w_1 = 0, w_{N_s} = 1$ の場合には、ペナルティーが少ない行為の価値が増加したことに加え、有効な行為の価値が減少したことが原因であると考えられる。 w_{N_s} は、規格化された Q 値の状態に対する平均値に対する重みであり、 p が大きくなることで、有効な行為の相対的な価値が減少する。その結果、ペナルティーの少ない様々な種類の行為が次世代に残りやすくなり、行為数が増加したものと考えられる。ゲインが増加した原因は、行為数の増加によりタスクの達成が容易になったためであると考えられる。従って、 w_{N_s} のように、規格化された Q 値の広い範囲の平均値に対する重みを大きくすることにより、少ない行為数でタスクを達成する行動の獲得を目標とする場合には、 p を十分小さくする必要があるといえる。なお、 p が最も有効に働くのは、タスクが達成されない場合に、ペナルティーを評価する場合であり、この場合には、規格化された Q 値の最大値が p となるため、0 から p における相対的な大きさが重要となり、 p そのものの絶対的な大きさは意味を持たないため、 p の大きさを変化させても、生き残

る個体の割合に影響を与えない。従って、 p を十分に小さい値に設定しても性能は劣化しないと考えられる。

5. 考察

本章では、従来の強化学習が抱えていた問題を提案手法がどのように解決しているかをまとめるとともに、その結果得られる提案手法の特徴について述べる。

・状態・行為空間の大きさに関する問題について

提案手法では、状態・行為空間の部分集合に対し Q 学習を行う。したがって、元となる状態・行為空間が大きい場合でも、Q 学習が実装可能な大きさの部分集合に対して Q 学習を実装すればよい。また、部分集合を更新しながら複数回繰り返すことにより広範囲な状態・行為空間における行動の獲得を可能としている。

・部分観測と一般化の問題について

提案手法では、Q 値をもとにして一般化行為と特殊行為を評価する。一般化行為は複数の状態で有効であるが、それは、異なる状態を同一と認識しているのではなく、異なる状態を別々に認識し、その異なる各状態で同一の行為を実行することを意味する。従って部分観測とはならない。これは、分類システムなどにおいて、異なる状態を同一の状態として認識することで一般化を行う方法と異なり、優れた性質であると考えられる。

探索空間の構成とルールの生成に関しては、提案手法では、内部状態は行為に伴って作られる。これは、行為をもとに実際に遷移可能な領域のみを対象に探索が行われることを意味し、学習効率がよい。また、提案手法では、行為と状態の組み合わせは Q 学習によって行われる。従って、スタートからゴールに至るまでの一連の各状態とその状態で有効な行為の組み合わせが獲得される。一方、分類システムでは、ルール (状態と行為のペア) は GA によってシンボリックに作られる。従って、あるルールによって遷移した状態に対応する有効なルールが存在するとは限らず、効率が悪い。

・外部状態の大きなタスクへの適用について

提案手法では、外部状態については動的な生成が行われず、極めて大きな外部状態を持つタスクへの適用は困難である。解決策としては、事前知識の利用や、タスクを複数のサブタスクに分割することなどが考えられる。提案手法における学習器は階層構造を有しており、階層構造を多段階化することで複雑なタスクを複数の簡単なタスクの結合体として学習することが可能である。階層構造の多段階化とタスクの分割については別途報告予定である。

・環境の変化に対する頑強性について

提案手法は、環境変化が探索空間内の変化である場合には、強化学習により迅速に適用可能であり、環境変化が探索空間外への変化である場合には、探索空間を自動的に再構成することで適応可能である。これは、環境変

化の大きさに応じて、異なるダイナミクスを用いて自動的に適応することを意味しており、強化学習と GA 双方の柔軟性を併せ持つ優れた性質であると考えられる。

・ 行為の数、学習に要する計算時間について

提案手法では、学習が進むにつれて行為数が必要な数まで減少し、学習によって行為数を獲得させることが可能であるといえる。必要な行為数は、タスクの複雑さや達成の難しさによって決まる。学習に必要な時間は、主に生成された行為数に依存するため、学習時間は、学習対象のもつ探索空間の大きさではなく、タスクの複雑さ、タスク達成の難しさによって決まると考えられる。

6. 結 論

本論文では、多自由度ロボットをはじめとする大規模な探索空間をもったシステムを対象に、階層構造をもつ学習機構を提案し、その学習アルゴリズムとして、GA により動的に探索空間を生成する Q 学習のアルゴリズムを提案した。提案手法を多数の冗長リンクからなるマニピュレータの障害物回避問題に適用し、シミュレーションにより有効な行動が獲得されることを確認した。

本手法は、多自由度ロボットのように、行為数、内部状態数の多い対象に対し、極めて有用であると考えられる。

現在我々は、アクチュエータを搭載した多数の均一なユニットの結合体からなり、環境やタスクに応じて様々な形態に分離・再結合することで様々なタスクを遂行可能な、超冗長メカニカルシステムを提案している [伊藤 00, 亀川 00]。今後の課題は、本手法を超冗長メカニカルシステムに対して適用することで、ハード・ソフト両面の柔軟性をそなえ、自己組織化することで環境に適応する、自律メカニカルシステムを構築することである。

