人工生命と学習

―― 生命に学ぶ新しい適応アルゴリズム ――

三上 貞芳，嘉数 侑昇

生命は実証されている優れた最適化システムとして，我々の学ぶべき教師である．特に生命における適応が，すべて個体に与えられる生死にかかわる賞罰信号のみに基づいた学習の結果として実現されていることは興味深い．ここでは生命の持つこのような「賞罰のみからの学習」のメカニズムを工学的手法に実現しようとする手法である，強化学習法を紹介する．さらに強化学習法の拡張として，生物が群として生存してきたことをシミュレートすることで，複数のエージェントによる協調動作の学習による獲得が可能であることを示す．例題として交通信号網の学習制御問題を取り上げ，シミュレーションにより有効性を検討する．

1．はじめに

人工生命とは生命の振る舞いを観測・模倣することで，新しい情報処理の規範を見出そうとする一連の研究の動きである．

生命は実証されている優れた最適化システムとして，我々の学ぶべき教師である．特に生命が適応してきた過程がすべて，外部からモデルを与えられることなく行われてきた点は重要な特徴である．生命における適応はすべて，個体の生死，栄養補充と飢餓状態などの，個体に与えられる何等かの賞罰信号のみに基づいた学習の結果として実現されている．これは汎用性の点で従来のモデル中心の工学的最適化問題にはなかった大きな特徴を示している．

機械・情報工学の夢の一つは「適応する機械」の実現であるといえる[2]．適応とは予期せぬ状況，すなわちモデルが構築されていない状況でも，与えられた目標を達成する手段を自ら見つけることのできる自律能力を意味する．

知的機械の分野では，複数の知的エージェントによる自律分散システムが，対故障性などから「適応する機械」として優れた能力を示すであろうことが認識されている[8]．それにもかかわらず，この分野ではいまだ有力なアルゴリズムは提案されていない．この理由は本来モデルの設定できない問題である，実世界での他のエージェントの振る舞いをもモデル化しようということ自体に無理がある点である．しかし計算機の記憶能力，処理能力が発達した現在，生命が長い時間と個体を費やして行ってきた「賞罰からの学習」メカニズムによる適応は，これを機械の上に実現することが現実の視野に入りはじめてきた．

本稿では，生命の学習メカニズムを工学的にまねるメカニズム，さらにはこれらを知的機械の各応用にどのように取り入れるかに関して，最近の研究成果，および応用例を交えて簡単に紹介する．

Sadayoshi MIKAMI, Yukinori KAKAZU 北海道大学工学部

2．適応する機械

実世界との相互作用を行う機械・ロボットには必ず，実世界の予測不可能性に対して，それに自律的に対処する能力が求められる．一例を挙げれば，

- 消火活動を行うロボット：火災現場は状況の予測ができず，かつ状況自身が動的に変化するような典型的な場合であり，進入，消火活動は試行錯誤の過程を必要とする．
- 配管検査作業ロボット：人間の入り込めないような狭い管内を歩く検査ロボットなどは，ロボット自身の故障に対しても可能な限り前進，脱出する能力が必要である．これは惑星探査ロボットなどにも共通する．
- 交通信号機ネットワーク：一交差点での交通流最適化は，必ずしも区域全体の交通流最適化とはならない．交通信号機ネットワークなどの協調動作を行う機械群に関しては，協調動作全体のモデルをあらかじめ用意することは規模などの点で困難である．

このような予測不可能性を扱うためには，従来の枠組みでは，原理的にあらゆる可能性に対してルールをあらかじめ用意することが必要となる．当然これは不可能なこととなる．一方で未知な環境で唯一確実に入手できる情報とは，環境への行動に対する結果の情報である．これを目的への賞罰の評価とすれば，機械の目標は「賞罰からの学習」となる．

3．環境適応アルゴリズムの成功例としての生命

生命は「生死」への接近・回避を唯一つの評価関数として，未知環境，動的環境への適応を成功させてきた．

適応への評価という立場でみた場合，生死に対する評価は次の２つの段階に分けることができる．

- 個体としての生死への評価
  目標：個体の満足
  評価：生に近づく度合，死に近づく度合
- 集団としての生死への評価
  目標：生命全体の満足
  評価：世代（＝生死）

集団としての生死を規範とする最適化手法のアナロジーは，進化を模倣した遺伝的アルゴリズムなどが挙げられる．一方で個体としての生死を規範とする最適化手法のアナロジーの一つは，強化学習[1]と呼ばれる形で工学手法に実現されている．

4．強化学習

強化学習は，確率的な試行錯誤を通じて，与えられた目標関数を最大化するような行動を選択するルールを獲得させる学習手法を指す．具体的には図１に示すように，センサ入力された状態から，行動を確率的に選択し，その結果（報酬，Payoff）を行動選択の政策(Policy)に反映させるサイクルからなる．目的は，以上のサイクルにより環境からの報酬を最大化させるような行動を選択する政策を獲得することである．

