複数の学習機構の階層的構築による行動獲得

大阪大学　○高橋泰岳　　大阪大学　浅田稔

Behavior Acuisition by Multi-Layered Reinforcement Learning

*Yasutake Takahashi : Osaka University　　Minoru Asada : Osaka University

Abstract— This paper proposes multi-layered reinforcement learning which can decompose the control structure into smaller transportable chunks, that make previously learned knowledge applicable to related tasks in a newly encountered situations. We apply the method to a simple soccer situation in the context of RoboCup, and show the experimental results.

Key Words: reinforcement learning, multi-layered control system, RoboCup

1. はじめに

強化学習は先見的な知識がほとんど必要なく適応的な行動が獲得できるという利点をもっているが，一つの学習機構で実世界の複雑でさまざまな行動を学習することには限界がある．これは対象とするタスクが複雑になるにつれて，学習機構が複雑膨大になる．すなわち学習時間も非現実的な程長くなり，また獲得された行動を再利用することも困難になるという理由による．

そこで複数の学習機構で層を作り階層化することで，それぞれの学習機構をコンパクトにし，関連するタスクを新たに学習する際に以前の学習した結果を再利用することができる可能性がある．これまで提案されてきた手法は，その階層構造やそれぞれの階層における役割，それぞれのノードへの分担など全てを人間が明示的に設計した上で，それぞれのモジュールを設計するもしくは一部を学習させるものが多い．最近 Tani and Nolfi[1] は明示的に階層ごとの役割を指定することなく，全ての階層において同じ構造のモデルを用いる階層構造を提案し，ロボットのセンサ・モータコマンドシーケンスの予測を可能にしている．この考え方は興味深いが，提案された手法は行動選択部分が欠落しており，このままではロボットによる自律的な行動獲得はできない．

本論文では複数の同じ学習機構の階層的構築による行動獲得をロボットに行わせる．提案する手法を実機に適用した結果を示す．

2. 階層学習機構

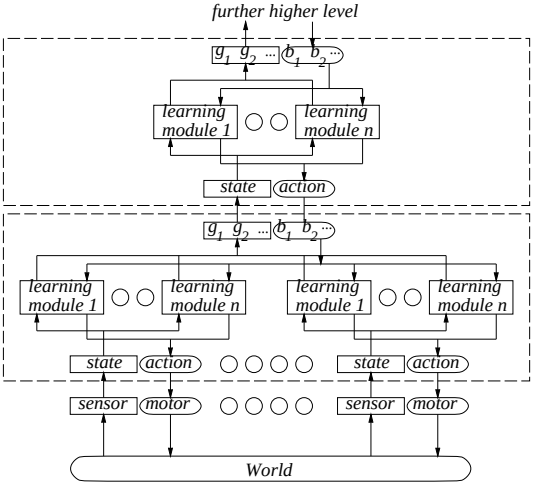

Fig.1 A hierarchical learning architecture

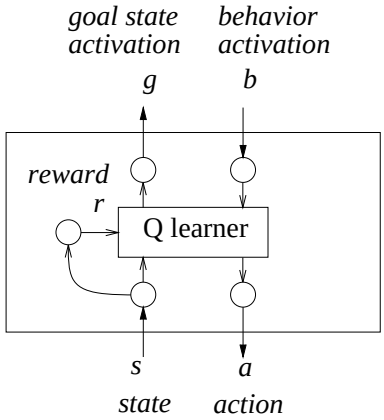

Fig.2 A behavior learning module

複数の同じ機構を持った学習機構の階層的構築による行動獲得をロボットに行わせる．学習機構を複数並べて層を作り，この層を階層的に構築する．下の層の学習機構はセンサ出力とモータコマンドを利用し低レベルの行動を学習する．上の層の学習機構は下の層の学習機構の学習結果を用いて，より抽象度の高い状況の遷移を学習する (Fig.1)．

それぞれの学習機構は状態と行動を認識し，それぞれのゴール状態により報酬を発生させ，これをもとに行動を学習する．学習アルゴリズムは広く利用されている Q 学習を連続値を扱うように修正した continuous valued Q learning[2] を用いる．また，学習機構は正規化した (ゴール状態を 1 とした) 状態価値 V を**ゴール状態活性度 (goal state activation)** g として上位に出力し，上位からはその学習機構が獲得した行動を出力するための指令**行動活性度 (behavior activation)** b が与えられる (Fig.2)．上の層の学習機構は下の層のゴール状態活性度/行動活性度を基に状態/行動空間を張る．

3. ゴール状態の自律的振り分け

学習中に複数の学習機構の間で互いに競合させることで，それぞれの担当領域 (ゴール状態) を自律的に決定させる．これにより従来の手法のようにサブゴールを設計者が前もって設計しておく必要がない．

ゴール状態に到達したときにのみ正の報酬が与えられ，他はゼロの報酬が割り当てられる場合，「学習の結果得られたある状態の Q 値の最大値 (状態価値) はその状態からゴールまでの距離の近さを表現している」と考えられることを利用する．それぞれのゴール状態を一様に振り分けるために，ある学習機構のゴール状態

