複数の学習機構の階層的構築による行動獲得

Behavior Acquisition by Multi-Layered Reinforcement Learning

○　高橋泰岳　(大阪大学)　　正　浅田　稔　(大阪大学)

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 navigation in the context of RoboCup, and show the experimental results.

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

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，一つの学習器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，学習器が複雑膨大になること，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．

そこで複数の学習器で層を作り階層化することで，それぞれの学習器をコンパクトにし，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，それぞれのモジュールへの分担など全てを設計者が明示的に構築した上で，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．

本論文では均一の学習器を複数用いて階層的に構築することによる行動獲得をロボットに行わせる．複数の学習器を同時に学習させ競合させることで，それぞれの学習器が担当するゴール状態を自律的に決定し，またこれを層にすることで下位層はより基本的な，上位層はより抽象度の高い行為を獲得する．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

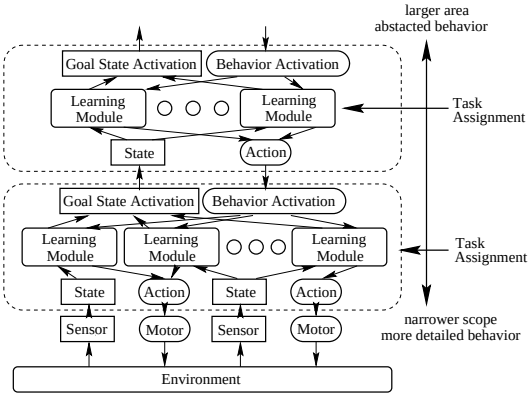

Fig.1 A hierarchical learning architecture

同じ種類の学習器を複数並べて層を作り，この層を階層的に構築する．下の層の学習器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．上の層の学習器は下の層の学習器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)．

それぞれの学習器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[1] を用いる．また，学習器は正規化した (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器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)．上の層の学習器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．

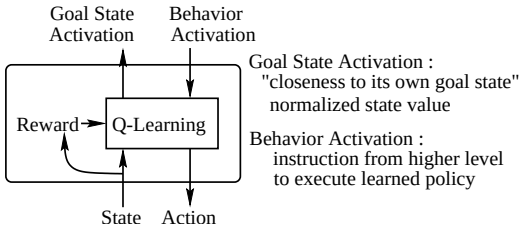

Fig.2 A behavior learning module

3. ゴール状態の自律的振り分け

それぞれのゴール状態を同一状態空間上に一様に振り分ける．このためにそれぞれの学習モジュール間の距離を知ることが必要だが，ゴール状態活性度をこれに利用する．それは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」と考えられるからである．

システムの学習中にある状態にあるとき，この状態におけるそれぞれの学習モジュールのゴール状態活性度を計算する．小さければこの状態付近を担当するモジュールがないと判断し新しい学習モジュールを足す．大きすぎればこの状態付近を担当するモジュールが多すぎると判断し，不必要な学習モジュールを削除する．このようにそれぞれの学習モジュールのゴール状態を自律的に決定させるので，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．

4. 階層型学習機構内での行動戦略

システムには任意のセンサ出力が目標状態として与えられる．階層型学習機構はまず一番下の層の状態空間の中で目標状態を認識し，この目標状態に一番近いゴール状態を持つ学習モジュールを探す．現在の状況がこのモジュールで学習済でありタスクを達成できるなら，担当の学習モジュールを起動する．

一番下の層の学習モジュールの担当範囲外であれば，一

つ上の層の中でゴール状態に一番近いゴール状態を持つ学習モジュールを探す．このモジュールでタスクを達成できるなら，担当の学習モジュールを起動する．このモジュールが下のモジュールを起動することで，下の層の学習モジュールで担当できる状態まで持っていき，あとは下の学習モジュールに任せる．さらにこの層の学習モジュールの担当範囲外であれば，更にこの手順を繰り返す．この手続きを踏むことでシステム内の階層構造の内容を分析することなく目標の状態を外部から与えることができる．

