例示の理解による階層型学習機構を用いた段階的行動学習

A Hierarchical Multi-Module Learning System based on Self-Interpretation of Instructions by Coach

○ 疋田 晃一 (阪大)　正 高橋 泰岳 (阪大, 阪大 FRC)　正 浅田 稔 (阪大, 阪大 FRC)

Koichi HIKITA, Osaka University, 2-1, Yamadaoka, Suita, Osaka
Yasutake TAKAHASHI, HANDAI Frontier Research Center, Osaka University
Minoru ASADA, HANDAI Frontier Research Center, Osaka University

We propose a hierarchical multi-module learning system based on self-interpretation of instructions by coach. The proposed method enables a robot (i) to decompose a long term task which needs various kinds of information into a sequence of short term subtasks which need much less information through it self-interpretation process for the instructions given by coach, (ii) to select sensory information needed to each subtask, and (iii) to integrate the learned behaviors to accomplish the given long term task. We show a preliminary result of a simple soccer situation in the context of RoboCup.

Key Words: reinforcement learning, multi-layered learning system, state space construction

1 はじめに

近年注目されている学習的アプローチの一つに強化学習があるが，考慮すべき情報が多い複雑な問題に対して強化学習を適用しようとした場合，必要となる計算資源の増大などの問題を解決する必要がある．これまでに学習時間の短縮に関して，一つのタスクに対して学習のスケジューリングを行なう研究[1]や複雑なタスクをサブタスクに分解し，それぞれに学習器を割り当てて学習を行ない，その複数の学習器を統合する研究[2]，教示を利用する研究[3]などがされている．また複数の学習器を階層的に構成することで計算資源を軽減する研究もされている．しかしこれらの研究の多くはサブタスクへの分解や階層構造の決定を設計者が行なっており，設計者の負担が大きい．

そこで本研究では階層型学習機構を用い，例示の理解を通して，複雑なタスクに対する行動獲得を段階的に行なう手法を提案する．抽象化を行ないながら階層を構成していくことで，単一の学習器で学習を行なう場合に生じる計算資源の増大化などの問題を回避し，扱う状態変数が異なる学習器を統合することで，結果として多数の状態変数を利用する一連のタスクを扱うことが可能となる．またコーチが目的を達成可能な行動系列を例示し，学習者がそれを自分なりに理解することで，これまでに獲得した学習器の有効性の判断や未学習であるサブタスクの発見及び新たな学習器の生成，階層構造の構築を自律的に行なう．本手法の有効性を検証するため，サッカーロボットのシミュレーション及び実ロボットを用い，ボール追跡，シュート，敵をかわしてシュートの順に 3 つのタスクを段階的に学習する．

2 基本的枠組

環境中には学習者とコーチが存在する．学習者にとって最適な行動系列を提示することは，コーチの負担が大きいため，コーチは学習者内部の学習機構に関する知識を持たず，目的を達成可能な行動系列を例示するものとする．学習者は例示された行動系列を基にこれまでに獲得した学習器が再利用できる部分と新たな学習器が必要な部分を判断し，新たな学習器を自律的に生成する．

タスクが与えられる度に，例示を基に既存の学習器の再利用可能性を判断し，再利用できないところでは新たな学習器を生成しながら学習し，新たな階層構造を構築する．そして学習の結果得られた下位層の学習器と階層構造自体を一つの学習器としたものをこれまでに獲得した学習器群に加えながら段階的に与えられるタスクを学習していく (**Fig.1**)．この過程において階層構造を持つ学習器を再利用することで，木構造を自律的に構成することができる．

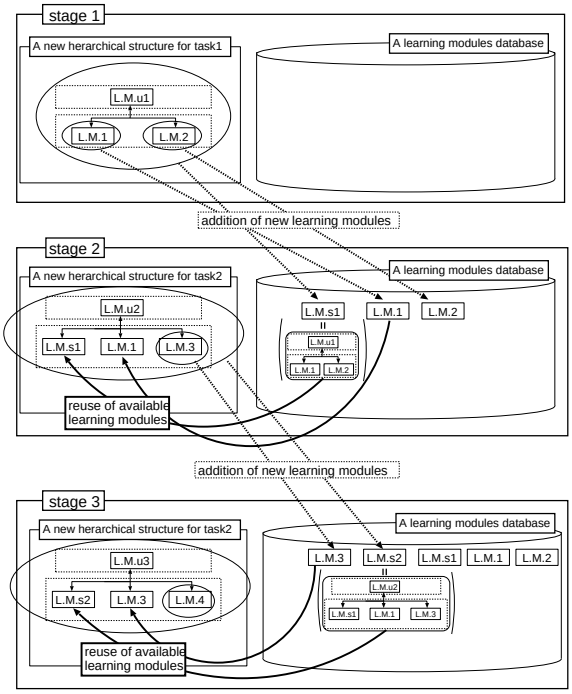

Fig.1 An example of a stepwise learning

3 階層型学習機構

Fig.2 に階層型学習機構を示す．基本的な構造は 2 層で，下位層はこれまでのタスクで獲得された学習器の中で新たなタスクに対して有効であると判断された学習器と新たに生成した学習器で構成される．下位層の学習器の状態空間と行動空間はそれぞれロボットのセンサ情報とモータコマンドにより構成される．上位層との情報のやりとりには抽象化を行なった情報を用いる．各学習器は上位層へゴール状態活性度 g を出力し，上位層からは行動指令 b を受け取る．ゴール状態活性度は状態価値を正規化したもので，下位層の各学習器がそれぞれのゴール状態にどれだけ近いかを表している．行動指令は下位層のどの学習器が獲得した行動を実際に出力するかを決

