時間軸スムージング学習

非会員 柴田 克成 (東京工業大学)

正員 岡部 洋一 (東京大学)

Temporal Smoothing Learning

Katsunari Shibata, Non-member (Tokyo Institute of Technology), Yoichi Okabe, Member (University of Tokyo)

In order to realize the mapping from spatial information to temporal information, Temporal Smoothing (TS) Learning is proposed. In this learning, the output of a learning unit, to which sensory signals are given as input, is trained to be smooth along time. In other words, the learning unit is trained so as that the second time derivative of the output itself becomes 0.

This learning can be applied to integrate local sensory signals into an analog spatial signal. It also can be used that an agent learn evaluation function in delayed reinforcement learning on behalf of TD Learning⁽⁶⁾⁽⁷⁾ when only one target state is chosen. When a neural network was employed as a learning unit and visual signals were given as inputs directly, the hidden neurons in the neural network represented spatial information and had a adaptability of changing the representation according to the agent's motion characteristics.

キーワード：時間軸スムージング学習, 強化学習, センサ信号統合, 空間-時間マッピング, 自律学習

1. はじめに

ここ数年, 自律学習やエージェントという言葉が良く使われるようになり, そこから生まれる機能の創発に注目が集まっている。自律学習においては, 与える知識に対する得られる機能の比が大きいことが得られた機能の柔軟性, 創発という点から重要である。このような観点から, エージェントにおいて用いる学習は, できるだけ汎用性が高いことが望まれる。つまり, 我々の存在する世界の普遍的特性を利用した学習を行うことが有効であると考ええる。

我々生物やエージェントは, センサから空間的な情報を手にいれ, 未来という時間に向かって適切な動作をしていかなければならない。従って, センサからの空間的な情報から, 現在の状態が時間という軸の中でどのような位置づけにあるかを知ることは, 次の状態の予測, また, エージェントがどのような行動をとるかを決定する上で重要である。

本論文では, 筆者らが, 局所センサ信号の統合⁽¹⁾⁽²⁾や強化学習における評価関数の学習⁽³⁾⁽⁴⁾に用いた時間軸スムージング学習 (Temporal Smoothing Learning) を, 空間から時間へのマッピングという観点から見直すとともに, 上記2つの応用の概要と新しい実験結果について述べ, この学習がエージェントにおける機能の創発に重要な役割を果たす汎用的な学習であることを示す。

2. 空間情報から時間へのマッピングの学習

もし, 我々の存在する世界の状態が微分方程式で表される力学系であり, 決定論的に変化すると仮定すると, その状態変化の軌跡は枝分かれすることがない。つまり, この軌跡は, 図1のように, 時間という一次元の軸の上に投射することができる。このことは, もしセンサからの信号が, 必要な状態を表すのに十分であると仮定すれば, センサの信号から現在の状態が時間的にどのような位置づけである

図1 センサから得られる空間情報から時間情報への変換

Fig. 1. Transformation from spatial information as sensory signals to temporal information

かを知ることが可能であるということを意味している。

エージェントの行動の学習則である強化学習では、通常、強化信号（報酬や罰） $r(t)$ の将来にわたる重み付け総和

$$\bar{r}(t) = \sum_{k=0}^{\infty} \gamma^k r(t+k) \dots\dots\dots (1)$$

ここで、 γ ：割引率 ($0 < \gamma \leq 1$)

を最大化させるという問題の定式化をしている。これは、さしあたって目指す報酬（目標）が1つに絞られているとすると、その報酬獲得までの時間を最短化することになる（第5章参照）。ここで、センサ信号から時間の情報へのマッピングが形成されていたとすれば、自分の動作を変化させることによって、その時点でより未来の状態と認識されている状態に到達できることが望ましい。そして、この動作を学習した後、学習した動作系列にあわせて再びセンサ信号から時間へのマッピングを学習させる。これを繰り返していくことによって、エージェントに柔軟性、適応性を持たせつつ、経路の最適化が進められる。以上のように、空間の情報から時間へのマッピングを学習によって形成することはエージェントにとって重要な機能である。そして、これを実現する学習が時間軸スムージング学習である。

