ニューラルネットを用いた強化学習による空間情報の抽象化と知識継承

大分大学　柴田克成

Spatial Abstraction and Knowledge Transfer through Reinforcement Learning with a Neural Network

Katsunari Shibata　Oita University

Abstract: In the real lives, "abstraction" seems to generalize the knowledge obtained through the past experiences and to be very useful to accelerate learning drastically. Abstraction can be considered as the process to discard useless information. The important subject, the author thinks, is the criterion to judge whether information is useful or not. The author has propounded an idea that necessary functions emerge through reinforcement learning using a neural network. In this paper, the idea that rewards and punishments are used as a rational criterion for abstraction and the internal representation in a neural network trained by reinforcement learning is considered as abstract information is introduced. Furthermore, it is shown in a two-sensor and two-motor combination task that the abstraction enables knowledge transfer that is difficult for the conventional techniques for acquisition of abstract information to realize.

1. まえがき

われわれ人間が、たった今、目や耳から得られるセンサ信号と全く同じセンサ信号を得ることは、過去にも将来にも恐らくないであろう。にもかかわらず、われわれはさまざまな状況において、同じ状況を過去に経験したかのように的確な行動をとることができる。一見この能力は、似た入力では似た出力となるニューラルネットの汎化能力で説明できるように思われる。しかしながら、たとえばわれわれは信号が赤から青に変わると前に進むが、信号の情報は、われわれが得る視覚信号のうちのほんの一部分であるし、信号の位置も場所によって異なるため、必ずしも過去の状況と入力信号が似ているから適切な行動がとれるとは言えない。にもかかわらず、信号の変化によって行動できるということは、信号の色が重要な情報であるということを知っているからに他ならない。また、信号の色が重要であるということは、生まれつき知っているとは考えにくい。したがって、何が重要で注目すべき情報なのかということも、学習によって培ってきたものではないかと考えられる。

このように、センサの情報から重要な情報を抽出する過程は一般的に「抽象化」と呼ばれる。「抽象化」とは、情報量を減らし、「重要な」情報のみを残すことであると考えられる。では、情報が「重要」であるかどうかの基準はどう考えれば良いのであろうか？Brooks も「抽象化は知能の本質であり，解決が困難なところである」と指摘している[1]が、ここが非常に重要なポイントであると筆者も考える。

一つの基準として、減らした情報から元の情報をいかに正確に再現させられるかが考えられる。主成分分析等の統計的手法を用いることも広い意味でこれの一種と考えることができる。しかしながら、膨大な入力信号のすべてを再現させることが本当に重要なことなのかという点で疑問が残る。

predictive representation[2]という考え方では、将来のセンサ信号を予測するために必要な情報を重要な情報とする。これを実現するための TD net[3]という方法も提案されている。これは非常に理にかなっているように見える。しかしながら、前述のように、センサ信号は膨大であり、今度は「予測すべき重要な情報は何か？」という問いにぶつかることになる。予測できるものを予測するという考え方[4]もあるが、何が予測できる情報かを知るためには、結局、あらゆる情報を予測してみなくてはわからないだろう。

また、action respecting embedding[5]という考え方では、センサ信号パターンの時間的な関係を、抽象化する情報空間の距離に反映させる仕組みを導入しており、物事の因果関係に応じた情報表現の獲得が期待される。ところが、システム本来の目的から来るトップダウン的な情報の抽象化を実現する仕組みはない。

筆者は、強化学習の自律学習能力とニューラルネット内部での適応的な情報表現の獲得能力に注目してきた。そして、人間が知識を付与して機械を知能化する従来の手法に対し、強化学習とニューラルネットを組み合わせることで、試行錯誤を通して、自律的に柔軟な行動獲得をするとともに、ニューラルネット内部に、認識、記憶などのさまざまな能力が獲得されることを示してきた[6]。また、筆者らは、ニューラルネットを教師あり学習させる際に、入力パターンが異なっていても、それに対する教師信号が近ければ、中間層表現も近くなるということを実験的に示し[7]、これが、目的に沿った情報の抽象化に大きな役割を果たしているのではないかと考えてきた。

本稿では、報酬や罰である強化信号を抽象化の基準とするという考え方を紹介する。そして、センサ信号をニューラルネットに入力して強化学習を行うことによって、空間情報としてのセンサ信号が、目的を達成するためという明確な基準の下に抽象化されることを、それが如実に現れる異種センサ間の知識継承の問題に適用し、その有効性を示す[8]。