を他のゴール状態から離すように，つまり他の学習機構たちの状態価値が小さくなる方向に移動させる．

4. 実験

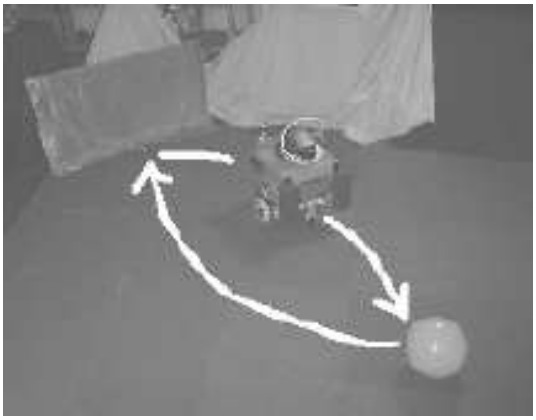

Fig.3 Experiment environment : A mobile robot, a ball and a goal

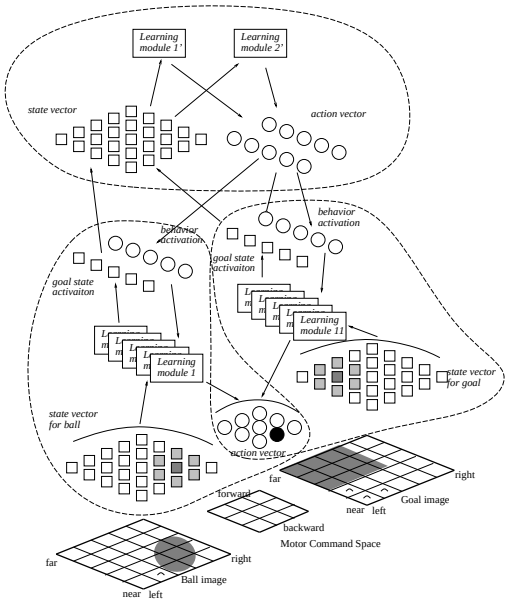

Fig.4 A hierarchy architecture of learning nodes

Fig.3 に示すロボットを使用する．このロボットはセンサとしてカメラを持ち PWS の移動機構を持つ．センサ出力として Fig.4 にあるように画像を縦横それぞれ 9 分割し，分割された領域の中心にボールやゴールの重心が近い程高い値を返すノードを用意する．またモータコマンドとして前進，後退など 9 通りを用意した．

この上に二層の階層的学習機構を作る．下の層には 40 の学習機構ノード (20 個はボール担当，他はゴール担当)，上の層には 2 つの学習機構ノードを作る．ただし，ボール/ゴールに一番近い状況および，見失った状況を担当する学習機構は前もって指定しておく．下位層の学習機構の状態数は 83($9 \times 9 + 2$)，行動数は 9．上位層の学習機構の状態数は 400(20×20)，行動数は 40($20 + 20$)．

タスクはボールに近付く，ゴールに近付くの二つ．学習中はこの二つのタスクを交互に経験させる (Fig.3)．

Fig.5，6 にボールに近付くタスクを遂行したときの学習後における下位層のそれぞれのノードのゴール状態活性度と行動活性度を示す．Fig.5 より下位層のノードがそれぞれの担当領域を決定し，タスク遂行中に移り変わるのがわかる．Fig.6 より上位層が階層のノードを順番に起動させることでタスクを遂行している様子がわかる．例えば 860 スッテプあたりはノード 11 の担当領域であり，ノード 12 を起動することで 860 スッテプ後半でノード 12 の担当領域へ遷移しており，これを繰り返すことでボールに近付くタスクを遂行している．

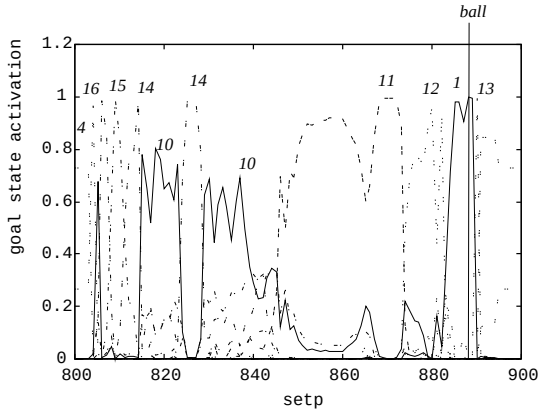

Fig.5 Goal state activation of modules at lower layer (ball)

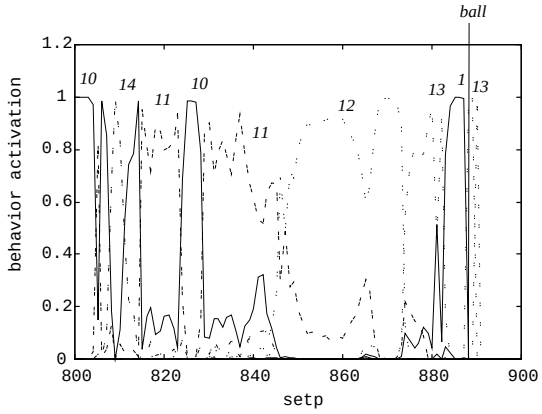

Fig.6 Behavior activation of modules at lower layer (ball)

5. おわりに

本論文では複数の同じ機構を持った学習機構の階層的構築による行動獲得をする手法を提案し，この手法を実機に適用した結果を示した．今後は階層をより多層化し，より複雑なタスクに適用しすることで本手法の有効性を確かめたい．また複数のタスクを学習するとき，以前の学習結果を再利用する事による有効性を確かめたい．本研究は科学技術振興事業団の戦略的基礎研究推進事業 (「脳を創る」プロジェクト) の援助を受けた．

参考文献

1) J. Tani and S. Nolfi: “Self-Organization of Modules and Their Hierarchy in Robot Learning Problems: A Dynamical Systems Approach” Sony CSL Technical Report, SCSL-TR-97-008 (1997).
2) Y. Takahashi, M. Takada and M. Asada.: “Continuous Valued Q-learning for Vision-Guided Behavior Acquisition” International Conference on Multisenso Fusion and Integration for Intelligent Systems (1999).