複数学習器を用いたマルチエージェント環境における行動獲得

枝澤 一寛 (阪大)　○高橋泰岳 (阪大, 阪大 FRC)　浅田稔 (阪大, 阪大 FRC)

Modular Learning System for Behavior Acquisition in Multi-Agent Environment

Kazuhiro Edazawa, *Yasutake Takahashi, Minoru Asada

Abstract— The existing reinforcement learning approaches have been suffering from the policy alternation of others in multiagent dynamic environments since other agent behaviors may cause sudden changes of state transition probabilities of which constancy is needed for the learning to converge. In this paper, we adopt the basic idea of the mixture of experts into an architecture of behavior acquisition in the multi-agent environment. Multiple modules are assigned to different situations and learn purposive behaviors for the specified situations which are expected as the consequence of other agent's behavior under different policies. However, it is hard to assign modules automatically in the multi-agent system that has highly dynamic ones. We show scheduling for learning is introduced to avoid the complexity in autonomous situation assignment.

Key Words: module-based reinforcement learning, multi-agent system

1. はじめに

これまで強化学習をロボットに適用する研究が多くなされてきた [1]．それらの多くでは学習者から見て環境の状態遷移確率が一定である，もしくはその変化が非常に遅いという条件が必要であった．これは学習者から見た状態遷移確率が変化する場合は学習が収束しないためであり，そのため他のエージェントの政策の変化により学習者から見た状態遷移確率が大きく変化するマルチエージェント環境下において目的の行動を獲得することは従来手法では困難である．

Asada et al.[2] はシステム同定の手法を用いて学習者と他のエージェントの状態ベクトルを推定し協調行動を獲得している．しかしこの手法は学習者は 1 体で，かつ他のエージェントは固定政策の必要がある．そのため他のエージェントの政策が変化する場合はこの手法は適用できない．Ikenoue et al.[3] はそれぞれの学習者の行動政策の切替を非常にゆっくりと行なうことにより複数ロボットの同時学習を可能にしている．しかしこの手法は学習時間が多くかかり，他のエージェントの政策が変化したときには再学習をする必要がある．マルチエージェント環境下では他のエージェントの政策は変化するものであり，学習者はこれに対処する必要がある．

Jacobs and Jordan[4] は複数の学習器を用い，各学習器の出力をゲートで重み付けしたものをシステム全体の出力とする Mixture of Experts と呼ばれる学習システムを提案している．各学習器の状況に対する適応度に応じて重み付けし全体の出力を求めるという考え方は，効率の良いシステムを作る上で広く適用できる．鮫島ら [5] や Haruno et al.[6] は，非線形・非定常なタスクの制御則をモジュール構造を用いて学習させるという MOSAIC(MOdular Selection and Identification for Control) を提案している．この手法は環境の予測性に基づいて複雑なタスクを時空間的に分割し，予測が正しく行なわれるモジュールに制御を行なわせるものである．単純な予測器が同時並行に状態を予測し，その予測の最も良い予測器と対となっている制御器が責任を持って環境を制御，また学習するという予測性を基準にしたスイッチング制御/学習方式である．彼らは比較的単純なダイナミクスを持った環境下での実験で成功している．マルチエージェント環境下では相手の政策によって状態遷移が動的に変化するため，このような刹那的な観測からある状況を表現するモデルを構築するために十分なデータを獲得することは困難であり，ある程度の観測を重ねることが必要となる．

本論文ではマルチエージェント環境下における競合行動の獲得にこのような複数の学習器を利用する際に学習のスケジューリングを行ない，学習者に状況を判断するための十分な試行を許すことで，学習者が状況の切り分けを適切に行なうことができ，結果的に良いパフォーマンスを出せることを示す．各学習器は計画器と予測器を持ち，学習者は環境の先験的な知識を持たずに，予測器は環境との相互作用のみから状態遷移モデルを構築し，計画器はそのモデルに基づいて強化学習の枠組で行動計画を行う．学習初期においては学習のスケジューリングをある程度設計者が行う．学習がある程度進んだ段階で学習者は相手の行動を観測することにより，どの学習器が現在の状況に最も適しているかを，予測器の予測誤差をある期間評価することにより判断し，自律的に状態遷移モデルの割り当てを行う．またその学習器の計画器を用いて行動する．相手の政策の変更により動的に変化する環境において，複数の学習器を使い分けることで行動が獲得できることを示す．

