モジュール型学習機構における 例示の理解に基づいた自律的なタスク分解

Self Task Decomposition for Modular Learning System through Interpretation of Instruction by Coach

○ 西 智樹 (阪大)　正 高橋 泰岳 (阪大, 阪大 FRC)　正 浅田 稔 (阪大, 阪大 FRC)

Tomoki NISHI, 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 method of self task decomposition for modular learning system based on self-interpretation of instructions given by a coach. The proposed method enables a robot (i) to decompose a long term task that needs much information into a sequence of short term subtasks that need much less information based on its self-interpretation process for the instructions given by the coach, (ii) to select sensory information needed for each subtask, and (iii) to integrate the learned behaviors to accomplish the given long term task. We show some results from 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，ロボットが日常生活の中で目にする機会が多くなってきている．それに伴い人が生活している環境中で人間の代わり，または人間と共に作業を行なうロボットの研究が注目されている．この様なロボットが作業を行なう環境は複雑であり，決まった環境で決まった作業を行なう産業用ロボットのように設計者が事前に起こり得る全ての状況を予測し，行動を記述することは事実上不可能である．そこでロボット自身が環境と相互作用することにより，状況に適した行動を獲得することができる強化学習などの学習的アプローチが注目されている．

一方，人間の生活環境で作業を行なうためには多くの情報を考慮する必要があり，そのためロボットが有するセンサの種類や数は増加の傾向にある．このようなロボットに単純な強化学習を適用した場合，センサの増加に伴い，状態数が指数関数的に増大し，学習時間の増加，膨大な計算資源の消費が問題となり，現実的な時間でロボットが状況にあった行動を獲得することが困難となる．このような問題を解決する手法として，学習のスケジューリングを行なう手法やタスク分解を行ないそれらを統合する手法を用いた学習が多く研究されている [1, 3]．これらの研究の多くはスケジューリング，タスク分解や統合を設計者が行なっているが，用いる学習器に適したタスク分解を設計者があらかじめ決定することは困難である．そのため学習者が自律的にタスク分解を行なうことが望ましい．

しかしながらタスクに関する知識を一切持たずにタスクに適したタスク分解を自律的に行なうことは困難である．そのため学習者がタスクの遂行に適したサブタスクや状態空間を自律的に構築するためには，何らかの形でタスクに関する知識を学習者が得る必要がある．学習者にタスクの知識を与える方法としては例示が考えられ，例示を強化学習に導入した研究として Whitehead[2] がある．この研究では 1 つの学習器で学習する場合についてのみ例示の有効性が示されているが，自律的にタスク分解を行ない複数の学習器によりタスクを遂行するシステムにおいても用いることができると考えられる．つまり例示を用いることによりタスク分解や状態空間の自律的発見を容易にすることができる．

そこで本研究ではいくつかの例示に基づき，状態空間のサイズが小さい学習器で学習できる範囲をサブタスクに設定し，学習，統合することによりタスクの遂行を実現する手法を提案する．またサッカーロボットのシミュレータを用い，シュート行動を学習させることにより本手法の有効性を検証した．

2 問題設定

Fig.1 A real robot

Fig.2 Captured camera images

環境は RoboCup の中型リーグの想定し，環境中には対象物としてボール，相手ゴール，味方ゴールが存在する．またロボットは中型リーグで用いられているものを用いる (図 1)．このロボットはセンサとして全方位カメラ，前方位カメラ，赤外線距離センサを有する (図 2)．

環境中には学習者と教示者が存在し，教示者は学習者が持つ学習機構に関する知識は持っていないものとする．

そのため教示者は明示的にサブゴールを与えたり，学習者にとって最適な行動を例示することはできないが，タスクを達成することができる行動系列は例示することができる．また学習者には生成できる学習器の大きさには制限があり，学習者はそれ以下の大きさの学習器の集合によりタスクを達成しなければならない．そのためタスクは数次元の状態空間を用いて実現できるエピソード的タスクの集合に分割できるとする．