ある時点で環境から報酬信号を得たとする．報酬を得ることができた原因は，過去にわたる状態・行動の遷移の系列が寄与している．したがって，これら関与する過去の行動・

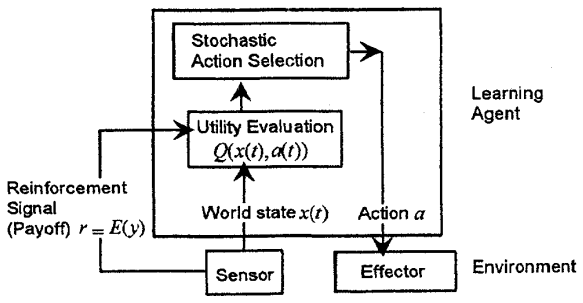

図１　強化学習

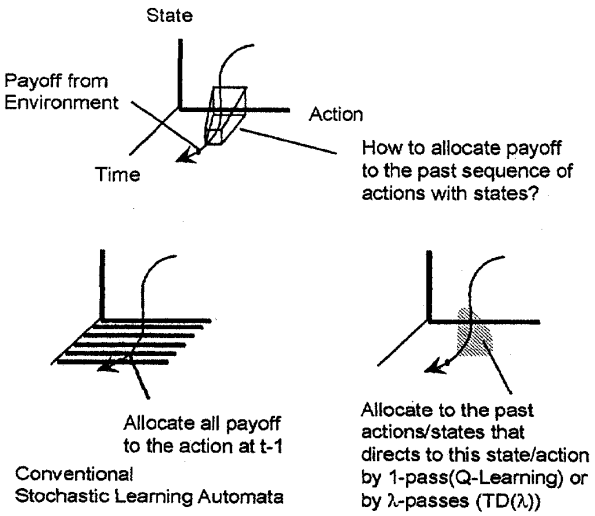

図2 報酬割当問題と各種強化学習法

状態は，報酬にしたがって評価されなければならない．しかし実際には厳密に過去にたどってきた系列のみが現在の報酬に寄与するのみならず，その近傍の行動をとることによっても同様に現在の報酬に寄与できる可能性を持っている．すなわち図2に示すように，過去の時間，状態，行動によるボリュームに対していかに報酬を割り当てるかが問題であり，その割当方法はその後の報酬予測能力に大きな影響を与える[2]．

強化学習は図2に示すように，報酬の割当方法と政策の表現方法により様々なものが提案されているが，これらはすべて以下のように定式化できる．

1．時刻 t で観測された状態を $x(t)$ とする
2．状態 $x(t)$ で可能な行動の集合を $A(t)=\{a_i(t)\}$ とする
3．t で行動 a_i を実行した結果，遷移した状態を $y(x,a_i)$ とする
4．関数 $E(y)\in\Re$ を，状態 y の報酬(profit)関数と呼び，y' が y より適切な行動である場合，$E(y')>E(y)$ であるものとする．
5．時刻 t における最適行動 $\hat{a}_{i(t)}$ とは，

$$V(x(t),\hat{a}_{i(t)})=\sum_{k=0}^{\infty}\gamma_k\cdot E(y(x(t+k),\hat{a}_{i(t+k)}),\quad\cdots\cdots(1)$$

としたときに，

$$\max_{\hat{a}_{i(t)}}\ V(x(t),\hat{a}_{i(t)}),\quad\cdots\cdots(2)$$

となるような行動 $\hat{a}_{i(t)}$ である．ここで γ_k は割引率と呼ぶ0から1までの定数を示す．式(1)，(2)の意味は，時刻 t で $\hat{a}_{i(t)}$ を選択した後，引き続き最適行動が選択され続けた場合，将来にわたる割引率付きの報酬の総和 $V(x(t),\hat{a}_{i(t)})$ が最大化されるように $\hat{a}_{i(t)}$ が選ばれていることである．割引率は将来の報酬の考慮の度合を与えるもので，たとえば $\gamma_0=1$ かつ $\gamma_k=0$，$k>0$ なるように設定することは，直後の報酬のみを考慮することを意味する．

これらの定義のもとに，強化学習は，$y(x(t-1),\hat{a}_{i(t-1)}),x(t)$ の観測から行動 $a_{i(t)}$ を

$$a_{i(t)}=P_t(x(t)),\quad\cdots\cdots(3)$$

なる $x(t)$ の関数(政策)として選択し，$\hat{a}_{i(t)}$ が適当な γ_k のもとでの最適行動 $\hat{a}_{i(t)}$ になるようにこれを繰り返すような，状態観測と実行のくり返しによる探索である．問題は，いかにVのよい近似を得るかであり，報酬割当問題は $E(y)$ のもとに政策 P_t を更新し，式(3)が最適行動に近づくようにすることである．

