解 説

強化学習法とロボットへの応用

Reinforcement Learning Method and its Applications to Robot

畝 見 達 夫* *創価大学工学部情報システム学科兼国際ファジィ工学研究所

Tatsuo Unemi* *Department of Information Systems Science, Faculty of Engineering, Soka University, and Laboratory for International Fuzzy Engineering Research

1. ま え が き

ここ数年、機械学習あるいはロボット工学の分野で研究が活発化しているテーマに強化学習法 (reinforcement learning method) がある。強化学習という用語自身は、もともと動物心理学あるいは動物行動学の分野で用いられた言葉であり、ラットなどの動物に、ある行動を起こしたときだけ餌などの報酬を与えるという操作を繰り返すと、その行動パターンが徐々に「強化」され、ついには、実際には報酬が与えられなくても、同様な状況に置かれるとその行動を起こすようになるといった一種の適応動作のことを指す。

機械学習の側からは、「強化学習」は、ある種の学習問題のクラスを指す言葉として用いられる。つまり、学習者はある環境の中で行動を起こすエージェント、たとえば、自律移動ロボットや動物個体が想定され、学習者は各時間ステップにおいて得られる感覚入力から行動を決定する。実際に取った行動に対して環境から報酬あるいは罰が与えられるが、報酬の大きさは多くの場合、過去数ステップの行動系列に対して決定される。学習の目的は、ある時間長さにわたる報酬の重み和を最大化することである。

人工知能あるいは機械学習における強化学習研究は、古くはコンピュータによる学習に関する先駆的研究として有名な、1950年代のSamuelのChecker Player[23]にまでさかのぼることができる。その後も、強化学習という名前こそ使われてはいないが1968年に発表されたD. MichieとR. A. Chambersの倒立振子の学習制御の論文[18]も、罰からの強化学習を扱った先駆的な研究である。もちろん、それ以前にもA. G. Barto, R. S. Sutton, C. W. Andersonらの研究[4], A. H. Klopffニューラルネットを用いた条件反射のモデルの研究[13], 遺伝的アルゴリズムを用

いたJ. Hollandの分類システムに関する研究[11]があるが、現在の強化学習研究の流れの中心は、1989年の機械学習国際会議[21]あたりから始まると考えてよからう。最近ではロボット学習の枠組としても拡張が試みられている[7]。

以下、強化学習の問題領域について述べた後、主な手法を紹介し、ついで、現状における研究課題について述べる。

2. 強化学習法の問題領域

強化学習法では、学習者は各時間ステップにおいて観測される状態から行動を決定する。ここで「状態」とは、学習システムにとっての外部からの入力であり、環境からの感覚入力や学習者の内部状態、あるいは、それらの組み合わせでもよい。実際に取った行動に対して環境から報酬あるいは罰が与えられるが、報酬の大きさは多くの場合、過去数ステップの行動系列に対して決定される。学習の目的は、ある時間長さにわたる報酬の重み和を最大化することである。報酬と罰を合わせて強化信号 (reinforcement) と呼ぶ。報酬を「正の強化 (positive reinforcement)」, 罰を「負の強化 (negative reinforcement)」と呼ぶこともある。ただし、負の強化という言葉は、行動科学では、報酬がない状況を意味する用語として用いられる場合もあるので、注意が必要である。

より一般的には、連続時間系を扱うことも考えられるが、現在の研究はすべて離散時間を前提にしている。状況は時刻の刻みごとに変化し、同時に学習者は、その刻みごとに行動決定を行なう。これはマルコフ決定過程 (Markovian decision process) に相当する。

学習者の目的を形式的に記述すればつぎのようになる。時刻 t における強化信号の大きさを r_t とすると、現在から未来にわたる強化信号の重み和：

$$v_t = \sum_{i=t}^{\infty} \gamma^{i-t} \cdot r_i \quad (1)$$

を最大化することが学習者の目的になる。 γ は $0 \leq \gamma \leq 1$ なる定数であり、報酬の割引率あるいは単に割引率 (discount rate) と呼ばれる。 $r_t > 0$ の場合には報酬、 $r_t < 0$ の

原稿受付 1994年11月21日

キーワード: Reinforcement Learning, Machine Learning, Markovian Decision Process, Robot Learning

*〒192 東京都八王子市丹木町1-236 〒231 横浜市中区山下町89-1 シイベルヘグナービル3階

*1-236 Tangi-cho, Hachioji, Tokyo 192, Japan. Siber Hegner Bldg., 89-1 Yamashita-cho, Naka-ku, Yokohama 231, Japan.

