選択的不感化ニューラルネットを用いた関数近似器による強化学習

新保 智之† 山根 健† 田中 文英† 森田 昌彦†
†筑波大学大学院システム情報工学研究科　〒305-8573 つくば市天王台1-1-1

1 序論

強化学習 [1] は、動物の行動学習をモデル化したものといわれているが，一般に学習効率が悪く，状態空間が広いと非常に長い学習時間を要する．そのため，状態を離散化して状態数を減らすか，状態空間の次元を非常に低く抑える必要があり，実空間で生きる動物との間には非常に大きな差が存在する．

これまで，強化学習を効率化するために，主に学習アルゴリズムに関する改良が図られてきた．しかし，このような改良は，一般に計算の複雑化や必要なメモリ量の増大を伴う．我々は，上述した動物との大きな差は，学習アルゴリズムの違いによるものではなく，状態の評価値を表す価値関数の近似器の違いに起因すると考えている．

通常，強化学習の価値関数は，事前に関数形がわからない上に非線形性が強く，部分的に不連続性をもつ．そのため，大域的近似手法は適さず，これまでほとんどが局所的近似手法に基づくものであった [1]．また，脳の情報処理の仕組みをモデル化したニューラルネットも一般的には適さないといわれている．

近年我々は，多層パーセプトロンを2変数以上の関数の近似に適用した場合，大域的な汎化は全く生じないこと，その原因が1対多対応による荷重の平均化にあること，そして選択的不感化という手法によってこの問題が解決されることを明らかにした [2]．

そこで，本研究では，選択的不感化法を用いたニューラルネットによって強化学習の価値関数の近似器を構成し，アクロボットの振り上げ課題に適用することで，冗長な状態変数を追加したときの学習効率や計算量の変化について調べる．

2 SDNNによる関数近似

ある神経素子の出力を，入力や内部電位に関係なく中立値にすることを「不感化」といい，素子群のうち不感化するものを別の情報に応じて決めることによって，2種類の情報を統合する手法を「選択的不感化法」という．これを層状のニューラルネットに適用したものが選択的不感化ニューラルネット（Selective Desensitization Neural Network : SDNN）である．本研究では，図1に示すようなSDNNによって価値関数の近似器を構成した．以下，文献 [3] と異なる部分を中心に説明する．

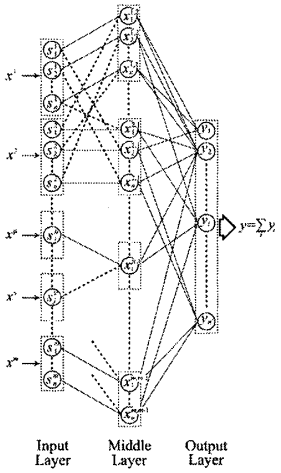


図 1: SDNNによる m 変数関数近似器の構成

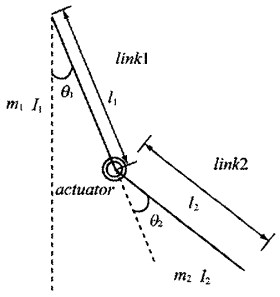


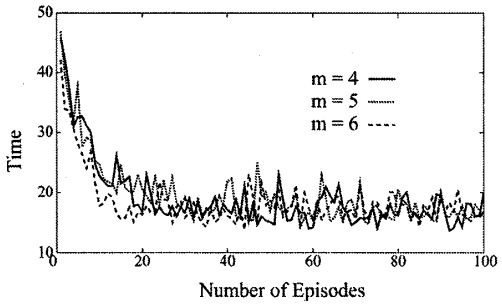
図 2: アクロボット

まず，入力層は m 個の状態変数に対応する m 個の素子群からなり，それぞれが状態変数の値を200次元の2値パターンによって表す．このパターンは，± 1 が常にほぼ同数であり，変数の値が連続的に変化するにつれて1と -1 が徐々に入れ替わるように符号化されている（具体的には，9個のパターンをランダムに作成し，それを環状または線状に並べてその間を補間することによって，すべての値に対応するパターンを作成した）．

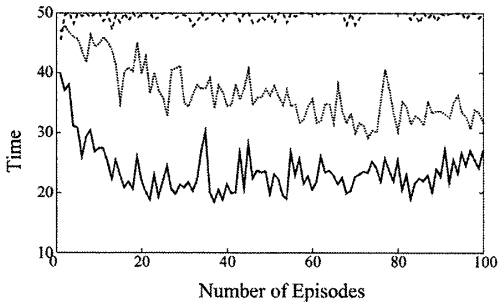
中間層は，それぞれ200個の素子からなる $m(m-1)$ 個の素子群からなる．各素子群は，入力層のある素子群の出力パターンをそのまま受けるとともに，別の素子群の出力を修飾パターンとした選択的不感化を受ける．出力層には，行動の種類に対応する複数の素子群があり，それぞれ600個の素子からなる．出力素子の出力の合計 $y = \sum_i y_i$ から評価値への変換には $Q = (y - 300)/30$，中間層から出力層への結合荷重は，一種の誤り訂正学習によって更新する．