近年，報酬関数と状態遷移がマルコフ的である場合に，最適行動を発見できるような政策更新手続きが提案されはじめている[1][3][5]．その代表的なものはQ-Learningとよばれる手法である[3]．以下ではQ-Learningを題材として強化学習の手続きを説明し，さらに強化学習の応用の実際例を示す目的で，学習する信号機ネットワークをQ-Learningにより実現した例を紹介する．

5. Q-Learning

式(2)に示すとおり，状態 $x(t)$ で選択する行動 $a(t)$ は，これ以降に得られる報酬関数の(割引率付きの)総和である式(1)を最大化するものでなくてはならない．仮に状態 $x(t)$ における行動それぞれに対して，十分に良い(1)式の近似が得られたものとする．これをQ値(またはUtility)とよび，$Q(x,a)$ と記述する．この場合状態遷移，および評価関数の値がマルコフ的であったならば，意志決定関数 $P_t(x(t))$ は，各時点での状態数 $|x(t)|$ の種類だけ用意すればよく，次のように書ける．

$$P_t(x(t))=P_{x(t)}. \quad \cdots\cdots(4)$$

すると最適行動 $\hat{a}_{i(t)}$ の選択である式(3)の政策は，$x(t)$ での最大Q値を与える a を選択することに帰着される．そして政策を与える関数の実態としては，各行動 a に対するQ値 $Q(x,a)$，$\forall a$ が保存されていることになる．

Watkinsらにより提案されたQ-learningは，状態観測と行動，その報酬の受け取りのサイクルにより，$Q(x,a)$ を $V(x,a)$ に近づけるようなQ値の更新式を与えている．これは次のように定式化されている（図3）．

いま状態 $x(t)=x$ で行動 $a(t)=a$ を行った結果，状態 y に遷移し，報酬関数値 $E(y)=r$ を得たものとする．この場合のQ値を Q' とすると，Q' は次のように書ける．

$$Q'(x,a)=r+\gamma\cdot\max_{a'}Q(y,a'). \quad \cdots\cdots(5)$$

従って，以前の x,a でのQ値 $Q(x,a)$ は式(5)に近づくよう修正されなくてはならない．これは適当な学習係数 $\alpha\in[0,1]$ を掛けて次のように行われる．

$$Q(x,a)\leftarrow Q(x,a)+\alpha\cdot(Q'(x,a)-Q(x,a)). \quad \cdots\cdots(6)$$

Q値は報酬の見積もりの近似値であるから，最大のQ値をもつ行動を常に選択することは誤った行動を固定することになり好ましくない．従ってQ値の大きさを反映して，これに確率的揺らぎを加えて行動を選択することが望ましい．これを実現する一例として，各行動 a_i に対応したQ値 $Q(x,a_i)$ から，ガウシアン分布に従った確率

$$\Pr(x,a)=e^{Q(x,a)/T}/\sum_{\forall a'}e^{Q(x,a)/T}, \quad \cdots\cdots(7)$$

を計算し，この確率に従って一つの行動 a を確率的に選択する機構が考えられている．ここで T は揺らぎの大きさをコントロールする係数で，温度パラメータと呼ばれる．以下のシミュレーションで例示するように，高速な収束を期待するには，試行の初期段階で温度を高くし広く探索を行い，徐々に温度を下げることでQ値の収束に近づいた段階での揺らぎをおさえるような，適切なアニーリングのスケジュールが重要となる．

以上より，式(3)の意志決定関数 P_x は，各行動 a に対するQ値 $Q(x,a)$，$\forall a$ を保管し，かつ式(7)による確率的行動選択により行動を返し，式(6)によりQを更新するような動作となっている．このことから次のような学習法としての特徴が導き出せる．

Q-Learningでは1試行で通過した状態，行動のパスに関するQ値が，その試行で得た報酬のみではなく，遷移先のQ値を含めた値で修正される．したがって遷移先のQ値がV

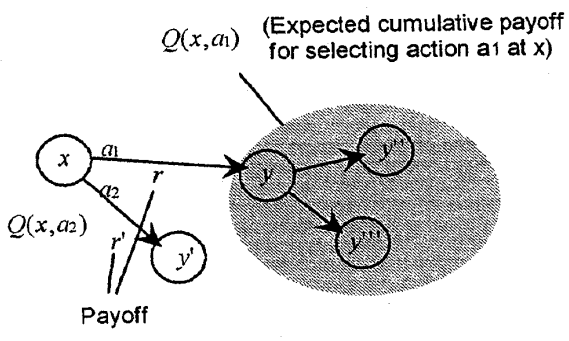

図3 Q値

値(式(1))の良い近似値となっていれば，実際の報酬信号rが，状態yでの真の報酬信号$E(y)$にゆらぎnを伴った$E(y)+n$で与えられたとしても，式(5)の係数γが大きく，Q値がnに対して十分大きい場合には，nの及ぼすゆらぎに敏感に影響されることはない．従ってQ-Learningは一般に，よい行動の系列から逸脱せずに探索を続けることができる特徴を有している．それゆえ強化学習法のクラスでは高速な収束性をあらわすことが知られている．

