小特集「最近の機械学習」

強化学習

Reinforcement Learning

畝見 達夫*
Tatsuo Unemi

* 創価大学工学部情報システム学科／国際ファジィ工学研究所
Dept. of Information Systems Science, Faculty of Engineering, Soka University./Laboratory for International Fuzzy Engineering Research.

1994年8月22日 受理

Keywords: reinforcement learning, machine learning, Markovian decision process, robot learning.

1. まえがき

強化学習(reinforcement learning)は，本来，動物心理学あるいは動物行動学の分野で用いられた用語である．ラットなどの動物に，ある行動を起こしたときだけえさ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．典型的には，このように報酬を契機として行動パターンを学習する場合に用いられる用語だが，広くは，罰による行動の抑制も含め，「条件づけ」といわれる一連の適応現象を実現する学習を「強化学習」と呼んでいる．

認知科学の立場からは，最も低いレベルの学習を支えるメカニズムとして「強化学習」を捉えることができる．このような単純な学習メカニズムの積上げだけで，論理や概念を対象とするような高次の学習が実現できるとは考えにくいが，比較的知能が高いとされる多くの動物に共通に観察される強化学習の機能が，人間の知能の基底部にあるとする直観も，さほど的はずれではあるまい．

機械学習の側から捉えなおすと，「強化学習」は，ある種の学習問題のクラスを指す言葉となる．すなわち，学習者はある環境のなかで行動を起こすエージェント，例えば，自律移動ロボットや動物個体が想定される．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．実際にと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[Samuel 59]にまでさかのぼることができる．差し手の系列としての対戦経験から，過去の対戦経験についての記憶にどのような得点を割り当てるかという問題が取り扱われており，強化学習の範囲と考えられる．コンピュータによる学習の研究のなかで強化学習という名前が最初に使われたのは，1965年のM. D. WaltzとK. S. Fuのプラントの学習制御を目標においた論文[Waltz 65]と思われる．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[Michie 68]も，罰からの強化学習を扱った先駆的な研究である．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[ML 89]あたりから始まると考えてよかろう．もちろん，それ以前にもA. G. Barto, R. S. Sutton, C. W. Andersonらの研究[Barto 83]，A. H. Klopfのニューラルネットを用いた条件反射のモデルの研究[Klopf 88]，遺伝的アルゴリズムを用いたJ. Hollandの分類子システムに関する研究[Holland 86]があるが，人工知能関連の国際会議で強化学習の論文が6件も発表されたのは，1989年が最初である．昨年は，機械学習国際会議の一環として強化学習ワークショップが開かれ，80名以上の参加者があった．現在は，ロボット制御への工学的応用と，計算論的な基礎研究が活発化している．

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

強化学習とは，ある種の学習問題のクラスを指す言葉である．学習者はある環境のなかで行動を起こすエージェント，例えば，自律移動ロボットや動物個体が想定される．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，それらの組合せでもよい．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う．これはマルコフ決定過程(Markovian decision process)に相当する．

形式的に記述すれば次のようになる．時刻 t における強化信号の大きさを r_t とすると，学習者の目的は，現在から未来にわたる強化信号の重み和

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

を最大化することである．ただし，γ は $0\leq\gamma\leq1$ 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，$r_t<0$ の場合には罰が与えられたものと考える．$\gamma=0$ 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目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ほうがよいことになる．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

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

を v_t の近似として利用する．

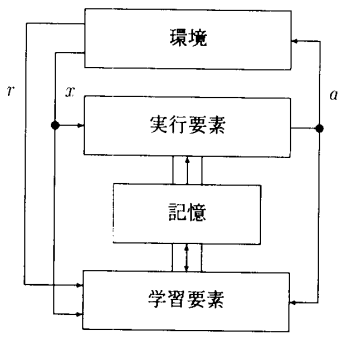
強化学習における問題の特徴は次のとおりである．

- システムが出力すべきデータが教師からは与えられず，システムが実際に行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．