5. 実験

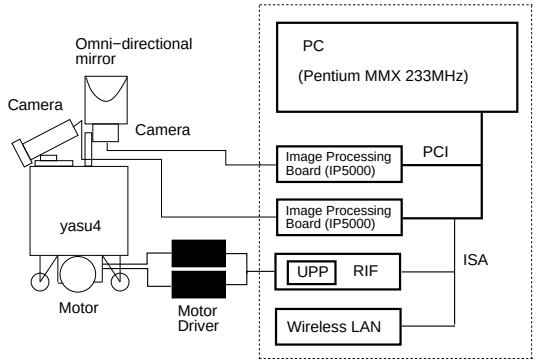

Fig.3 An overview of the robot system

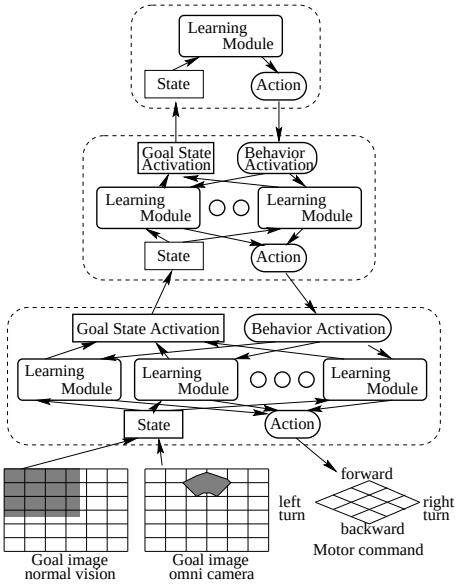

Fig.4 A hierarchy architecture of learning modules

Fig.3にロボットの簡単なシステムを示す．このロボットはセンサとして広角レンズを装着したCCDカメラと全方位ミラーを装着したカメラを持ち，二枚の画像処理ボードを使って実時間でボールやゴールの重心を抽出する．カメラの搭載位置のため，広角レンズを装着したカメラはロボット前方を，全方位ミラーを装着したカメラはロボットの側方と後方を観測することになる．また移動機構はPWSである．

Fig.4にあるように，最下位層の学習モジュールの状態空間は二つのカメラから得られた画像上のゴールの座標で構成され，それぞれの空間を9×9で離散化した．また行動空間は左右の車輪のモータコマンドに与える指令値で構成され，これも3×3で離散化した．それ以外の層の状態や行動は自分より一つ下の層の学習モジュールのゴール状態活性度と行動活性度によって構成され，それは学習モジュールが自律的に割り当てられるのでそれに依存する．タスクはゴールの画像中の位置座標によって与えられる．

ロボットは約3時間環境の中をランダムに移動し学習させた．Fig.4に示すように最上位層に一つの学習モジュールを持つ3層の階層型学習機構が得られた．

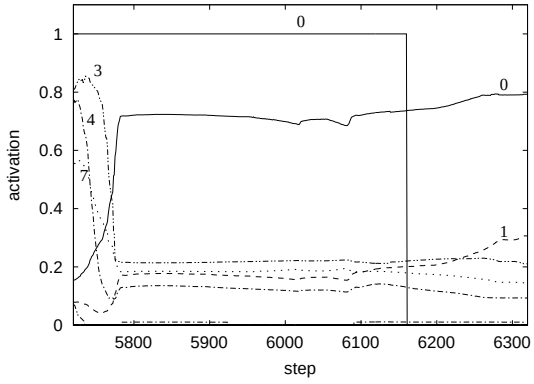

Fig.5 A sequence of the goal state/behavior activation of learning modules at middle layer

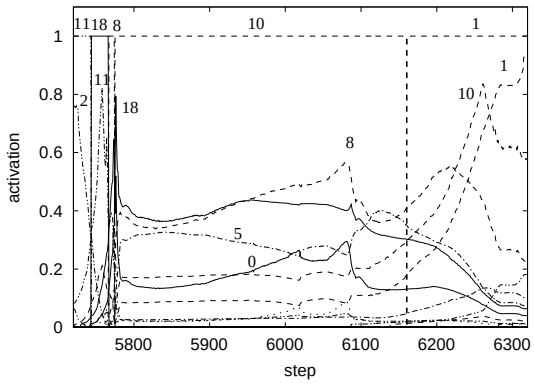

Fig.6 A sequence of the goal state/behavior activation of learning modules at lower layer

Fig.5，6にタスクとしてゴールを前方のカメラの中心でとらえる位置への移動を与えたときの，中間層と下位層のそれぞれのモジュールのゴール状態活性度と行動活性度を示す．0と1のステップ状に変化しているのが行動活性度であり，それ以外はゴール状態活性度である．5,720ステップではロボットはゴールから離れて反対方向を向いている．6,320ステップにおいてタスクを遂行し終っている．6,160ステップあたりまで中間層がモジュール0を起動し，下位の層のモジュール11，18，8，10を順番に起動することでゴール近くにまで移動し，その後は下位の層のモジュール1を起動して最終状態まで移動していることがわかる．

6. おわりに

本論文では均一の学習器を複数用いて階層的に構築すること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) 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).