場合には罰が与えられたものと考え、 $\gamma=0$ の場合は、現在の強化信号のみに着目し未来を無視することになる。つまり、行動の評価は極めて日和見的なものとなる。逆に $\gamma=1$ では、どんなに遠い未来でもよいから大きな報酬が得られる方がよいことになる。つまり、行動の評価は極めて長期的なものとなる。すなわち、 γ の値の大小によって、どのくらい先の未来までを考慮するかが決まる。しかし、未来の報酬は観測できないので、一般には過去から現在までの強化信号の重み和：

$$\bar{v}_t = \sum_{i=0}^t \gamma^{t-i} \cdot r_i \quad (2)$$

を v_t の近似として利用する。

強化学習における学習問題はつぎの2点で特徴づけることができる。

- ・システムが出力すべきデータが教師から直接与えられることはなく、システムが実際に行なった出力に対する評価という形で与えられる。
- ・システムの出力に対する評価が即座に与えられず、行為の系列に対する評価が遅れて与えられる。

$\gamma=0$ の場合は、つぎ以降のステップにおける報酬を無視することになるため、結果的に即座に与えられる報酬のみを考慮することとなる。この意味で、 $\gamma=0$ の場合を、あえて強化学習の問題として取り扱うことは意味がない。

強化学習システムは一般に状態 x_t に関する評価を決める部分である「学習要素」と、状態から次の行動 a_t を決定する「実行要素」に分けることができる。各要素と情報の流れの関係を図示すると図1のようになる。実行要素では、学習によって得られた状態の評価の見積りをもとに行動を決定するが、その時点の評価見積りを最大にするような行動選択が、必ずしも最適な決定となるとはかぎらない。なぜなら、強化学習では学習者の経験は学習者自身の行動に強く依存するからである。学習一般に関して、経験

図1 強化学習における処理要素と情報の流れ

の内容によって学習結果が大きく異なることは当然だが、強化学習では、学習者の行動選択が経験の内容を左右する。 v_t を真に最大化するには、環境に対して十分な探検 (exploration) を行なう必要がある。

状態の複雑さと行為の複雑さは、強化学習の問題をさらに分類する際の重要な要素となる。詳しくは、状態の表現複雑さ、状態変化の不確かさ、状態変化の文脈依存性、環境変動、行為選択肢の複雑さなどが問題とされる。もっとも単純な場合は、状態と行動選択肢が共に離散的かつ数え上げ可能で、すべての場合を表現するに十分な記憶が学習者側に用意でき、1つの状態に関して最適な行為は唯一であり、環境も安定であるというものである。実際の動物のモデルとして、あるいはロボット制御への応用を考えると、このような理想的な場合は想定しにくい。強化学習の特徴を捉えるための足掛かりとしては有効である。工学的には、信号の区間分類による記号化や問題設定の単純化によって有効に使える場合もある。

また、強化信号の見積りも重要である。あらかじめ、どのような強化信号がどのような分布で存在するかが分かっているならば、それに応じた学習方法をとることが可能な場合もある。たとえば、強化信号が0または-1であれば、-1が与えられるまでの時間をできるだけ引き延ばすような戦略が有効となる。また、 v_t の最大値が分かっているならば、その値に近い性能が得られるようになった時点で、探検を中止するという戦略をとることができる。

