視覚センサ信号を入力とした遅延強化学習

Delayed Reinforcement Learning when Visual Sensory Signals are Given as Inputs

柴田 克成(PY)、岡部 洋一　　東京大学 先端科学技術研究センター

〒153　東京都目黒区駒場 4-6-1　　email : shibata@okabe.rcast.u-tokyo.ac.jp

Katsunari SHIBATA(PY) and Yoichi OKABE

Research Center for Advanced Science and Technology (RCAST), Univ. of Tokyo

abstract - It is shown that a neural-network based learning system, which obtains visual signals as inputs directly from visual sensors, can modify its outputs by reinforcement learning. Even if each visual cell covers only a local receptive field, the learning system integrated these visual signals and represents spatial information through the learning.

１．はじめに

報酬や罰から適切な動作を学習する強化学習が自律学習の観点から最近注目を集めつつある。報酬や罰は、通常、一連の動作の後に得られるため、得られた報酬や罰を元に各状態に対する評価関数を経験から学習し、かつ、その評価関数を用いて動作を学習する方法（これを遅延強化学習と呼ぶ）が提案されている[1][2]。

従来、学習システム（ここでは階層型ニューラルネットワーク）に視覚センサの信号を入力し、適切な動作を学習させる場合、視覚センサ入力を学習しやすい形に人間が加工して学習をさせてきた。これは、視覚センサが、多数のセンサセルよりなり、かつ個々のセンサセルは局所的な受容野しか持たないため、これを直接学習システムに入力することは、無駄が多く、難しいと考えられてきたからである。

しかし、遅延強化学習は、現在の状態に対する適切な評価および動作を学習していくものであるため、原理的には視覚センサの情報を直接ニューラルネットへ入力しても学習が可能である。また、センサ信号から評価および動作へのマップが形成できれば、局所的な受容野しか持たない各センサセルからの信号を統合した空間的な情報がニューラルネットの中に形成されていると考えられる。そこで、実際にシミュレーションを行い、学習がうまくできること、さらに、中間層に空間情報がコーディングされることを確認した。

２．学習方法

図１のような遅延強化学習を行う学習システムを考える。動作生成部及び状態評価部は階層型のニューラルネットにより構成される。ただし、実際は、両者を１つのニューラルネットで構成する。このシステムは、試行錯誤の乱数成分 ***rnd*** を含む動作をしながら、評価値 □ と動作 ***m*** の学習を並列に行う。

評価関数の学習アルゴリズムとして、時間軸スムージング学習[3]を拡張した評価値の時間変化量一定化学習を用いた。単位時間あたりの理想とする評価値の変化量 □□ ideal を過去の最大所要時間 Nmax より

□□ ideal = □ amp / Nmax　　(1)

□ amp : 理想振幅、ここでは 0.9-0.1=0.8

Nmax[*i*] = N[*i*]　　if Nmax[*i*-1]<N[*i*]

= (1-1/□)Nmax[*i*-1]　　if Nmax[*i*-1]>N[*i*]　　(2)

N[*i*] : *i* 番目の試行時の目的達成に要した時間、□: 大きい定数

と求め（ニューロンの値域を０から１とする）、実際の評価値の変化量 □□(*t*)と比較し、１単位時間前の評価値 □(*t*-1)に対し、

□s(*t*-1) = □ (*t*-1) - □ (□□ ideal - □□ (*t*))　　(3)

□s : 評価値に対する教師信号、　□□ (*t*) = □ (*t*)-□ (*t*-1)、

□ : 学習のための定数

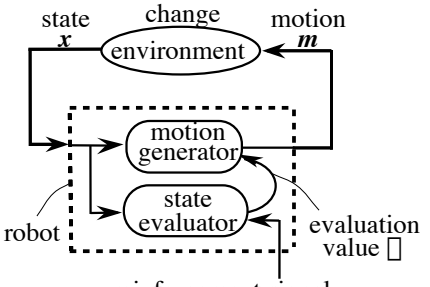

図１　遅延強化学習システムの構成

という教師信号を内部で生成し、評価値の時間変化を滑らかに、かつその変化量が一定になるように学習を行い、報酬が得られた（目的を達成した）時には 0.9 という教師信号で学習を行う。また、動作に対しては、

m *s* = ***m*** + □***rnd*** △□　　***m*** : 動作ベクトル（動作生成部の出力）　□: 学習のための定数　　(4)

という教師信号を内部生成し、より評価値の時間変化量が大きくなるように評価の学習と並列に学習する。

３．シミュレーション

図２のような２つの視覚センサを持った移動ロボットが target を捕らえるという問題を考える。視覚センサは、それぞれ２４個のセンサセルが１次元に配列され、180°の視野を有する。各センサセルの受容野は放射状に広がり、オーバーラップがなく、その中で target が投影される面積の割合を０から１の連続値で出力する。そして、このロボットは、target に到達した時のみ報酬が与えられる。また、学習の初期には、試行錯誤の乱数でしか動作できないため、target を近くに置き、学習が進むにつれて徐々に遠ざけた。さらに、視野から target がなくなった時には罰として 0.1 の教師信号を与えて学習を行った。ロボットが target に到達するかまたは視野からなくなるまでを１試行とし、その試行の終了後、target の位置を変えて再び動作と学習を繰り返した。ニューラルネットは、入力層ニューロンが４８個、中間層ニューロンが２０個、出力層は、評価用１つ、動作用２つの計３個のニューロンで構成し、学習は、バックプロパゲーション法を用いた。