2. 複数学習器による行動獲得

2·1 システム概要

提案システムを Fig.1 に示す．点線で囲まれた各学習器は予測器 (predictor) と計画器 (planner) を持ち，予測器は状態遷移確率モデルを構築し，計画器はその状態遷移確率モデルに基づいて動的計画法の手法で行動価値関数を推定する．ゲートは各学習器が予測する状態価値関数の値を基に，現在の状況を最も良く予測している学習器の計画器の計画する行動を選択し，状況

にあった行動をとる．

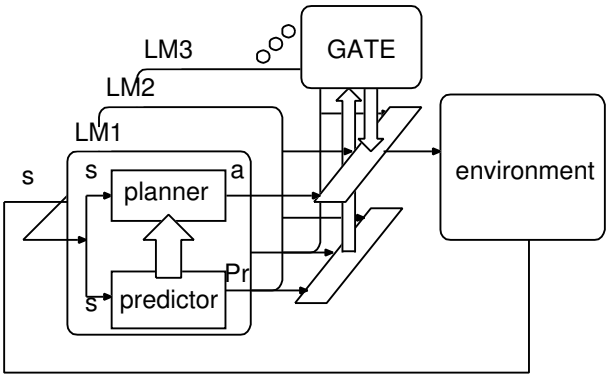


Fig.1 A multi-module learning system

2·2　予測器

各学習器の予測器は状態遷移モデルを持ち，ある状態 s，行動 a，次状態 s' となる確率

$$\hat{\mathcal{P}}^a_{ss'} = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \qquad (1)$$

を推定するモデルを持つ．このモデルは学習者が環境と相互作用することによって構築される．またシステムは状態遷移モデルだけでなくある状態 s と行動 a が与えられたときの次の報酬の期待値

$$\hat{\mathcal{R}}^a_s = E\{r_{t+1} | s_t = s, a_t = a\} \qquad (2)$$

を推定する報酬モデルも持つ．

2·3　計画器

予測モデルで計算された状態遷移確率 $\hat{\mathcal{P}}^a_{ss'}$ と報酬 $\hat{\mathcal{R}}^a_{ss'}$ が求まるとある状態，行動における行動価値関数 $Q(s,a)$ は

$$Q(s,a) = \sum_{s'} \hat{\mathcal{P}}^a_{ss'} \left[\hat{\mathcal{R}}^a_{ss'} + \gamma \max_{a'} Q(s',a') \right] , \qquad (3)$$

のようにして与えられる [7]．この値を基に動的計画法の枠組で行動を計画する．ここで γ は減衰係数を表す．

2·4　学習器の選択

学習器の切り換えには以下で定義する信頼度 g を用いる．各モジュールの信頼度 g_i の値は，ある一定期間 T の予測器の出力する予測確率が正しいほど大きな値となる．信頼度の大きな学習器を用いて行動することで，現在の状況に対して最適な行動が獲得できる．

ある状況 (状態 s で行動 a をとったときに次状態 s' になる場合) における予測確率は，各学習器の予測モデルに基づいて

$$p_t = \Pr\{s_t = s' | s_{t-1} = s, a_{t-1} = a\} \qquad (4)$$

のように予測される．この予測値を用いて，信頼度は

$$g = \prod_{t=-T+1}^{0} \exp(-\lambda p_t) \qquad (5)$$

のように計算される．ここで λ はスケーリングパラメータであり $\lambda = 0.2$ とした．

3.　サッカーロボットへの適用

3·1　タスクと仮定

本手法をサッカーロボットのタクスに適用する．ロボットが得られる環境からの情報は全方位視覚の画像情報のみである．タスクはパス行動の獲得とする (Fig.2 参照)．このタスクは味方にパスを行う学習者がディフェンスの動きに応じてボールまでアプローチを学習するというものである．相手が片方のパスコースをブロックした場合に学習者はもう片方のパスコースを狙ってボールにアプローチを行う行動を学習する．相手がどちらのパスコースをブロックするかという政策の変化から，行動に用いる学習器を切り換えて相手の行動に対応する．