3. 主な実現方法

強化学習は問題につけられた名前であるため、その実現方法は様々なものが提案されている。以下、代表的な実現方法について概説する。

3.1 Q 学習

C. J. C. H. Watkins によって提案された Q 学習[35]では、状態と行為の組に対する評価を見積もる。この評価を「Q 値」と呼び、状態と行為の組から評価の見積りを導く関数を「Q 関数」と呼ぶ。時刻 t において状態 x_t にあって、行為 a_t を選択した結果、状態は x_{t+1} となり、強化信号 r_t が得られたとすると、更新すべき Q 値の変更幅はつぎのように定義される。

$$\Delta Q(x_t, a_t) = \alpha(r_t + \gamma \max_b Q(x_{t+1}, b) - Q(x_t, a_t)) \quad (3)$$

α は学習率であり $0 < \alpha \leq 1$ なる定数である。 γ は割引率である。つまり、つぎのステップで最適と思われる行為を選択したときに得られると見込まれる評価の見積り $\max_b Q(x_{t+1}, b)$ を1段階だけ割引いた値と、そこで直接得られた強化信号 r_t の和に $Q(x_t, a_t)$ を近づける。これにより、 $\max_b Q(x_t, b)$ を最適な行動を取ったときの v_t (式(1)参照) に近づけることができる。

Q関数の実現方法は様々なものがある。もっとも単純なのは、すべての状態と行為の組み合わせについての表をつくる方法である[27]。表の内容として各Q値を記録しておき、上の式に従って、その値を直接更新する。もちろん、表の大きさが利用可能な計算機の記憶容量を越えるような問題領域には使えない。1つの対応策は階層型ニューラルネットを用いてQ関数を表現することである[15]。あるいはCMACを用いる方法[29]、記憶に基づく方法[22]、実例に基づく方法[31]なども有効である。

学習されたQ値をもとに行為を決定する方法としてWatkinsはつぎのようなBoltzmann分布に基づく確率的な行為選択法を提案している。すなわち、状態 x において行為 a を選択する確率を：

$$P(a|x) = \frac{\exp(Q(x, a)/T)}{\sum_{b \in \text{possible actions}} \exp(Q(x, b)/T)} \quad (4)$$

と定義する。ただし、 T は温度定数であり、値が大きいほど行為はよりランダムになり、積極的に探検を行なうことになる。逆に T を0に近づけると、わずかなQ値の差が行為選択に大きく影響することとなり、極限ではQ値を最大にする行為が選ばれることになる。つまり、

$$a = \arg \max_b Q(x, b) \quad (5)$$

となる。これを「貪欲な決定戦略 (greedy policy)」と呼ぶ。貪欲な決定戦略では、探検が行なわれにくいため、局所最適解に陥る可能性が高くなる。ただし、状態が複雑なために過去と同じ状態を訪れる可能性が低い場合には、この限りではない。

3.2 実例に基づく強化学習法

実例に基づく強化学習法 IBRL2[31]について概略を説明する。この方法では、エージェントの具体的な経験の記憶の集合が、そのエージェントの内部状態となる。学習要素は、記憶の単位として、感覚データ x_t 、行為データ a_t 、評価の見積り V_t 、および記憶の信頼性 R_t からなる4つ組 $\langle x_t, a_t, V_t, R_t \rangle$ を記憶にそのまま蓄える。意思決定関数 Policyは、現状に類似しかつ評価の高い記憶要素を記憶の中から検索することによって行動を決定する。十分に類似しかつ評価の高い記憶要素が見つからない場合には、ランダムな行動をとる。

時刻 t における評価はQ学習などと同様に過去から現在までの強化信号の重み和を用いる。個々の離散時刻における経験を長さ N の固定長の待ち行列に蓄え、0でない強化信号が得られた時点で、割引率 $\gamma (0 < \gamma < 1)$ を掛けながら記憶を過去にさかのぼり、個々の経験の記憶に付加された評価に割り引かれた強化信号の値を加える。すなわち、時刻 i の経験に対する評価 V_i に加える値は