なお、本研究は、文部省科学研究費（東工大 COE 形成基礎研究費スーパーメカノシステム）を使用して行われました。

◇ 参 考 文 献 ◇

- [Holland 86] Holland, J. H.: Escaping brittleness: the possibilities of general purpose learning algorithms applied to parallel rule-based system., in *Machine Learning II*, pp. 593-623 (1986).
- [Jakkola 94] Jakkola, T., Singh, S. P., and Jordan, M. I.: Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems, in *Advances of Neural Information Processing Systems 7*, pp. 345-352 (1994).
- [Sutton 88] Sutton, R. S.: *Reinforcement Learning: An Introduction*, The MIT Press (1988).
- [Svinin 99] Svinin, M., Ushio, S., Yamada, K., and Ueda, K.: Emergent systems of motion patterns for locomotion robots, in *Proc. Int. Workshop on Emergent Synthesis*, pp. 119-126 (1999).
- [Ushio 01] Ushio, S., Svinin, M., Ueda, K., and Hosoe, S.: An Evolutionary Approach to Decentralized Reinforcement Learning for Walking Robots, in *Proc. of the 6th Int. Symp. on Artificial life and Robotics*, pp. 176-179 (2001).
- [Watkins 92] Watkins, C. J. C. H. and Dayan, P.: Technical Note Q-Learning, *Machine Learning*, Vol. 8, pp. 279-292 (1992).
- [Westerdale 97] Westerdale, T. H.: Classifier systems - no wonder they don't work, in *Proc. the Second Annual Genetic Programming Conf.*, pp. 529-537 (1997).
- [Wilson 89] Wilson, S. W. and Goldberg, D. E.: A critical review of classifier systems, in *Proc. the Third Int. Conf. on Genetic Algorithms*, pp. 244-255 (1989).
- [Yamada 01] Yamada, K., Ohkura, K., Svinin, M., and Ueda, K.: Adaptive Segmentation of the State Space based on Bayesian Discrimination in Reinforcement Learning, in *Proc. of the sixth Int. Symp. on Artificial life and Robotics*, pp. 168-171 (2001).
- [伊藤 00] 伊藤, 松野: 強化学習によるヘビ型ロボットの制御, 第 18 回日本ロボット学会学術講演会予稿集, pp. 1261-1262 (2000).
- [亀川 00] 亀川, Chatterjee, R., 伊藤, 松野: 多様な移動形態を持つ 3 次元超冗長移動ロボットの開発と制御, 第 18 回日本ロボット学会学術講演会予稿集, pp. 1147-1148 (2000).
- [宮崎 99a] 宮崎, 木村, 小林: Profit Sharing に基づく強化学習の理論と応用, 人工知能学会誌, Vol. 14, No. 5, pp. 800-807 (1999).
- [宮崎 99b] 宮崎, 木村, 小林: Profit Sharing を用いたマルチエージェント強化学習における報酬分配の理論的考察, 人工知能学会誌, Vol. 14, No. 6, pp. 1156-1164 (1999).
- [高橋 99] 高橋, 浅田: 実ロボットによる行動学習のための状態空間の漸次的構成, 日本ロボット学会誌, Vol. 17, No. 1, pp. 118-123 (1999).
- [森本 00] 森本, 銅谷: 階層型強化学習を用いた実ロボットによる起立運動の獲得, 第 5 回ロボティクスシンポジウム, pp. 397-402 (2000).
- [中村 99] 中村, 黒山, 山田, 大倉, 上田: Instance-Based Classifier Generator による自律移動ロボットの行動獲得, 日本ロボット学会誌, Vol. 17, No. 3, pp. 371-379 (1999).
- [片山 99] 片山, 小林: TD() 学習の対数時間更新算法, 人工知能学会誌, Vol. 14, No. 5, pp. 879-890 (1999).
- [木村 00] 木村, 小林: Actor に適正度の履歴を用いた Actor-Critic アルゴリズム, 人工知能学会誌, Vol. 15, No. 2, pp. 267-275 (2000).
- [林 99] 林, 末松: POMDP 研究に基づいたハイブリッド分類システム, 人工知能学会誌, Vol. 14, No. 3, pp. 538-546 (1999).

〔担当委員: 和田俊和〕

2001 年 4 月 5 日 受理

—— 著 者 紹 介 ——

伊藤 一之 (学生会員)

1996 年 3 月 法政大学工学部電気工学科卒業。1998 年 3 月 同大学大学院工学研究科修士課程システム工学専攻修了。日清紡績株式会社を経て、1999 年 4 月東京工業大学大学院総合理工学研究科博士後期課程知能システム科学専攻入学。現在、同課程在籍中。主に知能ロボットに関する研究に従事。1997 年 The 5th International Ground Robotics Competition において優勝。日本ロボット学会学生会員。

松野 文俊

1957 年 7 月 26 日生。86 年大阪大学大学院基礎工学研究科博士課程修了。大阪大学基礎工学部助手、神戸大学工学部講師、助教授を経て 96 年より東京工業大学大学院総合理工学研究科助教授（知能システム科学専攻）となり、現在に至る。主に、ロボティクス・制御理論・地理情報システムに関する研究に従事。最近は、地理情報システムを基盤としたレスキュー活動支援システムに関する研究を行っている。89 年度日本ロボット学会研究奨励賞、93 年度システム制御情報学会論文賞、97 年度消防防災科学論文賞（消防庁長官表彰）、RoboCup Japan Open 2000 JSAI 賞、2001 年度計測自動制御学会論文賞・武田賞などを受賞。IEEE、日本機械学会、GIS 学会、情報処理学会などの会員。工学博士。