6．現在の研究状況

以上示してきたQ-Learningに代表されるような強化学習法は，汎用性の高い学習法であることが大きな利点であるが，一方でその収束性の遅さなどに関して多くの未解決部分が残されていることも指摘できる．これに対して現在，次のような様々な拡張が試みられている．

● R.Sutton, 1988(1)—TD(λ)による学習効率の改善．

Q-LearningやTDなどの学習更新式は，1回の試行で現在行われた行動・状態の直前の行動・状態群のみが影響を受ける．これに対して学習収束性を向上させる方法として，政策の更新を時間軸上で後ろに伸ばし，過去λステップまで政策の更新を繰り返す方法が提案されている．

● G.Tesauro, 1991(1)—ニューラルネットワークによる強化学習．

強化学習の最大の問題点は，学習収束速度が他の学習手法に比べて遅い点である．これを改善する一つの方法は，図2に示す報酬を与えるボリュームを大きくすることである．そこで政策をニューラルネットワークに記憶させることで，近傍も含めて学習の影響を与える手法が提案されている．ここでニューラルネットワークは例えばQ-Learningにおいて状態を入力素子にとり，アナログ出力をQ値として扱うことに相当する．問題空間での真の近さを正確に反映してはいないが，実験的にはニューラルネットワークの汎化能力は，大規模問題に対して十分な高速性を実現できるほど有効であることが確かめられている．

● S.Whitehead, J.Tenenberg, 1993(2)—複数の目的を持つ対象への強化学習．

今まで述べてきた学習はすべて目的関数が唯一つの報酬関数であらわされていた．しかし現実問題は最終的に複数の目標を達成することが必要なものがほとんどであり，この場合一目的関数のアプローチでは，目標間に優先順位をつける必要がある．しかしどのように優先順位をつけるか，若しくはどの目標から達成すべきかは，それに対する先験的知識が無い場合には問題として扱うことができない．

そこで目標の数に対応した複数個の強化学習エージェントを用意し，これらの出力Q値から，エージェント群全体での真のQ値を見積もることで目標を適応的に選択する方法が提案されている．もっとも簡単な見積はwinner-take-all方式であるが，将来の見積もりをも考慮した，より正確な方法も提案されている．

● L.J.Lin, 1993(3)—リカレントQ-Learningによる長シーケンスの獲得．

Q-Learningは行動の系列を強化する能力を持つが，長い系列を正確に記憶することにはまだ困難がある．そこでQ値をニューラルネットワークに記憶させ，さらにニューラルネットワークをリカレント型にすることで，内部状態をもたせ，系列を積極的に強化させる方法が提案されている．

7．応用例：交通信号機の協調動作の学習

7．1 問題定義

強化学習の応用例として，学習する交通信号機の実現をとりあげる．これは交通流のセンサ情報のみから，地域内の交通量を最大にさせるような現示の制御法を獲得させる問題である．

このような問題は実世界の協調動作学習に多く見られ，最適化問題として次のような難しい問題を含んでいる．

● （エージェント間に頻繁な通信はおこなえない．）

リアルタイム意志決定を必要とする問題，また不特定多数のエージェントによる協調問題で，例えば一行動単位で，すべてのエージェント間での通信による頻繁な情報交換をおこなうことは現実的ではない．たとえば広い地域での信号機からの情報を統合することは，通信容量の点から困難である．

● （環境がダイナミクスを有する）

ある時刻での行動の影響がダイナミクスにより時間遅れを伴って作用する．したがって次のエージェントの行動サイクルで観測する環境の状態は数サイクル前の行動を反映したものであり，通常の反射行動の学習ではなく，時系列学習が必要になる．この問題では，交通流のもつダイナミクス(車両の加減速，車両間隔などによる)による現示の影響の遅れがこれに相当する．

● （エージェントの行動が環境のダイナミクスを変化させる）

現示が交通流のダイナミクスに影響を与える．この場合，例えば迷路ナビゲーション問題などにみられるような，直面している状況を条件部とするルールを駆動させるプロトコルでは，協調動作は実現できない．

● （隠れ状態（Hidden States）問題が生じる）

エージェントが適切に反応するためには，ある２つの状態でエージェントの行動を切り替える必要があるが，エージェントのセンサではこれら状態を直接に識別できないばあいが現実問題にはしばしば見られる．これは隠れ状態問題と呼ばれている．隠れ状態の識別には状態の系列を観測する必用がある．交通信号問題では，前述のダイナミクスを有する点，また，これがエージェントの行動により変化する点で，現示の系列の学習が必要となる．