$$\Delta V_i = \gamma^{t-i} r_t \quad (6)$$

となる。実例に基づく方法は経験の丸暗記を基本とするた

め、最近の経験を待ち行列に蓄えること自身は特別な計算負荷にはならない。

強化学習に限らず、実例に基づく学習を成功させるための要点は、いかにして記憶要素の増加を防ぐかにある[2]。我々の方法では、新たな経験が既に記憶された過去の経験とある程度以上類似する場合には、つぎのいずれかの手段により記憶要素の増加を防ぐ。(1) 現在の経験の評価見積りの方が高ければ、過去の記憶を忘却し、新たな経験を記憶する。(2) 現在の経験の評価見積りの方が低ければ、過去の記憶の評価を下方修正し、新たな経験は記憶しない。さらに、(3) あらかじめ決められた記憶容量を越える場合には、最も信頼性の低い記憶要素を忘却する。信頼性は、(1) 失敗した場合は参考にした記憶の信頼性を下げる。(2) 成功した場合は参考にした記憶の信頼性を上げる。失敗とは、予想に反して低い強化信号しか得られなかった場合、あるいは、罰が与えられた場合である。詳細は文献[31]を参照されたい。負の強化信号のみから学習する方法については、倒立振子の制御への応用をとおして性能が確認されている[32]。

3.3 分類システム

遺伝的アルゴリズムの創始者としても有名な Holland らが提案した分類システムにも、強化学習を行なうための枠組が組み込まれている[11]。分類システムにおける実行要素は一種のプロダクションシステムである。各ルールには「強さ」が割り当てられており、条件部が満足された実行可能ルールの間の競合は、強さと条件部の具体性を参考に解消される。この強さは、ある意味でQ値のようなものであるから、Q学習の場合のように強さに比例した確率で実行ルールを選択することも考えられる。ある程度実行が進み、ルールの強さに十分な変更がなされた時点で、強さを適応度とみなして遺伝的アルゴリズムを適用する。これにより、むだなルールは淘汰され、有効なルールを組み合わせる新たなルールの生成が行なわれる。

強さの調整方法、つまり、報酬割当ての方法には主にバケツリレー・アルゴリズム (bucket brigade algorithm) [11] と利益共有 (profit sharing) 法[10]の2つがある。

バケツリレー・アルゴリズムでは、各ステップにおいて、活性化した、つまり、実行されたルールの強さを修正する。強さ修正のアルゴリズムはつぎのようになる。

- (1) 活性化したルールについて、その強さに比例した大きさだけ強さを減少させる。
- (2) 強さの減少分を、直前のステップで実行されたルールの強さに加える。
- (3) ルールの実行結果として得られた報酬を、そのルールの強さに加える。

得られた報酬は、ただちに過去のルールに伝播されず、次の実行の際に1段階だけ伝播する。この点は、前述

の Q 学習の場合の Q 値の変更の伝播と同様であり、学習は遅いが、各ステップにおける計算コストは小さい。

利益共有法では、実行されたルールの履歴を保存しておき、報酬が得られる度に、報酬の値を減じながら過去にさかのぼってルールの強さを修正する。過去に実行されたルールの強さも一時に修正するため学習は速い。報酬が稀にしか得られない場合には、それほど計算コストはかからないが、頻繁に得られる場合には高くつく。また、報酬の有無によって、そのステップの計算コストが大きく異なるため、ロボットの学習などに用いる場合には、実時間性の面で注意が必要となる。利益共有法における有効な報酬割当ての方法についての理論的考察も行なわれている[20]。

以上の他にも Q 学習の元になった TD(λ)[26]、確率的学習オートマトンによる方法[19]、ファジィルールとニューラルネットを組み合わせた方法[5]などがある。

4. 拡張の方向と課題

強化学習の枠組を自律移動ロボットや腕ロボットの制御に応用するには、いくつかの拡張が必要となる場合が多い。現在、一般に強化学習法研究に課せられた問題としては、探検戦略、階層的計画、多様な入出力データ構造、隠れマルコフ決定問題などが挙げられる。また、マルチエージェントへの応用も注目されている。