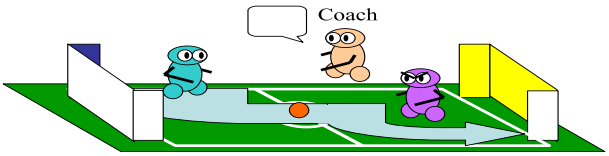


Fig.3 A coach given instruction to a learner

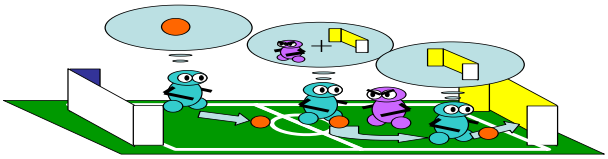


Fig.4 The learner follows the instruction and finds behaviors by itself

3　提案手法

3.1　タスク分解の流れ

学習の大まかな大まかな流れを図 5 に示す．但し x_i は状態変数を，赤線は例示を表している．また図中の番号は以下の手順の番号と対応している．

1. 教示者がいくつかの行動系列を例示.
2. タスク分解
   - ① 例示からサブゴール候補の選出.
   - ② サブゴール状態と状態変数の決定
   - ③ 学習器の生成とその有効範囲の決定
   - ④ 例示におけるみ学習領域の発見．あれば①へ.
     なければタスク分解を終了し 3 へ.
3. 生成された学習器を下位層に持つ階層構造を構築

3.2　学習器の有効性の判定

タスク分解を行なうためには獲得された学習器の有効範囲を決定する必要がある．しかし教示者は学習者が持つ学習機構に関する知識を持っていないため，学習者自身が学習器の有効範囲を決定しなければならない．そこで状態空間 S, 行動空間 A を持った学習器の状態 $s \in S$ における有効性は状態 s での例示行動 a_e と学習器の最適行動との Q 値に関しての類似性で判断する．具体的には以下の AE(式 (1)) が閾値 AE_{th} を越えているかどうかで判定する図 6.

$$AE(s, a_e) = \frac{Q(s, a_e) - min_{a' \in A} Q(s, a')}{max_{a' \in A} Q(s, a') - min_{a' \in A} Q(s, a')} \quad (1)$$

3.3　サブゴール候補の決定

新たなサブタスクはすでに獲得した学習器全ての $AE(s, a_e)$ が閾値 AE_{th} を下回っている領域であり，その領域の最後の状態をサブゴール状態候補とする (図 6).

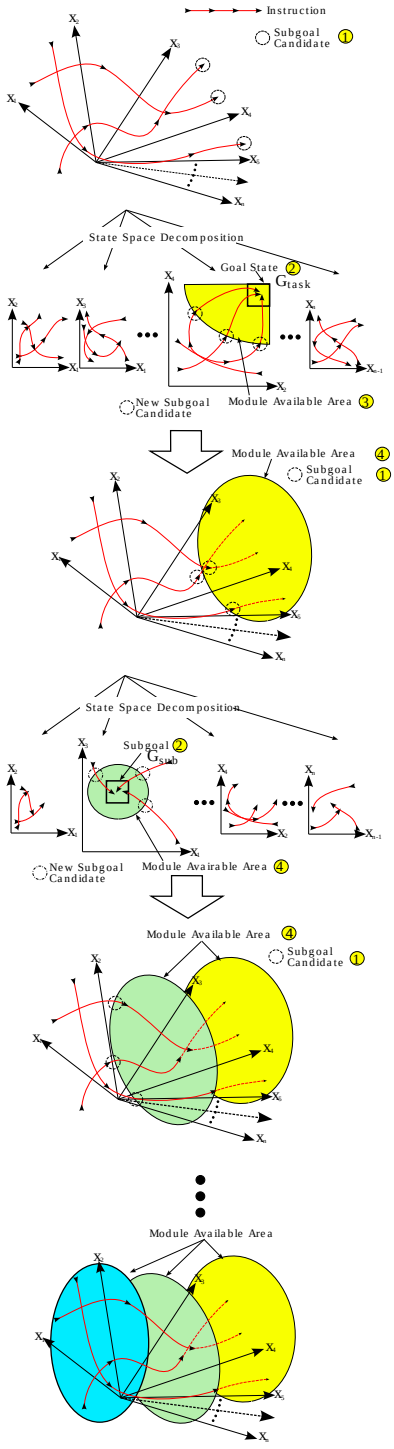


Fig.5 Rough sketch of the idea of task decomposition procedure

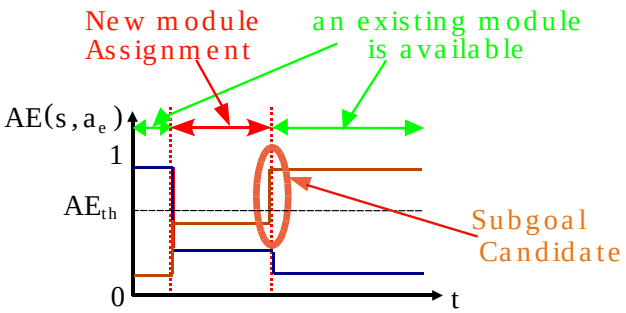


Fig.6 Subgoal candidate point

3.4 状態空間の自律的構成

3.4.1 状態空間の決定の基本アイデア

以下に示す 2 つの考えに基づき学習に用いる状態変数を選ぶ.