3. 学習アルゴリズム

センサから得られる空間情報と時間情報の対応付けを行うために、図2のように、センサ信号を入力とし、時間情報を出力する学習装置を考える。ここで、時間の情報を出力するとは、出力から、現在考慮している系列の中での入力した状態の時間的な位置づけがわかることであるとする。そのためには、出力の時間変化が滑らかに単調変化していることが必要である。また、出力 $x(t)$ の時間変化量 $dx(t)/dt$ を時刻 t によらず一定にするとすれば、学習装置の出力が時間の情報を線形変換した形、つまり、横軸を時間に縦軸を学習装置の出力とした場合に、その関係が図2中の点線のように直線で表されることが理想的な解となる。この時、時間軸に対する出力曲線の滑らかさは時間による2階微分値の自乗によって表わすことができる。そこで、出力曲線の2階微分値を0に近づけ、曲線の凹凸をなくし、直線に近づけるために、教師信号 $x_s(t)$ を

$$E(t) = \frac{\kappa}{2} \left(\frac{d^2 x(t)}{dt^2} \right)^2 \dots\dots\dots (2)$$

$$x_s(t) = x(t) + \kappa \frac{d^2 x(t)}{dt^2} \dots\dots\dots (3)$$

図2 時間軸スムージング学習
Fig. 2. Temporal Smoothing (TS) Learning

ここで、 $E(t)$ ：出力の誤差、 κ ：学習係数 (> 0)

とし、学習装置に対し教師あり学習を行う。ここで、 κ を0.5とすれば、教師信号は、差分近似により、

$$x_s(t-1) = \frac{x(t-2) + x(t)}{2} \dots\dots\dots (4)$$

と表わすことができる。つまり、2単位時間前の出力と現在の出力から1単位時間前の出力を学習する。この学習は、原則として毎単位時間行うが、その典型的な教師信号を図2中の矢印として示す（縦軸方向拡大）。この学習により、出力曲線は徐々に直線に近づき、時間の情報を表現できるようになる。ただし、直線が傾きを持つための学習を別途付加しなければならない。また、無限の時間を表現することはあまり意味がないため、現在着目している、または現在行おうとしている行動系列等の限定した時間の中で位置づけを行うべきものであると考える。これに関しては、具体的な適用例のところで示す。

ここでは、学習装置として階層型のニューラルネットを採用し、教師あり学習のアルゴリズムとして、バックプロパゲーション法⁽⁵⁾を用いた。また、通常の教師あり学習のように、オフラインで、予め用意したいくつかの入力と教師信号のペアで繰り返し学習するのではなく、オンラインで、その時に求めた教師信号で1回だけ小さい学習係数で学習させる。これによって、オンライン学習を実現し、適応性を持たせることができると共に、常に学習を行うことで、限られたデータにオーバーフィッティングし、汎化能力を失う過学習の問題⁽⁶⁾を防ぐことができる。

4. 局所センサ信号の統合の学習への適用

本章では、時間軸スムージング学習によって、局所的な受容野しか持たないセンサからの信号を統合し、連続値としての空間の情報を抽出できることを示す。

〈4・1〉空間情報の抽出の原理 我々の持つ視覚を始めとするセンサの多くは、局所的な受容野を持つたくさんのセンサセルにより構成され、広い範囲の情報を獲得できるようになっている。この時、この局所センサ信号から大域的な情報を知ることを考える。時間軸スムージング学習は、空間的な広がりを持つセンサ信号から時間へのマッピングを行うものであるが、空間の情報は時間的に滑らかにしか変化しないと考えると、エージェントが動作をしない場合には、時間的に滑らかに変化している学習装置の出力は、その時に変化している空間情報を抽出していることになる。

まず、図3のように、動く物体、視覚センサと階層型ニューラルネットからなるシステムを考える。視覚センサは、局所的な受容野を持つセンサセルを1次元に配列し、各センサセルは、受容野中に占める投射された物体の面積の割合を0から1の間の値として出力するものとする。ここで、物体が視野の中を左右に滑らかに動いているものとする。すると、各センサセルは局所的な受容野しか持たないため、そ

の出力は、中間値をあまりとらず、0 か 1 の離散値をとる場合が多い。ここでは、ニューラルネットの出力は入力信号だけの関数であり、かつ入力信号の中で物体の位置のみが変化している。従って、時間軸スムージング学習によって時間軸に対して滑らかに変化するように学習された出力は物体の位置の変化と共に滑らかに変化していく。そして、物体の位置と出力の間に1対1の関係が得られるようになる。

ただし、出力の値域を確保するために、値域拡大学習をあわせて行う。この学習は、1次遅れによって計算される過去の出力の平均に対する現在の出力値の偏差が大きい時にのみ適用される。ここでは、出力関数として0から1の値域を持つシグモイド関数を用いているため、偏差が正の時には0.9、負の時には0.1を教師信号として学習させる。