4.1 探検戦略

強化学習の枠組では、学習者は未知の状況に置かれたときも、情報を得るために何らかの行動を起こすことが必要となる。また、環境変化への対応や最適な行動の追求を行なわせるには、それまでの学習によって獲得した行動パターンに固執せず、ある程度の冒険をすることが必要となる。この2つの状況への対応方法は、学習システムの設計問題の一部であり、学習の能力に大きく影響する。しかし、この問題を正面から扱った研究はまだ少なく、今後の研究の発展が望まれる。

すべての強化学習機構には何らかの形で探検戦略が組み込まれている。たとえば Dyna のような参照表を用いる接近法では、絶えず冒険を好む戦略を採用することによって、比較的容易に最適な行動パターンを学習することができる。

しかし、出力の選択肢が増加した場合には、ランダムな戦略や、絶えず冒険を好むような戦略は無能であろう。出力データの構造を考慮した適切な探索戦略の開発も望まれる。

一般的に環境変動に対する適応戦略を考える際に、探検による未知知識の収集と既知の知識を利用した性能の高い行動の実行という矛盾する2種類の目的のバランスをどのようにかはかるかは、極めて重要な問題である。学習が十分行なわれた後では、探検よりも既知の知識に従って高い報酬が確実に得られる行動選択をすべきである。しかし、ある時点で学習が十分行なわれたかどうかを判断することは

難しい。先に述べたとおり、最高の性能が分かっているならば、その性能に近づくに従って探検を減らすことも考えられる。あるいは状態の多様性が分かっているならば、可能な状態全体について、どの程度網羅したかによって探検を減らすことも考えられる。大きな環境変動の直後では、既存の知識が役に立たなくなる可能性が大きいため積極的に探検すべきである。環境変動の度合は、性能の低下や未知状態の頻度、予測的中率などから判断することになる。

4.2 複雑なタスクへの適用

多くのサブゴールを含むような複雑なタスクを、そのまま強化学習の問題として取り扱うことは適切でない。なぜなら、上で述べたような枠組では学習の目標は強化信号の最大化という1つの尺度に集約されており、必ずしもタスクの達成度を強化信号に翻訳できるとは限らないからである。強化学習法自身は、問題領域に関する背景知識をできるだけ持たない状態から学習を開始することを前提としているが、複雑なタスクに適用するには、それなりに問題の構造を反映したヒントを学習者に与える必要がある。

1つの方法は、階層的な行動計画を可能とすることである。複雑な計画問題を抽象化あるいは分割することによって十分小さな部分問題に分け、それぞれの計画を統合することによって問題解決を行なうといった手法は人工知能や知識工学の分野では普通に行なわれている。現在の強化学習法の多くは、単一の評価基準にしか対応できず、副目標の集合を統一的に扱うような枠組はまだ確立されていない。あらかじめ設計者の手によって問題を分割しておき、それぞれの部分問題を強化学習法によって学習するといった試みはいくつかみられるが[15][12][24][16]、分割の方法自身を学習するところまでとはいっていない。この問題に関しては強化学習の枠組がそのまま使えるかどうか検討せざるを得ないであろう。

もう1つは経験の制約である。あるいは、教育カリキュラムの工夫といった方が分かりやすいかも知れない。すなわち、いきなり難しい問題を解かせようとしても、学習者が強化信号を得られるような行動を取らなければ、学習は進まない。そこで、簡単な問題から始めて、それぞれの段階で得た知識に基づけば容易に強化信号が得られるような問題へと徐々に問題を難しくする。これにより学習が誘導され、結果的に複雑な問題にも対応可能となる[3]。

強化信号からだけではなく、教示、つまり教師から直接与えられる出力例を利用する方法もある。これにはオンラインとオフラインの2とおりの方法が提案されている。オンラインとは、学習者が試行を繰り返しているさなかに教師が学習者の手を取って動作を変えることである。教師は学習者の動きを観察し、不適切な動作を取ろうとしたときに割り込んで、適切な動作を強制する。学習者は、教師に指示された動作を適切なものと仮定し、記憶に変更を加え

る[6]. オフラインの教示とは、前もって模範的な行動系列の情報を学習者に与えるものである。学習者は、模範行動に含まれる状態と近い状態に遭遇したときに与えられた模範行動を高い確率で実行する。行動選択を確率的にすることによって、模範行動よりも高い性能を獲得する可能性もある[15].