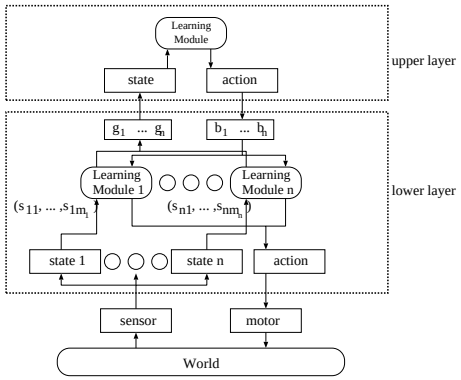

Fig.2 A hierarchical architecture

める上位層からの指令である．今回は行動を出力する学習器はどれか 1 つのみとする．

上位層は一つの学習器から構成される．状態空間は下位層の各学習器から出力されるゴール状態活性度をそのまま状態変数として構成する．そして行動は下位層のどの学習器を選択するかである．下位層からゴール状態活性度を受け取ると最適行動を計算し，どの学習器を選択するかを行動指令として下位層の各学習器に出力する．

4 学習アルゴリズム

学習の流れを以下に示す．

1. コーチがいくつかの行動系列を例示
2. これまでに獲得した学習器の有効性を判断し，有効な学習器を下位層に追加
3. 有効な学習器がない部分で新たな学習器を生成
   (a) 状態空間を決定
   (b) サブゴール状態の決定
   (c) 学習器の生成と行動価値関数 $Q(s,a)$ の計算
   (d) 学習器の評価と下位層への追加
   (e) 例示データ上を有効な学習器で覆うことができていれば 4 へ進む
   できていなければ (a) に戻る
4. 上位層を学習

4.1 学習器のタスクにおける有効性の判断

例示される行動は学習者にとって最適である保証はない．そのため例示された行動と学習器の獲得した最適な行動が完全に同一でなかったとしても，同じ目的を達成することのできる行動であれば，その学習器は有効と判断すべきである．そこで式 (1) で定義する $\overline{Q}$ を用いて，学習器の与えられたタスクに対する有効性を判断する．

$$\overline{Q}(s,a_e) = \frac{Q(s,a_e) - min_{a'}\, Q(s,a')}{max_{a'}\, Q(s,a') - min_{a'}\, Q(s,a')} \qquad (1)$$

$\overline{Q}$ が 1 に近ければ a_e はその学習器のゴール状態へ近づく行動であり，逆に 0 に近ければゴール状態から遠ざかる行動であると判断できる．そこで閾値 $\overline{Q}_{th}$ を設け，$\overline{Q}$ が $\overline{Q}_{th}$ より高い部分をその学習器が有効である部分とする．また例示される行動は最適である保証はないことから，一時的に $\overline{Q}$ が変動することが考えられる．そこである一定の時間 T_{th} 以下の変動は無視することとする．

4.2 サブゴール状態の決定

他のタスクに対しても有効であるような汎用性のある学習器を生成するため，次の経験則を導入する．

- ゴール状態は状態変数の値が最大，最小，中間値のいずれかであることが多い．そこでこの 3 つの値の内，新たな学習器が必要な部分の終りの状態に最も近い値をその状態変数のゴール状態とする．

またサブタスクを行なう上で必要であるが，ゴール状態には関係しない状態変数も存在する．そこで全ての例示データに対してゴール状態が同じ値とならなかった場合には，その状態変数はゴール状態 (報酬) には関係ない変数と考える．

5 実験

環境はロボカップの中型リーグのフィールドを想定し，ロボットは中型リーグで使われているロボットを用いる．タスクを

- タスク 1:ボールチェイス
- タスク 2:シュート
- タスク 3:敵をかわしてシュート

の順番で段階的に与え，学習する．学習時間を短縮するため，まずシミュレータ上で学習し，その結果を実ロボットに実装する．

段階的に学習を行なった結果，**Fig.3** のような行動が獲得され，階層構造は **Fig.4** のようになった．

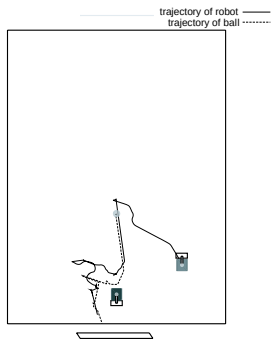

Fig.3 The acquired behaviors for task3

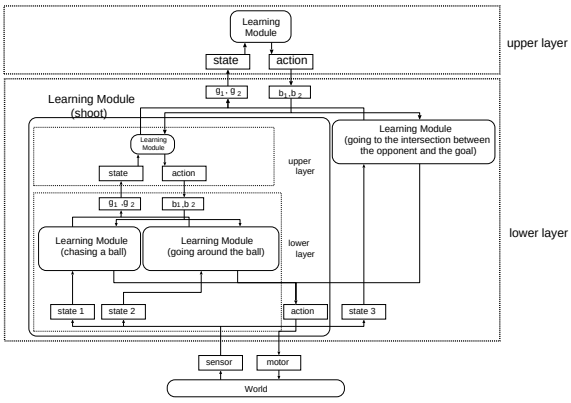

Fig.4 The acquired hierarchical structure for task3

参考文献

[1] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-based reinforcement learning for purposive behavior acquisition. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 146–153, 1995.

[2] Jonalthan H. Connell and Sridhar Mahadevan. Rapid task learning for real robots. In *ROBOT LEARNING*, pages 105–140. Kluwer Academic Publishers, 1993.

[3] Steven D. Whitehead. Complexity and cooperation in Q-Learning. *Machine Learning*, pages 363–367, 1991.