〈4・2〉 シミュレーション 図3に示した環境で、シミュレーションを行った。視覚センサセルの数は30個とし、ニューラルネットの中間層ニューロン数も30個とした。ニューラルネットの各結合の重み値の初期値は、-0.1から0.1の中から乱数によって決定した。従って、学習前の出力値は入力によらずほぼフラットであるが、多少の凹凸がある。物体は視野内を左右に単振動しているものとし、振動の周期は93.1単位時間とした。時間軸スムージング学習の教師信号は式(4)とし、値域拡大学習は約1000単位時間に一回の割合で適用された。学習前と学習後の物体の位置に対する出力の値を図4に示す。ただし、図4(a)は、学習前の出力曲線の凹凸がわかるように縦軸を100倍拡大してある。学習前には、凹凸のあった出力曲線が、学習後には、図4(b)のように、物体の位置と出力の間が滑らかにかつ1対1の関係が成り立つようになっていることがわかる。

次に、出力の分布を調べるため、物体が前後に単振動している場合と物体の大きさが単振動している場合についてシミュレーションを行った。いずれの場合も、視覚センサ上に投射された物体の大きさが変化するが、前者の場合、物体がセンサに近いほど投射された物体の大きさの変化が大きくなり、後者では、視覚センサ上の物体の大きさも単振動である。ただし、いずれの場合も、視覚センサ上で物体の大きさが変化する範囲は1から4となるように設定した。

図3 視覚センサにおける局所センサ信号を統合の学習を行うシステムの構成

Fig.3. Learning system for integration of local sensory signals from a visual sensor

図4 物体が視野内を左右に単振動している場合の物体の位置と出力の関係の学習による変化 (視覚センサセル30個)

Fig.4. Transition of the relation between the object location and the output when an object is simply oscillating in the visual field (Number of visual sensory cells is 30)

学習後の出力の様子を図5に示す。左の図(a)が物体が前後に単振動した場合、右の図(b)が物体の大きさ自体が単振動した場合である。また、左の図では、横軸に物体の前後の位置をとった場合と視覚センサ上の物体の大きさをとった場合を示した。2つのグラフを比較すると、視覚センサ上の物体の大きさを横軸とした場合、物体の大きさが単振動している場合はほぼ点対称な曲線であり、大きさが中間の2.5の時に出力もほぼ中間の値0.5になっている。一方、前後に単振動している場合は大きさが2.5の時に出力が約0.1となっており、横軸を同じにしても出力の分布が異なることがわかる。また、前者を物体の位置に対してプロットしたものはその位置が中間の2.5の時に出力が約0.6と中間に近い値をとっていることがわかる。これは、時間軸スムージング学習により、 dx/dt の値を一定値に近づけようという学習が働くため、物体の存在時間密度が高いところは拡大して表現されているためと考えられる。

図5 物体が視野内を前後に単振動している場合と物体の大きさが単振動している場合の出力分布の比較

Fig.5. Comparison of output distribution when the object is simply oscillating forth and back, and when the object size is simply oscillating

5. 強化学習への適用

〈5・1〉 目的達成までの所要時間による評価と時間軸スムージング学習 強化学習は、第2章で述べたように、目的とする報酬を一つに絞れば、その報酬を獲得する所要時間を最小化するという問題に置き換えることができる。そこで、時間軸スムージング学習を適用し、目的達成までの所要時間を予測するように学習し、これを評価値として用いることによって動作の学習を行う。

しかし、前述のように、時間軸スムージング学習では、出力曲線を直線に近づけることはできるが、その傾きを規定することはできない。また、各試行での所要時間を平等に評価するためには、評価値の傾きが常に一定でなければならない。そこで、時間軸スムージング学習を拡張した評価値の時間変化量一定化学習を用いる。この学習では、時刻 $t-1$ の評価値 $p(t-1)$ の教師信号 $p_s(t-1)$ を

$$p_s(t-1) = p(t) - \frac{P_{max} - P_{min}}{N_{max}} \dots\dots\dots (5)$$

ここで、 P_{max} , P_{min} : 評価値の理想値域の最大値、最小値、ここでは、 $P_{max} = 0.4$, $P_{min} = -0.4$,

と時刻 t での評価値 $p(t)$ を用いて表わし、この式の中で使われている目的達成までの最大所要時間 N_{max} を、

$$N_{max}(t) = \begin{cases} N(t) & \text{if } N(t) > \lambda N_{max}(t-1) \\ \lambda N_{max}(t-1) & \text{otherwise} \end{cases} \quad (6)$$