2. タスク設定と学習方法

本稿で行ったタスクおよび学習システムの概観を

Fig. 1 に示す。２種類のセンサと２種類のモータセットがあり、ともに、２種類のうちのどちらか片方を使って目的を達成できるようになっている。行うタスクは簡単で、5.0x5.0 の正方形の行動領域に、モータセットによって動かされる 1.0x1.0 の大きさの正方形の対象物体と、同じく 1.0x1.0 の正方形で中央に固定されているゴール領域がある。対象物体は毎回ランダムに配置され、それを動かして、中心がゴール領域に入ると報酬 0.7 が与えられる。物体は、行動領域から外に出ようとしても、壁にぶつかって出られないとする。

２種類のセンサはともに視覚センサで、5.0x5.0 の領域をカバーするが、センサ *s*1 ではセンサセルが 5x5=25 個配置され、センサ *s*2 では 7x7=49 個配置されている。各センサセルは、個々の受容野に対して映っている対象物体の面積の割合を 0 から 1 の間の連続値で出力する。ニューラルネットは、入力から出力への情報表現を徐々に変化させられるように、図のように５層のものを用いた。入力ニューロンは全部で 25+49=74 個用意し、２種類のセンサからの信号を別々に入力した。出力は６個用意し、最初の３個がモータセット *m*1 用、次の３個がモータセット *m*2 用とした。強化学習は actor-critic[9]を使用し、それぞれの３個の出力のうち、１個が状態評価を行う critic、残りの２個が動作出力である actor とした。各ニューロンの出力関数は-0.5 から 0.5 の値域のシグモイド関数とし、critic の出力は、ニューラルネットの出力に 0.4 を加えたものを用いた。２つの動作出力は、それぞれのモータセットで異なる図中の式にしたがって、対象物体の *x*, *y* 方向の速度である v_x, v_y を正確に制御できるものとした。また、それぞれの出力に試行錯誤（確率的動作）のための乱数 *rnd* を付加した。２つのセンサ、２つのモータセットのどちらを使うかは、各試行開始時にランダムに選択した。選択されていないセンサからの信号はすべて０とした。

学習は、強化学習のアルゴリズムに基づいて、教師信号を生成し、それを用いて教師あり学習を行った。まず、TD 誤差 $\hat{r}_{t-1}$ を

$$\hat{r}_{t-1} = r_t + \gamma P(\boldsymbol{o}_t) - P(\boldsymbol{o}_{t-1}) \qquad (1)$$

とする。ここで、r_t は時刻 t で与えられた報酬、γ は割引率(ここでは 0.96 を用いた)、$\boldsymbol{o}_t$ は時刻 t でのセンサ信号、$P(\boldsymbol{o}_t)$ は $\boldsymbol{o}_t$ を入力としたときのニューラルネットの選択されているモータセット用の critic の出力に 0.4 を加えたものとする。critic の教師信号 $P_{s,t-1}$ を

$$P_{s,t-1} = P(\boldsymbol{o}_{t-1}) + \hat{r}_{t-1} = r_t + \gamma P(\boldsymbol{o}_t)、 \qquad (2)$$

actor の教師信号ベクトル $\boldsymbol{a}_{s,t-1}$ を

$$\boldsymbol{a}_{s,t-1} = \boldsymbol{a}(\boldsymbol{o}_{t-1}) + \hat{r}_{t-1}\boldsymbol{rnd}_{t-1} \qquad (3)$$

と計算する。ただし、$\boldsymbol{a}(\boldsymbol{o}_{t-1})$ は $\boldsymbol{o}_{t-1}$ を入力としたときのニューラルネットの選択されている方の actor の出力ベクトル、$\boldsymbol{rnd}_{t-1}$ は時刻 t-1 に出力に加えた乱数ベクトルである。そして、$P_{s,t-1}$(実際には 0.4 を引いたもの)、$\boldsymbol{a}_{s,t-1}$ を用いて、$\boldsymbol{o}_{t-1}$ を入力とした時のニューラルネットに対し、選択されているモータセットの方の出力のみ１回だけ誤差逆伝搬(BP)法[10]で学習させた。

３. センサによらない状態表現の獲得

前述のように、本タスクでは、毎試行、センサとモータセットをランダムに選んで学習を行う。ただし、センサ *s*2 とモータセット *m*2 の組み合わせの場合にはニューラルネットの学習は行わず、単に行動を観察する。使用センサが *s*1 か *s*2 かによって入力は大きく異なるが、*s*1-*m*1 の組み合わせと *s*2-*m*1 の組み合わせでの学習を行うと、対象物体の位置が同じ場合、学習が進んだ理想的な状態では同じ出力が求められる。したがって、文献[7]で示したように、出力側に近い上位の中間層では、*s*1 を用いた場合も *s*2 を用いた場合も、両者は近い状態表現となることが期待される。そしてさらに、*s*1-*m*2 の組み合わせで学習させることで、上位の中間層での状態表現から *m*2 への学習が行われるため、結果的に、*s*2-*m*2 の組み合わせは学習しなくて