● （個々のエージェントの目的関数最大化は大域的目的関数の最大化とはならない．）

ある交差点で交通量を最大化することが，他の交差点での渋滞を容易に引き起こしうることは我々が日常体験することである．系列学習を用いれば，うまく地域全体の交通量を増すような状況に遭遇する場合に，その状況(状態と行動の系列)を獲得することにより以上の問題は解決できる．しかし，大域的評価関数を持たない場合には，全エージェントが同様な目的関数（自らの交差点での交通量最大化）を持つ以上，系列学習で獲得すべき状況に偶然遭遇することは少ない．したがって積極的なエージェントの役割分担が形成される必要がある．

7．2 強化学習による交通信号制御

Q-Learningによる強化学習を用いた交通信号制御の実現法は次のようになる．簡単化のため東西南北に直交する地域を想定し，現示は赤（０）と緑（青）（１）信号の２種類とする．各交差点に対して一つのコントローラ(エージェント)が用意されており，それは交

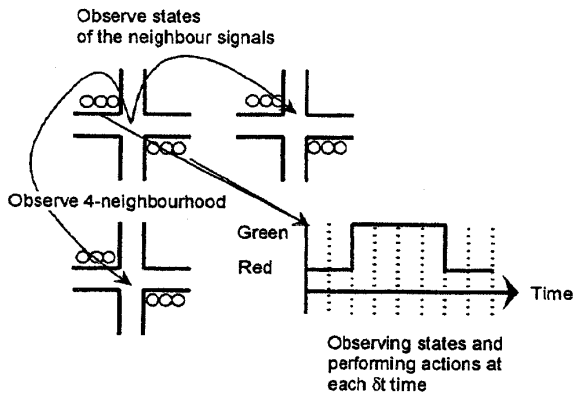

図4 学習する信号機の状態観測と行動

差点の4つの信号を次のような仕様で制御する.

● (状態観測) 各交差点iのコントローラは, N近傍 (例えば東西南北の4近傍) の交差点の信号の (南北方向の) 現示を回線を通じて知ることができるものとし, これにより状態

$$s(t)=\sum_{j=1}^{N} P_{n_j^i} 2^j, \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(8)$$

を得る. ただしn_j^iは交差点iのj番目の近傍の交差点の添え字とし, P_jはjの現示を示す.

● (行動) 交差点コントローラのとりうる行動は, 南北方向の信号の現示を基準としてそれを$a_0=0$ (赤) または$a_1=1$ (緑) の何れに設定するかの2種類とする.

● (基本動作サイクル) 各交差点のコントローラは独自のサイクルタイムTを持つ. サイクルタイム開始時点で, コントローラはその時点で観測されるセンサ情報$S(t)$を入力する.

その後状態$S(t)$にもとづいて政策関数から行動Aを選択し, これを直ちに実行し, 再び次のサイクルでセンサ入力を得る. 全エージェントのオフセット (基準時間から現示の開始までの位相) は0とする. サイクルタイム (現示の一サイクル) の最少単位はTとしている.

● (報酬関数) 交差点にはすべて車両センサが設置され, 交通量が測定できるものとする. 一サイクル終了時に, 交通量の増減から次のように交差点の評価をおこなう.

現在のサイクルをkとする. kでの南北方向の交通量をf_k^iとし, 東西方向の交通量をg_k^iとして, $b_1=f_k^i-f_{k+1}^i$, $b_2=g_k^i-g_{k+1}^i$とする.

(1) $b_1\geqq 0$ または $b_2\geqq 0$ なる場合に, 50%の確率で評価b=1を与える. 強化学習は一般に, "報酬信号"をえる機会がスパースである場合に, その収束が著しく遅れる問題点が指摘されている(1). この50%の報酬信号は, 目的関数の最大値周辺に低い丘を用意することで, スパースな探索にヒューリスティックを与える役割を果たしている.

(2) $b_1\geqq 0$ かつ $b_2\geqq 0$ なる場合には常に報酬信号b=1を与える. 1, 2以外の場合にはb=0を与える.

以上の準備のもとに, 局所的な交通流の最適化は, bを報酬関数としてQ-Learningを適用することにより行われる. Q-Learningによる系列の獲得能力のため, 交通流の持つダイナミクスにかかわらず, 最適行動の獲得が期待できる.

7. 3 世代学習による協調動作獲得

Q-Learningによる学習は, 個々のエージェント (信号機) の目的関数の最大化のみである. すなわちエージェント間の協調動作に関しては, 協調を行うことを指示する目標関数が与えられていない. したがってQ-Learningのみでは, 個々の報酬関数の和がエージェント群としての報酬関数である場合以外は, 明示的な協調動作は獲得させることができない.