ここで、 $\lambda = 1 - 1/\tau$, τ : 大きい時定数

とすることで、問題の複雑さと学習の進行状況に応じて評価値の時間変化量を適応的に変化させることができるようにした。そして、エージェントが目的を達成した時に教師信号を 0.4 として学習させる。ただし、ここでは、ニューラルネットの出力関数は -0.5 から 0.5 のシグモイド関数とした。この学習によって、単位時間当たりの評価値の時間変化量は $(P_{max} - P_{min})/N_{max}$ に近づく。一方、動作は、動作出力ベクトル $\mathbf{m}$ に試行錯誤のための微少な乱数ベクトル $\mathbf{rnd}$ を加えたものに従って行う。そして、それによって変化した所要時間の予測値 p の値から、動作信号に対する教師信号 $\mathbf{m}_s$ を

$$\begin{aligned} \mathbf{m}_s(t-1) \\ = \mathbf{m}(t-1) + \mathbf{rnd}(t-1)\{p(t) - p(t-1)\} \end{aligned} \quad (7)$$

として学習を行うことにより、評価値の時間変化量が最大となる方向への動作を学習していく。

図6に、本論文で仮定したロボット（エージェント）の構成を示す。ロボットは、センサ信号を受取り、その値を階層型のニューラルネットに入力する。ニューラルネットには、動作出力と評価出力の2種類の出力を設け、ベクトルである動作出力 $\mathbf{m}$ に従ってアクチュエータを動かす。一方、

図6 強化学習を行うロボットの構成

Fig.6. Structure of a robot that performs reinforcement learning

評価出力 p はスカラーであり、この値から式(7)のように動作出力の学習を行うと共に、式(5)に従って評価出力自身の学習も行う。この評価と動作の学習を並列に行うことにより、ロボットの経路が最短時間経路に近づいていく。

強化学習の有名なアルゴリズムの一つに、BartoらのTD学習⁽⁷⁾を用いた方法がある⁽⁸⁾。この方法では、学習装置はセンサ信号から現在の状態の評価値 p として式(1)で示した $\bar{r}$ を予測するように学習する。これによって、学習後の p は、強化信号が得られるまで時間と共に指数関数的に増大するようになる。

この学習において、当面の目標（報酬源）が1つに定まっているとすると、式(1)は

$$\bar{r} = \gamma^T r(T) \dots\dots\dots (8)$$

ここで、 T : 目的達成までの所要時間

と書くことができる。この時、 $\bar{r}$ は、目的達成までの所要時間 T と1対1の関係にあることから、目的達成までの所要時間を表わしていることになる。これを時間軸スムージング学習を用いた場合と比較すると、図7のように、TD学習では、時間を指数曲線を使って表現し、時間軸スムージング学習では、直線を使って時間を表現しているという点が違うことになる。また、別の見方をすると、TD学習

図7 強化学習の評価値の学習にTD学習を用いた場合と時間軸スムージング学習を用いた場合の評価曲線の比較

Fig.7. Comparison of evaluation curve when TD learning is employed and when Temporal Smoothing Learning is employed for the learning of evaluation value on reinforcement learning

における式 (1) の γ を 1 とし、かつ毎単位時間一定な微少なペナルティを受け取っているとすると等価になる。つまり、毎単位時間ペナルティを与えることで、時間の経過に対してペナルティを与えていることになる。

〈5.2〉 移動ロボットによるシミュレーション ここでは、視覚センサ付き移動ロボットが目標物を捕らえるという問題を取り上げた。従来、視覚センサ信号を強化学習に用いる場合には、視覚センサ信号を予め人間が設定した方法で離散的な状態に分け、その各状態に対して動作を学習させていた⁽⁹⁾。そして、その状態分けをいかに自律的に行わせるかが問題となっていた。ここでは、視覚センサ信号を直接ニューラルネットに入力して学習させることが可能であることを示す。