これが多くの帰納学習やパターン分類学習が扱う領域と大きく異なる点であり，自らの行為の系列が重要な意味を持つ．$\gamma=0$ の場合は，次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限らない．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，状態表現の複雑さ，状態変化の不確実さ，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



x：状態＝入力データ
a：行為＝出力データ
r：強化信号＝報酬または罰

図1 強化学習における処理要素と情報の流れ

とされる．最も単純な場合は，状態と行動選択肢がとも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，一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が，強化学習の特徴を捉えるための足掛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，どのような分布で存在するかがわかっていれば，それに応じた学習方法をとることが可能な場合もある．例えば，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の最大値がわかっていれば，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，その実現方法はさまざま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 TD(λ)

TD(λ)法[Sutton 88]は，現在の強化学習法研究の中心的存在の1人であるR. S. Suttonによって考案された一種の時系列予測手法である．

まず，通常の教師つきの学習を考えてみよう．時刻tで観察される状態をx_t，そのときの報酬をr_tとする．x_tは実数ベクトルであるものとする．状態xに対する報酬の見積りを$P(x)$とし，学習者は，関数Pを特徴づけるベクトルwを更新することにより学習を進めるものとする．もし，各時刻において報酬が与えられるとすると，r_tを見積もるためには

$$\Delta w_t=\alpha(r_t-P_t)\nabla_w P_t \tag{3}$$

とすればよい．ここで，αは学習率であり$0\leq\alpha\leq1$なる定数である．もし，$P(x)=w^Tx$なら上の式は，

$$\Delta w_t=\alpha(r_t-w_t^Tx_t)x_t \tag{4}$$

となり，これはいわゆるデルタ規則である．

しかし，強化学習で扱いたい問題は，x_tからr_tを予測するのではなく，その後，いつか得られるであろう報酬r_m $(m>t)$を予測することにある．これを実現するには式(3)を次のように変更すればよい．

$$\Delta w_t=\alpha(r_t+P(x_{t+1})-P(x_t))\sum_{k=1}^{t}\lambda^{t-k}\nabla_w P(x_k) \tag{5}$$

λは割引率である．$\lambda=1$の場合，TD(1)は，Widrow-Hoffの学習手続きとしてよく知られたアルゴリズムと等しくなる．詳しくは[Sutton 88]を参照されたい．

この手法は状態の善し悪しを見積もるだけであるから，行為選択にあたっては，各時刻においてどの行為を選ぶとどの状態へいくのかがわかっていないと意味がない．もちろん，2ステップ後のことはわからなくてもよい．もし，環境の地図がある程度わかっているのなら，探索による行動計画も可能となるが，上で述べた学習則は各時刻において局所的に行えるため，深い探索を時間をかけて計算する必要はなく，実時間性が要請される応用領域では有効な手段となり得る．見方を変えれば，行動系列を記憶するために十分な記憶容量がないエージェントが，自らの行動のみから環境の探索を行う場合に有効な手法とも考えられる．

3・2 Q 学 習

上述のTD(λ)法では，状態に対する評価を見積もるのに対し，C. J. C. H. Watkinsによって提案されたQ学習[Watkins 93]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値の変更幅は次のように定義される．

$$\Delta Q(x_t,a_t)=\alpha(r_t+\gamma\max_b Q(x_{t+1},b)-Q(x_t,a_t)) \tag{6}$$

αは学習率であり，$0<\alpha\leq1$なる定数である．γは割引率である．つまり，次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段階だけ割り引いた値と，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関数の実現方法はさまざまなものがある．最も単純なのは，すべての状態と行為の組合せについての表をつくる方法である[Sutton 90]．表の内容として各Q値を記録しておき，上式に従って，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超えるような問題領域には使えない．一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[Lin 93]．あるいはCMACを用いる方法[Tham 94]，記憶に基づく方法[Moore 93]，実例に基づく方法[畝見 92]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は次のような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

$$P(a|x)=\frac{\exp(Q(x,a)/T)}{\sum_{b\in\text{possible actions}}\exp(Q(x,b)/T)} \quad (7)$$

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，積極的に探索を行うことになる．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，

$$a=\arg\max_b Q(x,b) \quad (8)$$

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・3 分類子システム

遺伝的アルゴリズムの創始者としても有名なHollandらが提案した分類子システムにも，強化学習を行うための枠組みが組み込まれている[Holland 86]．分類子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応する．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て新たなルールの生成が行われる．

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)[Holland 86]と利益共有(profit sharing)法[Grefenstette 88]の二つがある．