4.3 多様な入出力データ構造

離散的かつ多様性の少ない入出力データを扱う問題領域では、強力で効率の良いアルゴリズムが多く提案されており、理論的な解析も行ない易い。しかし、現実的なロボット制御への応用や、人間の知能モデルとしての妥当性を考えるとき、連続値ベクトルや、様々に構造のことなるデータを同時にかつ統一的に扱うことが要請される。データの多様性が増すとともに学習が難しくなることは既に述べたとおりであるが、データの構造の性質を利用した効率の良い手法を考えなければ、今後の発展は難しいであろう。ニューラルネットなど、3.1節で述べたQ関数を表現する様々な手法は、主に入力として連続値ベクトルを扱う問題に対応するものである。しかし、出力側の多様性についてはまだ十分な研究がなされていない。人間は視覚や聴覚など、処理速度の異なる入力信号処理モジュールを統合的に扱う機構を備えており、そのような方向への拡張も検討する必要がある。

4.4 隠れマルコフ決定問題

自律移動ロボットや動物では、環境全体の状況を把握することは困難かつ無意味である。さらされる可能性のある環境の多様性は、行動主体の情報処理能力をはるかに越えており、人工知能の分野で盛んに議論されるフレーム問題を容易にひきおこす。実際、人間においても、各時点では自らの周囲の局所的な情報しか取り入れておらず、複雑な状況判断が可能なのは、ある程度の時間間隔における経験の履歴を元に思考しているからである。この能力に欠けると、たとえば、エレベータの前で何度も同じ動作を繰り返さなければならなくなる。どの階にいてもエレベータのドアを見る限り区別はできないが、自分がどこから来たかを考えれば、容易に判断がつく。もちろん、この状況も一種のフレーム問題をはらんでおり、目的地とは異なる階で降りてしまったにもかかわらず、気づかずにしばらく廊下を歩き続けてしまったような経験をお持ちの方も多かるう。ニューラルネットを用いた接近法では、リカレントネットを用いる方法[15]、実例に基づく方法では、実例を時系列として記憶する方法[30]が提案されている。

4.5 マルチエージェントと強化学習

複数のエージェントによる強化学習、あるいは、複数の強化学習エージェントによる集団行動の研究も、近年、活発化しつつある。先駆的なものとしては、分業化したマルチエージェントの協調問題解決を強化学習の枠組として捉

えようとする G. Weiß の研究[36]がある。個別の行動ルールを持ったエージェントが協調して1つのタスクを達成しようとするものであり、分類子システムの分散化とみることもできる。

強化学習を行なう複数のエージェントが協調的に問題解決を行なおうとする場合に、どのような種類の情報の共有が可能で、その特徴はどのようなものか、という問題について研究を行なったのは M. Tang である[28]。単純な追跡問題を用いて(1)入力情報の共有、(2)経験の共有、(3)行動戦略の共有の3通りの場合についてシミュレーション実験を行なっている。

人工生命的な立場からは、筆者の研究[33]がある。強化学習を行なうエージェントが複数集まった場合にどのような集団行動が発生し得るかという問題について考察を加え、実例に基づく強化学習を埋め込んだエージェントが衝突回避と求餌行動を行なう場合のシミュレーション実験について示している。また、社会性昆虫集団の競合状況について確率的学習オートマトンを用いた強化学習メカニズムを組み込んだ研究も久保ら[14]によって行なわれている。

シミュレーションではなく実物のロボットを使った研究としては、Mataric の研究[17]がある。4台のロボットが餌と仮定された物体を、巣と呼ばれる場所へ集める作業を行なう。問題領域に依存した単純な状態表現を用いることで、実物を使う難しさを軽減している。実物のロボットでセンサ情報をそのまま学習システムへの入力として使うことの難しさを感じさせる。