図8に視覚センサ付きの移動ロボットを用いたシミュレーションの環境を示す。このロボットは、両脇の車輪を独立に制御して回転させることによって前後および回転運動ができ、その車輪の上に2つの視覚センサを持つ。各視覚センサは、24個のセンサセルよりなり、各センサセルはオーバーラップのない放射状の局所的な受容野を持ち、全体として180度の視野を有する。そして、センサからの信号計48個をニューラルネットへの入力とし、評価の出力1個、動作の出力を2個の合計3個の出力を用意し、2つの動作出力に従って両脇の車輪を動かした。また、ニューラルネットは3層とし、中間層ニューロンは20個とした。目標物は図8中の平面内にランダムに置き、目標物の中心がロボットを通り抜けた時をロボットが目標物を獲得したものとし、報酬として0.4の教師信号で評価値を学習し、それ以外で目標物が視野から消えた、つまり、自分より後ろにいつてしまった場合は罰として、-0.4の教師信号によって学習を行った。そして、目標物に到達するかまたは見失った場合を1試行とし、再びロボットを元の位置に戻し、目標物を乱数によって決定された位置に置いて動作と学習を開始した。ただし、学習の初期には、目標物の初期位置を

図8 視覚センサ信号を入力としたロボットを用いたシミュレーションの環境
Fig.8. Simulation environment where a robot with visual sensors is used

ロボットに近い領域から選択し、学習の進行と共に徐々にその範囲を広げていった。

学習による評価関数およびロボットの経路を図9に示す。この図は、ロボット中心の座標で描いてあり、図中の X' 、 Y' はロボットから見た目標物の横方向および前方向の相対位置を表わす。そして、目標物の位置による評価の値を等高線で示してある。また、5ヶ所に目標物を置いた場合のそれぞれのロボットの経路（図中では、ロボット中心座標のため相対的に目標物が動くように描かれている）を示している。(a) が700 試行後、(b) が100000 試行後の様子である。ニューラルネット内の中間層から出力層の間のすべての結合の重み値の初期値を0としたため、試行を全く積んでいない場合には、評価関数曲面は完全にフラットであり、また、ロボットも乱数による動作しかできない。しかし、700 試行後には、評価関数曲面に凹凸ができ、ロボットは、確認した5ヶ所についてすべて目標物に到達することができるようになった。しかし、評価関数曲面には、放射状の尾根や谷ができていていることがわかる。これは、ロボットが経験を十分に積んでいないため、視覚センサセルの局所的な受容野が放射状に広がっていることの影響が出たためと考えられる。しかし、100000 試行後には、評価関数曲面は滑

図9 目標物の位置に対する評価関数の形状とロボットの経路の学習による変化
Fig.9. Transition of the evaluation surface and the locus of the robot through the learning

図 10 強化学習による中間層での情報のコーディングの変化をテストするためのニューラルネット
Fig.10. A neural network to examine the change of information coding in hidden neurons through reinforcement learning

らかになり、ロボットが目標物に到達する時間も短くなった。このように、時間軸スムージング学習に基づく強化学習によって、局所的な受容野しか持たない視覚センサセルの信号を統合し、滑らかな評価関数を形成し、かつ、ロボットの経路が最適に近いものとなっていることがわかった。

〈5・3〉 中間層ニューロンによる物体位置の表現 ここでは、時間軸スムージング学習に基づく強化学習を行った後の中間層ニューロンが、どのように空間の情報を表現しているかを示す。

まず、前節のように強化学習を行った後に、図 10 のように、ニューラルネットの出力層にニューロンを 1 個追加し、中間層ニューロンからそのニューロンとの結合の重み値を全て 0 にセットし、これに対して、教師あり学習を適用して、その出力の分布を調べた。具体的には、ロボットからの相対座標 (X', Y') が $(-5.0, 0.0)$, $(-5.0, 3.5)$, $(-5.0, 7.0)$ と $(5.0, 0.0)$, $(5.0, 3.5)$, $(5.0, 7.0)$ の計 6 か所に物体を置いた時のセンサ信号を入力し、教師信号を前半の 3 つに対して 0.4、後半の 3 つに対して -0.4 を与えバックプロパゲーション法によって教師あり学習を行った。この時、強化学習に用いた 3 つの出力層ニューロンは学習をさせなかった。また、ここでは、入力層と中間層の間は学習させたが、学習させない場合もほぼ同じ結果となった。

この学習後の物体の位置に対する学習させた出力の値の分布を図 11(a) に示す。また、比較のために、ニューラルネットの重み値の初期値を同じとし、強化学習を行わないで教師あり学習だけを行った場合の出力の分布を図 11(b) に示す。強化学習を行った方の出力分布は、 X' の値の大きさをきれいに分類するようになっているのに対し、強化学習を行っていない方は、出力の分布が放射状に広がっていることがわかる。このことから、強化学習を適用することによって、ニューラルネットの中間層に、物体の X' を区別するような成分ができあがっていたことがわかる。そして、この情報が、他の学習にも利用できる可能性を示している。