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次のようになる．

(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搬されずに，次の実行の際に1段階だけ伝搬する．この点は，前述のQ学習の場合のQ値の変更の伝搬と同様であり，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，報酬が得られるたびに，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．報酬がまれ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われている[宮崎 94]．

4. 拡張の方向と課題

実際の動物のモデル，あるいはロボット制御への応用として，さまざまな拡張が提案されている．本章ではその主な方向性について概説する．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あげられる．また，応用としてマルチエージェントに関する問題も注目されている．

4・1 探 検 戦 略

強化学習の枠組みでは，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．この二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．にもかかわらず，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．

とはいえ，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．例えば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

とることによって，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検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図るかは，極めて重要な問題である．学習が十分行われた後では，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，最高の性能がわかっていれば，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わかっていれば，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，既知の知識は無用となった可能性もあるので積極的に探検すべきだが，環境変動の度合いを直接観測することはできないから，性能の低下や，未知状態の頻度，予測の的中率などから判断するしかない．

4・2 階層的計画

外部からの入力に対して次の行為を決定するという反射的な行動の組合せだけで人間が行うような高度な行動が実現されるとは考えにくい．強化学習も，最終的には人工知能や認知科学が扱ってきたような，さまざまな知識と深い推論に基づく用意周到な行動計画の問題に結びつくものでなければ知能のモデルとはいえないし，また，複雑な作業を行うロボットへ応用することもできない．その第一歩は，階層的な行動計画を可能とすることであろう．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み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見られるが[Kaelbling 93, Lin 93, Mataric 94a, Singh 93]，分割の方法自身を学習するところまではいっていない．この問題に関しては強化学習の枠組みがそのまま使えるかどうかも検討せざるを得ないであろう．

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，理論的な解析も行いやすい．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，さまざまに構造の異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すで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．ニューラルネットワークなど，3・2節で述べたQ関数を表現するさまざまな手法は，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超えており，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もとに思考しているからである．この能力に欠けると，例えば，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は区別はできないが，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ろう．ニューラルネットワークを用いた接近法では，リカレントネットを用いる方法[Lin 93]，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[畝見 90]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みとして捉えようとするG. Weißの研究[Weiß 93]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一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見ることもできる．

強化学習を行う複数のエージェントが協調的に問題解決を行おうとする場合に，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ったのはM. Tangである[Tang 93]．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っている．

人工生命的な立場からは，筆者の研究[Unemi 93]がある．強化学習を行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[久保 94]によって行われている．

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，Matariの研究[Mataric 94b]がある．4台のロボットがえさ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

機械学習分野における強化学習研究を中心に述べてきたが，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よう．

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，例えば，[Ackley 92, Unemi 94]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が持たれており，例えば[Dayan 93, Fiechter 94, Singh 94]にあるように，比較的単純ないくつかのアルゴリズムについては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◇

[Ackley 92] Ackley, D. and Littman, M.: Interactions between Learning and Evolution, C. G. Langton, *et al.* (eds.), *Artificial Life II*, pp. 487–509, Addison Wesley (1992).

[Barto 83] Barto, A. G., Sutton, R. S. and Anderson, C. W.: Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems, *IEEE Trans. on Systems, Man and Cybernetics*, Vol. 13, No. 5, pp. 834–846 (1983).

[Dayan 93] Dayan, P.: The Convergence of TD(λ) for General λ, R. S. Sutton (ed.), *Reinforcement Learning*, pp. 117–138, Kluwer Academic (1993).

[Fiechter 94] Fiechter, C.-N.: Efficient Reinforcement Learning, *Proc. 7th Annual ACM Conf. on Computational Learning Theory*, pp. 88–97 (1994).

[Grefenstette 88] Grefenstette, J. J.: Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, *Machine Learning*, Vol. 3, pp. 225–245 (1988).

[Holland 86] Holland, J. H., Holyoak, K. J., Nisbett, R. E. and Thagard, P. R.: *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).

[Kaelbling 93] Kaelbling, L. P.: Hierarchical Learning in Stocastic Domains: Preliminary Results, *Proc. 10th Int. Conf. on Machine Learning*, pp. 167–173 (1993).

[Klopf 88] Klopf, A. H.: A Neuronal Model of Classical Conditioning, *Psychobiology*, Vol. 16, No. 2, pp. 85–125 (1988).