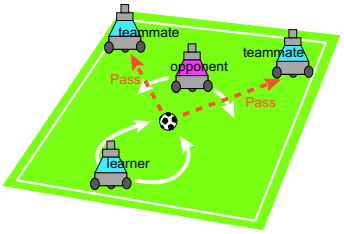
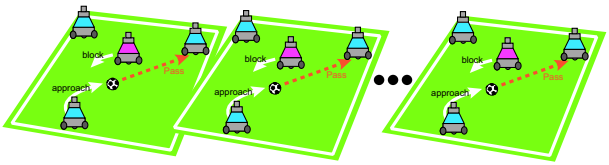


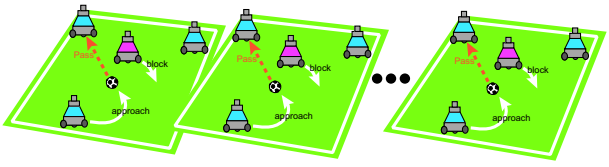
Fig.2 A task: 3 on 1

3·2　学習のスケジューリング

現在の相手に対してどの学習器の予測モデルを構築すべきかは学習器が未熟だと解らないので学習のスケジューリングを行う．その際に，最初は相手の行動を一定のものとして固定トレーニングを行う (Fig.3 参照)．また学習時の学習者の行動はそれまでの学習の結果を用いて ϵ-greedy 政策を行う．ある程度各学習器の学習が進んだ段階で，固定トレーニングを終了し，学習器の信頼度を基に学習すべき学習器を選択する．



(a) training pattern 1



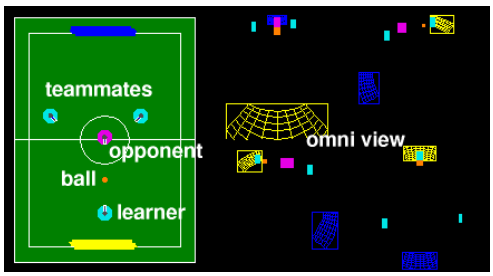
(b) training pattern 2

Fig.3 Fix trainings

3·3　実験設定

提案する手法の有効性を確認するために，シミュレーションにて実験を行った．ロボットは Fig.4(a) にあるような，センサに全方位カメラ，移動機構に全方位移動機構を持ったロボットを想定し Fig.4(b) に示すようにシミュレーションを行う．

(a) A real robot

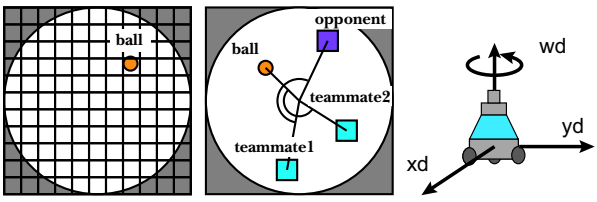


(b) A simulation environment

Fig.4 The robot and simulation screen shot

3·3.1　状態空間

強化学習に用いる状態集合 $\mathcal{S}$ には，全方位カメラ画像上におけるボールの位置，エージェントとボールとの角度を変数として用いた．(Fig.5(a)，Fig.5(b)) に示すようにボールの位置については 11×11 の格子状に離散化し，角度については 360° を 8 個に離散化した．



(a) state variables(position)　(b) state variables(angle)　(c) action variables

Fig.5 A state-action space

3·3.2　行動空間

行動集合 $\mathcal{A}$ は全方位移動機構における水平面の移動と鉛直方向の回転運動における目標出力 (x_d, y_d, w_d) を離散化したもので構成する．それぞれ 3 段階の離散化を行った (Fig.5(c))．