1. 吸収ゴール (abosorbing goal) をもつタスクを達成するためにはゴール状態をコンパクトに表現できる状態変数が少なくとも 1 つは必要である.
2. Q 学習は Q 値に基づいて行動を決定するため，状態空間は Q 値の予測が正しく行える空間である必要がある.

3.4.2 ゴール状態のコンパクトさ

ゴール状態のコンパクトさはサブゴール候補がどれくらい多く同じ量子化された領域に存在するかということとサブタスクの領域でその変数がどれくらい状態変化に敏感であるかということを表す評価値になっている.

各例示 k の量子化した各状態変数候補 x_i の軸上でサブゴール候補から同じ量子化された領域に存在するサブゴール候補の数のヒストグラムをとり, その値が最も大きい領域をその状態変数候補 x_i のゴール状態 (GR_i) とする. その時式 (2) に示す評価値を最大とする状態変数候補をゴール状態を最もコンパクトに表現する変数であるとする.

$$C_i = \sum_k l_{i,k} \tag{2}$$

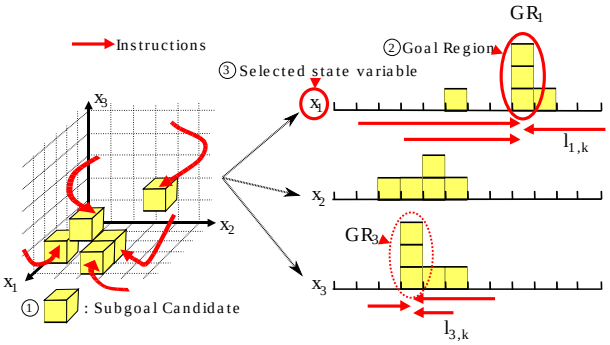


Fig.7 Specification of goal region and its state variable for a subgoal

3.4.3 状態空間の適正度

Q 学習は Q 値の予測に基づいて行動することから，適切な行動を獲得するためには Q 値を正しく予測することができるということは非常に重要なことである. そこで状態空間の Q 値を正しく予測する能力を反映した評価値である状態空間の適正度 QE(式 (3)) を導入する.

$$QE = \frac{1}{\sum_{\forall a \in A, \forall s \in S} e(s,a)} \tag{3}$$

$$e(s,a) = \sqrt{P(1-P)} \tag{4}$$

$$P = P(V(s') > V(s)|s,a) \tag{5}$$

$P(V(s') > V(s)|s,a)$ は状態 s で行動 a をとった時に次状態 s' の状態価値 $V(s')$ が現状態の状態価値 $V(s)$ よりも増加している確率を表しており，このように状態空間の適正度を状態価値の増加，減少という 2 値化した値を取る確率を用いることにより，経験の偏りによる V 値のばらつきの影響を緩和することができる.

QE が大きいほど全状態において，次状態の状態価値が現状態の状態価値よりも増加する，または減少するのどちらか一方の傾向が強いことを表し，反対に QE が小さいほどその傾向が弱いことを表す. そのため最も QE が大きいものを選択することにより，Q 値を適切に予測することができる状態空間を選択できると考えられる.

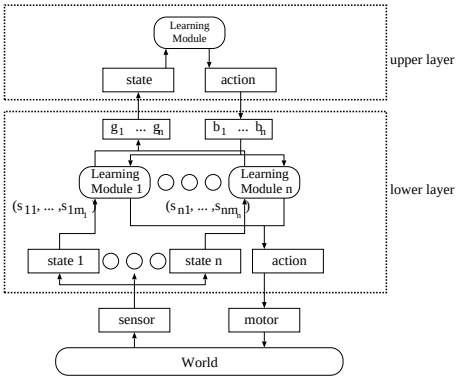


Fig.8 A hierarchical architecture

3.5 階層型学習機構

図 8 に階層型学習機構[1]を示す. 基本的な構造は 2 層で, 下位層はこれまでに獲得された学習器の中で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および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，下位層の各学習器の現状態と各ゴール状態との距離を反映している. また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.