次に、実際に、中間層がどのようなコーディングをしてい

図 11 強化学習による中間層での物体位置のコーディングの変化
Fig.11. The change of coding of the object location in hidden neurons through reinforcement learning

るかを調べた。恒等写像(秒時計型)ニューラルネットワーク⁽⁹⁾を参考に、図 12 のような真ん中の中間層のニューロン数を 2 つにしよった 5 層のニューラルネットワークを用意した。そして、強化学習を行い、2 つの中間層ニューロンが物体の位置をどのようにコーディングしているかを調べた。各層のニューロン数は、入力側より、48-20-2-10-3 とした。

図 13 に、中間層の 2 つのニューロンの値を縦軸と横軸にとり、図 8 中の格子に物体を置いた時の中間層の値をプロットし、格子を作成したものを示す。学習前はニューラルネット内の結合の重み値が微少な乱数で決定されているため、図 13(a) のように、2 つの中間層のニューロンとも真ん中の 0.0 近辺の値をとっている。しかし、図 13(b)(c) のように学習が進むにつれ、徐々に使用している中間層の値の領域が大きくなり、100000 試行後には、図 13(d) のように、中間層ニューロンのほぼ全域を使って物体の位置を表現していることがわかる。また、図中の数字付の○印は、図 8 中の○印に物体を置いた場合に相当する。これより、物体が車輪の近くに存在する時の状態が拡大して表現されていることがわかる。これは、この状態が、物体を取り込むか取り込まないかの境界となっており、ロボットにとって重要な(状態の分割を必要としている)領域であることを反映していると考えられる。また、物体の置かれた長方形の領域が、2 つの中間層ニューロンの軸で構成される正方形にそのまま収まるのではなく、45 度回転した形で物体の位

図 12 強化学習適用後の中間層での物体の位置のコーディングを調べるためのニューラルネット
Fig.12. A neural network to examine the coding of the object location in hidden neurons when reinforcement learning was applied

図 13 強化学習を行った際の中間層での物体の位置のコーディング
Fig.13. The coding of the object location in hidden neurons when reinforcement learning was performed

置が表現されている。これは、上記のような状態の再構成の結果、物体の置かれている領域を 45 度回転して表現することが最もバランスが良かったためと考えられる。

次に、図 14 のように、ロボットの右の車輪を動作信号（動作ニューラルネットの出力）の 3 倍した値に従って回転させ、左の車輪は 1 倍して回転させるという非対称な動作特性を導入し、対称な場合との比較を行った。

図 15 (a) に、学習後の目標物の位置に対する評価関数の

図 14 ロボットに適用した非対称動作特性
Fig.14. Asymmetrical motion characteristic applied to the robot

図 15 非対称動作特性を持たせた際の、学習後の目標物の位置に対する評価関数形状とロボットの経路
Fig.15. Evaluation surface and locus of the robot when the asymmetrical motion characteristic was applied

値と 5 ケ所に目標物を置いた場合のそれぞれのロボットの経路を示す。この図も、ロボットを固定した座標で描いてあるため、相対的に目標物がロボットに近づいてくるようになっている。この図より、評価関数が左右で非対称性であること、そして、ロボットの経路を見ると、目標物が右側にある場合でも、左前方に見えるように回転してから前進していることがわかる。これは、左側の車輪をあまりたくさん回転させることができないため、回らない車輪をたくさんまわして右前方へ行くよりもいったん回転してから右の車輪をたくさん回して左前方に進んだ方がロボットに

図 16 ロボットに非対称動作特性を持たせた場合の中間層における物体位置のコーディングの変化
Fig.16. The change of the coding of the object location in hidden neurons when the asymmetrical motion characteristics was applied

とってより良いということを学習した結果であると考えられる。また、目標物が近づいてくると、ロボットはそれを右側で捕らえようとしている。これも、右の車輪をたくさん回すことができるため、目標物がロボットの近くにある時には、右側に目標物があつた方が捕らえやすいということを学習した結果であると考えられる。

絶対座標におけるロボットの経路を図 15 (b) に示す。この図から、ロボットは、一見遠回りをしているように見えるが、所要時間を調べたところ、真っ直ぐに進むよりも早く目標物を捕らえることができることがわかった。また、目標物が右の方にある場合、ロボットは、まず回転しようとするが、右側の車輪がたくさん回ることから結果的にロボットが後ずさりをしていることがわかった。このことは、設計者自身が予測しなかったことであるが、強化学習による機能創発の利点を表していると言える。