3·4　状態遷移モデルの構築

各学習器の状態遷移モデルの構築には，相手の政策を固定してトレーニングを行った．相手が

1. 右にブロックする行動
2. 左にブロックする行動

という 2 種類の政策をとるとする．学習時における学習者の政策は ϵ-greedy 手法とし，相手とボールを観測し，環境と相互作用することで状態遷移モデルを構築する．ゴール状態に辿り着いた場合，エージェントがフィールドの外に出た場合，試行がある一定時間を越えた場合に試行をリセットするとした．また状態の観測は一定時間間隔でサンプリングを行っている．状態の観測においてサンプリングタイムは，実際のロボットの画像処理のフレームレートが 1/30 秒であると仮定し，シミュレーションにおいては 3 フレーム，つまり 0.1 秒毎に状態の観測を行うことを想定した．

3·5　ゴール状態の判定

相手の各行動におけるゴール状態の判定は，学習者がボールまで辿り着いたときに，相手が自分の正面にいなければゴールであるとして報酬 1 を与え，学習者は報酬モデルを更新する．

3·6　実験結果

パス行動獲得のタスクをシミュレーションにて行なった．学習者は敵の動きに応じてどの方向にパスするかを学習している．学習により獲得された行動の様子を Fig.6 , Fig.7 に示す．Fig.6 は相手が学習者に対して左側をブロックする行動を取ったときの行動の様子である．相手が左側をブロックしているので，学習者はボールに対して左側から近づいている (Fig.6③)．その結果右側の味方へのパスを成功させている (Fig.6④)．Fig.7 は相手が学習者に対して右側をブロックする行動を取ったときの行動の様子である．figfig:rel-L-block,Fig.8(b) はそれぞれ相手が左ブロック，右ブロックの行動を取ったときに，(5) 式で計算される各学習器の信頼度の変化を示している．Fig.8(a) を見るとタスク開始の 2 秒から 10 秒にかけて学習器 1 の信頼度が学習器 2 の信頼度よりも大きな値となり，学習器 1 を用いて行動計画を行っている．その結果 Fig.6 で見たようにパス行動を成功させている．同様にして Fig.8(b) においては学習器 2 の信頼度が学習器 1 の信頼度を上まわり，学習者は学習器 2 を用いて行動計画を行い，パス行動を成功させている．相手の行動が，左ブロック，右ブロックのとき，学習者は状況を判断してボールに対するアプローチを変えていることがわかる．またこのとき各学習器のモデルの更新も信頼度の値をもとに行っている．また複数学習器を用いたシステムと単一学習器を用いたシムテムについて，学習のスケジューリングを行った場合と行わない場合についての比較を行った (Fig.9)．

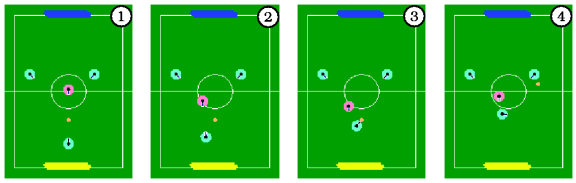


Fig.6 The acquired behavior for the left block policy of the opponent

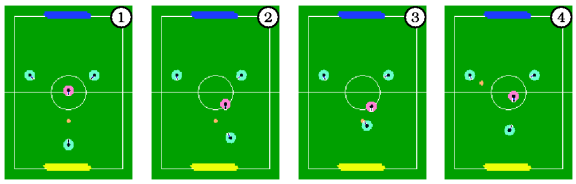
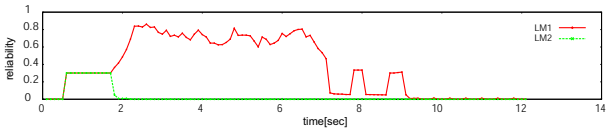
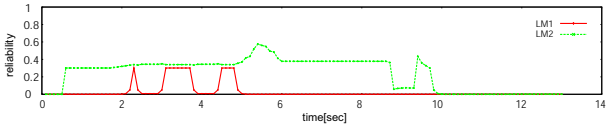