[久保 94] 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，情処学論，Vol. 35, No. 8 (1994).

[Lin 93] Lin, Long Ji: Scaling Up Reinforcement Learning for Robot Control, *Proc. 10th Conf. on Machine Learning*, pp. 182–189 (1993).

[Mataric 94a] Mataric, M. J.: Reward Functions for Accelerated Learning, *Proc. 11th Int. Conf. on Machine Learning*, pp. 181–189 (1994).

[Mataric 94b] Mataric, M. J.: Learning to Behave Socially, *From Animals to Animats—Simulation of Adaptive Behavior III* (1994).

[Michie 68] Michie, D. and Chambers, R. A.: Boxes: An Experiment in Adaptive Control, E. Dale and D. Michie (eds.), *Machine Intelligence*, Vol. 2, pp. 137–152, Oliver & Boyd, Edinburgh (1968).

[宮崎 94] 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, No. 4, pp. 580–587 (1994).

[ML 89] Spatz, B. (ed.): *Proc. 6th Int. Workshop on Machine Learning,* Morgan Kaufmann (1989).
[Moore 93] Moore, A. and Atkeson, C.G.: Memory-Based Reinforcement Learning: Converging with Less Data and Less Read Time, J.H. Connel and S. Mahadevan (eds.), *Robot Learning*, pp. 79-103, Kluwer Academic (1993).
[Samuel 59] Samuel, A.L.: Some Studies in Machine Learning Using the Game of Checkers, *IBM J. on Research and Development,* Vol. 3, pp. 210-229 (1959).
[Singh 93] Singh, S.: Transfer of Learning by Composing Solutions of Elemental Sequential Tasks, R.S. Sutton (ed.), *Reinforcement Learning*, pp. 99 115, Kluwer Academic (1993).
[Singh 94] Singh, S., Jaakkola, T. and Jordan, M.I.: Learning Without State-Estimation in Partially Observable Markovian Decision Processes, *Proc. 11th Int. Conf. on Machine Learning,* pp. 284-292 (1994).
[Sutton 88] Sutton, R.S.: Learning to Predict by the Methods of Temporal Differences, *Machine Learning,* Vol. 3, pp. 9-44 (1988).
[Sutton 90] Sutton, R.S.: Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming, *Proc. 7th Int. Conf. on Machine Learning,* pp. 216-224 (1990).
[Tang 93] Tang, M.: Multi-agent Reinforcement Learning: Independent vs. Cooperative Agents, *Proc. 10th Int. Conf. on Machine Learning,* pp. 330-337 (1993).
[Tham 94] Tham, C.K. and Prager, R.W.: A Modular Q-Learning Architecture for Manipulator Task Decomposition, *Proc. 11th Int. Conf. on Machine Learning,* pp. 309-317 (1994).
[畝見 90] 畝見達夫：予測と反省に基づく時系列の暗記学習，情処学会人工知能研報，70 4 (1990).
[畝見 92] 畝見達夫：実例に基づく強化学習法，人工知能学会誌，Vol. 7, No. 1, pp. 141-151 (1992).
[Unemi 93] Unemi, T.: Collective Behavior of Reinforcement Learning Agents, *Proc. 1993 IEEE/Nagoya University WWW on Learning and Adaptive System,* pp. 92-97 (1993).
[Unemi 94] Unemi, T.: Evolutionary Differentiation of Learning Abilities—a case study on optimizing parameter values in Q-learning by a genetic algorithm, to appear in *Proc. Artificial Life IV,* MIT Press (1994).
[Waltz 65] Waltz, M.D. and Fu, K.S.: A Heuristic Approach to Reinforcement Learning Control Systems, *IEEE Trans. on Automatic Control,* Vol. 10, No. 4, pp. 390-398 (1965).
[Watkins 93] Watkins, C.J.C.H. and Dayan, P.: Technical Note: *Q*-Learning, R.S. Sutton (ed.), *Reinforcement Learning,* pp. 55-68, Kluwer Academic (1993).
[Weiß 93] Weiß, G.: Learning to Coordinate Actions in Multi-Agent Systems, *Proc. 13th Int. Joint Conf. on Artificial Intelligence,* pp. 331-316 (1993).

「**著者紹介**」は，前掲(Vol. 9, No. 4, p. 523)参照.