ここで再び、図 12 のような構造のニューラルネットを適用し、中間層ニューロンのコーディングを調べた。すると、図 16 のように、ロボットの正面を表す線（図中の○3 から伸びている薄く太い線）が大きく曲がりくねっていることがわかる。これを良く観察すると、ロボットの近くではロボットの右側を、遠くでは左側を、つまり、ロボットの経路周辺を拡大して表現していることがわかる。これは、ロボットが右に回転するか、左に回転するか、前進するかの境界部分である。このように、周りの環境は全く同じでも、ロボットの動作特性を変化させるだけで物体の位置を適応的に表現する能力があることがわかる。またこの時、学習があまり安定しなかったため、比較的安定していた時の状態を図に示している。しかし、中間層ニューロン数を増やしたところ学習は安定した。このことから、非対称動作特性を導入した場合、物体の位置の表現が左右非対称となるため、2つの中間層ニューロンの空間である正方形の領域中に物体の位置をバランス良く配置することが難しく、学習が不安定であったと考えられる。

図 17 TD学習に基づく強化学習を行った後の目標物の位置に対する評価関数形状とロボットの経路
Fig.17. Evaluation surface and locus of the robot when the reinforcement learning based on TD learning was applied

6. 議 論

本論文では、時間軸スミージング学習に基づいて、局所センサ信号を統合することを直接学習させる方法と、強化学習によって副次的に統合を実現する方法を示した。前者は、値域拡大学習を適用しなければならない、自分自身は動作することができない等問題点が多く、自らの動作とも密接に関連している後者の方がより自律学習システムを形成する上で利用価値が高いものと考えられる。

また、時間軸スミージング学習では、状態の存在確率密度によってその状態が出力に占める密度が変化するという基本的な性質がわかった。これは、意味付けが非常にわかりやすく、また合理的であると考えられる。

前述の視覚センサを入力とした強化学習は、TD学習を用いても行うことができることを確認した。学習結果を図 17 に示す。この図と図 9 を比較すると、TD学習では、ゴールに近いほど評価関数の値の密度が高くなるという特徴を持つことがわかる。図中に、両学習に基づく強化学習で学習した場合の経路を示した。両者は、多少違いがあるものの比較的近い経路をとっていることがわかる。また、図に示した5箇所に目標物を置いた場合の平均到達時間は、共に193単位時間で全く同じであった。

また、本論文で示した、適応的に評価値の時間傾きの理想値を設定する方法も、単一報酬課題を前提とした場合、TD学習の割引率 γ の設定に応用することが可能である。ゴールへの到達ステップ数が最も多い場合に、評価値の値域全体を使用すると、

$$\gamma^{N_{max}} P_{max} = P_{min} \dots\dots\dots (9)$$

という式が成り立つ。これより、 γ の値を

$$\gamma = \left(\frac{P_{min}}{P_{max}} \right)^{\frac{1}{N_{max}}} \dots\dots\dots (10)$$

と学習の進行と共に適応的に変化させていくことができる。

ここで、TD学習との比較を行う。強化学習という応用面から考えると、単一報酬課題を想定する限りにおいては、前述のように両者とも、目的達成時間を評価していることになり、学習が誤差なく完全に終了した状態では、その経路は同一となると考えられる。ただし、実際には、学習装置の容量が限られている。そのため、(1) 動作は評価値の時間変化量 dp/dt を元に学習を行うが、TD学習では、評価値の時間変化は指数的であるため、目的から遠い状態では dp/dt の値が十分にとれず、学習の進行が鈍る可能性がある。しかし、時間軸スムージング学習ではこの値が一定のため、長い時間の動作系列を必要とするような問題では学習の高速化が期待できる。逆に、学習初期では、目的に近いところでの学習が主となるため、時間軸スムージング学習の方が、学習が遅くなる可能性もある。また、(2) 時間軸スムージング学習では、複数報酬源の場合への対応が困難である。これに関して、筆者らは、自らを振り返り、我々生物も、現時点での目標(報酬源)は、何らかの形で、前もって一つに絞っているのではないかと考えている。