Function Approximator for Reinforcement Learning Using Selective Desensitization Neural Networks

† Tomoyuki SHIMBO(shin@bcl.esys.tsukuba.ac.jp)
† Ken YAMANE
† Fumihide TANAKA
† Masahiko MORITA(mor@bcl.esys.tsukuba.ac.jp)
Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Ten-nodai, Tsukuba-shi, 305-8573 Japan (†)



(a) SDNN モデル



(b) RBFN モデル

図 3: 学習過程

3 実験方法

強化学習の性能を評価するためのベンチマークとして，アクロボットの振り上げ課題を用いる [1].

アクロボット（図 2）の物理パラメータは Sutton の実験例 [1] に合わせた．但し，選択できる行動は左右の 2 種類，制御の時間刻みは 0.1[s] とした．また，リンク 1 が 1 度未満の場合に負の報酬（罰），リンク 2 の先端が一定の高さを超えたときに正の報酬を与えた．また，50 秒を経過しても成功しない場合には，50 秒でエピソードを打ち切り，再スタートさせることとする．

強化学習のアルゴリズムは Q-learning を用いており，学習によって得られた行動価値関数との誤差が小さくなるよう近似器の出力を修正する．

価値関数の近似器として SDNN を用いたシステムを SDNN モデル，比較対象として，代表的な局所的近似手法である放射状基底関数ネットワーク（Radial Basis Funcion Network : RBFN）を用いたものを RBFN モデルと呼ぶ．RBFN の基底関数は標準偏差 $\sigma = 0.2$ のガウス関数とし，各次元について 0.2 の間隔で 11 点ずつ，合計 11^m 個を格子状に配置した．

4 結果と考察

学習過程を図 3（(a) : SDNN モデル，(b) : RBFN モデル）に示す．横軸がエピソード数，縦軸が継続時間であり，それぞれ乱数系列を変えた 10 試行の平均値がプロットされている．

冗長変数がない場合（$m = 4$）は，学習の初期段階において，両モデルにほとんど差はないが，最終的な平均到達時間は SDNN モデルの方が短く，第 81〜100 エピソードの平均値で比較すると約 6.5[s] の差があり，SDNN モデルの方がより適切に価値関数を近似できることを示している．

これに対して，各時刻に $[-1, 1]$ の乱数を 1 つ加えた場合（$m = 5$）は，RBFN モデルにおいて冗長次元の影響を強く受けており，学習性能が大きく低下した．乱数を 2 つ加えた場合（$m = 6$）では，さらにその傾向が顕著に表れた．一方，SDNN モデルの学習性能は，冗長変数がないときとほとんど変わらない．

また，次元の増加に伴って RBFN は計算量が指数関数的に増大するため，6 次元程度が現実的な限界である．一方 SDNN の計算コストの増加は m^2 のオーダーであり，より高次元空間での学習にも適用可能だといえる．

さらに，SDNN の優れた性質をいくつか挙げる．

- 多数の閾素子の出力パターンによって関数値を表現するため，部分的に不連続な関数を近似可能．
- 素子数を増やしても過剰適合が生じず，むしろ多いほど量子化誤差および統計的ノイズが減って近似精度が高くなる．
- 学習則は単純で計算量が少ない上に必要な繰り返しも非常に少なく，追加学習も容易であるため，オンラインでの使用に適している．

5 結論

選択的不感化ニューラルネットを価値関数の近似器として用いた場合，近似精度と学習効率が相反しない，冗長次元による性能の低下や計算コストの爆発が抑えられる，部分的に不連続な関数が近似可能，過剰適合が生じない，オンラインでの使用に適するなど実空間の強化学習に適した特徴があることを示した．

この成果は，実空間における自律ロボットや，膨大な情報の中から必要な情報を抽出する技術の開発に繋がる可能性がある．

参考文献

[1] R.S. Sutton and A.G. Barto : Reinforcement Learning, MIT Press (1998).

[2] 森田昌彦，村田和彦，諸上茂光，末光厚夫：選択的不感化法を適用した層状ニューラルネットの情報統合能力，電子情報通信学会論文誌，Vol.J87-D-II，No.12，pp.2242–2252 (2004).

[3] 新保智之，山根健，森田昌彦：冗長次元を含む状態空間における選択的不感化ニューラルネットを用いた強化学習，電子情報通信学会技術研究報告，Vol.108，No.383，pp.7–12 (2009).