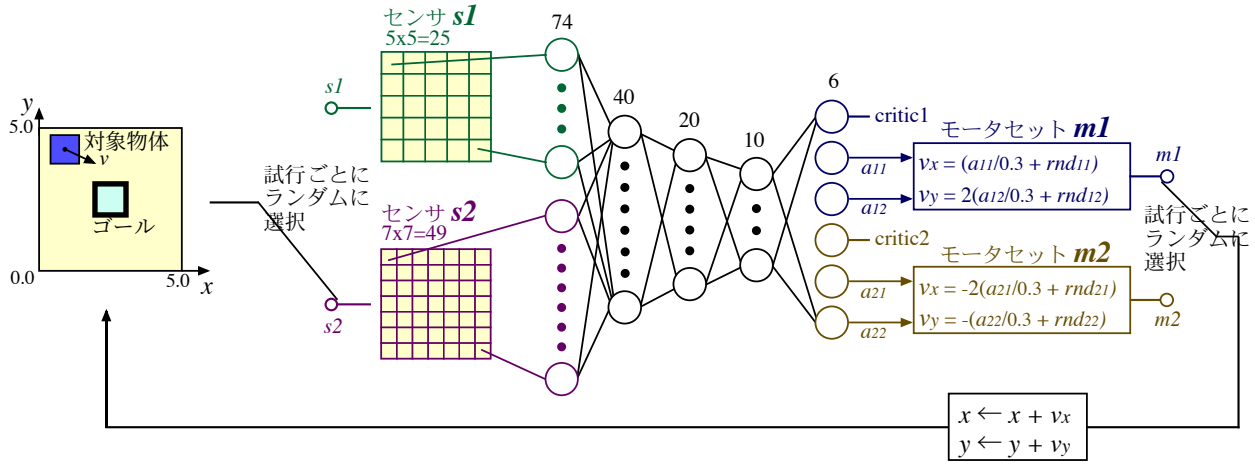

Fig.1 Two-sensor and two-motor combination task and learning system using a neural network.

も、ある程度タスクを成功させられることが期待される。しかしながら、*s*1-*m*1 か *s*2-*m*1 を学習しないと、センサによらない状態表現が形成されず、また、*s*1-*m*2 を学習しないと、モータセット *m*2 の学習ができないため、*s*2-*m*2 の組み合わせでは対象物体をゴールに持って行くことは困難であると考えられる。また、前述の action respecting embedding を始めとする情報の抽象化を目指した他の手法では、センサが違うと同一状態であることを知る術がないため、このような学習による知識継承の実現は困難と思われる。

4. シミュレーション結果

各組み合わせの場合のゴールまでの平均ステップ数を縦軸にとった学習曲線を Fig. 2 (a)に、比較のために、本来学習する 3 つの組み合わせのうち、どれか 1 つの学習をしなかった場合を Fig. 2 (b)(c)(d)に示す。ただし、100 ステップでゴールできない場合は、そこで試行を打ち切った。 (a)を見ると、学習した 3 つの組み合わせの場合の平均ステップ数は学習によって大きく減っており、学習していない *s*2-*m*2 の組み合わせでも、少し遅れはするものの、ゴールに到達できるよ