3.6 学習器の生成の流れ

新たな学習器を生成する流れを以下に示す.

1. ゴール状態のコンパクトさを計算.
2. 最もゴール状態をコンパクトに表現する状態変数を持つ学習器を生成
3. 生成された学習器で一定期間学習.
4. 生成された学習器のタスクの達成率が閾値を越えれば，学習器の生成を終了する. そうでなければ次へ.
5. これまでに選択された N_s 個の状態変数にその他の状態変数候補 1 つを付け加えた状態空間を $(N_a - N_s - 1)$ 個生成，およびその状態空間における状態遷移モデルと報酬モデルを構築. 但し N_a は全状態変数候補の数とする.
6. 生成された状態空間において状態空間の適正度を計算.
7. 最も有効だと判定された状態空間を持つ学習器を生成. 3 へ.

4 実験

4.1 実験設定

環境は RoboCup 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用いられているものを用いる. また選択可能な状態変数候補として，前方位カメラについては対称物 i の面積 A_{pi}，画像上での対称物 i の重心位置までの距離 D_{pi},θ_{pi},X 座標の値 X_{pi},Y 座標の値 Y_{pi} を用

[1] 今回用いた階層型学習機構は文献 [2] で提案されたものと同様のものを用いた

いる．また全方位カメラに関しては，対称物 i の面積 A_{oi}，画像上での対称物 i の重心位置までの距離 D_{oi},θ_{oi},X 座標の値 X_{oi},Y 座標の値 Y_{oi}, 対称物 i とボールとの相対距離 D_bi, 及び相対角度 θ_{bi} を用い，赤外線距離センサ j についてはセンサ出力 IR_j を用いる．つまり対称物 i としてボール b，味方ゴール mg，相手ゴール og が存在ので合計 39 のセンサ情報を用いる．また各状態変数候補は全て 11 状態に離散化している．行動は X,Y,θ 軸全てにおいて正の出力，負の出力，停止の 3 行動に離散化し，合計 27 通りある．

ボール追跡のような単純なタスクからシュートのような複雑なものへと段階的に例示を行ない学習することは段階的な行動系列の例示自体が暗にタスク分解を行なっていることになる可能性がある．そのため本実験ではシュート行動の例示のみからタスク分解およびシュート行動獲得を実現可能かを検証した．

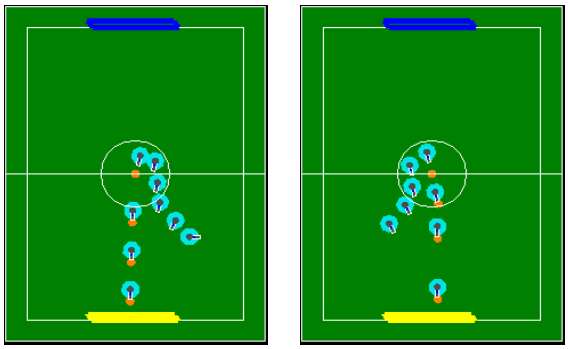


Fig.9 Instructed behaviors

4.2 実験結果

本実験では 4 つの行動系列を例示した．その一例を図 9 に示す．本提案手法により例示からタスク分解を行なった結果，4 つのサブタスクに分割され，各サブタスクを担当する学習器として学習器 LM[A_{ob}, X_{pb}],LM[θ_{omg}], LM[$\theta_{omg}, A_{ob}, \theta_{bmg}$] が順に生成された．但し学習器 LM[$A_{ob}, X_{pb}$] は全方位カメラのボールの面積 A_{ob} と前方位カメラのボールの X 座標 X_{pb} とを状態変数とする状態空間を持っていることを示している．各学習器の AE を図 12(a) に示す．

また獲得された学習器を階層型学習機構を用いて統合することにより獲得された行動と，その各時刻に選択される学習器の様子の一例を図 11 および図 12(b) に示す．