エージェントに明示的な協調の指示を与えるためには，エージェント全体の目標関数を最適化させればよい，

大域的最適化は全信号機にたいしてその変更がルールの適用に影響を及ぼすようなパラメータを最適化変数とすることで実現できる．ここではパラメータとして個々の信号機のサイクルタイム（各信号コントローラが意志決定をおこなうサイクル）を選ぶこととする．

a．信号機パラメータの遺伝子表現

パラメータの候補をたとえばm秒からm'秒までに限定すれば，すべての交差点にたいして一本のリスト $Q(j)=\{\forall iU_i(j)\}$ が得られる．これをj番目の遺伝子 $Q(j)$ とする．

b．分散・協調レベルの設定

中央に一台のコントローラがあり，全信号機と通信回線によりつながっていることとする．移動ロボットに代表される一般のロボットの問題では，各ロボットが非同期に中央のコントローラに通信をおこなう機能を仮定する．各ロボットまたは信号は，パラメータ $Q(j)$ を独自に要求し，ある一定の期間（T時間）このパラメータを用いて試行をおこなう．試行終了後，独自に中央に自らの評価関数の結果を通信回線により送信する．したがって，ここで最適化する大域的関数の形は各エージェント B_i の評価関数の関数 $F(G_i(Q(j))$ に限定される．信号問題では各交差点の交通量のT時間の総量の総和とする方法が考えられる．

もし試行期間Tを無限大時間にとればこれは完全な分散システムとなり，T＝各エージェントのサイクルタイムとすれば，これは完全な中央コントロールシステムとなる．したがってTとサイクルタイムとの比率を調整することは分散，協調間のレベルを設定することにほかならない．通信にかかわるコストはエージェントの個数に比例するため，大規模エージェント群ではTの比率を小さくすることは困難である．この方法で階層を導入する対象は以上の2点から限られたものとなるが，たとえば例示した交通信号制御問題など適する問題は多くあげられよう．

c．遺伝的アルゴリズムによるパラメータ探索

協調プランの獲得すなわち適合度を最大にする $Q(j)$ の探索は，試行を通じて広い探索空間を効率よく探す手法が必要となる．一般的な遺伝的アルゴリズムがこの目的に利用できる．具体的にはすべてのエージェントからT時間後の評価値が到着し，すべての $Q(j)$ の適合度が得られたならば，遺伝オペレータにより次世代の $\{Q(j)\}$ を用意し，T時間試行ごとに順次再びエージェントに送信するサイクルを構成する．詳細は省略する．

7．4 シミュレーションによる事例検討

a．シミュレーションの設定

簡単な交通シミュレーションを通じて，提案した個体学習・世代学習の効果を確認した．正確な交通シミュレーションの構築は一般に難しい問題とされているが[6]，ここでは未知環境を想定していることから以下に述べるような簡単なシミュレーションを用いることにする．ここでは行動選択に影響を強く与えるパラメータの一つである各交差点コントローラのサイクル T_i を対象として，遺伝的アルゴリズムによる世代学習の効果を確かめることを焦点においたシミュレーションを試みている．

シミュレーションパラメータを表1に示す．

地区形状は3×3の格子状であるとする．地区の境界はすべて外の地区と結合されているとし，車両はポアソン到着にしたがって境界部から地区に到着する．地区の境界に到着

表1 シミュレーションパラメータ

Number of intersections	3×3=9
Number of signals	4×9=36
Distance between signals	100m
Average speed of a car	30km/h
Accereration time	4 sec
Maximum number of cars	10, 30, or 50
Mean inteval of arrival of a car	2 sec until the number of cars saturates
Cycle time for signal controller	5 sec to 12 sec for learning and random controller, 10 sec to 24 sec for regular controller
Offset	0 sec
Split	Determined by controllers
Learning coefficients	$\alpha=0.2, \beta=0.01$
One trial for Genetic Search	40min
Population	6
Mutation ratio	0.01
Crossover ratio	0.5

した車両はシミュレーションから取り除かれるものとする.

各車両は交差点に到着した場合，交差点にあらかじめ割り当てられた確率にしたがって，その進路を確率的に変えるものとする．すなわち車両の次の進路を $\{d_i\}=\{$Right, Forward, Left$\}$として，交差点 i に東西南北のいずれかの方向 $k \in \{$N, E, S, W$\}$ から到着した車両には，確率 P_j^k により d_i の一つを選択するものとする．各交差点にはこの確率 P_j^k が固定されており，シミュレーション開始時にランダムに設定される.

信号待ちから進行可になった時点で，4秒後に定常速度30 km/h で進行を開始するように設定する．これにより加速の簡単なシミュレーションを導入する.

b．比較対象と結果

比較のため次の2種類のシミュレーションを用意した.

● (定時制御)各サイクルで現示を交互に変えるようなコントローラのみを用意する．この場合，サイクルはつぎのシミュレーションの倍の時間に設定する.

● (個体学習導入)ランダムに定めたサイクル時間により，強化学習により現示を最適化するコントローラから構成する.