これらの研究は、まだ緒についたばかりである。学習者集団の行動については、社会学的にも興味深い分野であり、今後の発展が期待される。

5. む す び

ロボット学習への応用の視点から強化学習研究について述べてきたが、動物心理学、ニューラルネットワーク、人工生命、学習理論などの視点からは、それぞれに異なった見方ができる。人工生命研究の立場からは4.5節で紹介した集団行動の研究だけでなく、たとえば、[1][34]のような動物をモデルとした学習能力の進化に関する研究も興味深いテーマである。また、計算論的な学習理論の視点からは、学習の収束性や学習可能性の問題に興味を持たれており、たとえば[8][25][9]にあるように、比較的単純なくつかのアルゴリズムについては学習の収束性などについての証明がなされている。

国内ではまだ研究者は少ないが、今後、応用と理論の両面にわたる発展を期待したい。

参考文献

- [1] D. Ackley and M. Littman: "Interactions between Learning and Evolution," C. G. Langton et al (eds.), *Artificial Life II*, Addison-Wesley, pp. 487-509, 1992.
- [2] D. W. Aha, D. Kibler and M. K. Albert: "Instance-Based Learning Algorithm," *Machine Learning*, vol. 6, pp. 37-66, 1991.
- [3] M. Asada, S. Noda, S. Tawaratsumida and K. Hosoda: "Purposeful Behavior Acquisition on a Real Robot by Vision-Based Reinforcement Learning," *Workshop on Robot Learning*, Rutgers Univ., 1994.
- [4] A. G. Barto, R. S. Sutton and C. W. Anderson: "Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 13, no. 5, pp. 834-846, 1983.
- [5] H. R. Berenji: "A Reinforcement Learning-Based Architecture for Fuzzy Logic Control," *International Journal of Approximate Reasoning*, vol. 6, no. 2, pp. 267-292, 1992.
- [6] J. A. Clause and P. E. Utgoff: "A Teaching Method for Reinforcement Learning," *Proceedings of the Ninth International Conference on Machine Learning*, pp. 92-101, 1992.
- [7] J. H. Connell and S. Mahadevan (eds.): "Robot Learning," Kluwer Academic, 1993.
- [8] P. Dayan: "The Convergence of TD(λ) for General λ ," R. S. Sutton (ed.), *Reinforcement Learning*, Kluwer Academic, pp. 117-138, 1993.
- [9] C. -N. Fiechter: "Efficient Reinforcement Learning," *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory*, pp. 88-97, 1994.
- [10] J. J. Grefenstette: "Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms," *Machine Learning*, vol. 3, pp. 225-245, 1988.
- [11] J. H. Holland, K. J. Holyoak, R. E. Nisbett and P. R. Thagard: "Induction," MIT Press (1986). 市川伸一ほか(訳), "インダクション—推論・学習・発見の統合理論へ向けて", 新曜社, 1991.
- [12] L. P. Kaelbling: "Hierarchical Learning in Stochastic Domains: Preliminary Results," *Proceedings of the Tenth International Conference on Machine Learning*, pp. 167-173, 1993.
- [13] A. H. Klopff: "A Neuronal Model of Classical Conditioning," *Psychobiology*, vol. 16, no. 2, pp. 85-125, 1988.
- [14] 久保正男, 嘉数侑昇: "蟻の餌争奪ゲームによるマルチエージェントシステムの協調動作評価", *情報処理学会論文誌*, vol. 35, no. 8, pp. 1555-1566, 1994.
- [15] Lin, Long-Ji: "Scaling Up Reinforcement Learning for Robot Control," *Proceedings of the Tenth International Conference on Machine Learning*, pp. 182-189, 1993.
- [16] M. J. Mataric: "Reward Functions for Accelerated Learning," *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 181-189, 1994.
- [17] M. J. Mataric: "Learning to Behave Socially," *From Animals to Animals—Simulation of Adaptive Behavior III*, 1994.