Fig.7 The acquired behavior for the right block policy of the opponent



(a) The opponent is blocking the left side



(b) The opponent is blocking the right side

Fig.8 The sequence of the reliabilities of the learning modules

3·7 学習のスケジューリング

スケジューリングありの学習とは 250 試行の固定トレーニングを行う学習であるのに対して，スケジューリングなしの学習では固定トレーニングなしで，交互に敵が政策を変えるという状況で学習を行う．スケジューリングありの学習も固定トレーニングが終了すれば，スケジューリングなしと同様に敵の政策は交互に変わる状況で学習を行う．

3·8 結果の比較

250 試行までのタスク成功率を見ると，それぞれ大差はなく成功率が上昇している．250 試行から 500 試行にかけては固定トレーニングを行っている単一学習器は失敗の連続である．それに対して複数学習器の成功率は以前の学習とは関係なく学習を進めるので成功率は上昇している．500 試行以降は，複数学習器を持つシステムでは自律的に割り当てを行い行動し，およそ

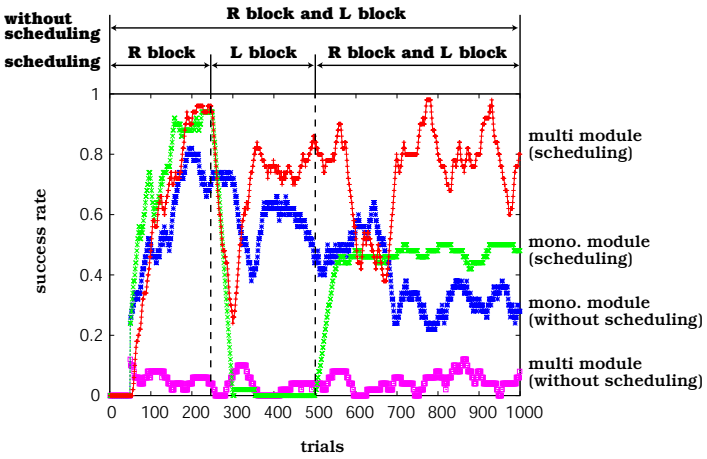


Fig.9 Curves of success rate

80%の成功率となっており，単一学習器を上回っているのがわかる．また複数学習器を用いて学習のスケジューリングを行わないと成功率は低い値となっている．

4. おわりに

本論文ではマルチエージェント環境下における競合行動の獲得に複数の学習器を利用する手法を実装し検証した．相手の出方に応じて状況を自律的に判断し，その状況に応じた行動を獲得できることを示した．またその状況と行動の学習に置いてある程度学習のスケジュールリングが必要であることを示した．

参考文献

[1] M. Asada, S. Noda, S. Tawaratumida, and K. Hosoda: “Purposive behavior acquisition for a real robot by vision-based reinforcement learning”, Machine Learning, 1996.

[2] M. Asada, E. Uchibe, and K. Hosoda: “Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and developmen”, Artificial Intelligence, Vol. 110, pp. 275–292, 1999.

[3] Shoichi Ikenoue, Minoru Asada, and Koh Hosoda: “Cooperative behavior acquisition by asynchronous policy renewal that enables simultaneous learning in multiagent environment”, Proceedings of the 2002 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems, pp. 2728–2734, 2002.

[4] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton: “Adaptive mixtures of local experts”, Neural Computation, Vol. 3, pp. 79–87, 1991.

[5] 鮫島和行, 銅谷賢治, 川人光男: “強化学習 mosaic:予測性によるシンボル化と見まね学習”, 日本ロボット学会誌, Vol. 19, pp. 551–556, 2001.

[6] M. Haruno, D. M. Wolpert, and M. Kawato: “Mosaic model for sensorimotor learning and control”, Neural Computation, Vol. 13, pp. 2201–2220, 2002.

[7] Richard S. Sutton and Andrew G. Barto: “強化学習”, 森北出版株式会社, 2000.