シミュレーションではこれらに世代学習を導入して，サイクルタイムを遺伝的アルゴリズムにより最適化させたときの，世代の変化に対する地域の総交通量の変化を調べている．世代を重ねるにしたがって総交通量が増加するばあい，世代学習による協調動作の効果が見られることを意味するになる．さまざまな交通の混雑さに対して比較をおこなうため，地域内の許容最大車両数を 10, 30, 50 の3段階に変化させてシミュレーションをおこなった．結果を図5から図7に示す.

スパースな交通流以外では，個体学習をおこなう場合が定時制御に対してより高い交通量を獲得できることがわかる．交通流が少ないときには世代学習による総交通量の改善はあまり見られないが，混雑した状況では，世代の変化により高い総交通量をあたえるパラメータへと収束に向かい，ほぼ8世代目（シ

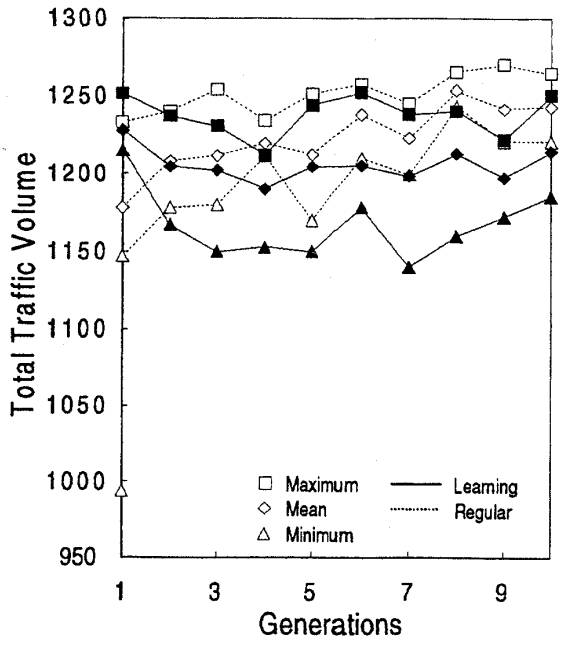

図5 スパースな環境でのシミュレーションの結果．区域内には最大10台の車両が同時に存在しうる.

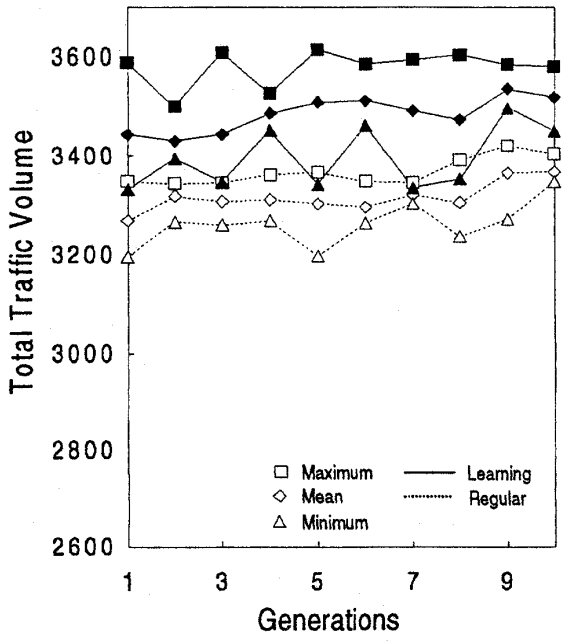

図6 中程度の交通量でのシミュレーション結果. 最大30台の車両が区域内に存在する.

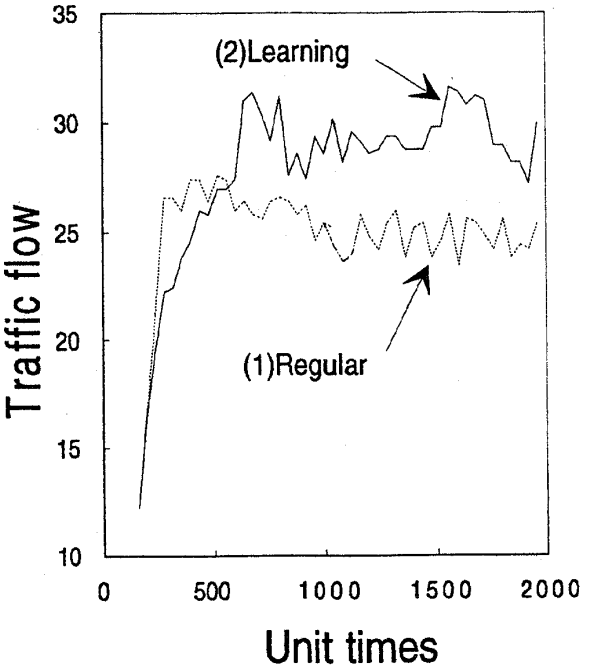

図8 一回の試行でのパフォーマンスの変化. 遺伝的アルゴリズム適用前でのパラメータ集団を用いた.