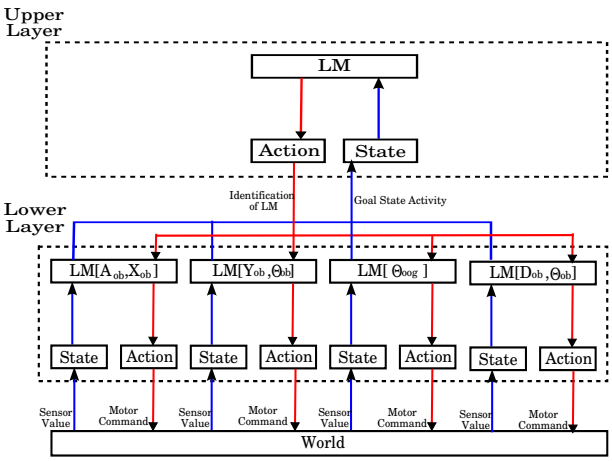


Fig.10 Acquired hierarchy for the shooting behavior

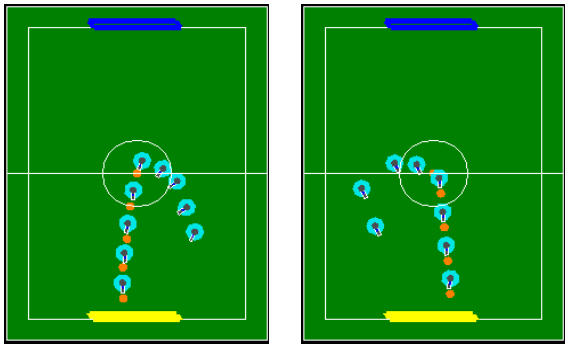
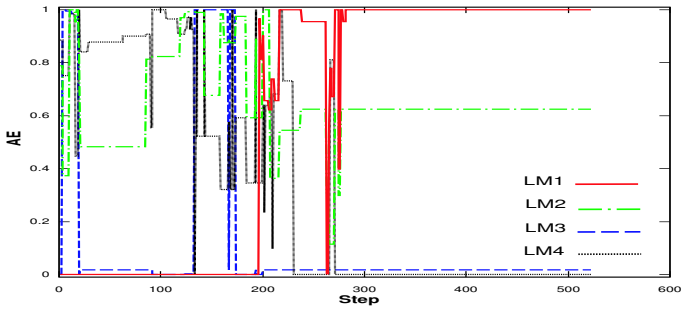
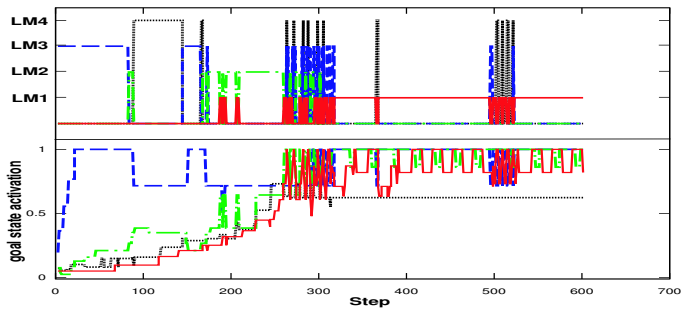


Fig.11 Acquired behaviors for shooting task



(a) Aquired LM in shooting task,wich are LM$_1$(A_{ob},X_{pb}), LM$_2$(Y_{ob},θ_{ob}), LM$_3$(θ_{oog}) and LM$_4$(D_{ob},θ_{ob})



(b) Sequences of the selected modules and variables which reflect distance to each LM's goal ,in shooting task

Fig.12 Experimental Result

5 おわりに

例示を学習者自身が理解し，サブタスクの発見および学習器の生成，さらには階層構造の構築を自律的におこなうことにより，学習者の能力あったタスク分解を可能とした．またこの手法の有効性をシミュレーションによりシュート行動を獲得する実験により示した．

参考文献

[1] Asada, S.Noda, S.Tawaratumida, and K.Hosoda. Purposive behavior acquisition for a real robot by vision-based reinforcement learning. In *Machine Learning*, Vol. 23, pp. 279–303, 1996.

[2] Steven D. Whitehead. Complexity and cooperation in q-learning. In *Proceedings Eighth International Workshop on Machine Learning(ML91)*, pp. 363–367, 1991.

[3] Y.Takahashi and M.Asada. Multi-controller fusion in multi-layerd reinforcement learning. In *IEEE/RSJ International Conferrence on Multisensor Fusion and Integration for Intelligent Systes(MFI2001)*, pp. 7–12, 2001.