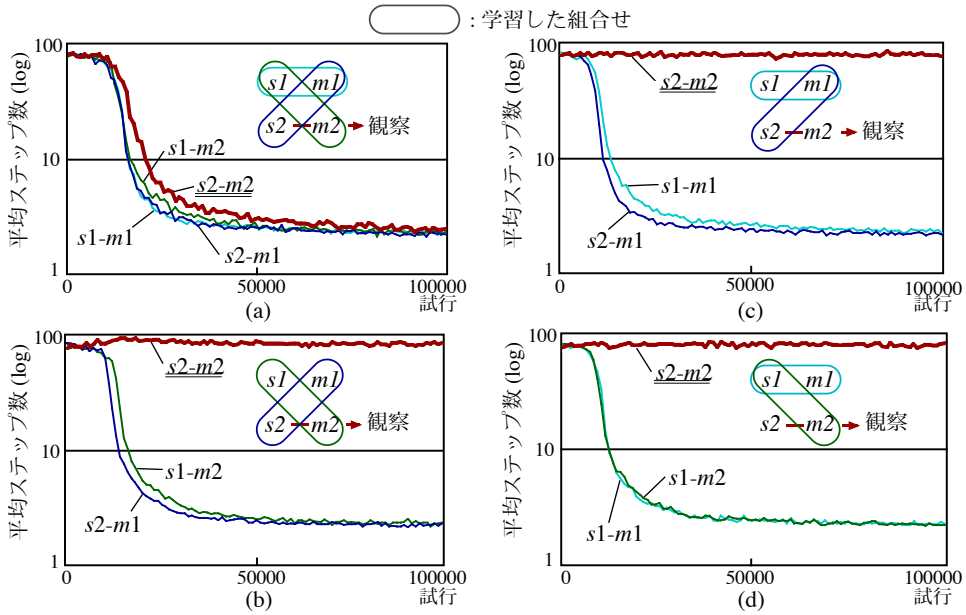

Fig. 2 Comparison of learning curves between 4 cases of learned combinations of sensor and motor set

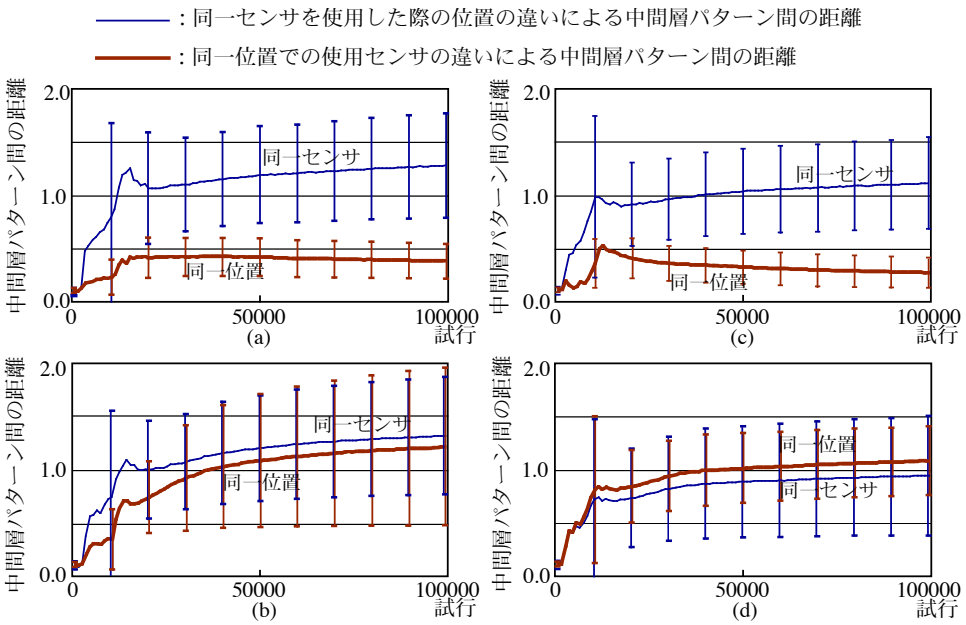

Fig.3 Comparison of the average distance between output patterns of the highest hidden layer during learning when the object location is varied with same sensor and when the sensor is different with the same object location. The vertical bar indicates standard deviation.

うになっていることがわかる。また(b)(c)(d)より、２つの組み合わせでしか学習させないと、*s*2-*m*2 の組み合わせではゴールに到達できないことがわかる。これは、前節での議論と一致する。

Fig. 3 にそれぞれの場合で、同一センサ間で対象物体の位置を 0.5 間隔で変化させたときの中間層ニューロンの出力パターンの距離（各ニューロンの出力の差の絶対値の和）と、同一位置で、使用センサを変えた場合の中間層ニューロンの出力パターンの距離の平均値と標準偏差をプロットしたものを示す。これにより、*s*1-*m*1, *s*2-*m*1 の組み合わせの学習を行った(a)と(c)の場合は、物体が同一位置にいる場合に、センサが異なって入力が全く違っても、中間層パターンは似ていることがわかる。しかし、*s*1-*m*1, *s*2-*m*1 のどちらかを学習していない(b)(d)の場合は、物体の位置が同じでもセンサが異なると中間層の表現が大きく異なることがわかる。この結果も、前節の議論と一致する。