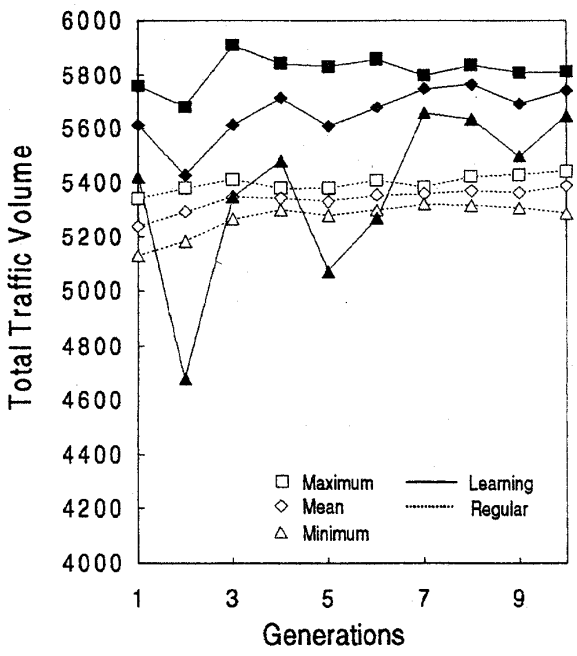

図7 混雑した交通量でのシミュレーション結果. 50台が区域内に許される.

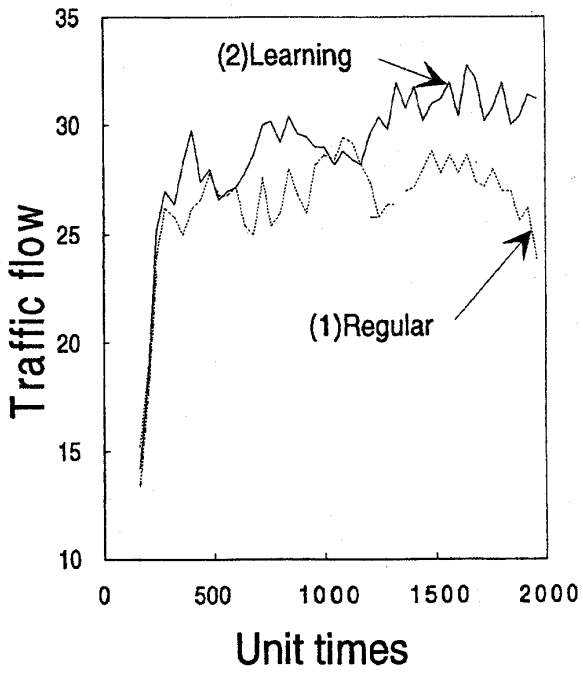

図9 一回の試行でのパフォーマンスの変化. 遺伝的アルゴリズムにより得られた最良のパラメータ集合を用いた.

ミュレーション上の時間で約32時間後)で収束していることがわかる. 混雑した状況が, より協調を必要とする状況であることを考えれば, これは世代学習が協調を獲得するのに有利であることを明示している.

図8と図9は，0世代目で最大評価値を与えたパラメータと，全世代を通して最大評価値を与えたパラメータを用いて，世代学習時と異なる乱数パラメータを用いた許容最大車両数50のシミュレーション下で，総交通量のサイクル時間の単位での変化をプロットしたものである．図から世代学習を経たエージェントがより早くから高い評価値を獲得していることが確認できる．

8．結　論

生命の持つ「賞罰のみからの学習」のメカニズムを工学的手法に実現しようとする手法として，強化学習法を紹介した．これが「実世界に対して自律的に適応する機械」の実現への足掛かりとなり得ることを示した．強化学習法の拡張として，生物が群として生存してきたことをシミュレートすることで，複数のエージェントによる協調動作の学習による獲得が可能であることを示した．例題として交通信号網の学習制御問題を取り上げ，シミュレーションにより有効性を確認できた．

参考文献

(1) Sutton, R.S.: *Reinforcement Learning*, 1, Kluer Academic Pub., (1993).
(2) Connell, J.H. and Mahadevan, S.: *Robot Learning*, 1, Kluer Academic Pub., (1993).
(3) Watkins, C.J.C.H. and Dayan, P.: Technical note: Q-learning, Machine Learning, Vol. 8 p.279, (1992).
(4) Lin, L.J., Mitchell, T.M., Reinforcement Learning with Hidden States, From animals to animats 2, 271, The MIT Press, (1993).
(5) Sutton, R.S., Learning to Predict by the Method of Temporal Differences, Machine Learning, Vol.3, p.9, (1988).
(6) Holland, J.H.: *Adaptation in Natural and Artificial Systems*, p.171, The MIT Press, (1992).
(7) Khisty, C.G.: *Transportation Engineering, An Introduction*, Prentice-Hall, (1990).
(8) （例えば）マルチエージェントロボットシステム特集，日本ロボット学会誌，Vol.10, No. 4号，p. 1，(1992)．