次に、空間情報から時間の情報への変換という観点から考えると、(3) 時間軸スムージング学習では、出力の変化量 Δx と時間の経過 Δt が

$$\Delta x = k \Delta t \quad \dots\dots\dots (11)$$

ただし、 k : 定数

と比例するため、変換が非常に容易であるが、TD学習では、 $\Delta t \rightarrow 0$ としても

$$\frac{dx}{dt} = kx \quad \dots\dots\dots (12)$$

と x が入って来てしまう。よって、空間から時間へのマッピングという大きな視点からは、時間軸スムージング学習の方が合理的である。また、TD学習では出力の時間変化量が一定でないため、第3章で述べた局所センサ信号統合の学習にうまく適用できないなど、強化学習以外での応用の可能性は、時間軸スムージング学習の方が上であると考えられる。しかし、今後、工学的側面、生理学的妥当性など様々な面からのさらなる検討が必要である。

7. ま と め

学習装置の出力の時間変化を滑らかにする時間軸スムージング学習をセンサ信号から得られる空間情報から、その状態の時間的な位置づけを知るための学習アルゴリズムとしての見方を紹介した。そして、この学習を局所センサ信号の統合の学習に適用し、多数の視覚センサセルの信号を統合し、物体の位置を1つのアナログ信号として表現することができることを示した。また、強化学習に適用することによって、目的達成までの所要時間の評価が実現できると共に、多数の視覚センサセルからの信号を直接入力とした場合でも、滑らかな評価関数を形成し、目的達成までの

最適に近い動作系列を学習することができた。また、その際に、学習装置として用いたニューラルネットの中間層ニューロンに空間的な情報がコーディングされ、かつ、その表現が適応的に変化する能力を持っていることを示した。

謝 辞

本学習を用いた強化学習とTD学習を用いた強化学習の比較に関して議論頂きました科学技術振興事業団川人学習動態脳プロジェクトの銅谷賢治氏に深く感謝致します。

(平成8年11月25日受付, 同9年3月21日再受付)

文 献

- (1) Shibata, K. & Okabe, Y.: "Unsupervised Learning Method to Extract Object Locations from Local Visual Signals", *Proc. of ICNN'94*, III, pp. 1556-1559 (1994)
- (2) 柴田克成, 岡部洋一: "時間軸スムージング学習を用いた局所センサ信号の統合と空間情報の抽出", 日本神経回路学会誌, 3, 3, pp. 98-105 (1996)
- (3) Shibata, K. & Okabe, Y.: "A Robot that Learns an Evaluation Function for Acquiring of Appropriate Motions", *Proc. of WCNN'94*, II, pp. 29-34 (1994)
- (4) 柴田克成, 岡部洋一: "視覚センサ信号を入力とした遅延強化学習", 日本神経回路学会全国大会第7回全国大会講演論文集, pp. 144-145 (1996)
- (5) Rumelhart, D. E., Hinton G. E. & Williams R. J.: "Learning Internal Representations by Error Propagation" *Parallel Distributed Processing*, The MIT Press, pp. 318-362 (1987)
- (6) 甘利俊一, 村田昇, Howard Yang, Klaus Müller: "過学習の理論-神話と事実-", 日本神経回路学会第6回全国大会講演論文集, pp. 28-29 (1995)
- (7) Sutton, R. S.: "Learning to Predict by the Methods of Temporal Differences", *Machine Learning*, 3, pp. 9-44 (1988)
- (8) Barto, A. G., Sutton, R. S. & Anderson C. W.: "Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems", *IEEE Trans. SMC*, 13, pp. 835-846 (1983)
- (9) 浅田稔, 野田彰一, 依積田健, 細田耕: "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, pp. 68-74 (1995)

柴田克成 (非会員) 1964年生まれ。1989年3月 東大大学院工学系研究科修士課程修了。1989年4月 (株) 日立製作所入社。1992年10月 同社退職。1993年9月 東大大学院工学系研究科博士課程中退。1993年10月 東大先端科学技術研究センター助手。1997年3月 東大大学院博士 (工学)。1997年4月 東工大大学院総合理工学研究科リサーチアソシエイト。主として、ニューラルネットを用いた自律学習システム、特に強化学習の研究に従事

岡部洋一 (正員) 1972年3月 東大大学院工学系研究科電子工学専攻博士課程修了。1972年 東大工学部電気工学科講師。その後、同助教授、IBM San Jose研究所客員研究員、教育用計算機センター助教授、電子工学科助教授、電子工学科教授、国際・産学共同センター教授など。現在同大学先端科学技術研究センター教授。主として、情報デバイス、特に、ニューラルネット、超伝導、脳磁計測などの

研究に従事