- [18] D. Michie and R. A. Chambers: "Boxes; An Experiment in Adaptive Control," E. Dale and D. Michie (eds.), *Machine Intelligence*, vol. 2, Oliver & Boyd, Edinburgh, pp. 137-152, 1968.
- [19] 三上貞芳, 嘉数侑昇: "確率的ルール学習による動的タスク割り当て問題へのアプローチ", *日本機械学会論文集*, vol. 58, no. 551, pp. 2276-2285, 1992.
- [20] 宮崎和光, 山村雅幸, 小林重信: "強化学習における報酬割当ての理論的考察", *人工知能学会誌*, vol. 9, pp. 580-587, 1994.
- [21] B. Spatz (ed.): "Proceedings of the Sixth International Workshop on Machine Learning," Morgan Kaufmann, 1989.
- [22] A. Moore and C. G. Atkeson: "Memory-Based Reinforcement Learning; Converging with Less Data and Less Read Time," in [7], pp. 79-103, 1993.
- [23] A. L. Samuel: "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal on Research and Development*, vol. 3, pp. 210-229, 1959.
- [24] S. Singh: "Transfer of Learning by Composing Solutions of Elemental Sequential Tasks," R. S. Sutton (ed.), *Reinforcement Learning*, Kluwer Academic, pp. 99-115, 1993.
- [25] S. Singh, T. Jaakkola and M. I. Jordan: "Learning Without State Estimation in Partially Observable Markovian Decision Processes," *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 284-292, 1994.
- [26] R. S. Sutton: "Learning to Predict by the Methods of Temporal Differences," *Machine Learning*, vol. 3, pp. 9-44, 1988.
- [27] R. S. Sutton: "Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming," *Proceedings of the Seventh International Conference on Machine Learning*, pp. 216-224, 1990.
- [28] M. Tang: "Multi-agent Reinforcement Learning; Independent vs. Cooperative Agents," *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330-337, 1993.
- [29] C. K. Tham and R. W. Prager: "A Modular Q-Learning Architecture for Manipulator Task Decomposition," *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 309-317, 1994.
- [30] 畝見達夫: "予測と反省に基づく時系列の暗記学習", *情報処理学会人工知能研究会報告*, vol. 70, no. 4, 1990.
- [31] 畝見達夫: "実例に基づく強化学習法", *人工知能学会誌*, vol. 7, pp. 141-151, 1992.
- [32] 畝見達夫: "実例に基づく強化学習法による失敗しない制御方法の学習", *人工知能学会誌*, vol. 7, pp. 1001-1008, 1992.
- [33] T. Unemi: "Collective Behavior of Reinforcement Learning Agents," *Proceedings of the 1993 IEEE/Nagoya University WWW on Learning and Adaptive System*, pp. 92-97, 1993.
- [34] T. Unemi: "Evolutionary Differentiation of Learning Abilities—a case study on optimizing parameter values in Q-learning by a genetic algorithm," to appear in *Proceedings of Artificial Life IV*, MIT Press, 1994.
- [35] C. J. C. H. Watkins and P. Dayan: "Technical Note; Q-Learning," R. S. Sutton, (ed.), *Reinforcement Learning*, Kluwer Academic, pp. 55-68, 1993.
- [36] G. Weiß: "Learning to Coordinate Actions in Multi-Agent Systems," *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 331-316, 1993.

畝見達夫 (Tatsuo Unemi)

1956年1月26日生。1978年東京工業大学工学部制御工学科卒業。1980年同大学院総合理工学研究科システム科学専攻修士課程修了。1981年同専攻博士後期課程中退。同専攻助手。1987年長岡技術科学大学工学部計画・経営系講師。1992年より創価大学工学部情報システム学科講師兼国際ファジィ工学研究所客員研究員。この間、自然言語処理、知識工学、機械学習などの研究に従事。現在は、遺伝的アルゴリズム、強化学習法、自律分散ロボットの研究を行なっている。情報処理学会、日本認知科学会、日本ソフトウェア科学会等会員。博士(工学)。