キーワード：遅延強化学習、視覚センサ信号、時間軸スムージング学習、ニューラルネットワーク

図３に、学習後のロボットと target の相対座標 (X', Y') に対する評価関数の値（等高線）、および target の位置を変化させた時のロボットの経路（ロボット中心の座標のため、相対的に target が動く）を示す。これより、視覚センサの各センサセルが放射状の広がりを持つ局所的な受容野しか持たないにも関わらず、それに依存しない滑らかな評価関数が学習によって形成されたことが分かる。また、ロボットは回転して target を正面に捉えてから前進するというほぼ最適に近い動作を獲得しており、視覚センサからの信号の代わりに (X', Y') を入力とした場合とほぼ同様の経路が形成されていることがわかった。

次に、学習後のニューラルネットの出力層を切り放し、新たに１つの出力層ニューロンを設け、中間層と出力層の間の結合を０にし、これに対し、バックプロパゲーション法による教師あり学習を行った。図４に示したロボット座標上の黒丸と白丸で表した６点に target を置いた場合について、それぞれ視覚イメージを生成し、その入力に対し、教師信号を白丸の部分では 0.1、黒丸の部分では 0.9 として学習を行った。学習後の target の位置に対する出力の分布を図４(a) に示す。入力層と中間層の間の結合を固定した場合もほぼ同じ結果であった。比較のため、強化学習を行う前のニューラルネットに対して同様の学習を行ったものを図４(b) に示す。この結果、強化学習を行った後のニューラルネットワークでは、target が右に見える場合と左に見える場合をきれいに分類できるようになったが、強化学習を行わないニューラルネットワークでは、出力値がきれいに分布せず、センサセルの受容野が放射状に広がっていることを反映した出力の分布となっていることがわかる。このことから、強化学習を行うことによって、中間層において target が自分の右に見えるか左に見えるかがコーディングされたことがわかった。

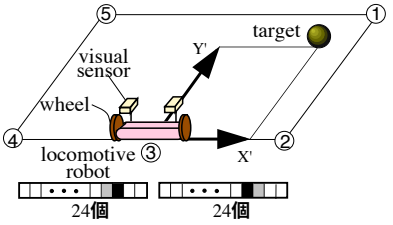

図２　シミュレーションの環境と視覚付きロボットおよび視覚センサ

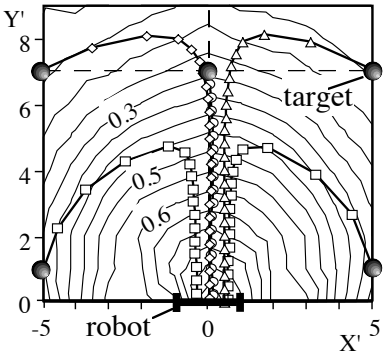

図３　学習後の評価関数とロボットの経路（ロボット固定の座標）

次に、ニューラルネットを、各層のニューロン数が 48-20-2-10-3 の５層とし、３層目の２個のニューロンに X', Y' の情報がどのようにコーディングされるかを調べた。学習は、上記と同様に行い、学習後の評価関数およびロボットの経路も上記とほぼ同じ結果が得られた。この時、(X', Y')の格子点に target を置いた時の視覚イメージを入力した時の中間層のニューロンの値の分布を図５に示す。x 軸 y 軸はそれぞれ２つの中間層ニューロンの値を表している。図中の丸付きの番号は、図２中の丸付きの番号に target を提示した場合を示す。これより、計４８個の視覚センサの信号を統合し、target の位置の情報を比較的きれいにコーディングしていることがわかった。ただし、格子の間隔は一様でないが、これは、target の位置の変化に対する評価や動作の変化が大きいところが拡大されているためである。この傾向は、入力を (X',Y') とした場合にも見られた。

４．結論

遅延強化学習において、視覚センサからの信号を直接入力とした場合でも適切な動作をうまく学習できることを示した。また、この時、中間層ニューロンにおいて、局所的な受容野しか持たない各センサセルからの信号が統合され、空間情報がコーディングされていることがわかった。

参考文献

[1] A. G. Barto, R. S. Sutton and C. W. Anderson, "Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problrms", IEEE Trans. SMC-13, pp. 835-846 (1983)

[2] K. Shibata and Y. Okabe : "A Robot that Learns an Evaluation Function for Acquiring of Appropriate Motions", Proc. of WCNN '94 San Diego, Vol. 2, pp. II-29 - II-34 (1994)

[3] 柴田克成、岡部洋一, “時間軸スムージング学習と局所センサ信号の統合”, 第７回日本神経回路学会全国大会講演論文集 (1996)

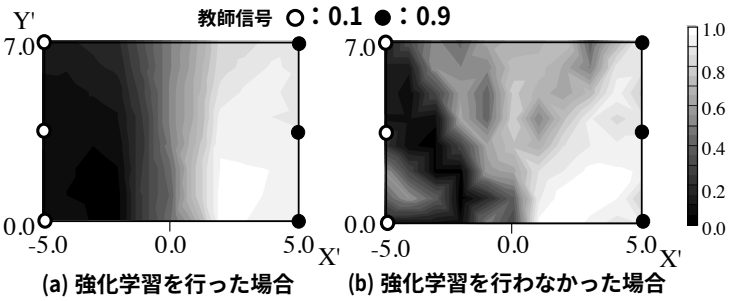

図４　強化学習を行った後に教師あり学習を行った場合と教師あり学習だけ行った場合の target の位置に対する出力値の分布

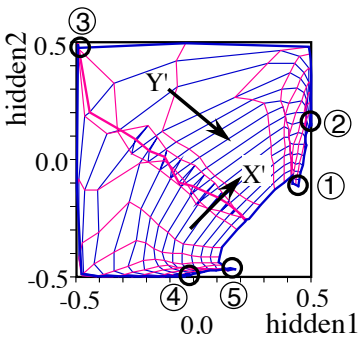

図５　中間層ニューロンにおける target の位置情報のコーディング