また、Fig. 4 には、前述の(a)と(b)の場合に、物体を２つの位置に置いてみたときのそれぞれについて、使用センサを変えた時に、10 個の上位中間層ニューロンの表現が実際にどう変化しているかを示す。上２つの(a)の場合はセンサが違っても、位置が同じであればパターンは似ているが、下２つの(b)の場合はセンサが違うと同一状態でも、状態表現が異なることがわかる。

また、ここでは、紙面の都合上省略するが、中間層ニューロン数を３層とも 40 個とした場合、中間層１層の３層構造(中間層ニューロン数 40 個)とした場合、いずれも、上記の場合よりも(a)で *s*2-*m*2 の場合の平均到達ステップ数の減少が遅くなった[8]。つまり，ニューラルネットの構造によって抽象化能力が異なる。

5. あとがき

本稿では、空間情報であるセンサ信号を報酬や罰を重要度の基準として学習することの合理性、および、それがニューラルネットを用いた強化学習で簡単に実現できることを紹介した。そして、２つのセンサと２つのモータセットから毎回ランダムに１つずつ選んで同一タスクを学習させる問題を取り上げ、異なるセンサを使っても、タスクの状態が同じであれば、学習を通してニューラルネットの中間層の表現が近くなることを示した。このようなことは、従来の合目的性が考慮されていない抽象化の手法では実現が困難と思われる。しかしながら、われわれ人間がこのような同一タスクの認識をする際は、もっとひらめきのような形で行われると思われる。本手法ではこのような能力の実現は困難であり、今後の課題である。

謝辞

本研究は、日本学術振興会科技学術研究費補助金基盤研究#15300064 および#19300070 の補助を受けた。ここに謝意を示す。

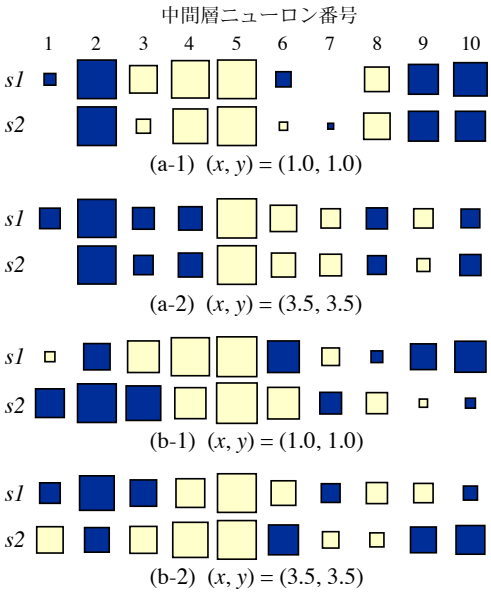

Fig. 4 Some samples of the output pattern of the highest hidden layer for the case of (a) and (b) in Fig. 2. The area of each square indicates the absolute value and the color indicates the sign (white is positive) of each output.

参考文献

[1] R.A. Brooks, "Intelligence Without Representation". Artificial Intelligence, Vol. 47, pp.139-159 (1991).

[2] M.L. Littman, R.S. Sutton and S. Singh, "Predictive Representations of State", In Advances in Neural Information Processing Systems, Vol. 14, pp. 1555-1561, MIT Press (2002)

[3] R.S. Sutton and B. Tanner, "Temporal-Difference Networks", In Advances in Neural Information Processing Systems, Vol. 17, pp. 1377-1384 (2005)

[4] J. Schmidhuber, "Exploring the Predictable", In Ghosh, S. Tsutsui, eds., Advances in Evolutionary Computing, pp. 579-612, Springer (2002)

[5] M. Bowling, A. Ghodsi and D. Wilkinson, "Action Respecting Embedding", Proc. of the 21st Int'l Conf. on Machine Learning , pp. 65-72 (2005).

[6] 柴田克成, ” 強化学習とニューラルネットによる知能創発“, 計測と制御, Vol. 48, No. 1, pp. 106-111 (2009)

[7] 柴田克成, 伊藤宏司, "階層型ニューラルネットにおける中間層での適応的空間再構成と中間層レベルの汎化に基づく知識の継承", 計測自動制御学会論文集, Vol. 43, No. 1, pp. 54-63 (2007)

[8] K. Shibata, "Spatial Abstraction and Knowledge Transfer in Reinforcement Learning Using a Multi-Layer Neural Network", Proc. of ICDL5 (Fifth Int'l Conf. on Development and Learning) (2006)

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

[10] D.E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning Internal Representation by Error Propagation", in D. E. Rumelhart, J. L. McClelland and the PDP Research Group, Parallel Distributed Processing (1986)